# Constraints and invariance in target detection

By Frederick Nicolls

Thesis presented for the Degree of Doctor of Philosophy,
Department of Electrical Engineering,
University of Cape Town.
Cape Town, February 2000

# Constraints and invariance in target detection

**Frederick Nicolls**
February 2000

## Abstract

The concept of invariance in hypothesis testing is discussed for purposes of target detection. Invariant tests are proposed and analysed in two contexts.

The first involves the use of cyclic permutation invariance as a solution to detecting targets with unknown location in noise. An invariance condition is used to eliminate the target location parameter, and a uniformly most powerful test developed for the reduced data. The test is compared with conventional solutions, and shown to be more powerful. The difference however is slight, justifying the simpler formulations. This conclusion continues to hold even when additional unknown noise parameters are introduced.

The second form of invariance involves invariance to data contained in linear subspaces of the observation space. Such invariance is proposed as a generic method of reducing mismatch resulting from overly-simple models, which cannot capture the full complexity of the data. The formulation involves a data-reducing projection, where components of the data which are difficult to model are discarded. The notion has relevance to both low-rank subspace interference and low-rank subspace mismatch.

Methods are presented for making estimates of invariance subspaces and covariance matrices for use in this invariant detection formulation. Using two different interpretations, necessary conditions on the subspace estimates are derived. These conditions can sometimes be solved exactly, but approximate methods are provided for the general case. It is shown that for use in invariant detectors, covariance matrix estimates have to be independent of data components contained in the invariance subspace. The EM algorithm can be appropriate for this estimation. Maximum likelihood methods for obtaining the estimates are presented under Toeplitz, circulant, doubly block Toeplitz, and ARMA constraints.

Using actual samples of x-ray data, results are presented which demonstrate that subspace invariant tests can be substantially more predictable than noninvariant tests. This is true in particular when models are highly constrained, as is commonly necessary in adaptive detection. The improved predictability usually comes at the cost of a decrease in performance, although this is not always so. The invariant detectors also avoid the need to preprocess data to better fit the assumptions.

# Acknowledgements


- The financial assistance of the National Research Foundation towards this research is hereby acknowledged. Opinions expressed in this work, or conclusions arrived at, are those of the author and are not necessarily to be attributed to the National Research Foundation.
- The financial assistance of De Beers Consolidated Mines Limited is also acknowledged.

# Contents

# Acronyms

| | |
|---|---|
| AMF | Adaptive matched filter |
| AR | Autoregressive |
| ARIMA | Autoregressive integrated moving average |
| ARMA | Autoregressive moving average |
| CFAR | Constant false alarm rate |
| DFT | Discrete Fourier transform |
| EM | Expectation maximisation |
| FFT | Fast Fourier transform |
| GLRT | Generalised likelihood ratio test |
| LRT | Likelihood ratio test |
| MA | Moving average |
| MLE | Maximum likelihood estimate |
| MVN | Multivariate normal |
| NAR | Noise-alone reference |
| ROC | Reciever operating characteristic |
| SMI | Sample matrix inversion |
| SNR | Signal-to-noise ratio |
| UMP | Uniformly most powerful |
| UMPI | Uniformly most powerful invariant |

# Chapter 1

# Introduction

In Section 1.1 a basic description is provided of the types of problem addressed in this work. This is followed in Section 1.2 by an overview of statistical hypothesis testing, which is the paradigm in which the concepts are presented.

The generalised likelihood ratio test is a common hypothesis testing formalism used in applied detection, and is presented in Section 1.3. Section 1.4 then discusses the general problem of noise modelling using parametric probability densities, which are required for developing tests.

The dominant theme in this dissertation is the use of constraints and invariance in hypothesis testing. Section 1.5 discusses these concepts in general terms, and outlines their use in this work. In Section 1.6 the specific problem of constraints and invariance in adaptive testing is outlined.

Section 1.7 provides a detailed outline of the topics presented in the remainder of the thesis, and Section 1.8 concludes with a summary of the notation used throughout.

## 1.1 Description of the general problem

Target detection constitutes an important component of signal processing. Perhaps the most traditional application in an engineering context is in the area of telecommunications, where it may be required to make a decision of whether a given sample of data contains just noise, or some target embedded in noise.

However, the use of target detection extends far beyond this simple application. In modern contexts it is common to see target detection applied to such diverse topics as radar and sonar, remote sensing, and image processing. Additionally, it is frequently applied to very complicated problems involving multiple observations, often using a variety of sensors which may operate using different modalities.

**Section 1.1: Description of the general problem**

When applied to these types of problem, the complexity of the models used to describe the observations can increase dramatically. For example, in image processing the models may have to take into account complex backgrounds arising from the imaging of natural scenery. The types of model required for such description are far removed from those typically used in basic telecommunications, or even in time-series analysis.

In particular, it becomes necessary to develop models which permit very long-range correlations between successive data samples. The traditional uncorrelated noise models are far too simplistic to accurately characterise the types of signal observed in these applications.

Unfortunately, complex models lead to complicated detection procedures. This in turn leads to computationally intensive implementations, usually in situations where they can least be afforded. Many engineering problems require real-time answers to problems which involve vast quantities of data. Thus there is a real need to extend the simple models to problems where they would not ordinarily be appropriate. Since simple models generally afford efficient implementations, if the extension is done appropriately then at least the computational requirement can be met.

In this work, one such extension is provided. It is appropriate in instances where a simple model is *almost* appropriate, but where certain well-defined deviations may occur. In particular, the use of subspace invariant detectors is proposed for problems where the actual observed data differ from the assumed model only within a restricted linear subspace of the observation space.

To some extent, the methods presented validate many ad hoc techniques that are often used to enhance the validity of simple models. For example, in nonstationary environments it is common to preceed a detection procedure with a preprocessing stage, which aims to reduce the mismatch between the model and the data. However, because the methods in this work are developed in a formal context, they provide an extension to such ad hoc techniques. Insights can therefore be gathered which may not be apparent from a simple analysis of the data.

Simple models can usually be considered to be constrained versions of more complex models. Therefore, one way of relaxing the rigidity of a model may be to simply lift some of these constraints. However, the methods presented relax the rigidity in a different way, namely by imposing an invariance condition on sets of observations. In a sense, this invariance imposes an equivalence on those components of the observations which are difficult to characterise using the assumed models.

There is a complex relationship between the role of the constraints and the role of the invariance group. Stronger invariance conditions can lead to more accurate modelling, but only at the cost of reduced detection performance. On the other hand, weaker constraints on model parameters can also lead to better modelling, this time at the cost of poorer estimation of the parameters.

The distinction between constraints and invariance becomes paramount when detectors have to be made adaptive. Usually some form of adaptivity is essential in complex detection problems, since

even the most complex models have difficulty in accurately characterising the data. However, as soon as adaptive detectors are considered, the problem of effective parameter estimation arises. It is particularly in this context that relaxing the constraints becomes problematic. The invariant detectors, however, can achieve better model accuracy *without* introducing additional adaptivity parameters.

Another important topic of this work is to place on firmer ground certain detection procedures which do not have a solid statistical foundation. In particular, the problem of detecting targets with unknown location is a common one, but statistical justification for the procedures used is not readily available. Some measure of validation is therefore provided for many of these procedures.

In this work a specific subset of detection problems is considered. Primarily, the situation is discussed where a single instance of a target from a known set of possibilities may be present in the observed data. This target is assumed to be additive, in that the presence of the target in no way affects the background that would have been observed were the target absent. This formulation is appropriate in applications such as transmission-mode imaging. More generally, it is also approximately valid whenever targets are transient and of short duration.

Although concerned with practical solutions, the emphasis of this work is in the concepts involved. Thus very few attempts have been made to optimise algorithms or provide efficient solutions. It is felt that these issues should be addressed *after* the concepts have been validated.

This thesis is not geared towards any specific practical application. Rather, it presents a set of methods which may be applied to general detection problems, in order to improve some aspect of the performance. The samples on which results are tested are taken to be x-ray images, but this does not imply that the conclusions are limited to these types of data.

## 1.2 Overview of statistical hypothesis testing

In the field of statistics, hypothesis testing is used to decide which of a number of predetermined candidate scenarios is true, based on a sample of observed data. The assumption is made that the sample is generated from one of a set of possible processes, each corresponding directly to an interpretation that is of relevance to the problem. In the simplest case, the hypothesis test is required to identify the process which generated the data. More generally, however, a number of different underlying processes may relate to the same hypothesis, thereby placing an equivalence on processes with different probability densities. The hypothesis test is then only required to identify the class of densities from which the data were generated. The class then corresponds to the hypothesis being tested, which effectively allows for unknown model parameters in the test formulation.

A hypothesis test is simply a function $t(\mathbf{x})$ of the observed data $\mathbf{x}$, which is chosen or designed to have certain useful or convenient properties with regard to the decision procedure. Since any invertible

transformation can be applied to this test statistic, without loss of generality it may be chosen so that only tests of the form

$$t(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{1.1}$$

are considered, where it is assumed that one of the two hypotheses $H_0$ and $H_1$ is in force. The notation used in this expression means that the decision of $H_1$ is made when $t(\mathbf{x}) > \eta$, with $H_0$ being chosen if $t(\mathbf{x}) < \eta$. Evidently the parameter $\eta$ could be fixed in advance and $t(\mathbf{x})$ scaled to yield certain desired properties, but it is usually preferable to lend an absolute interpretation to $t(\mathbf{x})$ in such a way that by varying $\eta$ a class of tests is obtained with different specific properties but similar overall characteristics. Thus all that is required to design a hypothesis test is a systematic method of arriving at a function $t(\mathbf{x})$ which results in a useful decision process.

To derive such a statistic a relative measure is required of how typical an observation is under a set of possibilities. In abstract terms this may be achieved by specifying a function $f_i(\mathbf{x})$ for each possible underlying process $i$, which reflects the extent to which $\mathbf{x}$ is typical of that process. The resulting set of functions should ideally have the property that, for each $i$, $f_i(\mathbf{x})$ tends to be larger than $f_j(\mathbf{x})$ whenever $\mathbf{x}$ more closely resembles the types of data generated by process $i$ over those generated by process $j$. Ostensibly there are many possibilities for the selection of such functions.

In the realm of parametric statistics the probability density is a natural candidate for providing this information: for each possible underlying process $i$, the conditional probability density function $p(\mathbf{x}|i)$ is a direct measure of how typical the observation $\mathbf{x}$ is when $i$ is in force. Thus for a given observation $\mathbf{x}$ the probability $p(\mathbf{x}|i)$ may be calculated, and the resulting value is an absolute measure of the extent to which there is evidence for the process $i$.

For the problem of a simple hypothesis versus a simple alternative, where there is only one possible underlying process under each hypothesis, the ratio $p(\mathbf{x}|H_1)/p(\mathbf{x}|H_0)$ directly expresses the relative evidence for the hypothesis $H_1$ over $H_0$. As such it is a good candidate for $t(\mathbf{x})$, and is used extensively in simple decision-making problems. This leads to what is commonly referred to as the likelihood ratio test (LRT), which is of the form

$$\frac{p(\mathbf{x}|H_1)}{p(\mathbf{x}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{1.2}$$

This test can be shown to have a number of strong optimality properties. Firstly, if $H_0$ corresponds to a target-absent condition and $H_1$ to the condition of target presence, then the LRT maximises the probability of detecting the target subject to a maximum false alarm constraint [79, p. 114]. Alternatively, it may be shown to minimise the false alarm rate under a constraint on the probability of detection. This is referred to as the Neyman-Pearson optimality criterion, and is prevalent in the radar literature [16]. On the other hand, under a Bayesian formulation the LRT is optimal in that it minimises the risk associated with the test [79, p. 63], where risk is defined to be the expected value of a loss function which assigns certain costs to particular types of errors. In this case the threshold parameter $\eta$ depends on both the loss function and the prior probabilities of each hypothesis being in

force.

The Neyman-Pearson criterion exhibits an asymmetry in that the two types of errors, namely a false alarm and a missed target, are not treated equivalently in the decision. Also, the quantities considered in the formulation are *conditional* errors, which depend on the hypothesis in force. It is the combination of these two properties that makes the Neyman-Pearson condition powerful and useful, and indeed the only sensible criterion in some instances.

Because only conditional densities and conditional errors are considered, the Neyman-Pearson formulation makes neither implicit nor explicit use of information regarding the relative probability of each hypothesis being in force. If such information is available, a Bayesian formulation is commonly appropriate, which minimises the risk associated with the test. In contrast to the Neyman-Pearson test, this risk is closely related to *absolute* frequencies of errors, which take the prior probabilities of the hypotheses into account. Thus, under a Bayesian formulation it can make sense to derive a test which minimises the *absolute* probability of an error occurring.

It is not always necessary to specify $p(\mathbf{x}|H_i)$ explicitly to find the hypothesis test. In some instances it may be convenient to assume a parametric form for the test statistic itself, and estimate the parameters using a sample of training data. Such a procedure is commonly used in artificial neural networks (ANNs) and support vector machines (SVMs), which can be formulated in such a way as to completely bypass the formal specification of conditional distributions. This has application to both the Bayesian and Neyman-Pearson hypothesis testing paradigms. In the Bayesian situation another possibility is to model the posterior distributions of the data instead of the conditional distributions, as done in logistic discrimination [115, p. 109]. Nevertheless, if it *is* possible to specify the conditional probability, then it is very useful for developing appropriate tests.

The optimality of the likelihood ratio test of Equation 1.2 is restricted to the case of the hypotheses being simple, where there are no unknown parameters in the conditional probabilities. For many problems this is too simplistic to accurately characterise the processes generating the data. Thus additional parameters have to be introduced to account for increased model complexity. The conditional densities used in the decision-making process are then of the form $p(\mathbf{x}|\boldsymbol{\theta}, H_i)$, which depend explicitly on the vector parameter $\boldsymbol{\theta}$.

In the case of Bayesian testing the introduction of such parameters does not constitute a problem: a probability density for the parameters is proposed, which may in general depend on the hypothesis, and the required density $p(\mathbf{x}|H_i)$ is calculated by integrating over $\boldsymbol{\theta}$. This quantity can then be used in the likelihood ratio test. The Neyman-Pearson test does not extend so easily, however. The ideal solution is to find a single test which is Neyman-Pearson optimal for *all* values of the unknown parameters, referred to as a uniformly most powerful (UMP) test. Such a test does not exist in general. A less powerful but more practical approach is to use the quantity $\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_i} p(\mathbf{x}|\boldsymbol{\theta}, H_i)$ as a measure of likelihood, and form the likelihood ratio using these quantities in place of the conditional densities $p(\mathbf{x}|H_i)$. This leads

to the generalised likelihood ratio test (GLRT). A third option is to redefine the optimality criterion to be less restrictive, and attempt to find an optimal test under the modified criterion. Locally most powerful (LMP) tests present one such possibility [79, p. 140]. Finally, the class of allowed tests may be restricted to those displaying some desirable characteristic, in the hope that an optimal test exists in this reduced class. Uniformly most powerful invariant (UMPI) and uniformly most powerful unbiased (UMPU) tests are relevant in this regard [34].

The Bayes test is often considered superior to all other tests, and to represent an ideal solution to a given decision problem. This is oversimplistic in two regards: firstly, it is only optimal in situations where minimum Bayes' Risk can be considered a sensible and useful optimality criterion. In general an acceptable loss function is usually fairly arbitrary, however, and is often chosen to have desirable computational properties rather than a meaningful interpretation. Secondly, specifying prior probabilities on the hypotheses is itself a difficult procedure, and the effects of errors in these specifications can be very significant. This is further hindered by problems where the prior probabilities may vary substantially over time, and have to be made adaptive.

Take for example the problem of detecting tumours in an x-ray image of a human torso. To specify the Bayesian detector, one is required to quantify the relative costs between, at the very least, a falsely detected tumour and a missed tumour. This is a difficult relationship to quantify, since it involves on the one hand an economic cost of subsequent testing (perhaps in the form of a biopsy), and on the other hand a more emotive and humanitarian cost. For this type of problem, prior distributions on the hypotheses are also difficult to specify. If one uses for instance the national average as a prior measure, it might happen that an entire group of people with different risk factors (based on say geographic or demographic differences) are completely misdiagnosed. In this instance an incorrectly specified prior can be disastrous. Similar observations apply to the use of a prior distribution on, for instance, the size of the tumour: it is not inconceivable that some unknown factor may result in a group of people with unusually large tumours. If the detector considers these to be atypical of tumours and therefore disregards them, the consequences are severe.

For the tumour detection problem, and for many others, conditional loss functions based on a likelihood formulation make considerably more sense. Under this interpretation, one would fix the probability of detecting a tumour in advance, and choose the test that minimises the resulting false alarm rate. Because the two types of conditional errors are fundamentally different, the asymmetric nature of the testing criterion better suits the requirements. Similar conclusions apply to most radar problems — it is more sensible to think about the probabilities of detecting or of missing a target which is actually present (that is, conditional probability of detection), than to consider overall error rates. An example of where the Neyman-Pearson criterion is indicated in radar problems is in the task of large-scale detection and clearing of land-mines: a sensible specification for the problem is to require in advance that say 95% of mines be found, and to minimise the false alarm rate subject to this constraint.

In this work, the Neyman-Pearson criterion is used for designing detectors and assessing performance. This essentially lends a likelihood (rather than a Bayesian) interpretation to the detection methodology. The methods presented are therefore appropriate for problems where conditional errors in the Neyman-Pearson formulation are relevant.

## 1.3 The GLRT as a testing philosophy

As discussed in Section 1.2, the likelihood ratio is an important quantity in hypothesis testing. Under the Neyman-Pearson criterion it is a sufficient statistic for the optimal test under simple hypotheses. The fact that it appears in optimal formulations in various other cases serves to highlight its importance.

In general the testing problem becomes considerably more complicated when the hypotheses are no longer simple. Optimality conditions have to be redefined, and often become so restrictive that best tests do not exist. As mentioned in Section 1.2, if the hypotheses are composite then it may be reasonable to use the quantity $\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_i} p(\mathbf{x}|\boldsymbol{\theta}, H_i)$ as a measure of how typical an observation is under each hypothesis, where $\boldsymbol{\Theta}_i$ is the set of possible values of $\boldsymbol{\theta}$ under $H_i$. This expression measures the highest probability of observing the data vector $\mathbf{x}$, over all possible values of the parameter vector permitted within the hypothesis. The ratio of the maximum probability under the assumed hypotheses $H_1$ and $H_0$ can then be used directly as a test statistic. In the engineering literature, this test is called the *generalised* likelihood ratio test, or GLRT [126, p. 92]. In the statistics literature, it is commonly just referred to as the likelihood ratio test for composite hypotheses [81, p. 240].

The GLRT formulation is often used as a semi-automatic procedure for developing tests in instances where no optimal test exists. Once the conditional probability densities under the hypotheses are specified, the GLRT is determined up to an unknown threshold parameter. The principle should not be applied indiscriminately, however, and numerous warnings appear in the literature indicating that the resulting test may not always be sensible [81, p. 263] or may not exhibit certain expected properties [134]. Nevertheless, in many practical problems it provides a good solution, and in some instances the resulting test may be optimal [120].

In this work, the GLRT is used extensively as a reference for comparing alternative optimal or suboptimal detectors. For example, in Chapter 3 it is compared with the uniformly most powerful invariant test for detecting a known target with unknown location in certain types of noise. For this particular problem, the GLRT is a desirable solution on account of its intuitive properties, and the fact that the implementation is simple. Thus it is important to know the extent to which the test is suboptimal, to provide justification for its use. In contrast, in Chapter 6 the GLRT is presented as an effective but nevertheless complicated detection procedure for finding a partially-known target in noise with unknown correlation properties. In this context the performance of the GLRT is good, but its implementation is inefficient. Thus it is used as a baseline for comparing other suboptimal but nevertheless more tractable

tests.

In a useful and general form, the composite likelihood ratio testing problem can be formulated in terms of two sets of parameters: those which are relevant to the specification of the hypotheses (the test parameters $\boldsymbol{\theta}_r$), and those which are not (the nuisance parameters $\boldsymbol{\theta}_s$). The probability density of the observation $\mathbf{x}$ is then $p(\mathbf{x}|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$, where the value of $\boldsymbol{\theta}_r$ determines the hypotheses and the value of $\boldsymbol{\theta}_s$ is assumed irrelevant. As discussed by Kendall and Stewart [81, p. 240], a reasonable test for $H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}$ versus $H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r_0}$ is to form the ratio

$$l(\mathbf{x}) = \frac{\max_{\boldsymbol{\theta}_r, \boldsymbol{\theta}_s} p(\mathbf{x}|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)}{\max_{\boldsymbol{\theta}_s} p(\mathbf{x}|\boldsymbol{\theta}_{r_0}, \boldsymbol{\theta}_s)} \tag{1.3}$$

and use it as a test statistic. Letting $\hat{\boldsymbol{\theta}}_r$ and $\hat{\boldsymbol{\theta}}_s$ be the maximum likelihood estimates of $\boldsymbol{\theta}_r$ and $\boldsymbol{\theta}_s$, and $\hat{\boldsymbol{\theta}}_{s|r_0}$ be the MLE of $\boldsymbol{\theta}_s$ conditional on $\boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}$, this can be written as

$$l(\mathbf{x}) = \frac{p(\mathbf{x}|\hat{\boldsymbol{\theta}}_r, \hat{\boldsymbol{\theta}}_s)}{p(\mathbf{x}|\boldsymbol{\theta}_{r_0}, \hat{\boldsymbol{\theta}}_{s|r_0})}. \tag{1.4}$$

They go on to prove that under some regularity conditions on the maximum likelihood estimates, the procedure of choosing $H_1$ when $l(\mathbf{x}) > \eta$ constitutes an asymptotically optimal test. That is, for a sufficiently large number of observations the test performs as well as if the nuisance parameter values were known in advance. Furthermore, the asymptotic distribution of the test statistic is derived (p. 246), as well as the asymptotic test power (p. 247).

In the field of statistical signal processing a slightly modified terminology has arisen. Specifically, in [126, p. 92] Van Trees provides a formulation of a hypothesis test for the null hypothesis that $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus the alternative $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, where $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ form a disjoint partition of the parameter space. With the conditional density of the observation $\mathbf{x}$ given by $p(\mathbf{x}|\boldsymbol{\theta})$, the test is

$$l(\mathbf{x}) = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1} p(\mathbf{x}|\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} p(\mathbf{x}|\boldsymbol{\theta})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{1.5}$$

Van Trees calls this a *generalised* likelihood ratio test (GLRT). Although the distinction between the testing parameters and the nuisance parameters is not explicit here, the test can be seen to include the LRT of Equation 1.3: if $\boldsymbol{\theta}$ is taken to be the combined set of parameters $(\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$, then the hypothesis $H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}$ corresponds to the choice $\boldsymbol{\Theta}_0 = \{(\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)|\boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}\}$. Similarly, the hypothesis $H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r_0}$ corresponds to $\boldsymbol{\Theta}_1 = \{(\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)|\boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r_0}\}$. Under this interpretation it is apparent that the expression for the likelihood ratio $l$ in Equation 1.3 is a special case of that presented in Equation 1.5.

In this work the signal processing terminology will be used. Thus the term "likelihood ratio test" will be reserved for problems where both hypotheses are simple. For composite hypotheses, the corresponding test will be referred to as a generalised likelihood ratio test.

In some cases it may be valid to assume that the targets and noise add to give the observation $\mathbf{x}$: that is, $\mathbf{x} = \mathbf{s} + \mathbf{n}$, where $\mathbf{s}$ is the target and $\mathbf{n}$ the noise. In this case the distribution of the observation when a target is present will differ from that where the target is absent only through a change in location of the distribution. The noise probability density and the target vector are then all that is required to fully specify the GLRT statistic.

Such additivity of targets and noise is quite restrictive. It is invalid in most conventional light-optical images, where if targets are present they generally occlude the background. In most transmission-mode imaging systems, however, the additivity assumption is at least approximately valid. For example, in x-ray images a tumour manifests itself as a compact region which exhibits higher attenuation than surrounding regions. The background anatomical detail is still present, and appears superimposed on the tumour. Under suitable preprocessing of the data, usually a logarithmic transformation of the actual photon counts, approximate additivity of the tumour data values and the anatomical data values can be imposed. This conclusion does however ignore the fact that the presence of a dense target results in increased photon noise at the target location, a factor which may be negligible if the target is small and exhibits low attenuation.

The GLRT has no absolute optimality properties. Furthermore, for detection in typical engineering environments it has a number of disadvantages which may restrict its usefulness. Most of the disadvantages stem from the fact that the test is *entirely* specified by the noise density and the chosen test threshold. Firstly, the density is often difficult to specify with a high degree of accuracy, particularly when unknown parameters are included in the formulation. This is discussed further in Section 1.4. Secondly, once the relevant probability density functions have been specified, the GLRT statistic is fixed. If the resulting test is intractable or exhibits some undesirable properties, the only recourse is to modify the noise densities and try again. There is no general systematic procedure for performing these modifications. Thirdly, the GLRT statistic often has an intractable or parameter-dependent sampling distribution, making it difficult to set the free test threshold parameter meaningfully. Finally, the fact that the test has only one degree of freedom means that it is impossible to specifically customise the test, for example when multiple targets may be present with varying degrees of detectability.

In spite of these problems, the GLRT plays an important role in the design of hypothesis tests in practical problems. Specifically, it can provide insights which can guide the search for a more appropriate or suitable test. Modifications to the GLRT are not entirely without precedent: for a specific class of problems, Kendall and Stewart [81, p. 259] demonstrate that the GLRT can be made unbiased by adjusting the maximum likelihood parameter estimates so that they are unbiased. Since the test has no optimality properties, there is no reason to avoid deviating from the details of the testing principle.

## 1.4 Specifying the conditional density of the noise

The development of the likelihood ratio or the generalised likelihood ratio for the hypothesis test requires explicit specification of the noise density. In general, for the noise model to be accurate, it may be necessary to include unknown parameters in the formulation. Thus an appropriate form for the noise density $p(\mathbf{x}|\boldsymbol{\theta})$ needs to be provided, which is accurate yet restrictive enough to be useful, and which is of a form convenient for subsequent processing.

In this work, the noise may be comprised of more than just simple statistical random noise. There is usually also clutter present in the signal, which may have long-range correlations in all directions. For example, for detection in x-ray images of people, the anatomical structure constitutes the clutter field. This clutter is distinct from other noise sources in the system, such as photon noise in the x-ray imaging process or thermal noise in the detector electronics. If on the local scale the clutter can be considered stationary, however, it can conceptually be lumped together with the other noise sources into an overall noise density $p(\mathbf{x}|\boldsymbol{\theta})$.

Specifying the noise density is difficult. In some respects it is the most important stage of the detector design, since often this conditional density fully determines the form of the hypothesis test. The easiest case occurs when the density can be derived from an explicit and detailed knowledge of the signal formation process. The parameter $\boldsymbol{\theta}$ then describes the possible states of the underlying process, and for each value of this parameter a density for the noise may be developed based on physical principles. The detector follows accordingly, and insofar as the modelling is accurate, the performance should be good.

In general, however, the conditional noise density cannot be justified purely on physical grounds: the signal formation process is too complicated. A difficulty then arises in how to determine which parameters should be included into $\boldsymbol{\theta}$. In abstract terms there are at least two reasons for introducing unknown parameters into the conditional density:

- Parameters should be introduced when they *significantly* improve the match between actual data and assumed probability density functions. They can then result in improved detection performance by means of more accurate model specification. The improved accuracy has to be weighted against the fact that the parameter ultimately has to be estimated, which tends to reduce the resulting performance.
- Parameters may be introduced to take into account any peripheral knowledge which may not be adequately represented in the training data. For example, for detection in x-ray images of human torsos one may anticipate that different people will exhibit different average attenuation, depending on their size. Even if the training set contains only people of similar size, it would make sense to introduce a parameter to accommodate this possible variability.

It is difficult to quantify the improvement made by the introduction of a parameter into the conditional density. The main reason for this difficulty is in that the density estimate is only used to derive the subsequent hypothesis test. It is quite conceivable that a bad density estimate will still result in a good test. Also, it might happen that the hypothesis test is essentially independent of a parameter that is seemingly required for better accuracy in the density. Modelling for detection and for estimation are clearly very different.

In practice, especially for data with high dimensionality, it is seldom possible to only add parameters as they are required. The primary cause is the near impossibility of joint density estimation for even fairly small sets of data. Thus it is more common to use one of the standard "off-the-shelf" models. Within the restriction of multivariate normal densities, useful assumptions and resulting densities which are used in this work are the following:

- The noise samples are often assumed independent, especially as a first attempt at an approximate solution. This corresponds to a multivariate normal (MVN) process with a covariance matrix equal to some scaling of the identity matrix. In its most simple generalisation, the variance may be considered an unknown parameter.
- A less restrictive assumption is that of spatial or temporal stationarity. Under this condition, the statistics of the noise are assumed to be identical at all locations in the observed data. The stationarity assumption is considerably more general than the white noise assumption, but also has *many* more parameters.
- Between the white noise and the stationarity assumption are the autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models which are common in time series analysis. The number of parameters in each case may be chosen to provide the desired degree of flexibility.

It is unreasonable to expect that these simple and restricted assumptions will be valid in even a small set of practical problems. At the very least, even the stationarity assumption is usually questionable. However, it is often possible to use these models under the assumption that they are only partially valid. Images, for example, typically have the property that the statistics are approximately constant in regions that are close in space to one another. Thus the stationary constraint may be considered approximately valid on the short term, and the stationary model used in this context. For purposes of implementation, a preprocessing stage is often used which transforms the data into a form where the model assumptions are more valid.

As an aside, if the restricted models are valid then they may permit estimates to be made of quantities which could otherwise not be estimated in a less restricted model. This is a consequence of the fact that there is information contained in the model constraint, which can be used *in addition* to any observed

data samples. A particular aim of this thesis is to assess whether constraints can be used to permit estimates to be made which are invariant to certain transformation groups.

From the detection point of view, every possible state of the parameter represents a different probability density that has to be considered. In general, if the parameter value is not known, the detector either implicitly or explicitly makes an estimate of this parameter in the course of the detector calculation. On account of the finite and nonzero variance of this estimate, there is added uncertainty in the estimated distribution of the observed data. Thus, in general, if there are more possible states then the estimation is more difficult, which in turn results in less precise testing. A model with a very high number of states has the capacity to encompass a high degree of complexity in the data. However, if the complexity of the model is higher than required, there will be a decrease in the overall performance.

This introduces the classical bias/variance tradeoff in the context of statistical hypothesis testing in the likelihood framework: if the parametric specification of the conditional density is too restrictive, then there is not enough capacity in the model to capture the data complexity. In that case there is no value of the parameter vector which corresponds to a probability density that accurately represents the data. The density estimate will then be biased, and the test performance will suffer. On the other hand if there are a large number of parameters encompassing too large a number of possible states, then the parameter estimation is compromised and the density estimate has high uncertainty. This too causes a degradation in detection performance. Qualitatively, therefore, the model should be flexible enough to capture the complexity of the data, while at the same time not being so complex as to include states which are not required for effective modelling.

## 1.5 Constraints and invariance in testing

The previous section discussed some restrictive parametric models which may at times be appropriate for modelling a noise process. All the models are stationary, in that they assume the statistics independent of the location of the origin of the coordinate system in which the values are represented. Each of the models may be considered to be constrained versions of a more general process, namely that of all MVN random processes of the appropriate order. In this context, the constraints restrict the allowed variability of the models, thereby permitting improved performance through added knowledge.

### 1.5.1 Constraints in hypothesis testing

The primary reason for introducing constraints is to reduce the number of parameters. As discussed in the previous section, fewer parameters can be estimated better, and reduce the set of possible models which have to be considered viable in the detection procedure. Insofar as the constraints are valid, these factors should result in improved overall performance. It was suggested earlier that additional

parameters should be introduced whenever they are required for a better match between the data and the model. In the context of constraints another possibility is indicated: start with a general model, and introduce constraints as required.

As with the process of adding parameters, adding constraints is difficult in the context of detection mainly on account of the indirect nature of the relationship between the probability density and the resulting test. There is also a similar bias/variance tradeoff associated with the model constraint interpretation: if more constraints are added, there are effectively fewer free parameters. If the constraints are valid, this results in better estimates which in turn lead to improved detection. However, if too many constraints are added the model becomes overly restrictive, and the potential for bias increases.

There is a possible misconception that constraints simplify the subsequent processing. For the detection problem, such simplification may be seen to occur when moving from an unconstrained model to the assumption of a white noise constraint. However, this is by no means a consistent finding — in general constraints are difficult to work with, and in the worst case may not afford useful closed-form solutions to components of the detection procedure. On the other hand, added constraints may be used to develop algorithms with reduced computational requirements, but these are often complicated, difficult to derive, and do not extend readily from 1-D signals to 2-D images.

A difficult issue regarding constrained models in detection (and in general regarding the use of conditional densities in deriving detectors) is that in some respects the models really only have to accurately represent probabilities in the vicinity of the decision boundary. These are the regions where most of the samples that correspond to difficult decisions will lie. In the framework developed in this thesis there appears to be no way of utilising this observed characteristic. Perhaps the most reasonable modern paradigm which explicitly deals with this problem is that of support vector machines (SVMs), which are essentially only concerned with observed samples which are near to the required decision boundary.

An additional advantage of a well-specified constraint is that it can be used to condition solutions to problems where there are otherwise not enough data. An analogy may be drawn to a system of equations with too many unknowns — if the amount of data is fixed, then one has to reduce the number of effective unknowns, usually by adding knowledge regarding relationships between them. Conceptually this can be achieved through the introduction of constraints, which provide this additional information. The constraints may be considered to reduce the number of free parameters or variables, by specifying relationships between them which may not be violated.

Thus, for statistical problems, constraints may enable estimates to be made of quantities which may otherwise not be estimated. Consider for example a problem where two variables are postulated, each from a Gaussian distribution with known mean but unknown variance. If the variables are assumed independent and the variances are allowed to differ (that is, a fairly unconstrained model), then observing the value of the first variable provides no information regarding the variance of the second. However, under the assumption that the two variances are equal (that is, a more constrained

model) this is no longer the case: now the value of the first variable contains as much information regarding the unknown variance as does the value of the second. Thus through the use of well-chosen and accurate constraints it may be possible to calculate parameter estimates based on only partial data.

### 1.5.2 Invariance in hypothesis testing

Whereas constraints imply a relationship between the model parameters, it is also possible to enforce relationships between sets of observations. This introduces the topic of invariance. As applied to the problem of hypothesis testing, invariance is a formal procedure whereby certain observations or classes of observations are considered equivalent for purposes of detection. That is, invariance arguments are used to effectively equate a set of distinct observations which differ in a very precise manner from one another.

For the invariance requirement to be reasonable, it either has to be demonstrated that the precise differences between observations that are assumed equivalent are not important to the specific problem, or that the equivalence condition provides other advantages such as more accurate model specification. In both cases it may be expected that the end result could be an improvement in detection performance. Additionally, in some instances the application of an invariance criterion may be considered a data-reduction mechanism, which maps the high-dimensional data to lower-dimensional equivalence classes.

The use of invariance is in certain cases quite well established in the literature. The most common application is in detection using observations which are corrupted by low-rank additive interference. As an example, suppose that all observations of a random process are assumed to be subject to a constant arbitrary offset with unknown magnitude. It then makes sense to require equivalence between the observation $\mathbf{x}$ and the observations contained in the set $\{\mathbf{x} + \alpha\mathbf{1}, \alpha \in \mathbb{R}\}$, where $\mathbf{1}$ is the vector of ones. In general the interference subspace may have a higher dimensionality than that used in this example, but similar principles apply.

When used in this context, invariance can be used to justify optimality of certain tests. Scharf [119, pp. 127–153], for example, discusses a number of specific cases where nuisance parameters (which are parameters that are required to accurately model the data density, but are of no further interest in the detection problem) are eliminated through the application of a reasonable invariance class. If these nuisance parameters are the only unknowns in the formulation, then the resulting invariant test has a very strong optimality property, namely that it is uniformly most powerful in the class of all tests which share the same invariances. Such a test is referred to as a uniformly most powerful invariant (UMPI) test. Insofar as the invariance restriction is reasonable for the problem, the test is optimal.

It is not always possible to eliminate nuisance parameters through the application of an invariance class. However, when it can be done, it is an ideal method for dealing with these parameters. Firstly,

it overcomes the need to estimate these parameters before testing, as is done in the GLRT formulation. This has the advantage that the added variance brought about by the estimation is avoided, resulting in more powerful tests. Secondly, invariant statistics often have tractable sampling distributions, on account of the nuisance parameters having been entirely eliminated. This is in contrast to the GLRT or similar formulations, which require a combined analysis of conditional densities and sampling distributions of the parameter estimates. Thus it is easier to specify the threshold needed to yield the desired properties in the test.

Formally, invariance is expressed in terms of a transformation group on the observation space. A group $\{\mathcal{G}\}$ of functions is proposed, each with domain and range being the observation space itself. The elements of this group are precisely the set of functions to which the data are assumed invariant. That is, if $\mathbf{x}$ is an observation vector and the vector function $\mathbf{f}$ is any element of $\{\mathcal{G}\}$, then the observations $\mathbf{x}$ and $\mathbf{f}(\mathbf{x})$ are equivalent under the transformation group. The requirement that $\{\mathcal{G}\}$ be a group is needed to ensure strict equivalence between sets of observations.

Because invariance implies equivalence between observations, subsequent decision-making and estimation need only be concerned with the set to which a given observation belongs. Thus each set may be indexed in some manner, and resulting densities derived for the probability of occurrence of each set. Only the set to which the observation belongs is of relevance in the subsequent calculations; the original observation is not important. In the context of statistics, the concept of a maximal invariant is useful for labelling these sets. A maximal invariant statistic is a statistic which takes on a unique value for each equivalence class, and as such is sufficient for the invariant detection or estimation.

In the literature, invariance is typically discussed in terms of group theory and the problem of group representations [55]. Interpreted in this way, invariance may be considered a formal method of including knowledge of the symmetry of a problem into the mathematical representation and subsequent search for solutions. For an invariant test to be optimal, however, the problem has to be *completely* symmetric with respect to the nuisance parameters. This is sometimes difficult to assert: again using the example of vector samples with constant unknown offset, the parameter $\alpha$ must be completely unknown before the maximal invariant statistic may be considered sufficient for the problem. However, this parameter is seldom entirely unknown: usually it is possible to impose limits on the values that it may take on, based on say the saturation values of the electronics from which the samples are obtained. Thus the optimal invariant test is strictly only optimal in a mathematical sense: optimality only holds insofar as assumptions are valid.

In this work, invariance is used in three ways. The first application is in the problem of detecting a known target with unknown location in additive noise. The target location is considered to be an unknown nuisance parameter, and an equivalence requirement is used to eliminate it from the problem. The reduced problem is then shown to admit an optimal test, which is therefore UMPI.

The other applications of invariance are not presented in an optimality framework, but attempt to

improve the performance of a detector by improving the match between the data and the models used. The first application in this regard relates to the subspace interference problem discussed earlier. However, instead of assuming the interference subspace known, as is required in the optimal test formulation, a candidate interference subspace is estimated from actual samples of data. The resulting estimate is used to derive an invariant detector for the problem. Because the interference subspace has been estimated, and because the subspace interference assumption may not be entirely appropriate, the resulting test can no longer be claimed optimal. Nevertheless, by including the assumption of a subspace interference component, it is sometimes possible to achieve better detection performance through more accurate modelling. That is, a noise plus subspace interference model may be more appropriate than a simple conditional probability model. This is particularly true for highly-constrained models, which may not have the capacity to accurately represent the data.

The noise plus subspace interference model is limited in that it can only accommodate certain types of deviation from the random component of the model. The final application of invariance addresses this issue. In particular, invariance is proposed as a general method of enhancing modelling accuracy. This is a unique invariance application, and does not appear to have been addressed elsewhere in the engineering literature. The assumption is made that the data are in conflict with some convenient model, but that this mismatch may be either partially or completely eliminated by placing an equivalence on sets of observations which are responsible for the inaccuracy. For example, invariance to portions of the data contained in particular linear subspaces of the observation space is proposed as a mechanism for ignoring data components which are in conflict with the convenient assumption of spatial stationarity. That is, by ignoring the parts of the data which constitute the causes of nonstationarity, the simple models can be made more appropriate.

The concept is applicable to models having tighter constraints than just stationarity, however. The validity of low-order autoregressive, moving average, or autoregressive moving average models can similarly be extended through application of a judicious invariance condition. Then it is not just the components of the data which violate the assumption of stationarity that are ignored, but also those components which violate the constrained model. Using this approach it is possible to make the simple standard models discussed in the previous section appropriate in instances where they would otherwise be invalid.

Although the concept has general applicability, it is presented in the context of invariance to linear subspaces. That is, the actual data are considered to be in conflict with a convenient assumed model only in terms of the values taken on in some linear subspace of the observation space. By ignoring this subspace through the application of an invariance requirement, better modelling accuracy is achieved. This improvement does come at the cost of reduced detectability, since portions of the target may have been eliminated through the invariance requirement. However the more accurate model specification can at least partially compensate for this effect.

When used in the context of enhancing model accuracy, the invariant tests have no claim to optimality. Nonetheless, their performance may be better than noninvariant tests for the same problem, which simply ignore the fact that the assumed models are inappropriate for the data to which they are being applied.

The development of invariant detectors quite naturally leads to a requirement for invariant estimators. This need arises because, even in the ideal case, the parameters of invariant detectors still have to be estimated from samples of real data. These samples are generally subject to the same interference or modelling mismatch that made the invariant test desirable. Thus the estimates should be invariant to the same transformations that are used in developing the invariant detector. For example, in the subspace interference problem the parameter estimates should not depend on the components of the observation which lie in the interference subspace, since these components do not conform to the assumed models.

Invariant estimation appears to have received very little attention in either the statistical or the signal processing literature. However, the need for such estimates in the problems discussed here is evident. In the framework of maximum likelihood estimation, the expectation-maximisation (EM) algorithm is ideally suited to the subspace invariant estimation problem. The EM algorithm is a procedure by which maximum likelihood estimates of parameters can be obtained in the presence of missing data. Since the data contained in the invariance subspace can simply be considered unavailable for estimation, the missing data interpretation is a natural one. Thus it can be used as a generic method of solution.

The issue of how to specify a constrained probability density model along with a suitable invariance subspace is a difficult one. Evidently as the dimension of the invariance subspace increases and more components of the data are ignored, a given distribution with a moderate degree of capacity should be able to model a given set of data with a higher degree of accuracy. On the other hand, there will be a corresponding decrease in the ultimate detectability on account of an increased likelihood that the target to be detected will lie in the invariance subspace. This work provides a number of methods for estimating invariance subspaces and probability models, mostly phrased in a maximum likelihood context.

## 1.6 Subspace invariance in adaptive detection

Adaptive detectors are important in applications where a single probability density cannot capture the complexity of the observed data samples. In this case additional parameters are introduced which describe a set of possible densities, with the implication being that any one member of this set is in force when the sample is generated. Under a likelihood formulation, the particular density in force is considered completely unknown. The observed data are then used to estimate the parameter value (either explicitly or implicitly), while at the same time testing for the presence of a possible target.

**Section 1.6: Subspace invariance in adaptive detection**

In most cases adaptive detectors are required when the noise density changes from one sample to the next, in a manner that cannot be easily characterised. It is then not reasonable to expect that a single simple probability model will accurately characterise the variability that occurs in the data. To some extent this is a result of poor modelling, although as discussed in earlier sections more accurate models may be impossible to develop and implement effectively. In particular, adaptivity becomes essential when models are simple but data samples are complex.

There are many paradigms for adaptive detection. One of the more convenient paradigms makes the assumption that apart from the data which are to be tested for target presence, additional samples are available. In the simplest case the noise in all the samples are assumed to be realisations of the *same* random process, with the same underlying parameter values in force. Conceptually, a reasonable approach is then to use the additional data to make an estimate of the noise density, and this estimate is used to resolve the details of the detector to be used for the decision process.

The assumption of additional data samples is useful, because it allows explicit relationships to be established between the amount of data used for the estimation and the power and accuracy of the resulting test. However, in many instances an adaptive formulation is required in cases where it is not appropriate to assume the availability of such additional data. The possibility then exists to use the *same* data that are to be tested for target presence to make noise parameter estimates. For this approach to be useful, very tight constraints are required on the assumed noise models. This is because the amount of data available for estimation is low.

Thus, depending on the amount of data which can be considered available for estimation, different degrees of constraints have to be imposed on the associated models. The presence of constraints however increases the risk of bias in the estimated probability densities. Therefore the invariance concepts outlined in the previous section become important for purposes of reducing model mismatch.

Particularly in the context of adaptive detectors, there are two different criteria which a detection statistic must meet. Firstly, it should exhibit good discrimination between the conditions of target absence and target presence. That is, the test statistic values resulting from just noise being present should differ as much as possible from the values obtained when there is a target present. This is referred to as the *detectability* criterion. However, there is an additional requirement that these sets of values be characterisable in an absolute sense. Thus when a target is absent, the possible values taken on by the test should be *predictable*, at least in a statistical sense.

In this context, constant false alarm rate (CFAR) tests are important. These are tests which have the property that the distribution of the test statistic under the noise-only assumption is independent of the unknown model parameters in force at the time. Then, when the test statistic is compared with a threshold value, the resulting test has a false alarm probability which is not a function of the noise parameters. In general the detection probability *is* however permitted to depend on these parameters.

If a CFAR test cannot be found, it is sometimes possible to adapt the test threshold according to the estimated noise parameters in such a way that an approximately CFAR test results. As the noise parameter estimates improve, the test becomes truly CFAR and may in some instances tend towards optimality. For this approach to be appropriate, the estimated noise density has to be very accurate. An invariance can be useful in two regards. Firstly, it can reduce the model mismatch. Secondly, it can enhance the validity of certain assumptions over greater data lengths, so that more data can be considered valid for parameter estimation. These factors can result in improvements in both the detectability and predictability of resulting detectors.

In this work, the subspace invariance methods outlined in the previous section are applied to various formulations of adaptive detectors. These invariances can be used to extend the validity of the simple models which are required in an adaptive formulation, on account of implementability requirements. By improving model accuracy, tests with better detectability and predictability are obtained.

## 1.7 Thesis outline and description

One of the primary topics of this thesis is the role of invariance in certain hypothesis testing problems. Invariance is applied in two contexts. The first regards the detection of targets with unknown location in a noise sample. In this case the appropriate invariance criterion is equivalence between observations which differ from one another by an unknown cyclic shift. The second invariance application is that of subspace invariance, where observations which differ from one another by some unknown additive component in the invariance subspace are considered equivalent.

The two types of invariance are very different, and are treated separately. Chapter 3 is therefore restricted to a discussion of detecting targets with unknown location in noise. The remainder of the work, Chapters 4 to 6, deals with the topic of invariance to subspace components.

Although the discussion of the unknown target location problem is almost equally important, the treatment of subspace invariant detectors constitutes the bulk of this dissertation. The presentation is broken down into three chapters, each dealing with a different aspect of the associated detection problem.

Chapter 4 provides a formulation of subspace invariant detectors, and provides some methods for calculating the required quantities when the noise parameters and invariance subspace are known. Methods are then presented for estimating candidate invariance subspaces under both the interpretations of low-rank subspace interference and low-rank model mismatch. These methods all require that the assumed model parameters be specified in advance, and use actual samples of data to arrive at the required estimates.

Chapter 5 discusses the dual problem, namely the estimation of noise parameters when the invariance

subspace is known. A missing data interpretation is proposed, and the EM algorithm is used to obtain subspace invariant estimates of the required parameters. Detailed methods are then presented for the subspace invariant estimation problem under various covariance matrix constraints. A partial solution is also provided for simultaneously estimating the noise parameters and the invariance subspace.

Since constraints are ultimately used to regularise parameter estimates based on limited data, Chapter 6 discusses subspace invariant detection in the context of adaptive detectors. Several detectors are discussed, and it is demonstrated in each case how an invariance subspace can be included in the formulation. Justification for doing so is also presented.

The following four sections provide a more detailed outline of what is presented in each chapter. The discussion is quite general, and is intended as an overview. Specific references and examples are provided in the detailed discussions in the relevant chapters.

### 1.7.1 Detection of targets with unknown location

The first form of invariance, discussed in Chapter 3, involves the use of invariance to cyclic permutations of the data in the problem of detecting a target with unknown location in noise. The second, discussed in the remaining chapters, presents the use of subspace invariance for increasing accuracy of highly-constrained data models.

The problem of detecting a target with unknown location or unknown arrival time is common [106, 38]. The GLRT formulation is often applied, and its performance is assessed by comparing it to an optimal detector designed for the known target location problem. Under certain assumptions the solution can be implemented in a sliding detection window context, where a fixed detection statistic is calculated for a data window centred on each possible target location. The decision of target presence is then made if any of the resulting values exceeds some test threshold.

This solution has some measure of intuitive appeal and often affords an efficient implementation. However, the procedure is not optimal, even if the test applied at each window location is itself optimal for the known target detection problem. This suboptimality arises because the target location is unknown, and effectively has to be estimated from the data before the final test statistic is calculated. Since the location estimate is subject to variation, a loss in performance results.

The fact that the target location is unknown is often ignored when tests are developed. These tests are then applied in a sliding window framework, with little justification provided for the validity of this extended procedure. Chapter 3 addresses this issue by formulating an optimal test for some specific cases of detecting targets with unknown location.

In particular, a test for detecting a known target with unknown location in zero-mean noise with a known circulant covariance matrix is derived. For the required symmetries to apply, an assumption is

required that the target wraps around the observation interval as it shifts within the observation. This test is uniformly most powerful in the class of tests which are invariant to cyclic shifts of the data values. As such, it can be considered the optimal test for the detection problem.

The conditions under which the test is optimal are fairly restrictive, although practical scenarios are easy to formulate where the development is entirely appropriate. Nonetheless, the true contribution of the optimal test is in assessing the extent to which the GLRT is suboptimal for the same problem. At least for a specific class of problems, this provides a definitive indication of how good the GLRT is for these types of problem. Since the GLRT is often related to the sliding detection window paradigm, this procedure can also be validated.

For the case of the noise distribution being known, it is demonstrated in Section 3.3 that the difference between the optimal detector and the GLRT is marginal. This result is obtained using statistical simulation for some simple cases, since the test statistics are in general not tractable.

The optimal test differs from the GLRT in that in the former an invariance criterion is applied which entirely removes the unknown target location parameter from the problem. This parameter therefore does not have to be estimated, resulting in reduced variability in the final detection statistic. The elimination of nuisance parameters is the primary reason for introducing invariance into hypothesis tests.

In Section 3.4 the solution to the unknown target location problem is extended to the case where the noise is still circulant, but may contain nuisance parameters which are deterministic but unknown. In this case the invariant testing formulation does not provide an optimal solution, in that no reasonable invariance class can be formulated which eliminates all the nuisance parameters. Nonetheless, at the very least the unknown target location parameter *can* be removed, by using the same invariance condition that was applied to the known noise case. The remaining unknown parameters can simply be handled by applying the GLRT to the modified problem.

In this case, the presence of the invariance constraint again improves performance by eliminating the location parameter. In Section 3.5 results are presented comparing the performance of the invariant test to the GLRT, this time for the case of white noise with unknown variance. The invariant detector developed for this problem requires invariant estimates of the remaining noise parameters after the application of the invariance condition.

### 1.7.2 Subspace invariance in detection

The second form of invariant hypothesis testing concerns the use of tests which are invariant to data components contained in a linear subspace of the original observation space. The concept appears to be most useful when the invariance subspace has low rank.

**Section 1.7: Thesis outline and description**

Subspace invariant hypothesis tests have been presented by Scharf [119], Scharf and Friedlander [120], and Behrens and Scharf [9]. In all cases discussed, the tests are developed for the problem of detecting targets in an environment of noise and subspace interference, where subspace interference is defined to be a completely unknown additive contribution which is required to lie in a reduced rank subspace of the observation space. As such, the subspace invariant detector is uniformly most powerful in the class of all tests which are invariant to the components of the observation lying in the interference subspace. The noise mean, covariance matrix, and interference subspace are assumed known, although some partially-known target formulations are permitted. Although interesting from a theoretical perspective, no results are provided which indicate the extent to which the ideas are useful in practice. It is not clear whether the subspace interference formalism is useful for modelling any real data, and no methods are provided for estimating either the noise parameters or interference subspace.

Chapter 4 begins with a preliminary discussion of the mathematical concepts required when dealing with subspace invariance. This is followed in Section 4.2 by a derivation of the invariant detector for a known target in noise which is multivariate normal and has an additive subspace interference component. It is demonstrated that the resulting test is optimal under the conditions described, and that the projection of the data out of the interference subspace constitutes a maximal invariant statistic for the detection problem. For reference, the noninvariant detector is also provided.

In certain real problems, a reasonable interference subspace to introduce into the modelling may be apparent. However, for the subspace invariant detector to be useful, a method is required of estimating candidate interference subspaces from actual samples of data. Section 4.3 addresses this estimation problem.

The discussion begins with a description of the effect that subspace interference can have on the sample mean and sample covariance of the corrupted data. A maximum likelihood formulation is then proposed for estimating the interference subspace from actual data samples, under the assumption that the distribution of the noise before the interference is added is known. The invariant likelihood function is presented in Section 4.3.2, and using this function a necessary condition on the maximum likelihood interference subspace estimate is obtained. The derivation uses results presented in Appendix 4.6.

Unfortunately, in general the necessary condition on the estimate does not appear to have a tractable closed-form solution. Nonetheless it can be solved exactly in some simple circumstances. These cases are discussed in Section 4.3.3. Specifically, under the assumption that the noise samples are independent and identically distributed, the ML interference subspace estimate is shown to be the span of the eigenvectors corresponding to the largest eigenvalues.

The solution to the white noise case is extended to provide an approximate solution for the general covariance case. This extension is the subject of Section 4.3.4. A whitening transformation is applied to the data and the interference subspace estimated in this transformed coordinate system. The required estimate is then obtained by transforming this estimate back into the original coordinates. The resulting

estimate no longer constitutes a maximum likelihood estimate, but in Section 4.3.6 it is demonstrated that the procedure has merit for at least one simulated case.

As discussed in earlier parts of this introduction, the use of subspace invariant detectors is not restricted to the problem of detection in noise and subspace interference. Rather, an invariance subspace can additionally be used to partially or completely eliminate model mismatch, if indeed this mismatch happens to lie in a restricted linear subspace of the observation space. Section 4.4 provides a general discussion of this use of invariance in testing, and in Section 4.5 a method is presented for actually estimating such mismatch dimensions from actual data samples. For this problem a form is assumed for the covariance matrix of the data, and the invariance subspace serves to minimise the difference between this covariance and the sample covariance actually observed.

Section 4.5.2 provides an example of how the general invariant detector formulation can be applied to a detection problem in real data. It is shown that the resulting detector is indeed better than the corresponding noninvariant detector, both in terms of overall detectability and predictability.

### 1.7.3 Subspace invariant covariance estimation

Chapter 5 discusses the dual problem to that of Chapter 4, namely that of noise parameter estimation when the invariance subspace is known. In particular, the topic of invariant covariance matrix estimation is presented, specifically for the case where constraints may be present.

It is argued that in general when there is an invariance subspace present, the noise parameter estimate should be made invariant to data contained within this subspace. This is because the estimation data are themselves subject to either subspace interference or low-rank model mismatch, and therefore have components which are invalid in the invariance subspace. If this factor is ignored, the estimates can become corrupted.

In Section 5.1 the need for subspace invariant parameter estimates is discussed for two invariant detector formulations. Under the assumption that secondary noise-only data samples are available, an outline of simple plug-in and GLRT solutions are presented for the problem of detecting a known target in noise and subspace interference. It is demonstrated that in both cases invariant maximum likelihood estimates of the noise parameters are required, which maximise the likelihood of the data in the subspace orthogonal to the invariance subspace.

When the covariance matrix of the noise is completely unknown and no constraints are enforced, it turns out that for specifying the invariant detector the parameter estimates are not explicitly required to be invariant to data in the invariance subspace. This is discussed in Section 5.2. However, this situation is unique in that the components which are incorrectly estimated due to the presence of the invariance subspace are in fact ignored in the invariant detector. In general this is not true.

**Section 1.7: Thesis outline and description**

Since the data in the invariance subspace should not be used, one way of addressing the estimation problem is to assume the corresponding portions of the observations missing, and use formal statistical principles for dealing with missing data. The EM algorithm is proposed as a fairly generic method of obtaining parameter estimates under these conditions. In Section 5.3 the EM algorithm is outlined, and a fairly general formulation is presented for parameter estimation in the presence of covariance constraints, when portions of the data are assumed unavailable or invalid for estimation. The expectation and maximisation stages are explicitly provided for this problem.

Section 5.4 uses this general formulation to derive an algorithm for finding the invariant maximum likelihood covariance matrix estimate from a sample of data, under the constraint that the matrix be circulant and the estimate be invariant to the data components contained in a known linear subspace. The section begins with some mathematical preliminaries required for the characterisation of circulant matrices, and the relevant expectation and maximisation stages are then presented for both the real and complex cases. Some simple examples are given of the invariant estimates as applied to samples of data.

The circulant constraint is interesting in itself, but is primarily useful in the context of invariant covariance matrix estimation under a Toeplitz constraint. In Section 5.5 this maximum likelihood estimation problem is discussed, first for the case of no invariance subspace, and then for the case of a known invariance subspace. Again examples are provided of the estimates obtained for some real samples of data.

Section 5.6 outlines the subspace invariant estimation problem for some alternative constraints. The case of estimation under doubly block circulant and doubly block Toeplitz covariance constraints are presented, which have application in the context of cyclostationary and stationary 2-D signals. The problem of invariant covariance matrix estimation under ARMA constraints is then discussed, and a method presented for obtaining these estimates.

Section 5.7 addresses the problem of simultaneous estimation of the noise covariance matrix and the invariance subspace. Unfortunately, because the interference subspace and invariance subspace estimates presented in Chapter 4 are not true maximum likelihood estimates, no ideal general solution has been found to this problem. Nevertheless, under the assumption that the noise obeys an ARMA model of known order, such simultaneous estimates can indeed be obtained using direct maximisation over a modified likelihood function. The solution obtained approximately maximises the likelihood function over both the invariance subspace and the covariance matrix. It is shown for some artificial situations that the results for this simultaneous estimation are good.

Finally, in Section 5.8 some practical examples are provided of the improvements that may be brought about by the use of an invariance subspace in some real detection scenarios. Section 5.8.1 models the data using a first-order autoregressive model, with the assumption that interference is present in the observations. It is demonstrated that by incorporating invariance to an estimated interference subspace,

both the detectability and the predictability of the resulting test are increased. Even more importantly, it is shown that the invariant test generalises better, providing considerably better results when applied to a set of testing data than the noninvariant test. Another example, presented in Section 5.8.2, shows that improved predictability can be expected from a test which includes an invariance subspace intended to reduce model mismatch.

### 1.7.4 Adaptive detection with invariance

The methods presented in the first two subspace invariance chapters are useful for applying nonadaptive detectors to data which are difficult to model. A single detection statistic is obtained which can be applied to a wide variety of situations, since it tends to ignore those aspects of the data which contribute most of the variability.

The notion of subspace invariance for enhancing model match is however most appropriate when the models are highly constrained. It is in these cases that the invariance subspace can extend the applicability of the models. A situation where constrained models are essential is in adaptive detection. Here it may be required to make parameter estimates based on relatively small amounts of data, and constraints are required to regularise these estimates.

The invariance subspace is a high-dimensional quantity, and a large number of samples are required to estimate it effectively. This property does not make it conducive to adaptivity. However, the covariance matrix estimate can certainly be made adaptive, resulting in an overall adaptive test statistic. An adaptive testing paradigm that seems appropriate in this case is therefore to estimate the invariance subspace off-line, using what is considered to be a representative sample of noise data. The invariance subspace is then assumed fixed and known, and subsequent detection can be made adaptive while all the time being invariant to data contained in the invariance subspace.

For the adaptive detection problem, the role of different components of the data becomes important if the processes involved are to be fully understood. Section 6.1 presents a decomposition of the detection problem into different subspaces, thus demonstrating how the observation components enter into the detection problem. In particular, it is shown that the portion of the data which contains only noise under each hypothesis contributes information regarding noise in the target-carrying subspace. This information reduces the uncertainty in the target subspace, resulting in improved detection. When unknown noise parameters are present, there is an additional mechanism by which the noise-only components can improve detection, namely by providing information regarding the value of the unknown noise parameters.

Section 6.2 presents a number of adaptive detectors, all of which assume that additional noise-only observations are available on which to base covariance matrix estimates. These detectors resemble certain tests discussed in the literature, but are reformulated to better model the particular types of

problem under consideration. The tests are shown to have a CFAR property when no constraints on the covariance matrix are imposed. Methods of incorporating invariance subspaces into these tests are also presented.

In Section 6.3 the specific case of adaptive detection with covariance constraints is presented, for the same set of detectors discussed previously. It is demonstrated that although the CFAR property is no longer strictly valid, it is still asymptotically appropriate. Thus as the parameter estimates improve, the resulting tests become CFAR. This conclusion continues to hold when an invariance subspace is introduced. The particular case of using the same data sample to simultaneously estimate noise parameters and perform detection is then discussed in detail, and the resulting tests are shown to have similar properties to those discussed previously.

Finally, in Section 6.4 a number of results are presented showing the performance of the various adaptive detectors. It is shown that the presence of a invariance subspace can improve the predictability of the detectors dramatically, usually at the cost of a decrease in overall detectability.

## 1.8 Notation and conventions

A brief description of the conventions used in this document is provided in this section.

The basic notation for numerical entities is as follows:

- A bold-font upper case character (Greek or Roman) denotes a matrix.
- A bold-font lower case character denotes a vector.
- A normal-font lower case character denotes a scalar.

When relationships exist between entities, the same character will generally be used to represent them (but perhaps with a change in weight or face). Additional relationships are indicated through use of superscripts and subscripts, using these conventions:

- A superscript in round brackets denotes elements of a sequence, particularly for purposes of iteration. Thus $\mathbf{S}^{(p)}$ denotes the $p$th iterate of the quantity $\mathbf{S}$.
- The transpose of the matrix $\mathbf{S}$ is denoted by $\mathbf{S}^T$, and the conjugate transpose by $\mathbf{S}^\dagger$.
- Integer arguments in round brackets are used to denote elements of an entity. Thus the $m \times n$

matrix $\mathbf{S}$ has elements $\mathbf{S}(j,k)$, arranged as follows:

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}(1,1) & \mathbf{S}(1,2) & \cdots & \mathbf{S}(1,n) \\ \mathbf{S}(2,1) & \mathbf{S}(2,2) & \cdots & \mathbf{S}(2,n) \\ \cdots & \cdots & \ddots & \cdots \\ \mathbf{S}(m,1) & \mathbf{S}(m,2) & \cdots & \mathbf{S}(m,n) \end{pmatrix}. \tag{1.6}$$

Similarly, the $m$-dimensional vector $\mathbf{s}$ is given by

$$\mathbf{s} = \begin{pmatrix} \mathbf{s}(1) \\ \mathbf{s}(2) \\ \vdots \\ \mathbf{s}(m) \end{pmatrix}. \tag{1.7}$$

- To be concise, this above form for matrices is often compromised in lieu of double subscripts on scalar values. Thus an alternative form for $\mathbf{S}$ is

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \ddots & \cdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{pmatrix}. \tag{1.8}$$

  This relationship may be denoted by the shorthand assertion $\mathbf{S} = (s_{ij})$.

- Single subscripts are used to specify the elements of an ordered set. For example, $\mathbf{s}_k$ denotes the $k$th element of a set of vectors, and $s_k$ the $k$th element of a set of scalars.

- It is often useful to consider the columns of a matrix to be a collection of vectors. Generally these vectors will be denoted by the same character representing the matrix (but in lower case), with a subscript providing the ordering. Thus $\mathbf{S}$ above is comprised of $n$ $m$-dimensional vectors such that

$$\mathbf{S} = \begin{pmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_n \end{pmatrix}. \tag{1.9}$$

- The linear subspace spanned by the columns of the matrix $\mathbf{S}$ is represented by $\langle \mathbf{S} \rangle$. In conjunction with the previous point it may also be denoted by $\langle \mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_n \rangle$.

- Any subscripts which are not scalar variables are simply used to augment the name-space of the symbol or draw attention to some property it possesses. For example, $\mathbf{S}_{\text{alt}}$ might be used to indicate a matrix which is related to (but different from) $\mathbf{S}$. In the same spirit, dimensions of matrices may be indicated by the subscripted form $\mathbf{S}_{m \times n}$, which indicates a $m \times n$ matrix.

In most cases the first element of any entity will be numbered starting from one. However, in some

instances it is more convenient to start from zero. In this case the fact will be pointed out explicitly.

Certain symbols are almost always reserved for entities of a particular form:

- Generally the symbols $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ are used to denote a diagonal matrix, where all the off-diagonal elements are zero.
- $\mathbf{I}$ denotes an identity matrix. A subscripted form may be used to emphasise dimension, particularly if the matrix is not square. $\mathbf{I}_{m\times n}$ therefore represents a $m \times n$ matrix of zeros with ones along the main diagonal.
- $\mathbf{U}$ denotes a matrix which has orthonormal columns, so $\mathbf{U}^\dagger\mathbf{U} = \mathbf{I}$. The subscripted form $\mathbf{U}_S$ has columns which span the same subspace as the columns of $\mathbf{S}$.
- $\mathbf{U}_I$ is used to denote a $p$-dimensional invariance subspace, and $\mathbf{U}_H$ the $(n - p)$-dimensional complementary subspace. These matrices are unitary, and have dimensions of $n \times p$ and $n \times (n - p)$ respectively.

The notation $\mathbf{x} : N[\mathbf{m}, \mathbf{C}]$ implies that the random vector $\mathbf{x}$ has a multivariate normal distribution, with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{C}$. The notation also carries over to random scalars, where appropriate.

Since it is virtually impossible to specify a system of notation which is consistent, concise, and versatile, occasional deviations from these outlines might be necessary. Additionally, in some cases specific requirements for notation each have to be dealt with on their own merit. In both cases the usage is described in the relevant portion of the text.

# Chapter 2

# Review of relevant literature

This chapter discusses some of the literature relevant to the issues addressed in this dissertation. The majority comes from the field of statistical signal processing, although the classical statistical literature is also largely applicable.

Section 2.1 deals with general literature that has proved useful in researching this work. Relevant subjects range from general statistics through to more specialised topics such as time-series analysis and spectrum estimation.

Sections 2.2 and 2.3 provide a fairly detailed overview of literature where the detection of targets in array environments are considered. This is quite a specialised application, in that the assumption can be made of additional data samples, which share the distribution of the noise in the data to be tested for target presence. From a theoretical perspective it is of immense value, since it provides a framework in which adaptive detectors can be developed and assessed.

Section 2.4 discusses some of the literature which concentrates on the importance of predictability in detection performance. In particular, constant false alarm rate detectors are discussed.

In Section 2.5 the specific topic of invariance in hypothesis testing is reviewed. Although vast quantities of literature make use of this concept, very few of them are specifically dedicated towards it. However, a few papers which *do* concentrate on this aspect are presented.

Section 2.6 discusses some topics of importance which do not fall directly into the categories of detection papers discussed in previous sections. In particular, the problem of simultaneous detection and estimation from the same sample of data is reviewed. This is followed by some more general formulations which permit less detailed descriptions of the targets to be detected.

Finally, Section 2.7 presents some of the available and relevant literature on the subject of constrained covariance matrices and how to estimate them. Discussion centres around the cases of both Toeplitz

matrices and covariance matrices associated with general ARMA processes.

## 2.1 General literature

Statistical signal processing and detection theory as an engineering discipline has been growing rapidly since around 1970. Probably the most seminal work in this regard is the book by Van Trees [126], although Helstrom [59] and others had written about many of the ideas before the publication of this book.

The early literature deals primarily with detection and estimation in a continuous-time context. Commonly, solutions then involve expressing the observed signals in terms of an infinite set of discrete coefficients, which can be dealt with in an approximate fashion as necessary. In modern practice there appears to be little or no advantage in developing methods for continuous signals, particularly with digital implementations: signals are almost always discretised, and can be worked with as such. Furthermore, the continuous-time solutions lend no further insights into the structure and nature of solutions to problems. Nevertheless, sometimes continuous-time methods are presented, particularly in the older literature, and it may be necessary to translate them to discrete time. In practice this seldom presents any difficulty, and the conversion is generally simple. In fact, the discrete formulations are usually simpler than their continuous-time counterparts, which sometimes require knowledge of fairly advanced applied mathematical concepts.

Statistical signal processing has grown out of the field of statistics, and as a subject has evolved quite far from its origins. Nonetheless, the statistical literature is still particularly relevant for more advanced results, and should not be ignored. Statistics as a subject is highly developed, and one often finds that certain very specific applied problems have been dealt with at length by researchers in the field.

From the point of view of general statistics, the book by De Groot [25] is clear and useful. Elementary multivariate statistical methods are nicely presented by Morrison [98], and for more advanced results the book by Muirhead [99] is comprehensive. The collection of works by Kendall and Stewart [81] is indispensable in almost all aspects of statistics. Also, Papoulis [103] provides a treatment of probability and statistics which is useful in practice.

Books on general statistics and applied statistics are useful up to a point, but as far as detection theory is concerned are quite limited. Lehmann [87] presents a comprehensive account of advanced statistical decision theoretic principles in a parametric context, with emphasis placed on the Neyman-Pearson criterion as an optimality condition. Ferguson [34] has similar content matter, although Bayesian detection formulations are more dominant. Berger [10] also provides a very readable text which is specifically directed towards detection theory.

Not all statistical formulations are useful for practical detection problems in engineering. For example,

many principles in statistics have been primarily developed under conditions of approximate independence, either between observations or between samples within observations. For signal processing problems this emphasis is often inappropriate. An exception may however be found in the subject of time-series analysis, which specifically addresses the problem of characterising random signals with significant nonzero correlation between samples. The most valuable reference for models of this type is the book by Box, Jenkins, and Reinsel [15], which has had numerous revisions since its first edition. More accessible for non-specialists are the works by Brockwell and Davis [17, 18], and Fuller [41]. Additionally, although geared towards econometrics, the book by Judge et al. [69] is comprehensive. Most time-series analysis problems concentrate on situations where available data are severely limited, and a stationarity assumption is in almost all cases assumed.

The statistical literature is useful, and as has been mentioned may at times prove immensely invaluable if the problem under investigation has indeed been treated by statisticians. However, there *is* a difference between the types of problem looked at by statisticians and by engineers, and aspects of a solution which have relevance to the one group may be insignificant to the other. The engineering literature is valuable in the sense of being geared towards providing methods for solving real-world problems, particularly of the types encountered in radar and image processing.

Whenever a statistical problem formulation is being used, at some point the process has to be characterised. Under the assumption of stationarity, this implies a form of spectrum estimation. The historical article by Robinson [117] is interesting in this regard. More useful for practical application is the review article by Kay and Marple [78], which discusses in detail many fairly advanced concepts. It does, however, preceed the more recent findings on multi-window and multi-taper spectrum estimation [125], which explicitly address many of the more arbitrary aspects of spectrum estimation which have caused much concern. The book by Kay [75] elaborates on many concepts discussed in [78], and the algorithms provided for time-series parameter estimation are useful, even if not optimal in a likelihood context. More recently, books have become available which address time-series models in an engineering framework, such as the work by Porat [104]. Finally, the review paper by Makoul [90] on linear prediction principles is useful and enlightening.

From the point of view of detection theory in an engineering context, the book by Scharf [119] has a good balance between the different formulations and philosophies. It appears to draw heavily on work by Lehmann [87] and Ferguson [34]. Kazakos and Papantoni-Kazakos [79] also provide a good reference for general applied detection problems, particularly in instances where optimal tests cannot be found or robust formulations are desirable. Kay [77] provides a general treatment of statistical signal processing techniques, although almost exclusively confined to estimation rather than detection problems.

The fields of nonparametric (or semi-parametric) detection and pattern recognition cannot be ignored in hypothesis testing problems. For providing a link between parametric and semi-parametric statistical

methods, the book by Ripley [115] is unequalled. Bishop [11] is similarly indispensable for neural network modelling and training. For more recent developments such as support vector machines, Haykin [58] provides a coherent introduction. The book by Fukunaga [40] is useful from the point of view of parametric pattern recognition and clustering. McLachlan [93] provides a similarly useful account, more from the perspective of formal statistics.

When working in the field of multivariate statistics, a good set of references regarding matrix algebra and calculus are indispensable. This is particularly true when constraints are incorporated into the discussion. Most useful in this regard from a theoretical point of view is the book by Basilevsky [6]. Also useful for practical implementation is the work by Golub and Van Loan [47]. On a slightly less advanced level, Searle [123] presents a number of useful results, which are aimed specifically at solutions to mathematical problems which commonly arise in statistics.

## 2.2 Basic parametric detection with multiple observations

There is a large body of literature available on the subject of detecting a deterministic target in known or partially-known noise when multiple observations are available. Typically the additional samples are assumed to have been obtained from different sensors, each sampled simultaneously in time. In its simplest form, this problem assumes that all the observations are governed by a common underlying random process, with possible target presence manifesting itself in just one of the observations. The noise-only observations can then be used to estimate the statistical properties of the noise, which are incorporated into an hypothesis test on the potential target-carrying data.

Within this basic framework, there are a number of variations which can be permitted. The target and the covariance matrix may be assumed fully or only partially known, the additional noise samples may be assumed independent or correlated with one another, and the target may be assumed to be present in either one or many of the observations. With such a rich set of possible problem definitions, it is hardly surprising that much literature has been generated on the subject.

The assumption that multiple data samples are available is met, at least partially, in quite a few applications. Perhaps the most simple case is where a scene or signal is captured using different imaging modalities, or receivers which are sensitive to different frequency bands. In that case, multiple observations are obtained which are in perfect registration, and which are often closely related to one another. If target presence is essentially restricted to one band, then the conditions described in the introductory paragraph are at least partially met.

Most of the literature on detection using multiple observations, however, is presented in the context of array processing [57, 84]. In this application it is assumed that there are a large number of receiver elements, usually with a known and fixed relative spatial orientation. Readings are simultaneously

obtained from each of these receivers, resulting in multiple observations which are correlated on account of the proximity of the sensors.

The array processing context is particularly amenable to the placement of constraints on the received signals, and is therefore of particular interest in this work. For example, when a plane wave passes over a sensor array, there are strong relationships between the signals received at each sensor. These relationships can be conveniently and accurately accounted for by means of constraints between the random processes used to model each sensor.

Under an assumption of short-term stationarity, where the statistical properties of the process generating the signal are assumed to be at most slowly-varying, the multiple-observation assumption can also be appropriate. In this case if targets do not occur frequently and are localised in time or space, then the portion of the observed signal adjacent to the location being tested for target presence may be assumed to contain only noise. In this sense, the neighbouring samples may conceptually be considered to constitute additional noise-only data. Such an approach is amenable to the inclusion of constraints into the problem, if only to enforce the stationarity assumption.

The problem of detection using multiple observations is important from a theoretical point of view, because it directly addresses the issue of how to incorporate estimated parameters into hypothesis testing formulations. It is apparent from the literature that the plug-in classifier philosophy [115, p. 28], where estimates of unknown parameters are plugged into a detection statistic designed for known parameters, does not in general yield the best solution. Rather, the detection statistic should be derived for the problem as a whole, using a model which simultaneously describes the entire set of observations.

Another important aspect of this particular problem is that it is closely related to the problem of adaptive detection. The additional target-free observations can be used to adapt the effective detection statistic applied to the data to be tested. This is particularly apparent when considered in connection with the quasi-stationarity assumption discussed earlier.

In the remainder of this section, some of the more fundamental papers on this topic are discussed. In all cases, it is assumed that there are multiple observations available. Where appropriate, using the terminology of Kelly [80], if target presence may only occur in one of these vectors then this vector will be referred to as the primary data. The additional noise-only vectors then constitute the secondary data. The section following this one continues with the discussion of detection with multiple observations, but presents more advanced topics and formulations.

In [111], Reed, Mallett, and Brennan provide the first really useful result on this topic by quantifying the loss in detectability brought about by using a plug-in classifier formulation for the problem of detecting a completely known target in unknown noise. The target, if present, is assumed to occur in the primary data, while the remaining secondary data vectors are assumed to contain independent but identically distributed samples of MVN random noise. The noise covariance is estimated directly using

the sample covariance of the secondary data, and plugged into the optimal detection statistic for the primary data. This optimal detection statistic is derived based on the assumption that the covariance is known. Because the resulting test statistic is a function of the inverse of this sample covariance, they refer to the technique as sample matrix inversion (SMI).

The most important result of this paper is that it provides a quantitative relationship between the number of noise samples used for covariance matrix estimation, and the ultimate discrimination ability of the plug-in detection statistic. To arrive at this result a signal-to-noise ratio (SNR) measure is presented which quantifies the fractional loss in discrimination ability brought about by the detection statistic being based on the estimated covariance matrix, rather than the true one. The expected value of this loss is independent of the true noise covariance. Their results, however, give no indication of the actual detectability that can be achieved using the estimate. Khatri and Radhakrishna Rao [82] address this issue by deriving a probabilistic lower bound on the actual attained detectability. They also consider the problem of detection using the estimated covariance matrix when several alternative targets may be present in the data.

Kelly [80] extends the results of [111] by solving the problem instead using the GLRT, and introducing an unknown multiplicative scaling factor into the target description. Again it is assumed that there is a set of independent secondary noise vectors available which share the noise characteristics of the primary vector on which detection is to be performed. Kelly makes the observation that the SMI method of Reed, Mallett, and Brennan suffers from a problem in that a test threshold cannot be meaningfully set, since it would depend on the unknown covariance matrix. The GLRT, on the other hand, has a distribution (under both hypotheses) which is independent of the actual noise covariance matrix. The distribution of the test statistic under $H_0$ is therefore completely known, so the test has a CFAR property.

Fuhrmann [39] provides the first indication of the effect that appropriate constraints can have on improving detection performance for this detection problem. More specifically, he imposes a Toeplitz constraint on the covariance matrix, corresponding to the assumption of stationarity. Simulation results are provided demonstrating the improvement in detectability loss ratio (as defined and used in [111]) and test performance (when using the AMF test [116]) brought about by this constraint, both for the case where the stationary condition is in fact valid. The improvement is significant, and without loss of performance effectively allows the number of secondary data vectors to be reduced in proportion to the ratio of the number of free parameters in each case. The results all assume that the secondary data samples are independent of one another, and of the primary data vector.

Whereas Reed, Mallett, and Brennan [111] assumed the target to be detected completely known, Robey et al. [116] assume it known only to within some unknown scaling. They derive the GLRT for this problem under the assumption that the noise covariance matrix is known, and plug in the sample covariance of the secondary data to yield the resulting detection statistic. They call this test the adaptive

matched filter (AMF). It is demonstrated that the AMF has a CFAR property, and that in some instances it may outperform Kelly's GLRT for the same problem.

In [80], Kelly asserted that the GLRT solution is preferable to the SMI technique [111] because it provides a test which has a CFAR property. The result of Robey et al. [116], just described, suggests that the CFAR property comes about not because of the use of the GLRT, but rather as a consequence of including an unknown scale factor into the target specification. It is the estimate of this parameter that provides the necessary normalisation for the test to be CFAR. Whether this property can in general justify the inclusion of the unknown parameter into models of situations where the target amplitude is in fact known is difficult to answer, and is addressed in more detail in Section 6.2.

The papers discussed up to this point are mostly conceptual. Those which follow address some of the more practical aspects of the multiple-observation detection problem.

Margalit, Reed, and Gagliardi [92] initially assume the noise covariance matrix known, and discuss in detail the optimal LRT for detecting a known target in the case where multiple noise observations are available. If the additional samples are independent of the noise in the primary data then they cannot aid in the detection. However, the assumption is made that the different observations have a known and nonzero cross-correlation, and in this case detection performance can indeed be improved. Methods are then presented for estimating the relevant covariance and cross-covariance parameters from the data, which are included into the detection statistic using a plug-in approach.

The assumption of a nonzero correlation between the noise observations makes the analysis well-suited to detection in multiband scenarios, where such correlations may be expected to occur. However, the assumption that the noise covariance is known means that the secondary data only aid in test performance by reducing the effective noise variance in the primary data. That is, the noise in the primary data conditional on having observed the secondary data has lower variance than if the secondary data had not been observed at all. It is demonstrated in Chapter 6 that for the problem of detection where the covariance matrix contains unknown parameters, this reduced conditional variance is only one mechanism by which the secondary data can improve performance. Thus the assumption that the covariance is known limits the utility of the results somewhat.

Also in [92], some interesting observations are made regarding the properties of many signals that occur in practice. It is demonstrated that images can quite successfully be modelled as realisations of a Gaussian process with a rapidly space-varying mean and a slowly-varying covariance. Furthermore, the covariance matrix is often approximately diagonal. Using this property, the detector proposed for the general case is modified for use under these assumptions. Specifically, an estimate of the image mean is subtracted from each image, and the covariance and cross-covariance of the residual modelled as specific scalings of the identity matrix.

Chen and Reed [22] discuss the same problem as [92], except that an unknown scaling factor is included

into the target magnitude. Again the assumption is made that after removal of the mean, the covariance matrix of the process being modelled is some scaling of the identity matrix. A difference in this paper, however, is in the assumption that the cross-covariance matrix between the observations is an unknown deterministic parameter. The GLRT formalism is used to derive a detector which is shown to have a CFAR property, with the distribution of the test statistic under the null hypothesis being only a function of the number and dimension of the observations, not of the unknown parameters. Once again, the assumption that the additional observations are correlated with the primary data makes the method suited to detection in multiband images, where independence between the bands cannot be assumed.

Reed and Yu [112], acknowledging that target presence is seldom limited to one spectral band, extend the test in [22] to allow for target contributions in the secondary data observations. The target is assumed to have the same profile in each image scene, but the amplitude in each case is modelled as a different deterministic unknown. The GLRT is derived for the resulting problem, and it is shown to be a CFAR test. The similarity between the specific problem and multivariate analysis of variance (MANOVA) in the statistics literature is also highlighted.

The results of Margalit, Reed, and Gagliardi [92], Chen and Reed [22], and Reed and Yu [112] all rely on the assumption discussed earlier, namely that after subtracting an estimate of the mean from the data, the covariance of the residual is approximately diagonal. Also, in these papers methods are presented which make use of overlapping data samples to obtain covariance matrix estimates. These assumptions and processing mechanisms require the assumption of some measure of statistical stationarity.

## 2.3 Advanced parametric detection with multiple observations

The previous section presented some results regarding the detection problem when multiple observations are available. Some papers concerning issues related to practical uses of the detectors were also discussed. In this section papers dealing with more advanced or specialised instances of the problem are presented. Some of these generalise the detection formulations, while others provide results pertaining to optimality of the resulting tests.

El Ayadi [32] uses a plug-in classifier formulation for the multiple-observation detection problem, for the case where the noise samples corresponding to each observation are independent of one another. The covariance of the noise within each observation is assumed unknown. The target may be present in all of the observed data vectors, but the form in which it appears is assumed known. Instead of using a maximum likelihood estimate for the noise covariance, however, an estimate is developed which is as far as possible independent of target presence in the data. This estimate uses components of the observations which are known to contain only noise under each hypothesis, and is therefore referred to as a noise alone reference (NAR) covariance estimator. The resulting test is called the NAR test. This testing paradigm is then extended to the case where only the linear subspace containing the target

is known, and a corresponding NAR estimate derived.

The covariance estimates used in [32] are based on a generalised class of covariance matrices, comprised of a weighted sum of the outer products of each set of observed data vectors. This is convenient from the point of view of NAR estimator selection, but is somewhat arbitrary. It is shown in Chapters 3 and 5 (in two different contexts) that estimates with the required properties can be obtained using an invariant maximum likelihood estimation criterion, which places the estimate on similar ground to that used in the GLRT. Furthermore, through an appropriate formulation it is sometimes possible to impose additional constraints on these estimates, while still affording a tractable method of solution.

Kirsteins and Tufts [83] discuss the problem of detecting a target in noise plus low-rank interference. Again multiple observations are assumed to be available. Under the assumption that the noise and target energies are weak, the interference subspace is taken to be the principal components of all the observed data. The resulting detector simply steers the observations out of the estimated interference subspace, and makes a decision based on the residual. Using a GLRT formulation, the solution is extended to the strong target case: the target present in each data snapshot is modelled as an unknown linear combination of a known set of basis vectors, and the noise covariance is taken to be of the form $\mathbf{Q} + \sigma^2\mathbf{I}$. Here $\mathbf{Q}$ is a low-rank matrix corresponding to the interference, and $\mathbf{Q}$ and $\sigma^2$ are assumed unknown. The GLRT is derived for the problem under the assumption that the rank of $\mathbf{Q}$ is fixed.

The problem of detecting a target with unknown scaling in noise with covariance $\mathbf{R} = \mathbf{\Psi}\mathbf{B}\mathbf{\Psi}^\dagger + \lambda\mathbf{R}_0$ is discussed by Bose and Steinhardt [12]. Here $\mathbf{\Psi}$ and $\mathbf{R}_0$ are both assumed known. This corresponds to the assumption that the interference subspace is known, but the noise covariance within this subspace is unknown. An invariance transformation group is defined, and a 4-D maximal invariant statistic found. The specific case of $\mathbf{R}$ being unstructured is then discussed, and it is shown that under the proposed transformation group the AMF test statistic of Robey et al. and the GLRT statistic of Kelly together constitute a maximal invariant. Since a UMPI test does not exist for the problem, a locally most powerful invariant (LMPI) test is proposed, which is optimal in the limit of zero SNR. The test requires a time-consuming matrix operation in its implementation. Gau and Reed [44] provide a method whereby this computational complexity can be reduced for the case where the interference is large in comparison to the thermal noise.

For the case where independent secondary target-free vectors are available, Raghavan, Qiu, and McLaughlan [110] address the problem of detecting a completely unknown target in zero-mean noise with unknown correlation properties. They propose that invariance to arbitrary nonsingular linear transformations is a meaningful condition to impose on the test for this problem, and demonstrate that a uniformly most powerful test exists under this restriction. This test has a CFAR property.

In [13], Bose and Steinhardt discuss the case of detecting an unknown rank-one target for the case where it is permitted to appear in each of the multiple observations. Within all the observations, the target is assumed to be of a specific form. Apart from this constraint, however, the target is assumed

unknown. An invariance argument is applied, and it is shown that the GLRT shares the required invariances. The resulting test has a CFAR property.

Burgess and Van Veen [20] generalise many detection scenarios with multiple observations are available by considering the GLRT for a generalised multivariate analysis of variance (GMANOVA) problem, where it is required to test between $\mathbf{B} = \mathbf{0}$ and $\mathbf{B} \neq \mathbf{0}$ in the model $\text{vec}(\mathbf{X}) : N[\text{vec}(\mathbf{ABC}), \mathbf{R}_x \otimes \mathbf{S}^*]$. Here $\mathbf{R}_x$ is unknown and $\otimes$ is the matrix Kronecker product operator. This is similar to the model proposed in [13], but without the restriction on the rank of the target. The model can be shown to subsume many of the multiple-observation detection problems, including those where targets are only partially known and targets which may appear in each of the multiple observations.

A particularly significant aspect of the analysis in [20] is that, for situations where there are a large number of unknown parameters, the data in a well-chosen subspace can provide much better detection performance than if all the data are used. The reason for this improvement is a reduction of the number of free parameters, especially important when the training or estimation data are limited. Subspace projection mappings are therefore advocated prior to estimation and detection, which reduces the number of free parameters. In short, the reduced detectability brought about by the projection can be countered by a gain in statistical stability, which tends to increase performance.

Another way of obtaining this increase in stability is through the introduction of constraints. Because of the tradeoff discussed, there is an optimum balance between the number of free parameters and the amount of estimation data. In this thesis the constraint approach is emphasised.

Many of the papers discussed use an interference subspace as part of the model. For example, Kirsteins and Tufts [83] discuss detection in white noise with low-rank subspace interference, and Bose and Steinhardt [12] consider a similar problem with a more general covariance matrix. However, in most cases at least the interference subspace or the covariance matrix is assumed known, and no indication is provided of how to estimate the unknown quantities. In Chapter 5 a method of simultaneous estimation is presented, at least for certain classes of covariances. This estimation is made possible by phrasing the problem in a maximum likelihood context.

In the papers discussed, the term “optimal” is often used quite liberally with regard to the derived hypothesis tests. As always, any claim of optimality has to be clarified by a precise description of the conditions under which the property holds. For the problem of detection in an array environment, the arrival time of a target is invariably unknown. In most cases, however, the solutions and optimality claims are presented in the context of targets with known arrival times. The implication is then that the derived procedure is applied at each time interval that target presence may occur. The hypothesis tests derived for the known arrival time problem are however in general *not* optimal for the case where the arrival time is unknown.

## 2.4 Predictability of detectors

It has been mentioned in the introduction that it is important for a detector to exhibit performance which is close to the design performance. For this property to be achieved, a number of conditions have to be met. Firstly, the criterion by which the detector is designed and selected should suit the practical requirements for the problem. This was discussed in the introduction, where it was demonstrated that the Neyman-Pearson formulation is more appropriate in certain instances than the Bayesian detector.

Secondly, the performance characteristics should ideally be independent of any unknown noise parameters or target parameters present in the formulation. If the test characteristics depend on parameters which are unknown and vary from one observation to the next, the detection and false alarm rates will also vary. If the false alarm rate does *not* depend on the unknown parameters, then the test is said to have a constant false alarm rate (CFAR) property. Such tests have long been considered important in the radar and sonar literature.

Thirdly, it is desirable that the performance characteristics of the test be robust to deviations of the data from the assumed probability density. This draws on the topic of robust statistical formulations in hypothesis testing, which is increasingly gaining attention in recent literature. Unfortunately many of the useful formulations in this regard are related to the Bayesian testing paradigm, where tests are made robust to errors in the specification of the prior densities.

This section presents some of the literature related to CFAR detection, as well as less formal adaptive normalisation methods that have been used to provide tests with predictable performance characteristics. Much of the literature related to CFAR testing has already been presented in the previous two sections.

The problem of robust detection is not explicitly addressed in this work, although it is demonstrated by example in Section 5.8 that the notion of subspace invariance can aid in this regard. A good review article of robust methods in signal processing problems is provided in [72].

Some of the early treatments of CFAR tests in applied detection are due to Nitzberg [100], Dillard [29], Gupta et al. [53], and Rickard and Dillard [114]. These constitute just a small portion of fairly dated literature regarding the topic of adaptive thresholds in testing. In all instances the problem is that of detecting a very short duration target in a potentially nonstationary noise and clutter environment. The target presence is assumed to affect the value of only a single sample data point.

The basic detector structure proposed is that of a preprocessing stage (often a square-law transformation), followed by a tapped delay line. The output of the tap corresponding to the point to be tested is then compared with an adaptive threshold derived from the output of the remaining taps. If this threshold is exceeded, a detection is registered. The problem, then, is to decide how to calculate the required adaptive threshold value from the various tap outputs.

In particular, Nitzberg [100] discusses an adaptive threshold which is based on the assumption that the logarithm of the noise and clutter power is a polynomial function of certain tap outputs. The order of this polynomial is assumed known, but the coefficients unknown. Dillard [29], using a different preprocessing stage, considers a simpler case where the samples neighbouring the one to be tested are simply averaged to find an appropriate test threshold. The resulting test is analysed in terms of its performance in stationary white noise with different variances. Gupta et al. [53] modify the detector structure slightly, and allow for longer duration targets which may span more than one tap. The successive noise samples are assumed independent, with a covariance matrix that is known except for an overall unknown scaling factor. An adaptive test is then derived which has the approximate CFAR property. Finally, Rickard and Dillard [114] build on the detector presented in [29], considering the implications on detectability caused by one or more interfering target returns within the set of taps used to make the mean level estimate. A remedy is suggested and analysed which uses ranking and censoring techniques to make an adaptive detector which is more robust to interfering targets.

These papers present some of the very basic principles of applied CFAR testing. In some respects, all that is required to generalise the results to targets which may span several samples is a rotation in the observation space.

As mentioned, many of the papers discussed in the previous sections discuss CFAR detectors in the context of array processing. For example, the GLRT of Kelly [80] and the AMF of Robey et al. [116] achieve a CFAR property by modelling the amplitude of the target as an unknown deterministic parameter. A similar situation arises in [22], where the test of Margalit, Reed, and Gagliardi [92] is made CFAR by a similar amplitude parameter introduction. Raghavan, Qiu, and McLaughlan [110] develop a CFAR test for detecting unknown targets in noise with unknown covariance. Gau and Reed [44] develop a CFAR test for a known target with unknown scaling in white noise and subspace interference.

Constant false alarm rate tests have also been discussed specifically in the context of detection in clutter. For example, in [61] a CFAR detector is presented for a spatially correlated background environment, and in [52] CFAR detection in clutter with a lognormal distribution is considered. These methods are however presented in a particular radar context, and the techniques used do not extend well to the type of detection problem treated in this work. They are therefore not discussed further.

Another interesting use of the NAR estimates discussed in [32] is presented by El Ayadi and Picinbono [33], where the problem of detecting a known target with unknown scaling in noise with known correlation but unknown power is addressed. A minimum variance unbiased estimate of the noise power which is invariant to the presence of target is derived, and it is shown that it is equivalent to estimating the power from the noise-only subspace. When used in a test this property has implications on predictability, mainly because the noise estimate is accurate in both the cases of target absent and target present.

The papers outlined represent just a small fraction of the work that has been done on the development of predictable detectors. They are interesting either because the detectors discussed are intended to operate under a wide variety of conditions, or because they explicitly use invariance concepts in their formulation. Further references can be found in [60].

## 2.5 Invariance in detection problems

One of the main topics of this work is the use of invariance in general hypothesis testing, either in a cyclic permutation context or a linear subspace context. The topic of invariance has received much attention in the signal processing literature. Nevertheless, it has its roots in the theory of group representations [55], which is a branch of applied mathematics which deals formally with problems relating to equivalence classes within data.

Invariance arguments have been used extensively as a method of justifying optimality of certain classes of detectors. In a statistical context, the books by Ferguson [34] and Lehmann [87] provide a comprehensive treatment of this aspect of hypothesis testing. In short, if an optimal test does not exist for a given problem, then it is sometimes reasonable to restrict the class of possible tests which are considered, and attempt to find an optimal test within this smaller class. The requirement that tests be invariant to certain transformation groups represents one possible restriction.

Scharf and Lytle [121] provide one of the first discussions of invariant hypothesis testing in the engineering literature. Specifically, they discuss the problem of detecting a known target in Gaussian noise which has a known covariance structure, but with an unknown overall energy. They propose a reasonable invariance class for this problem, and find the uniformly most powerful detector within this class. Since the test is invariant to the unknown noise power, the distribution of the test statistic under $H_0$ is independent of this parameter. Thus the test is CFAR.

These results are generalised in subsequent literature, particularly the book by Scharf [119]. Here a number of problems are considered where invariance arguments can be used to derive tests and establish optimality. For example, invariance to subspace interference is used to develop and justify tests when this type of interference is present in the observations. Similarly, justification is provided for invariance to subspace-constrained rotations in certain problems. In addition, cases of targets completely known and only partially known are addressed. The book is essential reading as far as invariance in applied hypothesis testing is concerned.

The paper by Scharf and Friedlander [120] (along with corrections in [118]) addresses the optimality properties of the GLRT for a similar class of problems to those presented in [119]. In particular, the detection of a class of subspace-restricted targets in a background of known or partially-known coloured noise and possible low-rank subspace interference is considered. In each case the GLRT is

derived, and shown to be optimal within the class of detectors which are invariant to the same sets of transformations to which it is itself invariant. This procedure establishes the optimality of the GLRT for the problems. As an aside, the results presented in [120] assume that the signal subspace and the invariance subspace are linearly independent, although they are not required to be orthogonal. Behrens and Scharf [9] demonstrate that the required detectors can be conveniently interpreted using oblique projection operators, yielding further insight into the relationship between the relevant subspaces.

Another interesting application of invariance is presented by Raghavan, Qiu and McLaughlin [110]. For the problem of detecting a completely unknown target in noise with a MVN distribution but an unknown covariance matrix, they argue that invariance to arbitrary nonsingular linear transformations is reasonable. The results assume the presence of secondary noise-only observations. The test is developed further by Raghavan, Pulsone, and McLaughlin [109], who show that the result only holds when the unknown target is completely unspecified. In particular, if a subspace constraint is placed on the target then the test is no longer optimal. An interesting and important analysis of the relationships between the components of the vectors lying in relevant subspaces is also provided.

Uses of invariance in the literature are growing steadily, particularly as a means of justifying optimality of certain tests. There are however other applications of invariance in practical problems. Lenz [88], for example, discusses group theory as applied to pattern recognition and image processing.

## 2.6 Other detection literature

Detection in an array environment constitutes the largest portion of the literature relevant to the topics addressed in this thesis. However, there are certain issues which appear to be completely ignored in those discussions. This section presents additional ideas in detection theory which are important. In particular, papers which discuss the problem of detection when only a single observation is present are discussed, as well as some more general treatments of detection when targets are partially unknown.

Bowyer, Rajasekaran, and Gebhart [14] make the distinction between white thermal noise and clutter in an additive background signal, and discuss the problem of adaptive filtering for clutter rejection. The clutter is modelled as an autoregressive random process, and an adaptive prewhitening Kalman filtering approach is used for the parameter estimation. (More specifically, Kay [73] and Gingras and Masry [46] discuss the subject of autoregressive spectral estimation in additive noise, providing some references and background to difficulties that arise.) They then use a simple matched filter for white noise as the detector component.

Kay [74] places the approach taken in [14] on firmer statistical ground by addressing the problem using the GLRT. He assumes the noise to be autoregressive with known order, and the target to be detected completely known. Approximate maximum likelihood estimates of the noise are then made

under each hypothesis, yielding an approximation to the required test. It is demonstrated that as the data records become large, the test approaches the optimal matched filter for known parameter values.

This previous paper is significant in that it formalises the process of detection in unknown noise, for the case where only a single data sample is observed. By using the autoregressive assumption, the need for multiple samples of correlated data is overcome. Thus the same sample of data is used for estimating the unknown noise parameters and detecting the target. The resulting test however does not have a constant false alarm rate.

For the testing problem just described, but with the target amplitude assumed unknown but deterministic, Baggenstoss and Kay [5] demonstrate that the GLRT is asymptotically CFAR and optimal. They then use the Rao test to derive an alternative detector which is asymptotically equivalent to the GLRT, but which is simpler to calculate. Performance results are provided which show that the loss in using this modified test is negligible for a large class of problems.

The use of the GLRT in signal processing is not restricted to the assumption that the noise distributions be Gaussian. In an abstract context, Kay [76] demonstrates that for the problem of detecting a known target with unknown amplitude in independent and identically distributed noise samples, the GLRT is asymptotically optimal whenever the noise probability density function is symmetric. In [77], Kay and Sengupta apply this result to the problem of detecting a target with unknown amplitude in noise of a fairly general form, again with an assumption that components of the density function governing this process be symmetric. The Rao test is developed as a reduced complexity solution for small target amplitudes, and some computer simulation results provided. Other extensions in a non-Gaussian context are presented in [124].

In a different vein, there are a number of papers which deal with the detection of certain types of transient targets in various types of assumed noise. The fact that the targets are transient represents quite a lot of information, and these papers attempt to consider detection under these circumstances while including as little additional information regarding the target as possible. The methods are not as powerful as those which assume the target profile known, but exhibit greater tolerance to mismatch in the target model and are applicable to a larger variety of problems.

Porat and Friedlander [105] discuss the problem of detecting an unknown target in white noise with known variance. It is assumed that the target can be modelled as the impulse response of a rational transfer function filter with unknown coefficients. A modified GLRT is proposed, which avoids the problems associated with singularities in the MLE by effectively substituting more convenient maximum a posteriori (MAP) estimates.

Friedlander and Porat [35] consider the role of data-reducing time-frequency and time-scale transforms in transient target detection. Initially assuming that the target lies in a known low-rank linear subspace (but is otherwise unknown), they analyse the use of the short-time Fourier transform (STFT) and the

wavelet transform for purposes of detection in white noise. The GLRT is derived for the detection problem, and its distribution calculated. They acknowledge that targets seldom lie precisely in a known subspace, and using a robustness condition discuss the implications to the detection of a modelling error in the target specification. The analysis is then extended to the case where the target subspace is parameterised by an unknown vector parameter.

The analyses of [35] are extended in two subsequent papers by the same authors. In [106] the case of the target subspace being completely known is discussed in more detail, with specific reference to a framework in which different levels of unknown parameters can be included into the detection problem. On the other hand, in [107] the case of the target subspace being parameterised by an unknown is discussed in more detail, and it is demonstrated that this parameter is not estimable under $H_0$. This invalidates one of the conditions required for the GLRT to be asymptotically optimal.

Frisch and Messer [36] discuss the detection of a partially-known target in noise, where the relative bandwidth and the time-bandwidth products are assumed known. They demonstrate that the wavelet transform provides a useful representation, in that the GLRT is easily calculated. It is demonstrated that the performance is better than for the case where the target is assumed completely unknown.

In [37], Frisch and Messer define a transient target to be one which is localised in some vector representation. They use this definition to abstractly discuss the GLRT for detecting a deterministic partially-known target in known noise for the three cases of arrival time and duration being known, arrival time and duration being unknown, and duration being known but arrival time unknown. It is demonstrated that as more prior information regarding the nonzero coefficients is included, the power of the GLRT increases.

El Ayadi [31] uses the GLRT formalism to derive a detector for a known target deformed by an unknown linear transformation in white noise with unknown power. As in [32] the use of a noise-alone reference estimate of the power is proposed, which is independent of the presence of a target in the signal.

The general detection literature is vast, and the references provided in the last few sections can in no way be considered complete. The review article edited by Hero [60] gives a more comprehensive list of references, particularly in the context of statistical signal processing.

## 2.7 Constrained covariance modelling and estimation

The assumption of spatial or temporal stationarity leads to covariance matrices which have a Toeplitz-type structure. For 1-D signals, the covariance matrices are Toeplitz, with equal elements along diagonals parallel to the main diagonal [65, p. 25]. For stationary images, the structure of the covariance matrix is doubly block Toeplitz. That is, the matrix is comprised of Toeplitz blocks which are themselves arranged in a Toeplitz structure.

Toeplitz matrices are of importance in this work because they represent the least restrictive constraint which is applied to the data in the problems discussed. The constraint that the covariance matrix have this structure is required to enforce statistical stationarity of the associated model. However, tighter constraints within the class of stationary processes still have the property that their covariance matrices are Toeplitz.

The Toeplitz matrix structure can be exploited in many instances, allowing reduced-complexity solutions to be obtained for problems in which they appear. Durbin's algorithm for the solution of the Yule-Walker equations in linear prediction is one such instance [48, p. 194], and provides a general method of solving linear systems of equations with a Toeplitz coefficient matrix. Because of their structure, Toeplitz matrices can also in general be inverted more simply than unstructured matrices. Trench's algorithm is a classical method of inverting these matrices [135]. The extent of the improvement brought about by these methods is a reduction in complexity from $O(n^3)$ to $O(n^2)$, where the matrix is $n \times n$ [48, p. 195].

Detailed mathematical aspects of Toeplitz matrices are provided by Widom [131], and Grenander and Szegö [51]. A useful property of these matrices is that they are closely related to Szegö polynomials [68, 67], which can be used to conveniently obtain solutions to certain types of problem. Using the relationship between Toeplitz matrices and these polynomials, Justice [70] presents an algorithm for the inversion of positive definite matrices with this structure. In [71] he extends this result to the inversion of doubly block Toeplitz matrices, by using the bivariate analogues to the Szegö polynomials. Other references that discuss the inversion of block Toeplitz and doubly block Toeplitz matrices are [2] and [130].

Another aspect of Toeplitz-structured covariance matrices that is of interest is methods for calculating their eigenvalues and eigenvectors. Many efforts have been made to characterise these quantities for Toeplitz [50, 91, 64] and doubly block Toeplitz matrices [127]. Iterative methods have also been presented for actually performing either a full or partial decomposition [8, 101].

Maximum likelihood estimation of covariance matrices under a Toeplitz assumption is quite complex. Because there are a large number of free parameters, simple search strategies are unreliable on account of the presence of local maxima in the search space. Burg, Luenberger, and Wenger [19] presented the first detailed solution to the estimation problem, finding conditions on the gradient of the likelihood function which are required for an extremum. They then presented an algorithm, termed inverse iteration, for solving the required gradient equations.

Subsequently the expectation-maximisation (EM) algorithm of Dempster, Laird, and Rubin [28] has become the preferred method of solving this constrained optimisation problem. Initially proposed by Miller and Snyder [96], it involves embedding the Toeplitz covariance estimation problem into a higher-order problem, where the constraint is converted to one of the covariance matrix being circulant. They prove that the sequence of Toeplitz matrices obtained using this technique increases in likelihood.

**Section 2.7: Constrained covariance modelling and estimation**

The convergence properties of the algorithm are only partially known, however. This topic is discussed further in Chapter 5, where additional references are provided. For a readable introduction to the EM algorithm, the reader is referred to [97].

Tighter constraints than just statistical stationarity are applicable in this work. In particular, autoregressive moving average models are of interest. There are a number of approximate maximum likelihood ARMA parameter estimation schemes available in the literature, but surprisingly few for obtaining exact MLEs. Part of the reason for this is because ML estimates of ARMA parameters are not efficient in the sense that they do not attain the Cramer-Rao bound [15]. For most applications approximate maximum likelihood estimates therefore suffice, and are considerably easier to calculate [75, 104].

For strict application in the GLRT, true maximum likelihood estimates are required. Also, in this work extensive use is made of the EM algorithm, which requires MLEs to guarantee an increasing likelihood value as the iterations progress. This motivated the choice of exact likelihood methods.

With regard to maximum likelihood estimation, a useful paper is that by Ansley [3], which provides an efficient method of calculating the exact likelihood of a sample from an ARMA process. This is required in any search for the maximum likelihood estimate. An algorithm for the implementation of the method, which uses an innovations decomposition to calculate the required likelihood, is provided by Mélard [94]. In this work the optimisation is simply performed by inserting the routine into a Nelder-Mead simplex procedure [108]. If necessary, better speed and performance could probably be obtained using results presented by Burshtein [21], which allows both the likelihood and the likelihood gradient to be calculated.

# Chapter 3

# Detection of targets with unknown location

For some detection problems, the location of a candidate target in a sequence of observations may be known. In data communication across a synchronous channel, for example, the boundaries between successive symbols are precisely known: the start of the symbol may be triggered through a separate mechanism, or more commonly the source and receiver are synchronised and kept that way by a phase locked loop. One therefore knows exactly when a symbol is required to occur. It is then possible to gather a sequence of data which represents precisely one of a set of symbols, and to trivially apply a hypothesis test to it. In many respects the statistical formulation is perfect for this problem, and results in a powerful detector.

Maintaining the communications application, the problem in the preceeding paragraph may be contrasted with the detection of asynchronous transient transmissions of particular symbols. An interplanetary space probe, as a simple example, might transmit a predetermined signal back to an earth receiver under some error condition. Since it is unknown when an error might occur, the receiver has to consider every time interval to potentially be carrying the known target transmission. In that sense, the location of the target transmission in the sequence of samples has to be considered unknown, and the test developed accordingly.

The synchronous detection problem is in many respects contrived. Thus it will generally occur in situations which have actively been designed to be synchronous. Most naturally occurring problems seldom exhibit such synchronisation, and unknown location parameters need to be introduced in the modelling. The problem of detecting targets with unknown location is therefore an important one.

Quite often the location parameter may not be entirely unknown. For example, in detecting instances of ships in a satellite image or aerial photograph, the set of possible locations can be constrained

to correspond to sea or water. Within this constraint, however, every possible location effectively needs to be tested for the presence of a ship. There may also be prior information on the location parameter (ships are more likely to be found in shipping lanes), although this information can prove to be misleading in some applications (ships carrying contraband will perhaps *not* be using shipping lanes).

In practice, from a computational perspective all data are discretised. If the sampling frequency is sufficiently greater than the dominant frequencies present in the target, then a small shift in the location of the target will not significantly alter the sampled signal. Thus the set of candidate target locations may almost always be considered finite. This suggests that detection may be performed by applying some appropriate operation at each possible location, and perhaps for each possible target. Such approaches have led to a collection of ad hoc detectors for dealing with the problem of detection with unknown location.

For example, the use of the magnitude of the Fourier transform as a shift-invariant statistic is well known, and one sometimes sees it used in shift-invariant detection. However, because this statistic is not maximal, it is invariant to more than just translations. In fact, Hayes [56] has shown for the discrete case that, even if the equivalence class includes circular shifts, time reversal, and change of sign of the sequence, the statistic is not maximal. The result is suboptimality of any detector based on this quantity.

Section 3.1 begins with a discussion of the generalised likelihood ratio test for the problem of detecting targets with unknown location in noise with a known multivariate normal (MVN) probability distribution. Two different types of shifts are considered, namely linear and circular, corresponding to different problem formulations.

Section 3.2 then presents the uniformly most powerful invariant (UMPI) test for the unknown circular shift case, in the event that the noise is MVN with a circulant covariance matrix. An invariance group is proposed for the problem, and is shown to be appropriate. A maximal invariant statistic is found, its distribution under each hypothesis calculated, and the resulting optimal invariant test derived. It is demonstrated that under an identity covariance matrix assumption this test can be simplified, and a method of prewhitening the more general problem is presented.

The GLRT and the UMPI tests for the detection problem are then compared using simulations. Results are presented for detecting three different targets, each with three different energies, in white noise. It is seen that the UMPI is uniformly better than the GLRT, as it indeed must be. However, the difference is slight, lending justification for the use of the conventional GLRT solution to the problem.

Where the earlier sections assume the noise covariance matrix known, Section 3.4 discusses the problem of detecting targets with unknown location in noise which is only partially known. The invariant formulation still requires the assumption that the covariance matrix be circulant. It is argued that for

use in the invariant test, the estimates of the unknown noise parameters are required to be invariant to the same group of transformations as the detection statistic, and an invariant maximum likelihood estimation procedure is proposed for finding such estimates. The specific case of the noise covariance matrix being some unknown scaling of the identity matrix is discussed in detail.

Section 3.5 presents results comparing the performance of two detectors for this problem of a target with unknown location. The noise is assumed to be white, with unknown variance. The conventional GLRT solution, which assumes both the location and variance parameters unknown, is compared with the invariant GLRT which only assumes the noise parameter unknown. It is demonstrated using simulations that the invariant test is again superior to the GLRT, this time by a more significant margin.

Section 3.6 concludes with some remarks concerning the implications of the results to actual detection problems.

Finally, to resolve any confusion, it should be noted that the term "location" is often used differently in the statistics literature to the way it is used here. In particular, for real 1-D random variables, "location" describes the position of the probability density along the real line. For example, if $x : N[m, \sigma^2]$, then $m$ is referred to as a location parameter.

## 3.1 GLRT formulation for circular and linear shifts

In this section, the generalised likelihood ratio test (GLRT) is presented for the problem of detecting a known target with unknown location in MVN noise with known mean and covariance. Some ambiguity arises in the specification of the unknown location, however — it is not apparent how targets which lie at the edges of the observation interval should be dealt with. One possibility is to assume that targets may translate completely out of the interval; an alternative is to assume that they wrap around at the edges.

The two sections which follow deal with the GLRT for each of these instances in turn. It is demonstrated that, apart from a change in the definition of the candidate targets, the two cases may be identically formulated and lead to very similar tests.

### 3.1.1 Formulation for linear shifts

In principle, a sequence of measured real data may often be considered to be continuous and extend infinitely in both directions in time or space. For purposes of computation it is necessary to sample these data and restrict the observation interval. This restriction causes complications in the specification of the detection problem, in that a target may be present which is not entirely included in the observation interval. This is an artifact of the sampling process, and arises in the mapping of the actual problem to

a useful representation. Probably the most reasonable way of dealing with this problem is to simply test for all possible instances of targets as they appear in the sampled data. Thus targets which appear at the edges of the observation interval are truncated, and are therefore less detectable. This truncation partially reduces the symmetry of the problem.

Figure 3.1 depicts instances of a target at five possible locations in the sampled noise-free data. These

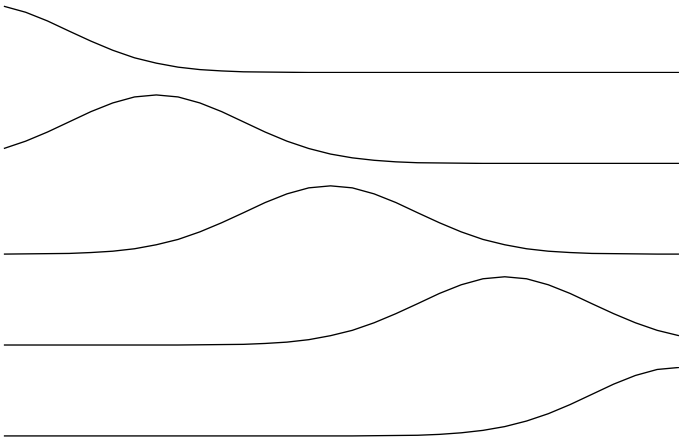


Figure 3.1: Instances of five possible targets (Gaussian bumps) which may occur in the observed data interval. The targets differ in terms of their location, but targets near the edges are truncated.

targets are said to be related by a linear shift in their location, which just means that they can translate out of the observation interval. Each of these instances have to be considered possible by the detector.

For convenience and clarity, only the detection of 1-D targets in signals is considered, although the extension to target detection in images is in principle trivial. Proceeding with the GLRT formulation for this particular unknown location problem, let the prototype target to be detected be represented by the column vector $\mathbf{s}$. As discussed previously, with sufficiently close sampling only a finite number of possible instances of targets need be considered in the observation interval. Usually the candidate locations are then taken to be the sample locations themselves, although apart from convenience there is no need to require this. Assuming that $k$ possible target-present situations may occur, the target instances may be represented by the elements of the finite set $\{\mathbf{s}_1, \ldots, \mathbf{s}_k\}$. Each element $\mathbf{s}_i$ is a linearly-shifted version of the prototype target $\mathbf{s}$, where the shift is such that vacated elements of the vector are replaced with zeros. The example targets in Figure 3.1 will all be elements of this set.

If the noise is assumed to have a multivariate normal distribution and the target and the noise are assumed additive, then the GLRT may be written as

$$t_{\mathrm{GLRT}}(\mathbf{x}) = \ln \frac{\max_{i\in[1,k]}\, p_i(\mathbf{x})}{p_0(\mathbf{x})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{3.1}$$

where $\mathbf{x}$ are the observed data, $p_0(\mathbf{x})$ is the distribution of $\mathbf{x}$ under $H_0$, and $p_i(\mathbf{x})$ is the distribution under $H_1$ with the target $\mathbf{s}_i$ present. If the noise is $N[\mathbf{m}, \mathbf{C}]$, this test is equivalent to

$$t_{\text{GLRT}}(\mathbf{x}) = \max_{i \in [1,k]} \left\{ \mathbf{s}_i^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{m}) - 1/2 \mathbf{s}_i^T \mathbf{C}^{-1} \mathbf{s}_i \right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{3.2}$$

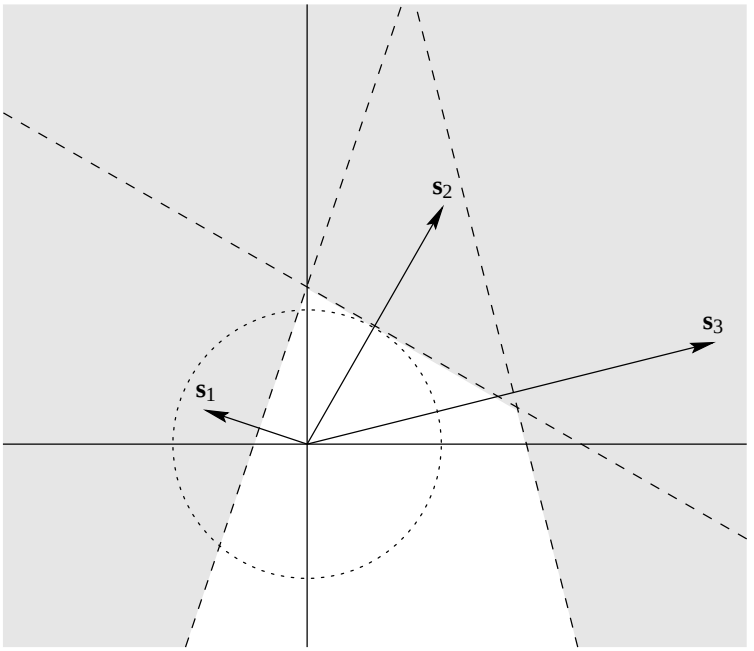
Thus the GLRT is obtained by calculating $\mathbf{s}_i^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})$, comparing with the threshold $\eta + 1/2 \mathbf{s}_i^T \mathbf{C}^{-1} \mathbf{s}_i$, and deciding $H_1$ if exceeded for any $i$.

The GLRT has one free parameter, $\eta$, which may be chosen to provide certain desired test characteristics. Because there is only one degree of freedom, the extent to which the test can be tuned to have any required properties is limited. A problem with including truncated targets into the target-present hypothesis is that the decision regions in the resulting test tend to become asymmetric with respect to the decision boundaries, particularly when targets have radically different energies.

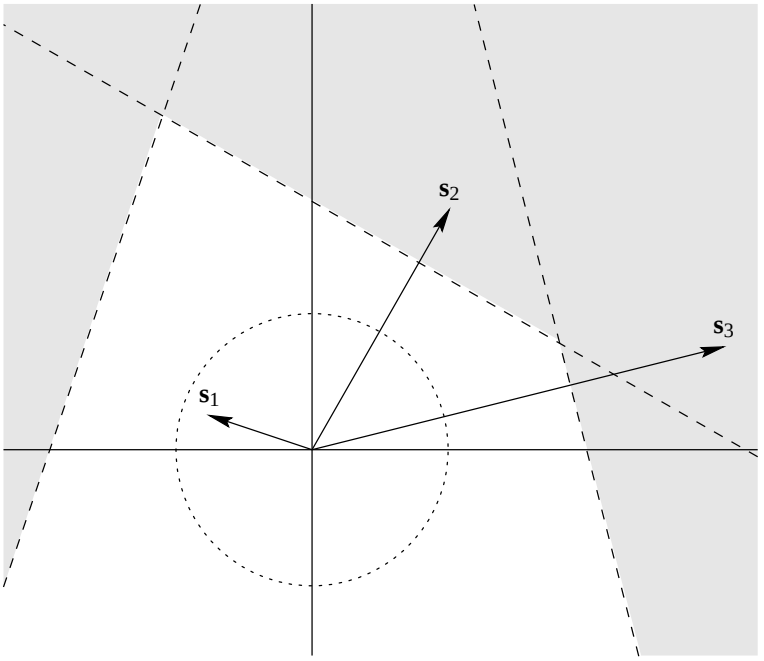
Figure 3.2 demonstrates conceptually the effects of such asymmetry on the GLRT decision regions for the simple case of detecting one of three targets in a two-dimensional observation vector. The targets are chosen to have energies which differ considerably from one another. The first diagram corresponds to the choice $\eta = 0$, with the noise distribution $N[\mathbf{0}, \mathbf{I}]$. The second diagram corresponds to some fixed positive value chosen for $\eta$, under the same noise condition. It is evident that the choice of $\eta$ has little effect on the portion of the decision boundary corresponding to the high-energy target $\mathbf{s}_3$. This is in contrast to the boundary related to $\mathbf{s}_1$, which is highly dependent on $\eta$.

Because the possible targets differ so significantly, one cannot hope to find a test which has performance which is independent of the unknown location parameter — the low energy targets are inherently less detectable than the higher energy ones. More significantly, perhaps, the marked differences in the targets also tends to indicate that no uniformly most powerful (UMP) test will exist for the problem. That is, it will not be possible to find a single test which can be said to be optimal in any respect for all possible values of the unknown location parameter.

If the preceeding GLRT formulation does not produce a reasonable test for any choice of $\eta$, the only recourse (aside from deviating from the GLRT formulation) is to modify the set of candidate targets to be tested for. For the problem of detection of transient targets with unknown location, it may be reasonable to include only those targets which are completely contained in the observation interval. At the very least, this has the effect of making the energies of all candidate targets equal, resulting in greater similarity between the locations of the component decision boundaries with respect to changes in $\eta$. At the same time, however, it does introduce a mismatch between the assumed distributions and the observed samples in cases where targets are in fact only partially present in the observation interval.




(a) Case of $\eta = 0$.




(b) Case of $\eta > 0$.

Figure 3.2: Decision regions of GLRT for detection of three different targets in zero-mean white noise. Each figure corresponds to a different value of the threshold parameter $\eta$.

### 3.1.2 Formulation for circular shifts

In the previous section it was argued that by grouping targets which are fundamentally different from one another into the target-present hypothesis, the chances of there being a UMP test are reduced. The symmetry in the problem specification for the detection of targets with unknown location is restored if they are considered to wrap around the edges of the observation interval, rather than to shift out of it. A signal which is shifted in such a manner that it wraps around the edges of the interval is said to be *cyclically* (or *circularly*) *shifted* or *permuted*. Figure 3.3 depicts instances of four possible targets corresponding to different locations in the sampled data, under the assumption of such circular shifts.

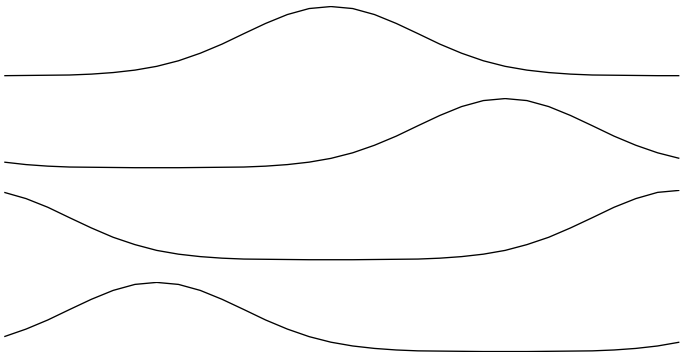


Figure 3.3: Instances of four possible targets (Gaussian bumps) which may occur in the observed data interval. The targets differ in terms of their location, and are assumed to wrap around the observation interval.

From a mathematical perspective, a circular shift is quite natural for discrete signals on a finite interval. This stems from the shifting property of the discrete Fourier transform: it is well-known that a circular shift in the discrete-time domain can be effected by multiplication in the discrete Fourier transform (DFT) domain [102]. The DFT can therefore be used for efficient implementation of circular convolutions in discrete time. The relationship is analogous to linear shifts and linear convolutions with the continuous Fourier transform.

For some problems the assumption that shifts are circular is completely valid from a modelling perspective. This is particularly true of man-made signals [42, 133]. For example, under some failure condition a piece of equipment may be designed to send out a constantly repeating signal with a fixed and known period and predetermined waveform. A detector designed to recognise this condition will not know the phase of the transmitted signal, which has to be considered an unknown. If the signal is observed for an integer number of periods, however, an integer number of waveforms will be received under the failure assumption. The target waveform is therefore known *exactly* to within an arbitrary cyclic shift. Many other scenarios can be formulated where cyclic permutations are the natural method of shift for a signal.

**Section 3.1: GLRT formulation for circular and linear shifts**

In this section the GLRT is derived for the detection problem where the target is known only to within some cyclic permutation. Again the assumption is made that the noise is additive and multivariate normal. In principle the test is exactly the same as for the linear shift case, with the set of candidate targets taken to be cyclically-permuted versions of a prototype target. However, the equations can be simplified somewhat by introducing notation which is specific to the cyclic case.

Continuing with the formal development, if the prototype target is the $n$-dimensional vector $\mathbf{s}$ then it is assumed that the possible targets which may be present are $\mathbf{s}_0, \ldots, \mathbf{s}_{n-1}$, where

$$\mathbf{s}_\theta = \mathbf{P}^\theta \mathbf{s}, \tag{3.3}$$

and $\mathbf{P}$ is the cyclic permutation matrix, with

$$\mathbf{P} = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{3.4}$$

This matrix has the property that premultiplying the column vector $\mathbf{s}$ by $\mathbf{P}$ cyclically permutes the elements one position downwards:

$$\mathbf{Ps} = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} s_1 \\ \vdots \\ s_{n-1} \\ s_n \end{pmatrix} = \begin{pmatrix} s_n \\ s_1 \\ \vdots \\ s_{n-1} \end{pmatrix}. \tag{3.5}$$

Therefore premultiplying by the matrix $\mathbf{P}^\theta$ will result in a circular shift downwards by $\theta$ positions. Note that this transformation is reversible, so the matrix $\mathbf{P}$ is invertible. Thus it makes sense to consider the matrix $\mathbf{P}^\theta$ for $\theta$ negative, which will just effect a cyclic shift of a premultiplied vector by $|\theta|$ positions upwards. Furthermore, if $\mathbf{s}$ is $n$-dimensional then the vector $\mathbf{P}^n\mathbf{s}$, which has been cyclically permuted $n$ times, will be identical to $\mathbf{s}$.

If it is required to detect instances of such cyclically-permuted targets in MVN noise with mean $\mathbf{m}$ and covariance $\mathbf{C}$, the GLRT is essentially the same as that of the previous section. However, there are now considered to be $n$ possible targets, corresponding to elements of the set $\{\mathbf{P}^0\mathbf{s}, \ldots, \mathbf{P}^{n-1}\mathbf{s}\}$. Although cyclic shifts of greater than $n$ positions are possible, they do not result in any new targets. Therefore the GLRT can be written as

$$t_{\text{GLRT}}(\mathbf{x}) = \max_{\theta \in [0,n-1]} \left\{ \mathbf{s}^T \mathbf{P}^{-\theta} \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m}) - 1/2 \mathbf{s}^T \mathbf{P}^{-\theta} \mathbf{C}^{-1} \mathbf{P}^\theta \mathbf{s} \right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{3.6}$$

where use has been made of the property that $(\mathbf{P}^k)^T = \mathbf{P}^{n-k} = \mathbf{P}^{-k}$. In the event that $\mathbf{C}$ is circulant, a

further simplification can be made:

$$t_{\mathrm{GLRT}}(\mathbf{x}) = \max_{\theta\in[0,n-1]} \left\{ (\mathbf{P}^{\theta}\mathbf{s})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}) - 1/2\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s} \right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{3.7}$$

This stems from the result that for $\mathbf{C}$ persymmetric (which defines symmetry about the northeast-southwest diagonal, and is true for both Toeplitz and circulant matrices), the relationship $\mathbf{P}^k\mathbf{C}\mathbf{P}^{-k} = \mathbf{C}$ holds for all $k$ [23].

Finally, a special case often of interest is that of detection in white noise, where $\mathbf{C} = \sigma^2\mathbf{I}$. Under these conditions the test becomes

$$t_{\mathrm{GLRT}}(\mathbf{x}) = \frac{1}{\sigma^2} \max_{\theta\in[0,n-1]} \left\{ (\mathbf{P}^{\theta}\mathbf{s})^T (\mathbf{x}-\mathbf{m}) - \frac{1}{2}\mathbf{s}^T\mathbf{s} \right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{3.8}$$

The first term on the left of the test inequality is simply the inner product between the observation $\mathbf{x}-\mathbf{m}$ and some cyclic permutation of the target $\mathbf{s}$. Furthermore, it is apparent that the statistic $\max_{\theta\in[0,n-1]}(\mathbf{P}^{\theta}\mathbf{s})^T(\mathbf{x}-\mathbf{m})$ is sufficient for the decision process.

### 3.1.3 Discussion

In both tests presented, selecting the value of $\eta$ based on required detection or false alarm rate criteria generally has to be performed by simulation or by calculations performed on actual data samples, since the distribution of the test statistic $t_{\mathrm{GLRT}}(\mathbf{x})$ does not have a simple closed form. This is true even in the case of detection in white noise with known variance. The same observation applies in the task of assessing test performance.

The two tests presented represent different assumptions on the candidate targets which may be considered present in the observation. Clearly when circular shifts are assumed, the problem formulation no longer accurately represents a situation where continuous data are sampled and windowed. Nonetheless, it can be argued that the differences between the linear and circular shift case are small, particularly when the observation interval is long and the targets are transient. The differences in the problem specifications then only occur near the edges of the data interval.

Irrespective of the precise formulation used, a question remains of how good the GLRT solution is to the unknown target location detection problem. One way of addressing this issue is to compare the performance of the GLRT to a simple detector which assumes that the target location is precisely known. This is sometimes referred to as a clairvoyant detector. The comparison then has to be made for each possible value of the location parameter, since in general the power of the test will differ for targets at different locations. For all values of $\theta$ the detector which "knows" the location of the possible target will outperform the GLRT, since it is the optimal solution for what then amounts to a simple hypothesis versus a simple alternative. The comparison is somewhat unreasonable, however,

since the problem of detection when the target location is known is fundamentally different from that of the location being unknown.

To make a reasonable assessment of the performance of the GLRT, therefore, an alternative formulation is required. In the section which follows, it is demonstrated that when the noise has a circulant covariance matrix (of which white noise is a special case), a uniformly most powerful invariant (UMPI) test exists for the problem of detecting a target known only to within some arbitrary cyclic permutation. This test is optimal in the Neyman-Pearson sense out of the class of *all* tests which are cyclic permutation invariant. Since the GLRT derived for the same problem is also invariant to cyclic permutations, its performance is necessarily worse than that of the UMPI test for *every* value of the unknown location parameter. By comparing the performance of the two tests it is therefore possible to ascertain the degree to which the GLRT is suboptimal.

Using this approach the tests may only be compared quantitatively under the assumption that shifts are circular, and for certain values of the noise covariance matrix. Nevertheless, if the GLRT performs well under these particular conditions, it may reasonably be expected that it will perform well under most conditions with shifts defined differently. This conclusion becomes more valid as transient targets become shorter and observation intervals longer. In essence, insofar as the problems of unknown linear shifts and unknown cyclic permutations of the target are similar, a comparison under these restricted conditions can provide an indication of the performance for the related cases.

## 3.2 UMP cyclic permutation invariant detection

This section presents a derivation of the uniformly most powerful invariant test for detecting a target which has an arbitrary and unknown cyclic permutation. An invariance condition is imposed on the testing problem to eliminate the dependence on the unknown shift parameter. The required invariance that is imposed on the hypothesis test is reasonable for the structure of the problem, so the resulting optimality is strong. The application of cyclic permutation invariance to the unknown target location problem appears to be unique, so the derivation is presented in detail.

As with the presentation for the GLRT, it is assumed that $n$ samples of data are observed, represented by the vector $\mathbf{x}$. Furthermore, it is assumed that any targets which may be present are observed in zero-mean additive noise with a known circulant covariance matrix — a more general case of the covariance matrix having certain unknown parameters is discussed in Section 3.4. The noise component $\mathbf{n}$ is therefore assumed to be distributed according to $N[\mathbf{0}, \mathbf{C}]$, with $\mathbf{C}$ circulant. Under hypothesis $H_0$ the distribution of $\mathbf{x}$ is $N[\mathbf{0}, \mathbf{C}]$. Under hypothesis $H_1$ some cyclically-permuted version of the prototype target signal $\mathbf{s}$ is added to the noise samples, where the order of the permutation is considered completely unknown. Thus under $H_1$ the distribution of $\mathbf{x}$ is $N[\mathbf{P}^\theta\mathbf{s}, \mathbf{C}]$ for some $\theta$. The shift parameter $\theta$ is an unknown deterministic quantity, and without loss of generality can be restricted to take on any integer

value from 0 to $n-1$.

A UMP test does not exist for the problem described here. If such a test existed, it would have to perform just as well as if the target location was known, which is clearly unreasonable. There is however a very natural equivalence class for the problem. An invariant test formulation therefore applies in this instance. Furthermore, a test which is UMP within the restricted class of invariant tests can be found. In the remainder of this section the invariance of the hypothesis testing problem is discussed. A maximal invariant statistic is found, and its distribution calculated under each hypothesis. The optimal invariant test is then formulated, and finally a link is provided between detection in white noise and detection in noise with an unknown circulant covariance matrix.

### 3.2.1 Invariance of the hypothesis testing problem

For the problem of detecting a target with unknown circular shift in white noise (or in general in any noise with a circulant covariance matrix), it is reasonable to require that the same decision be made for arbitrarily cyclically-permuted instances of any given observation. For example, a test should produce the same result for all three observations in Figure 3.4.

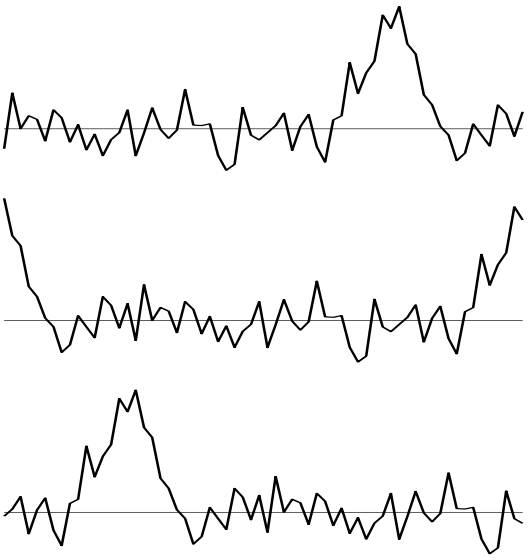


Figure 3.4: Three equivalent observations under cyclic permutation invariance. An invariant detector should make the same decision regardless of which instance is observed.

The transformation group relevant to the problem is therefore

$$\mathcal{G} = \{g_k(\mathbf{x}) = \mathbf{P}^k\mathbf{x},\, k = 0, \ldots, n-1\}. \tag{3.9}$$

This places an equivalence on the observations $\{\mathbf{P}^0\mathbf{x}, \ldots, \mathbf{P}^{n-1}\mathbf{x}\}$, which is natural on account of the symmetry of the elements of the observations under each hypothesis. Thus the observations $(x_1, \ldots, x_{n-1}, x_n) \equiv (x_n, x_1, \ldots x_{n-1}) \equiv \cdots \equiv (x_2, \ldots, x_n, x_1)$ are all considered equivalent by the detector. By the way the hypothesis testing problem has been formulated, enforcing this equivalence is in no way restricting the form of the test.

The testing problem can be seen to be invariant to the group $\mathcal{G}$. Consider the distribution of the observation $\mathbf{x}$ under each hypothesis: in both cases $\mathbf{x}$ is MVN with covariance matrix $\mathbf{C}$. Under $H_0$ the mean is $\mathbf{0}$, and under $H_1$ it is one of the elements of the set $\{\mathbf{P}^0\mathbf{s}, \ldots, \mathbf{P}^{n-1}\mathbf{s}\}$. An element $g_k(\mathbf{x}) = \mathbf{P}^k\mathbf{x}$ of the group $\mathcal{G}$ now represents a linear transformation of $\mathbf{x}$. The distribution of $\mathbf{y} = g_k(\mathbf{x})$ is therefore $N[\mathbf{P}^k E\mathbf{x}, \mathbf{P}^k\mathbf{C}(\mathbf{P}^k)^T]$, where $E\mathbf{x}$ is the expected value of $\mathbf{x}$. Noting that $(\mathbf{P}^k)^T = \mathbf{P}^{n-k} = \mathbf{P}^{-k}$ and that $\mathbf{P}^k\mathbf{C}\mathbf{P}^{-k} = \mathbf{C}$ for $\mathbf{C}$ persymmetric,

$$\mathbf{y} : N[\mathbf{P}^k E\mathbf{x}, \mathbf{C}]. \tag{3.10}$$

Thus under $H_0$ the mean of the transformed vector $\mathbf{y}$ is $\mathbf{0}$, and under $H_1$ it is an element of the set $\{\mathbf{P}^k\mathbf{P}^0\mathbf{s}, \ldots, \mathbf{P}^k\mathbf{P}^{n-1}\mathbf{s}\} = \{\mathbf{P}^0\mathbf{s}, \ldots, \mathbf{P}^{n-1}\mathbf{s}\}$. The transformation $g_k$ therefore preserves the form of the distribution, and retains the partition of the parameter space under each hypothesis. Since this conclusion applies to all elements $g \in \mathcal{G}$, the hypothesis testing problem is invariant-$\mathcal{G}$.

### 3.2.2 Maximal invariant statistic for the problem

To find invariant tests, it is useful to find a maximal invariant statistic for the problem. This is a statistic which has the required invariances, but which also manages to retain the maximum amount of useful information contained in the observation regarding the decision process. One such statistic can be obtained by defining $\mathbf{P}_{\max}(\mathbf{x})$ to be that function which cyclically permutes the elements of $\mathbf{x}$ until the element with the maximum value is in the first position of the vector. Note that for the distributions being considered here $\Pr\{x_i = x_j\} = 0$ for $i \neq j$, so the maximum element of $\mathbf{x}$ will be unique with probability 1.

The statistic $\mathbf{P}_{\max}(\mathbf{x})$ is invariant to the group $\mathcal{G}$: since one of the elements of $\mathbf{x}$ is always maximum and elements of $\mathcal{G}$ simply permute the observation $\mathbf{x}$ cyclically, $\mathbf{P}_{\max}[g(\mathbf{x})] = \mathbf{P}_{\max}[\mathbf{x}]$ for all $g \in \mathcal{G}$. Additionally, for the same reasons, the condition $\mathbf{P}_{\max}[g(\mathbf{x}_1)] = \mathbf{P}_{\max}[g(\mathbf{x}_2)]$ means that $\mathbf{x}_1$ and $\mathbf{x}_2$ must be related to one another through a cyclic shift, so $\mathbf{x}_2 = g(\mathbf{x}_1)$ for some $g \in \mathcal{G}$. Thus the statistic $\mathbf{P}_{\max}(\mathbf{x})$ is maximal.

As explained by Lehmann [87] or Scharf [119], the significance of this result is that only functions of the maximal invariant statistic need to be considered when looking for a test which is invariant to $\mathcal{G}$.

### 3.2.3 Distribution of the maximal invariant statistic

The method described by Hogg and Craig [62, p. 142] in relation to order statistics provides a means of determining the distribution of the maximal invariant. Firstly it is reasserted that two elements of $\mathbf{x}$ are equal with probability zero, so the joint probability density of $\mathbf{x}$ can be defined to be zero at all points which have at least two of their coordinates equal. The set $\mathcal{A}$ where the probability density of $\mathbf{x}$ is nonzero can therefore be partitioned into $n$ mutually disjoint sets:

$$\begin{aligned} \mathcal{A}_1 &= \{\mathbf{x} | x_1 = \max(x_1, \ldots, x_n)\} \\ &\vdots \\ \mathcal{A}_n &= \{\mathbf{x} | x_n = \max(x_1, \ldots, x_n)\}. \end{aligned} \tag{3.11}$$

Thus $\mathcal{A}_i$ is the set of all points in $\mathbb{R}^n$ which have no elements equal, and have $x_i$ as their largest element.

Consider the function $\mathbf{y} = \mathbf{P}_{\max}(\mathbf{x})$. This defines a $1-1$ transformation of each of $\mathcal{A}_1, \ldots, \mathcal{A}_n$ onto the same set $\mathcal{B}$, where it so happens that $\mathcal{B} = \mathcal{A}_1$. For points in $\mathcal{A}_i$, the transformation $\mathbf{y} = \mathbf{P}_{\max}(\mathbf{x})$ cyclically permutes the elements of $\mathbf{x}$ upwards by $i-1$ positions. Thus the inverse function is

$$\mathbf{x} = \mathbf{P}^{i-1}\mathbf{y} \tag{3.12}$$

which simply rotates them back downwards by the same amount.

Letting $J_i$ be the determinant of the Jacobian of the inverse transformation corresponding to $\mathcal{A}_i$, it can be seen that

$$J_i = |\mathbf{P}^{i-1}|. \tag{3.13}$$

By the structure of $\mathbf{P}^{i-1}$, it is always possible to obtain an identity matrix by means of a number of row exchanges. Thus it must be the case that $J_i = +1$ or $J_i = -1$. Denoting the probability density of $\mathbf{x}$ by $f_x(\mathbf{x})$, the results of this section can be combined to yield the probability density $f_y(\mathbf{y})$ of $\mathbf{y} = \mathbf{P}_{\max}(\mathbf{x})$ [62, p. 143]:

$$f_y(\mathbf{y}) = \begin{cases} \sum_{k=0}^{n-1} f_x(\mathbf{P}^k\mathbf{y}) & y_1 = \max(y_1, \ldots, y_n) \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

This expression can finally be used to find the distribution of the maximal invariant statistic under each hypothesis. Under $H_0$,

$$f_x(\mathbf{x}) = (2\pi)^{-n/2}|\mathbf{C}|^{-n/2}e^{-\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}}. \tag{3.15}$$

Therefore the distribution of $\mathbf{y} = \mathbf{P}_{\max}(\mathbf{x})$ is

$$f_y(\mathbf{y}) = \begin{cases} (2\pi)^{-n/2}|\mathbf{C}|^{-n/2} \sum_{k=0}^{n-1} e^{-\frac{1}{2}(\mathbf{P}^k\mathbf{y})^T\mathbf{C}^{-1}(\mathbf{P}^k\mathbf{y})} & y_1 = \max(y_1, \ldots, y_n) \\ 0 & \text{otherwise.} \end{cases} \tag{3.16}$$

Once again using the relation $(\mathbf{P}^k)^T = \mathbf{P}^{-k}$, and noting that if $\mathbf{C}$ is persymmetric then $\mathbf{C}^{-1}$ is also persymmetric, this can be simplified to

$$f_y(\mathbf{y}) = \begin{cases} n(2\pi)^{-n/2}|\mathbf{C}|^{-1/2} e^{-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}} & y_1 = \max(y_1, \ldots, y_n) \\ 0 & \text{otherwise.} \end{cases} \tag{3.17}$$

When $H_1$ is in force, the situation is slightly more complex: now the mean of the observation takes some value in the set $\{\mathbf{P}^\theta\mathbf{s}, \theta = 0, \ldots, n-1\}$. The probability density of $\mathbf{x}$ is therefore

$$f_x(\mathbf{x}) = (2\pi)^{-n/2}|\mathbf{C}|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{P}^\theta\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{x}-\mathbf{P}^\theta\mathbf{s})}, \tag{3.18}$$

where $\theta$ is some integer in the range 0 to $n-1$. Substituting into the expression for $f_y(\mathbf{y})$ gives

$$f_y(\mathbf{y}) = \begin{cases} (2\pi)^{-n/2}|\mathbf{C}|^{-1/2} \sum_{k=0}^{n-1} e^{-\frac{1}{2}(\mathbf{P}^k\mathbf{y}-\mathbf{P}^\theta\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{P}^k\mathbf{y}-\mathbf{P}^\theta\mathbf{s})} & y_1 = \max(y_1, \ldots, y_n) \\ 0 & \text{otherwise.} \end{cases} \tag{3.19}$$

The first case in this expression needs to be looked at in more detail: for the condition $y_1 = \max(y_1, \ldots, y_n)$,

$$\begin{aligned} f_y(\mathbf{y}) &= (2\pi)^{-n/2}|\mathbf{C}|^{-1/2} \sum_{k=0}^{n-1} e^{-\frac{1}{2}[\mathbf{P}^k(\mathbf{y}-\mathbf{P}^{\theta-k}\mathbf{s})]^T\mathbf{C}^{-1}[\mathbf{P}^k(\mathbf{y}-\mathbf{P}^{\theta-k}\mathbf{s})]} \\ &= (2\pi)^{-n/2}|\mathbf{C}|^{-1/2} \sum_{k=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^{\theta-k}\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{P}^{\theta-k}\mathbf{s})}, \end{aligned} \tag{3.20}$$

where once again the persymmetry of $\mathbf{C}$ has been used. Noting that the summation in this final expression is over terms corresponding to all unique cyclic permutations of $\mathbf{s}$, it must be the case that

$$\sum_{k=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^{\theta-k}\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{P}^{\theta-k}\mathbf{s})} = \sum_{l=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^l\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{P}^l\mathbf{s})}. \tag{3.21}$$

Thus the density of $\mathbf{y}$ under $H_1$ is

$$f_y(\mathbf{y}) = \begin{cases} (2\pi)^{-n/2}|\mathbf{C}|^{-1/2} \sum_{l=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^l\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{y}-\mathbf{P}^l\mathbf{s})} & y_1 = \max(y_1, \ldots, y_n) \\ 0 & \text{otherwise.} \end{cases} \tag{3.22}$$

Finally, a form that is sometimes useful is

$$f_y(\mathbf{y}) = \begin{cases} (2\pi)^{-n/2}|\mathbf{C}|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}+\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s})} \sum_{l=0}^{n-1} e^{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{P}^{-l}\mathbf{y}} & y_1 = \max(y_1,\ldots,y_n) \\ 0 & \text{otherwise,} \end{cases} \tag{3.23}$$

where use has been made of the relation $\mathbf{P}^{-l}\mathbf{C}^{-1} = \mathbf{C}^{-1}\mathbf{P}^{-l}$.

Under both cases the density of the maximal invariant is independent of the unknown parameter $\theta$. Thus the unknown permutation parameter has been removed from the problem by the invariance requirement.

### 3.2.4 Optimal invariant likelihood ratio test

Once the observation $\mathbf{x}$ has been mapped onto the corresponding maximal invariant statistic, a simple likelihood ratio test can be performed on this quantity. The likelihood ratio for the problem is

$$\begin{aligned} l(\mathbf{y}) &= \frac{(2\pi)^{-n/2}|\mathbf{C}|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}+\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s})} \sum_{l=0}^{n-1} e^{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{P}^{-l}\mathbf{y}}}{n(2\pi)^{-n/2}|\mathbf{C}|^{-1/2} e^{-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}}} \\ &= \frac{1}{n} e^{-\frac{1}{2}\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} \sum_{l=0}^{n-1} e^{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{P}^{-l}\mathbf{y}}. \end{aligned} \tag{3.24}$$

The log-likelihood ratio is therefore

$$L(\mathbf{y}) = -\ln n - \frac{1}{2}\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s} + \ln \sum_{l=0}^{n-1} e^{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{P}^{-l}\mathbf{y}}. \tag{3.25}$$

The best invariant test is to compare this ratio to a threshold $\eta$, and decide $H_1$ when exceeded:

$$\ln \sum_{l=0}^{n-1} e^{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{P}^{-l}\mathbf{y}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta + \ln n + 1/2\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}. \tag{3.26}$$

This test is uniformly most powerful out of all tests which share the same invariances. Thus no other test which is invariant to cyclic permutations of the observations can perform as well, regardless of the value of the unknown parameter $\theta$. Since the invariance is a perfectly reasonable one for the problem, the test may be considered optimal.

It is worth noting that the estimation of the parameter $\theta$ is explicit in the GLRT of Equation 3.7. Thus the most likely location of the detected signal is also provided by that test. For the UMPI test, however, the dependence on the parameter is completely eliminated by the invariance condition. At no point does this test make use of any estimate of $\theta$, either implicitly or explicitly.

### 3.2.5 Prewhitening in the invariant detection problem

The previous results were based on detection in a zero-mean additive Gaussian noise environment. For the special case where the noise is white and the elements identically distributed, $\mathbf{C} = \mathbf{I}$. In this case the UMPI test simplifies to

$$\ln \sum_{l=0}^{n-1} e^{\mathbf{s}^T \mathbf{P}^{-l} \mathbf{y}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta + \ln n + 1/2 \mathbf{s}^T \mathbf{s}. \tag{3.27}$$

It may be observed that the test in Equation 3.26 may be written in the form

$$\begin{aligned} &\ln \sum_{l=0}^{n-1} e^{\mathbf{s}^T \mathbf{C}^{-1/2} \mathbf{C}^{-1/2} \mathbf{P}^{-l} \mathbf{y}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta + \ln n + 1/2 \mathbf{s}^T \mathbf{C}^{-1/2} \mathbf{C}^{-1/2} \mathbf{s} \\ \Longrightarrow &\ln \sum_{l=0}^{n-1} e^{(\mathbf{C}^{-1/2}\mathbf{s})^T \mathbf{P}^{-l} (\mathbf{C}^{-1/2}\mathbf{y})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta + \ln n + 1/2 (\mathbf{C}^{-1/2}\mathbf{s})^T (\mathbf{C}^{-1/2}\mathbf{s}), \end{aligned} \tag{3.28}$$

where use has been made of both the symmetry and persymmetry of $\mathbf{C}^{-1}$, and the consequent symmetry and persymmetry of $\mathbf{C}^{-1/2}$. This may be seen to be precisely the test for the case of detection in white noise given in Equation 3.27, using instead the whitened target $\mathbf{C}^{-1/2}\mathbf{s}$ and the whitened observation $\mathbf{C}^{-1/2}\mathbf{y}$.

Finally, it is noted that the components of a random vector with a circulant covariance matrix can be diagonalised by means of the discrete Fourier transform. That is, it is generally possible to find both $\mathbf{C}^{-1/2}\mathbf{s}$ and $\mathbf{C}^{-1/2}\mathbf{y}$ using the DFT. This can provide a fast method of calculating the required test statistic in the general case of the covariance matrix being circulant.

## 3.3 Comparison of GLRT and UMPI tests

The UMPI test presented in the previous section is optimal for the problem of detecting targets with unknown cyclic permutation in noise with zero mean and a known circulant covariance. This condition is even stronger than that of a stationary constraint, which only requires that the covariance be Toeplitz.

It has been argued that this optimal test for cyclic shifts is useful from the point of view of assessing the extent to which the GLRT for the same problem is suboptimal. That is, if the GLRT performs well for the case of cyclic shifts and circulant covariances, then it may be expected to work well for shifts defined differently and more general covariance structures. No optimal test exists for these cases, but they are sufficiently similar to the UMPI test problem for general comparisons to apply. This section therefore presents results which compare the UMPI test performance to that of the GLRT for the unknown cyclic permutation detection problem.

The UMPI test statistic for the circulant detection problem involves calculating the sum of the elements of the set $\{e^{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{P}^{-l}\mathbf{x}}, l = 0, \ldots, n-1\}$. Each of these elements has a lognormal distribution, since the exponents are each normally distributed. Furthermore, since the exponents are correlated, the lognormal elements are dependent on one another.

A closed-form solution does not exist for the distribution of the sum or logarithm of the sum of independent lognormally distributed random variables. There is a fair bit of literature on the subject of finding approximate distributions, however [7]. The standard approach is to assume univariate lognormality of the sum, and estimate the parameters using a moment or cumulant matching approach [66, 122]. The general case of the lognormal variates being correlated seems similarly intractable, although the same approaches for approximating the distribution seem to be applicable [54, 1].

For purposes of assessing the performance of the hypothesis test, a problem with this approximation approach is that it is not certain under which conditions it is valid and accurate. Any apparent insights gained into the performance of the test therefore have to be confirmed as a matter of course. Furthermore, attempting to determine the conditions under which the approximations apply is not trivial, particularly with the large number of different combinations of correlations between the elements of the sum. In the interests of simplicity and accuracy, therefore, the test performance is assessed directly using statistical simulations.

Similar observations can be made with regard to the distribution of the GLRT statistic. In this case the component of the test which depends on the observation is given by $\max_{\theta\in[0,n-1]}(\mathbf{P}^\theta\mathbf{s})^T\mathbf{C}^{-1}\mathbf{x}$. Under $H_0$, with $\mathbf{x} : N[\mathbf{0}, \mathbf{C}]$, an assessment of the performance of the GLRT therefore involves finding the distribution of the maximum of the set of correlated normal random variables. As in the case of the UMPI test, the distribution of the test statistic has no simple closed form. Once again it therefore seems reasonable to quantify the performance of the test based on direct simulation.

The required simulations can be performed very efficiently on account of the fact that the primary components of the detection statistic may be calculated using the DFT. It is therefore possible to make performance assessments based on large numbers of generated samples, ensuring accurate results.

The remainder of this section provides a comparison between the GLRT and the UMPI tests for the problem of detecting the basic targets shown in Figure 3.5. Various scalings of these targets are considered, corresponding to energies of 2, 4, and 8. In all cases, the additive noise is comprised of 64 uncorrelated samples of zero-mean Gaussian noise with unit variance. In terms of method, there is no loss of generality in assuming uncorrelated noise, since the prewhitening transformation may always be used to cast the problem into this form. In that case all that is required is to modify the target signal to be analysed.

Figure 3.6 shows distributions for each of the test statistics for the specific case of the targets of energy 8. In each case the distribution to the right corresponds to $H_1$. These histograms were generated

**Section 3.3: Comparison of GLRT and UMPI tests**

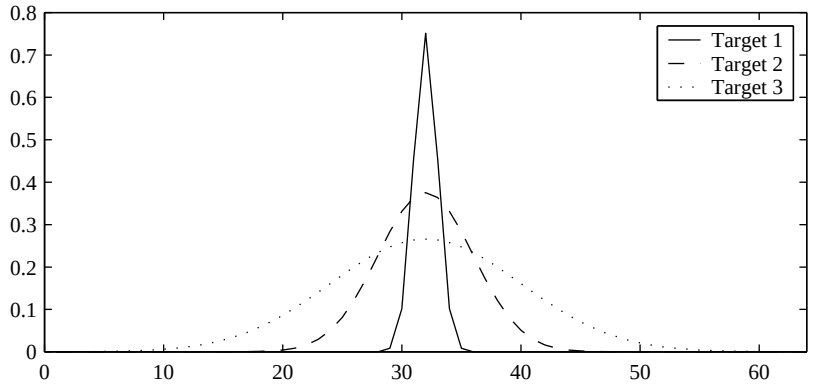



Figure 3.5: Unit energy test targets used in Monte-Carlo simulations: the targets are sampled Gaussian functions centred on 32, with standard deviations of 1, 4, and 8.

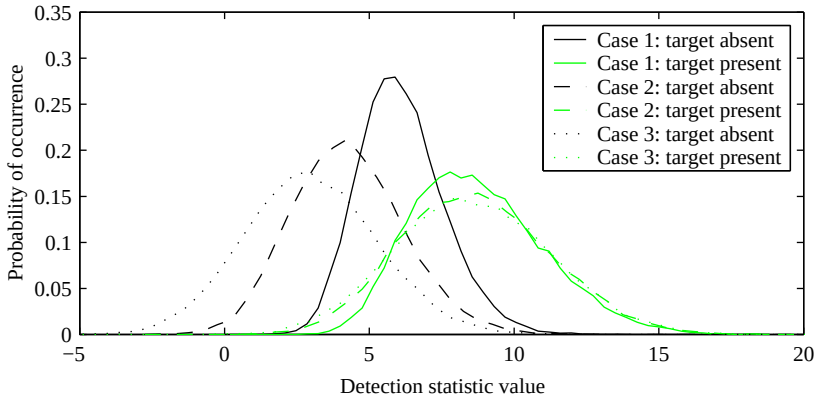
using 25000 noise samples under each hypothesis and 50 bins. Note that the curve styles in the plot correspond directly to those used for plotting each of the targets in Figure 3.5.

The test performance is not apparent from the histogram plots under the various hypotheses. To demonstrate this information, Figure 3.7 shows receiver operating characteristic (ROC) curves for the various target detection scenarios. The first plot demonstrates the relative performance of the two tests for the three targets in Figure 3.5, under the condition that they have been scaled to have an energy of 2. The remaining two plots show similar results for targets scaled to have energies of 4 and 8 respectively. Since the GLRT and UMPI test are both invariant to cyclic permutations and the latter is UMP in this class, it exhibits uniformly better performance than the GLRT in all cases.

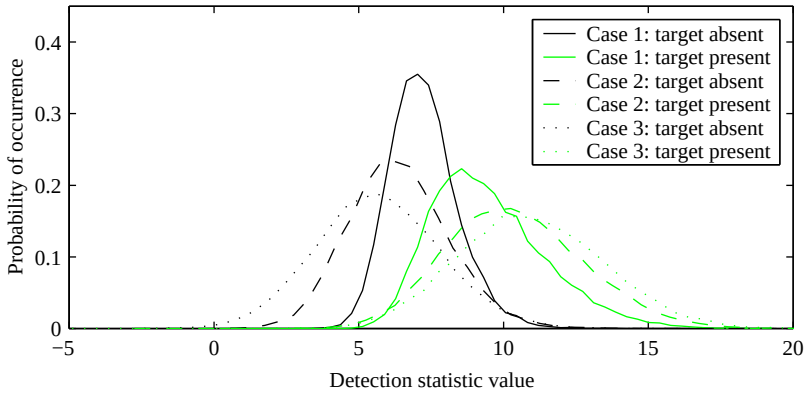
It is clear from the plots that the difference in performance between the tests is marginal for the cases which were analysed. A closer inspection of the results indicates that the difference does however depend on the properties of the target being detected:

- The difference in performance between the GLRT and the UMPI test becomes smaller when the energy of the target to be detected becomes higher. This may be attributed to the fact that the target location estimate used in the GLRT improves under these conditions, resulting in better relative performance.
- The difference becomes smaller as the targets become less localised. That is, the performance of the GLRT is better relative to the UMPI test for the case of detecting target 3 in Figure 3.5 than for target 1. A possible explanation is that for the wider targets the location estimate used in the GLRT is less important, with errors having less effect on the overall test.

On the whole, the suboptimal GLRT performs exceedingly well for the specific detection problems analysed. The conclusion that may be drawn is that the GLRT is effective for dealing with the problem




(a) GLRT statistic.




(b) UMPI statistic.

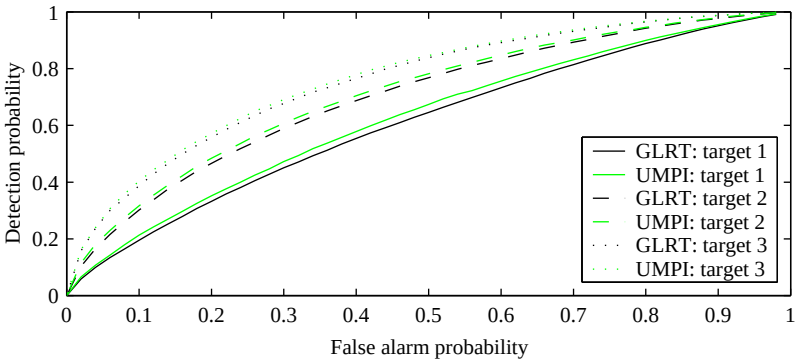
Figure 3.6: Distributions of the GLRT statistic and UMPI statistic under both $H_0$ and $H_1$ for each of the targets tested. The targets were rescaled to have an energy of 8.

of detecting a transient target with unknown location in additive Gaussian noise. This observation becomes especially true when shifts are defined circularly, and the noise covariance is circulant.

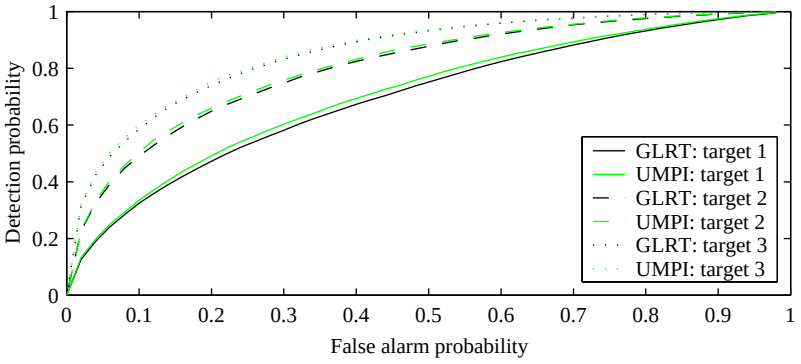
## 3.4 Cyclic permutation invariant detection with additional nuisance parameters

Up to now, the results in this chapter have assumed the noise distribution to be completely known. That is, the only unknown parameter in the problem formulation is the target location.

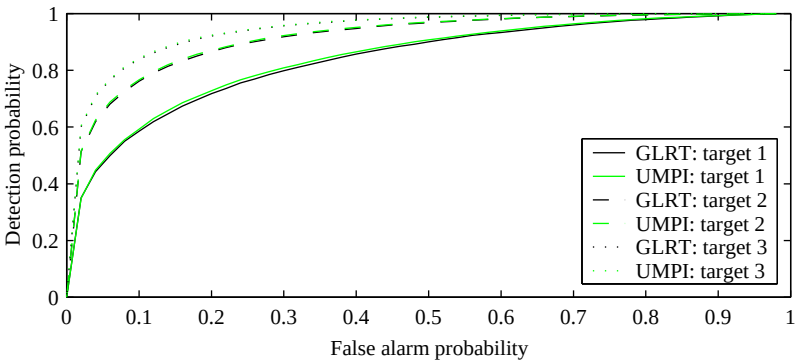
**Section 3.4: Cyclic permutation invariant detection with additional nuisance parameters**




(a) Case 1: Targets of energy 2.




(b) Case 2: Targets of energy 4.




(c) Case 3: Targets of energy 8.

Figure 3.7: ROC curves for detection of the test targets scaled to have energies of 2, 4, and 8.

Often there may be other nuisance parameters defining the noise process. These parameters have to be dealt with by the test, usually through some estimation procedure. The GLRT, for example, estimates *all* the unknown parameters under each assumed hypothesis by maximising the associated likelihood. The likelihood ratio is then taken to be the ratio of the maximised likelihood under each hypothesis. In general the nuisance parameter estimates will be different under each assumed hypothesis.

The results of the previous section demonstrate that instead of estimating unknown parameters, it may be more effective to eliminate them from the problem using an appropriate and sensible invariance criterion. For the specific cases discussed, the resulting test has a greater power, and is therefore preferable. In this section the invariance approach is extended to the case where additional unknown parameters are present in the distributions.

In general it will not be possible to eliminate all nuisance parameters by imposing absolutely reasonable equivalence classes onto a problem. However, *some* of the parameters may still be dealt with in this manner. For example, if it is required to detect targets with unknown location in white noise with unknown power, the location parameter may still reasonably be eliminated by requiring invariance to cyclic permutations, as discussed in the previous section. The noise power represents an unknown parameter which is not removed by the invariance condition, however, and some alternative method of dealing with it has to be adopted.

The invariant hypothesis testing framework introduces the possibility of making *invariant* estimates of these remaining nuisance parameters under each hypothesis. That is, estimates are used which are invariant to the same group of transformations that have been deemed reasonable for the testing problem. The reason for requiring this is so that the resulting test is indeed invariant to the group operation. One way of arriving at such estimates is to find the parameter values that maximise the likelihood function of the maximal invariant statistic under each hypothesis. Because these estimates are based on the maximal invariant, they are clearly invariant to the required transformation group. Additionally, because of the maximality property of this statistic, as much valid information as possible regarding the unknown parameters is incorporated into the statistic.

In practice, if such invariant maximum likelihood estimates are used, then the resulting test is equivalent to performing the GLRT on the maximal invariant statistic. The distribution of this statistic may already be independent of some of the nuisance parameters (through the action of the invariance class), but in general it may still depend on other unknowns. Formally, suppose an invariant problem formulation admits a maximal invariant statistic $\mathbf{y}$, which has a density $p_{H_0}(\mathbf{y}|\boldsymbol{\theta}_0)$ under $H_0$ and $p_{H_1}(\mathbf{y}|\boldsymbol{\theta}_1)$ under $H_1$. The parameters $\boldsymbol{\theta}_0$ are those parameters which are not eliminated from the distribution of $\mathbf{y}$ under $H_0$, and $\boldsymbol{\theta}_1$ are the parameters which remain under $H_1$. If each of the parameter vectors are permitted to take on values in the sets $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ under the two respective hypotheses, then the invariant GLRT is given by

$$l(\mathbf{y}) = \frac{\max_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} p_{H_1}(\mathbf{y}|\boldsymbol{\theta}_1)}{\max_{\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}_0} p_{H_0}(\mathbf{y}|\boldsymbol{\theta}_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{3.29}$$

**Section 3.4: Cyclic permutation invariant detection with additional nuisance parameters**

The invariant maximum likelihood estimates of the nuisance parameters under each hypothesis are those values that maximise each invariant likelihood function.

Any estimate which is based purely on the value of the maximal invariant statistic is guaranteed to be invariant. Thus the use of maximum likelihood estimates represents only one possibility of many. In principle, any preferred nuisance parameter estimate could be made, and substituted into the relevant invariant likelihood ratio. The maximum likelihood estimates, however, are convenient from a computational perspective, and from the point of view of detection are familiar from the role they play in the GLRT.

In the section which follows, details of an invariant formulation are presented for the problem of detecting a known target with unknown circular shift in certain instances of parametrically-known noise. The unknown location parameter is effectively eliminated by requiring cyclic permutation invariance in the solution. The covariance matrix of the noise is assumed to be circulant, but depends on unknown parameters which have to be dealt with by the test. These parameters are estimated by maximising the invariant likelihood function of the maximal invariant statistic under each hypothesis.

The specific case of the noise covariance being some unknown scaling of the identity matrix is then discussed in detail. It is demonstrated that cyclic permutation invariant estimates of the unknown noise power can be made under each hypothesis, by maximising the related invariant likelihood functions. The ratio of the likelihood values taken on by these functions then constitutes an invariant test statistic, which can be compared to a threshold in the usual manner. Finally, the performance of this resulting test is compared to the standard GLRT for the problem, where both the location parameter and the noise power are assumed unknown, and no invariance condition is imposed.

### 3.4.1 Invariant estimation of certain circulant covariance matrices

In this section the equations determining the cyclic permutation invariant maximum likelihood estimate are derived for the problem of estimating unknown covariance matrix parameters. As with the invariant detection problem, for the invariant formulation to apply the problem has to exhibit some specific structure. It is therefore assumed that the covariance matrix is circulant, although some of its parameters are unknown.

If $\mathbf{C}$ is full-rank and positive definite, then it can always be expressed in the form $\mathbf{C} = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T$, where $\mathbf{A}$ is unitary, and $\mathbf{\Sigma}$ diagonal and positive definite. A special property of circulant matrices is that this decomposition may always be performed using the *same* matrix $\mathbf{A}$, and in Chapter 5 it is shown that for the complex case $\mathbf{A}$ may be taken to be the unitary DFT matrix. Thus there exists a known matrix $\mathbf{A}$ such that $\mathbf{A}^T\mathbf{C}\mathbf{A}$ is diagonal for *every* circulant $\mathbf{C}$ of interest.

A sample $\mathbf{x}$ from an assumed $N[\mathbf{P}^\theta\mathbf{s}, \mathbf{C}]$ process is available on which to base the covariance estimate.

As explained, the covariance matrix $\mathbf{C}$ depends on certain unknown parameters. The class of problems considered here involves the case where the unknown parameters of $\mathbf{C}$ may be expressed directly in terms of the diagonal elements of $\mathbf{A}^T\mathbf{C}\mathbf{A} = \mathbf{\Sigma}$. In this instance, $\mathbf{C}$ is completely specified by the diagonal elements of $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, so $\mathbf{C}$ may be considered to be a function of these elements.

The covariance parameter estimate is required to be invariant to the group of transformations $\mathcal{G}$ defined in Equation 3.9, namely that of cyclic permutations of the elements of the observation vector. As shown in Section 3.2.3, the vector $\mathbf{y} = \mathbf{P}_{\max}(\mathbf{x})$ is a maximal invariant for this class, and if $N[\mathbf{P}^\theta\mathbf{s}, \mathbf{C}(\sigma_1^2, \ldots, \sigma_n^2)]$ then

$$f_y(\mathbf{y}) = (2\pi)^{-n/2}|\mathbf{C}(\sigma_1^2, \ldots, \sigma_n^2)|^{-1/2} \sum_{l=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^l\mathbf{s})^T\mathbf{C}(\sigma_1^2,\ldots,\sigma_n^2)^{-1}(\mathbf{y}-\mathbf{P}^l\mathbf{s})} \tag{3.30}$$

for $y_1 = \max(y_1, \ldots, y_n)$. The invariant likelihood function under the target present assumption is therefore given by

$$L = -n/2\ln(2\pi) - 1/2\ln|\mathbf{C}(\sigma_1^2, \ldots, \sigma_n^2)| + \ln\sum_{l=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^l\mathbf{s})^T\mathbf{C}(\sigma_1^2,\ldots,\sigma_n^2)^{-1}(\mathbf{y}-\mathbf{P}^l\mathbf{s})}. \tag{3.31}$$

Using the decomposition for $\mathbf{C}$ and adopting the notation $\mathbf{z}_l = \mathbf{A}^T(\mathbf{y} - \mathbf{P}^l\mathbf{s})$, this likelihood may be written as

$$\begin{aligned} L &= -n/2\ln(2\pi) - 1/2\ln|\mathbf{\Sigma}| + \ln\sum_{l=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^l\mathbf{s})^T\mathbf{A}\mathbf{\Sigma}^{-1}\mathbf{A}(\mathbf{y}-\mathbf{P}^l\mathbf{s})} \\ &= -n/2\ln(2\pi) - 1/2\ln\prod_{k=1}^{n}\sigma_k^2 + \ln\sum_{l=0}^{n-1} e^{-\frac{1}{2}\mathbf{z}_l^T\mathbf{\Sigma}^{-1}\mathbf{z}_l}. \end{aligned} \tag{3.32}$$

Finally, denoting by $\mathbf{z}_l(k)$ the $k$th element of the vector $\mathbf{z}_l$, this invariant likelihood may be written as

$$L = -n/2\ln(2\pi) - 1/2\sum_{k=1}^{n}\ln\sigma_k^2 + \ln\sum_{l=0}^{n-1} e^{-\frac{1}{2}\sum_{k=1}^{n}\frac{\mathbf{z}_l^2(k)}{\sigma_k^2}}. \tag{3.33}$$

To arrive at the invariant MLE of the covariance matrix, it is required to maximise this likelihood with respect to the parameters $\sigma_1^2, \ldots, \sigma_n^2$, subject to any desired constraints on these elements.

The expression for the invariant likelihood may be maximised under several conditions on the elements of $\mathbf{\Sigma}$. One case which may be of interest is when *all* the parameters $\sigma_1^2, \ldots, \sigma_n^2$ are considered unknown and independent, which corresponds to the assumption that the noise is circulant, but otherwise completely unknown. A necessary condition on the invariant MLE is then that the derivative $dL/d\sigma_i^2$ be identically zero for each $i$, which leads to $n$ simultaneous equations with $n$ unknowns. The equations

are not easily solved, however, since the desired parameters are implicit in the equations.

A simpler case involves the assumption that a sample is provided from a $N[\mathbf{P}^{\theta}\mathbf{s}, \sigma^2\mathbf{I}]$ process, where $\theta$ and $\sigma^2$ are assumed unknown. Through the action of the invariance class, a maximal invariant is obtained which is independent of $\theta$. However, it is still required to make an invariant estimate of the parameter $\sigma^2$, which parameterises the unknown component of the covariance matrix. It is easy to see that this scenario corresponds to the choice of $\mathbf{A} = \mathbf{I}$ in the previous discussion, along with the condition that $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$ in the diagonalised matrix.

Under these conditions, the invariant likelihood function simplifies to

$$\begin{aligned} L &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln\sum_{l=0}^{n-1} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}\mathbf{z}_l^2(k)} \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \ln\sum_{l=0}^{n-1} e^{-\frac{1}{2\sigma^2}\mathbf{z}_l^T\mathbf{z}_l}. \end{aligned} \tag{3.34}$$

A necessary condition on the invariant maximum likelihood estimate $\hat{\sigma}^2$ of the noise power is that

$$\left.\frac{dL}{d\sigma^2}\right|_{\sigma^2=\hat{\sigma}^2} = 0. \tag{3.35}$$

This yields the condition

$$\hat{\sigma}^2\sum_{l=0}^{n-1} e^{-\frac{1}{2\hat{\sigma}^2}\mathbf{z}_l^T\mathbf{z}_l} = \frac{1}{n}\sum_{l=0}^{n-1}\mathbf{z}_l^T\mathbf{z}_l e^{-\frac{1}{2\hat{\sigma}^2}\mathbf{z}_l^T\mathbf{z}_l}, \tag{3.36}$$

which can be solved using numerical techniques.

### 3.4.2 Invariant detection using invariant estimates of nuisance parameters

Once the invariant estimates of the nuisance parameters have been found, they can be substituted into the invariant likelihood function, and the maximum likelihood value obtained. The ratio of the maximum likelihood under each assumed hypothesis may then be used as a test statistic.

Consider the case of invariant covariance matrix estimation discussed in the previous section. To use these estimates in a detector, estimates of $\mathbf{C}$ under both the null hypothesis and the alternative are required. These estimates can be used in the invariant likelihood ratio statistic, which is then given by

$$t(\mathbf{y}) = \frac{\max_{\mathbf{C}} p_{H_1}(\mathbf{y}|\mathbf{C})}{\max_{\mathbf{C}} p_{H_0}(\mathbf{y}|\mathbf{C})}, \tag{3.37}$$

where the maximisation is implied to be over the class of circulant covariance matrices which are considered valid for the problem. If the invariant maximum likelihood estimates of the covariance are

$\widehat{\mathbf{C}}_0$ under $H_0$ and $\widehat{\mathbf{C}}_1$ under $H_1$ (that is, these are the values that maximise the invariant log-likelihood of Equation 3.31), then the resulting test statistic is

$$t(\mathbf{y}) = \frac{|\widehat{\mathbf{C}}_1|^{-1/2} \sum_{l=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^l\mathbf{s})^T \widehat{\mathbf{C}}_1^{-1}(\mathbf{y}-\mathbf{P}^l\mathbf{s})}}{|\widehat{\mathbf{C}}_0|^{-1/2} \sum_{l=0}^{n-1} e^{-\frac{1}{2}\mathbf{y}^T \widehat{\mathbf{C}}_0^{-1}\mathbf{y}}}. \tag{3.38}$$

The log-likelihood ratio is therefore

$$\ln t(\mathbf{x}) = -1/2 \ln |\widehat{\mathbf{C}}_1| + \ln \sum_{l=0}^{n-1} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{P}^l\mathbf{s})^T \widehat{\mathbf{C}}_1^{-1}(\mathbf{y}-\mathbf{P}^l\mathbf{s})} + 1/2 \ln |\widehat{\mathbf{C}}_0| - \ln \sum_{l=0}^{n-1} e^{-\frac{1}{2}\mathbf{y}^T \widehat{\mathbf{C}}_0^{-1}\mathbf{y}}. \tag{3.39}$$

The suggested detector is therefore to take the observation $\mathbf{x}$, calculate the maximal invariant statistic $\mathbf{y}$, make invariant estimates of $\mathbf{C}$ under each hypothesis using $\mathbf{y}$, and apply the test

$$\ln t(\mathbf{y}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.40}$$

to the results.

A convenient simplification once again arises when the noise covariance is assumed to be some unknown scaling of the identity matrix. The noise power $\sigma^2$ then represents an unknown parameter. Under $H_0$, the invariant MLE of the variance parameter is simply $\hat{\sigma}_0^2 = 1/n \mathbf{y}^T\mathbf{y}$. Under $H_1$, the relation provided in Equation 3.36 may be used to make a corresponding invariant estimate of $\hat{\sigma}_1^2$.

Under the assumption that $\widehat{\mathbf{C}}_0 = \hat{\sigma}_0^2\mathbf{I}$ and $\widehat{\mathbf{C}}_1 = \hat{\sigma}_1^2\mathbf{I}$, the log-likelihood ratio in Equation 3.39 becomes

$$\ln t(\mathbf{y}) = -n/2 \ln \sigma_0^2 - \frac{1}{2\hat{\sigma}_1^2}(\mathbf{y}^T\mathbf{y} + \mathbf{s}^T\mathbf{s}) + \ln \sum_{l=0}^{n-1} e^{\frac{1}{\hat{\sigma}_1^2}\mathbf{s}^T\mathbf{P}^{-l}\mathbf{y}} + n/2 \ln \sigma_1^2 - \ln n + \frac{1}{2\hat{\sigma}_0^2}\mathbf{y}^T\mathbf{y}. \tag{3.41}$$

The terms $\mathbf{s}^T\mathbf{P}^{-l}\mathbf{y}$ in this test statistic are again the elements of the circular convolution between the target $\mathbf{s}$ and the maximal invariant $\mathbf{y}$. Also, since $\mathbf{y}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}$ and

$$\sum_{l=0}^{n-1} e^{\frac{1}{\hat{\sigma}_1^2}\mathbf{s}^T\mathbf{P}^{-l}\mathbf{y}} = \sum_{l=0}^{n-1} e^{\frac{1}{\hat{\sigma}_1^2}\mathbf{s}^T\mathbf{P}^{-l}\mathbf{x}}, \tag{3.42}$$

this test statistic can be calculated directly from the observation $\mathbf{x}$.

## 3.5 Performance of GLRT with invariant estimates

In this section performance results are presented for the problem discussed previously, namely that of detection of a known target with unknown location in noise with unknown power. It is assumed that the noise samples are independent and identically distributed, so the covariance matrix of the additive

noise process is some unknown scaling of the identity matrix.

Equation 3.41 in the previous section provides one test statistic for this problem. As explained, the test is based solely on the maximal invariant statistic $\mathbf{y} = \mathbf{P}_{\max}(\mathbf{x})$ for the problem, both in terms of the test statistic formulation and the invariant estimation of the unknown parameters. As has been mentioned, the resulting test is precisely the same as performing the GLRT on the maximal invariant statistic. For this reason, the test is referred to as the invariant GLRT. Summarising, the invariant GLRT is given by

$$-n/2 \ln \sigma_0^2 - \frac{1}{2\hat{\sigma}_1^2}(\mathbf{y}^T\mathbf{y} + \mathbf{s}^T\mathbf{s}) + \ln \sum_{l=0}^{n-1} e^{\frac{1}{\hat{\sigma}_1^2}\mathbf{s}^T\mathbf{P}^{-l}\mathbf{y}} + n/2 \ln \sigma_1^2 - \ln n + \frac{1}{2\hat{\sigma}_0^2}\mathbf{y}^T\mathbf{y} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{3.43}$$

where $\hat{\sigma}_0^2 = 1/n\mathbf{y}^T\mathbf{y}$ and $\hat{\sigma}_1^2$ is determined using the relationship in Equation 3.36.

The invariant GLRT is a somewhat unorthodox solution to the detection problem with unknown nuisance parameters, and makes explicit use of an invariance group in its formulation. A more common approach to the problem would simply be to consider both the parameters $\theta$ and $\sigma^2$ unknown, and use the standard GLRT test on the resulting problem. This leads to the test statistic

$$t_{\text{GLRT}}(\mathbf{x}) = \frac{\max_{\theta,\sigma^2}(2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{P}^\theta\mathbf{s})^T(\mathbf{x}-\mathbf{P}^\theta\mathbf{s})}}{\max_{\sigma^2}(2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x}}}. \tag{3.44}$$

Under these circumstances it can be shown that the numerator is maximised for the choice

$$\hat{\theta} = \arg\max_{\theta} \mathbf{s}^T\mathbf{P}^\theta\mathbf{x} \tag{3.45}$$

and

$$\hat{\sigma}_1^2 = 1/n(\mathbf{x} - \mathbf{P}^{\hat{\theta}}\mathbf{s})^T(\mathbf{x} - \mathbf{P}^{\hat{\theta}}\mathbf{s}). \tag{3.46}$$

The denominator is maximised for $\hat{\sigma}_0^2 = 1/n\mathbf{x}^T\mathbf{x}$. Substituting into the test statistic and taking the logarithm yields the resulting GLRT:

$$-n/2 \ln(2\pi\hat{\sigma}_1^2) + n/2 \ln(2\pi\hat{\sigma}_0^2) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{3.47}$$

It is this test to which the invariant GLRT of the previous paragraph is to be compared. Note that this test is also invariant to transformations in the group $\mathcal{G}$, since the estimates $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are invariant to cyclic permutations of their arguments. However, the resulting estimates do *not* maximise the invariant likelihood function. This test is therefore referred to simply as the standard GLRT for the problem.

Figure 3.8 shows distributions of the log-likelihood ratio of each of these tests for the problem of detecting the targets shown in Figure 3.5 in white noise with unit variance. The targets were scaled to have an energy of 8 before the test statistics were calculated. In all cases the location of the target (in terms of cyclic shifts) is assumed unknown, as is the noise variance. Note that these results may be

compared directly with those of Section 3.3, since the same targets are used and the noise variances are actually equal. Thus the differences between the results in this section and those of Section 3.3 differ only in that here the noise variance is additionally assumed unknown.

The ROC curves for both the GLRT and the invariant GLRT are shown in Figure 3.9, again for the targets shown in Figure 3.5. The amplitudes of the targets have been rescaled to result in energies of 2, 4, and 8 for each plot in the figure. In each case the invariant GLRT can be seen to outperform the standard GLRT, in some cases by quite a significant margin. The difference is more pronounced for shorter duration targets.

The reason for the better performance of the invariant GLRT relative to that of the standard GLRT can be attributed to the action of the invariance condition on the problem. The invariance has effectively eliminated the unknown location parameter, which no longer has to be estimated from the data. This reduces the uncertainty in the calculated test statistic, thereby improving the performance.

## 3.6 Concluding remarks

The results of this chapter partially validate the use of the GLRT for detecting targets with unknown location in noise. For known noise parameters, the detection statistic for this test simply involves the calculation of the set of inner products $\{\mathbf{s}_1^T\mathbf{C}^{-1}\mathbf{x}, \ldots, \mathbf{s}_k^T\mathbf{C}^{-1}\mathbf{x}\}$, where $\mathbf{x}$ are the observed data and $\mathbf{s}_1, \ldots, \mathbf{s}_k$ are shifted versions of the prototype target which may occur in the data. The GLRT compares the maximum value of this set to a threshold, and a decision of $H_1$ is made if exceeded.

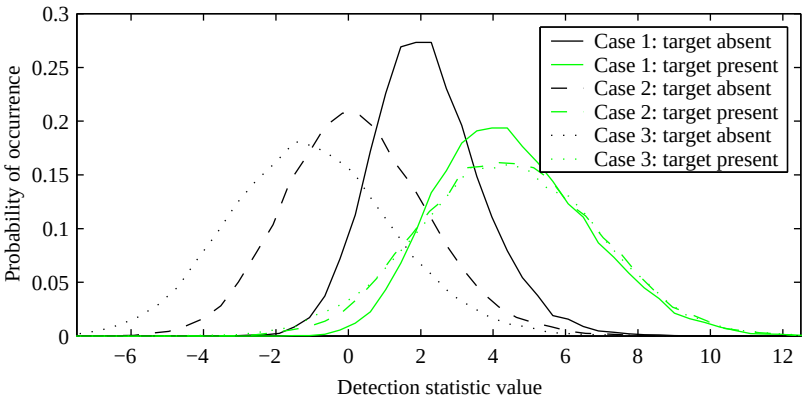
As shown in this chapter, the GLRT is not optimal for this problem. Unfortunately it is not possible to assess the extent of its suboptimality for all possible detection cases, but such an assessment can indeed be made under the assumption of circular shifts and circulant covariances. For this case an optimal UMPI test can be derived, which is most powerful in the class of all tests which are invariant to cyclic permutations of the data. When this UMPI test is compared with the GLRT, the performance difference is seen to be negligible for some common targets.

It must be emphasised that the tests presented are only appropriate for testing the possible presence of a *single* target in the observed data. The presence of multiple targets could be dealt with in an ad-hoc capacity using the proposed tests as a basic structure, but complications may arise if this approach is adopted. The most serious difficulty in this regard is the estimation of the number and relative location of targets present, which is required for meaningful detection to take place. A better approach is to develop explicit models for the multiple target situation, and derive the GLRT accordingly.

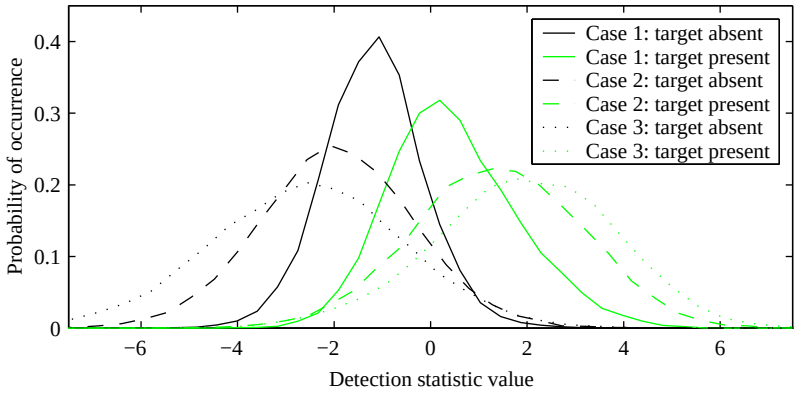
For detecting a target with unknown location in a long sequence of data, it is common to slide a window across the sequence and perform some predetermined detection operation at each point. For transient targets with limited duration, this results in a set of values which are usually designed to

become larger as the likelihood of a target being present increases. The GLRT formulation supports this implementation method insofar as the test statistic is concerned, although the selection of the threshold for the test may be nontrivial. This is particularly true if the test threshold is to be obtained through statistical analysis procedures.

Also demonstrated in this chapter are ways in which an invariance formulation can be applied to detection problems where additional nuisance parameters are present. The use of an invariance subspace to eliminate some of the parameters has been proposed, with invariant estimates being made of those parameters which are not eliminated via the invariance condition. It has been shown that even if the invariance constraint does not eliminate all the parameters, better test performance is nevertheless obtained through partial parameter elimination. The conclusion may be drawn that if a reasonable invariance group exists for a problem, then detection performance will benefit by transforming the data to a maximal invariant. Such a transformation can entirely eliminate unwanted parameters, which then no longer have to be either estimated or considered any further.



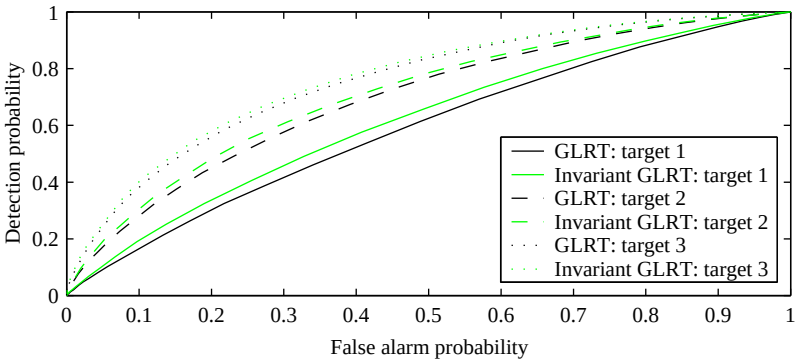

(a) GLRT statistic.



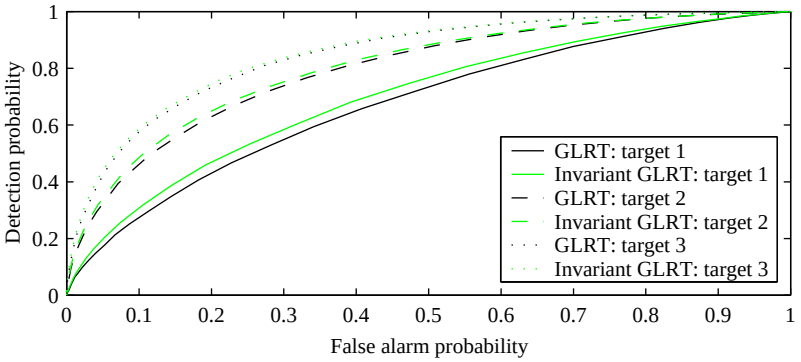

(b) Invariant GLRT statistic.

Figure 3.8: Distributions of the GLRT statistic and invariant GLRT statistic under both $H_0$ and $H_1$ for each of the targets tested. The targets were rescaled to have an energy of 8.

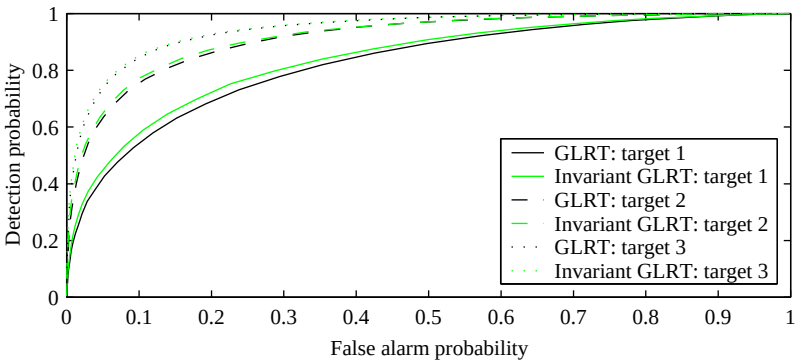
## Section 3.6: Concluding remarks




(a) Case 1: Targets of energy 2.




(b) Case 2: Targets of energy 4.




(c) Case 3: Targets of energy 8.

Figure 3.9: ROC curves for detection of the test targets scaled to have energies of 2, 4, and 8.

# Chapter 4

# Subspace invariance in detection

When considering use in practical situations, certain statistical models have significant advantages over others. These advantages can take the form of easy and convenient model description, efficient parameter estimation, and closed-form solutions for specific applications.

This is particularly true of the assumption of a Gaussian probability density. It can be specified using a relatively small number of simple parameters, which are often easily estimated and have well-documented sampling distributions. The assumption of a normal distribution often leads to linear models which are completely specified by first- and second-order statistics, resulting in desirable computational properties.

In some cases it is possible to relax the rigidity of a convenient assumption without completely compromising the resulting simplicity. For example, the autoregressive integrated moving average (ARIMA) models of Box, Jenkins, and Reinsel [15] extend the validity of conventional ARMA modelling to nonstationary random signals by means of a simple differencing operation. This presents very little added complexity in terms of theoretical analysis and practical implementation, but opens up a whole new range of applications in which ARMA models can be used.

In this section an alternative method of relaxing or extending a preferred statistical characterisation is presented. The application of the ideas is with regard to simple detection of a known target in noise. A nominally MVN random process is used for modelling the noise, although in principle the ideas are not restricted to this case. In particular, the situation is discussed where each noise observation is assumed to be a realisation of a random process which is only valid within a restricted linear subspace of the original observation space. The complementary portion of the observation is assumed either to have been corrupted by unknown interference, to not be conducive to simple statistical modelling, or to exhibit statistical behaviour that differs from the assumed distribution. Although these situations lead to identical detectors, the formulations have different statistical justifications and have to be treated

separately.

The first case considered is that of subspace interference. Here it is assumed that there is an underlying process $\mathbf{z}$ generating samples on which it is required to perform detection, but that these samples are not directly observable. Instead, samples of the form $\mathbf{x} = \mathbf{z} + \mathbf{U}_I\mathbf{c}$ are observed, where $\mathbf{c}$ is an unknown deterministic parameter. Thus there is an additive interference component $\mathbf{U}_I\mathbf{c}$ corrupting the data observation. It is easy to show that the interference necessarily lies in the subspace $\langle \mathbf{U}_I \rangle$ spanned by the columns of $\mathbf{U}_I$. This is a linear subspace of the original observation space, and is referred to as the interference subspace.

A property of the subspace interference formulation is that successive or separate observations are associated with *different* values of the interference parameter $\mathbf{c}$. Thus, apart from the subspace constraint, nothing is known about the interference component.

In Section 4.2 the problem of performing detection in the presence of such interference is discussed. The paradigm used is that of invariant detection. Because nothing is known about the interference component for any given observation, it is reasonable to require invariance to these components. It is demonstrated that within the implied invariance class a uniformly most powerful (UMP) test exists, which can be claimed to be optimal in the Neyman-Pearson sense.

The use of invariance in dealing with subspace interference is useful for modelling processes which exhibit statistical predictability, but where low-rank additive contributions are present which may vary considerably. For example, some time-series processes are almost stationary, but exhibit low-frequency trends which are sufficient to cast the stationarity assumption in doubt. Rather than trying to characterise these trends, they may be modelled as subspace interference, thereby eliminating them from the decision process.

To use the subspace invariant detector in a subspace interference context, it is necessary to identify the interference subspace. Often this information is provided by peripheral knowledge (such as through physical modelling of the processes involved), or may be fairly evident just from looking at the data. However, to be truly useful in a general context some method of identifying interference subspaces from samples of real data is required.

Section 4.3 deals with this identification for the case where the underlying process ($\mathbf{z}$, as described above) is multivariate normal with known parameters. A maximum likelihood formulation is presented for estimating the interference subspace, under the assumption that the rank of the interference is known. Unfortunately a closed-form general solution has not been found, but a number of approximate methods are proposed which appear to have merit. Furthermore, exact solutions are provided for two simple cases. Finally, using statistical simulations, the performance of the estimates are assessed.

The invariant test has application beyond that of detection in subspace interference. In general, subspace invariance can be used whenever the actual distribution of the data is partially in conflict with some

convenient assumed model. Although such a conflict may arise as a result of subspace interference, it is by no means always the case. In fact, subspace invariant detection can be used in any situation where mismatch between the data and the model lies mainly within a linear subspace of the original observation space. Used in this context the detector has no optimality properties. However, the inclusion of an invariance subspace can reduce model mismatch, thereby improving overall performance.

Section 4.4 discusses this form of invariance in hypothesis testing. Some examples are given of problems which might benefit from the inclusion of an invariance subspace. From a practical perspective, the most valuable use of the procedure is in extending the validity of simple or tractable models for use in situations where they would otherwise be inappropriate.

The problem of identifying candidate invariance subspaces for the model mismatch problem is discussed in Section 4.5. Since a maximum likelihood solution is difficult to formulate, an alternative method is developed. In Section 4.5.2 this method is applied to a problem involving real data, and it is shown that the invariant detector has better detectability and predictability than the noninvariant detector.

## 4.1 Linear subspaces

This chapter develops ways in which invariance to linear subspaces can aid in detector performance. To understand the concepts involved, the notion of a linear subspace has to be understood. This section provides an overview of the relevant linear algebraic concepts. For a detailed treatment of linear algebra, the books by Deif [26] and Horn and Johnson [63] are recommended. For a treatment which is more specific to the subspace invariant detection problem, the introductory chapters of the book by Scharf [119] should be sufficient.

The notion of an abstract vector space is useful, but is not required in this discussion. Instead, a simpler vector space can be defined in terms of sets of ordered $n$-tuples, either with real or complex elements, that are organised into an $n$-dimensional vector $\mathbf{x} = (\mathbf{x}_1 \ \cdots \ \mathbf{x}_n)^T$. If the elements of $\mathbf{x}$ are real, then we say that $\mathbf{x} \in \mathbb{R}^n$; if complex, then $\mathbf{x} \in \mathbb{C}^n$. Under the usual definitions of vector addition and multiplication by a scalar, each of the spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ constitute vector spaces. Considering for instance the set $\mathbb{R}^n$, the assertion follows because the following two properties hold:

$$\begin{aligned} \mathbf{x} \in \mathbb{R}^n \implies a\mathbf{x} \in \mathbb{R}^n \text{ for any real scalar } a, \text{ and} \\ \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^n \implies \mathbf{x} + \mathbf{y} \in \mathbb{R}^n. \end{aligned} \tag{4.1}$$

Under the circumstances described, these two conditions are all that are required for the set in question to be a vector space.

**Section 4.1: Linear subspaces**

An important property of a vector space is that it is closed under vector addition and scalar multiplication. Thus these basic operations, when applied to elements of the space, cannot yield a resulting vector which lies outside of the space.

This closure property of $\mathbb{R}^n$ and $\mathbb{C}^n$ is familiar, and merits no further discussion. Nonetheless, closure as an abstract concept is useful for defining a subspace of a vector space. Specifically, given a vector space $V$, $W$ is a subspace of $V$ if it is contained in $V$ and if it is also closed under vector addition and scalar multiplication. Thus $W$ is a vector space in its own right, but it is defined in the context of the larger space $V$.

A vector space (and therefore a vector subspace) always has a set of linearly independent basis vectors. The number of vectors in this basis is called the rank or dimension of the space, and is equal to the minimum number of scalar descriptors needed to uniquely describe an arbitrary point in the space. Thus if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ form a basis of $\mathbb{R}^n$, then every $\mathbf{x} \in \mathbb{R}^n$ can be written as

$$\mathbf{x} = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n, \tag{4.2}$$

for some set of real coefficients $a_1, \ldots, a_n$. Using the basis, every point in $\mathbb{R}^n$ is therefore associated with a unique set of $n$ real-valued coefficients.

The basis forms a spanning set of the vector space, since the totality of vectors that can be generated from the basis (through the operations of vector addition and scalar multiplication) is exactly the vector space itself. Thus for the real basis $\mathbf{v}_1, \ldots, \mathbf{v}_n$ described earlier, the set

$$\{a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_n | a_1, \ldots, a_n \in \mathbb{R}\} \tag{4.3}$$

is identically the vector space $\mathbb{R}^n$.

Every set of linearly independent vectors spans a linear space. A linear space can therefore be described in terms of a spanning set of vectors. In this work, the space spanned by the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is represented using the notation $\langle \mathbf{v}_1, \ldots, \mathbf{v}_k \rangle$. If $k < n$ and the component vectors are linearly independent, then this space will be a $k$-dimensional subspace of $\mathbb{R}^n$, and will be isomorphic to $\mathbb{R}^k$. In some instances it is convenient to arrange the basis vectors into a matrix

$$\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k \end{pmatrix}, \tag{4.4}$$

in which case the space spanned by the vectors will be denoted by $\langle \mathbf{V} \rangle$. Every vector of the form $\mathbf{Vc}$ can then be written as

$$\mathbf{Vc} = \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} = c_1\mathbf{v}_1 + \cdots + c_k\mathbf{v}_k, \tag{4.5}$$

and is therefore contained in $\langle \mathbf{V} \rangle$. Conversely, every $\mathbf{x} \in \langle \mathbf{V} \rangle$ can be written in the form $\mathbf{Vc}$ for some unique vector $\mathbf{c}$.

From the properties of vector spaces, every space has to contain the origin of the coordinate system, which is the point corresponding to the vector $(0 \ \cdots \ 0)^T$. Thus every subspace must also contain this point. Consider now the three-dimensional Euclidian space $\mathbb{R}^3$. A 1-D subspace of this space must contain the origin, and has to have a spanning set of a single vector. Thus every 1-D subspace of $\mathbb{R}^3$ is comprised of points lying along a straight line through the origin, extending to infinity in both directions. The totality of all such lines is the totality of 1-D subspaces of $\mathbb{R}^3$.

Similarly, the 2-D subspaces of $\mathbb{R}^3$ are identically the set of planes in 3-D space passing through the origin. These planes can have any orientation whatsoever. In general, a $k$-D linear subspace of $\mathbb{R}^n$ can be associated with a $k$-dimensional hyperplane in $\mathbb{R}^n$ which passes through the origin of the coordinate system.

As discussed, every vector space has a linearly independent set of spanning vectors or basis vectors. Nevertheless, certain sets of basis vectors are preferable to others. In particular, when basis vectors are of unit length and mutually orthogonal, then the description of points with respect to the basis is considerably simplified. Such a basis is called an orthonormal basis. The difference between the two types of bases is depicted in Figure 4.1 for the case of a 2-D subspace of $\mathbb{R}^3$. In general there is seldom

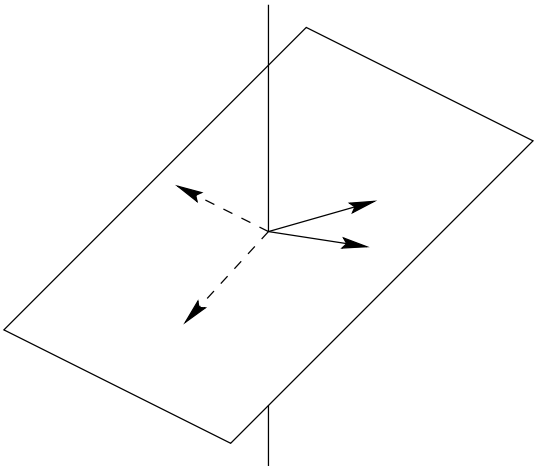


Figure 4.1: Sets of spanning vectors for a 2-D subspace of $\mathbb{R}^3$. The two linearly independent solid vectors span the required subspace, but are not orthogonal. The dashed vectors, however, constitute an orthogonal basis.

any reason to prefer a nonorthogonal basis if an orthogonal one can be found.

It is useful to be able to orthogonally project a given vector into a reduced-rank subspace of the original space. If the subspace is spanned by the columns of the matrix $\mathbf{V}$, such a projection operator is given

by

$$\mathbf{P}_V = \mathbf{V}(\mathbf{V}^T\mathbf{V})^{-1}\mathbf{V}^T. \tag{4.6}$$

Since from the form of this operator $\mathbf{P}_V\mathbf{x} \in \langle\mathbf{V}\rangle$ for all $\mathbf{x}$, it is apparent that the range of this transformation is the required subspace. Additionally, the operator is idempotent, since $\mathbf{P}_V\mathbf{P}_V = \mathbf{P}_V$. These two properties are sufficient to define a projection into $\langle\mathbf{V}\rangle$. That this projection is orthogonal will not be shown — see [119, p. 46] for details. Finally, if the columns of $\mathbf{V}$ constitute an orthogonal basis for the required subspace, then $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, so $\mathbf{P}_V = \mathbf{V}\mathbf{V}^T$ yields the required orthogonal projection.

## 4.2 Detection in subspace interference

This section discusses the problem of detecting a known target in noise which is comprised of two additive components: a random component from a process with a known probability density, and a deterministic interference component which is unknown but constrained to lie within a linear subspace of the observation space. An invariant detector is derived which is optimal in the sense that no other detector which shares the same invariances can provide better performance. A compelling argument can therefore be made that this detector is best within the class of all reasonable tests for the problem. The use of this detector in real situations is outlined, and a form for the test presented which allows the relevant detector components to be calculated.

Stated formally, each noise observation is assumed to be of the form $\mathbf{x}_i = \mathbf{z}_i + \mathbf{U}_I\mathbf{c}_i$, where $\mathbf{z}_i$ is a realisation of some $N$-dimensional random vector with known probability distribution, and $\mathbf{c}_i$ is a *completely* unknown deterministic parameter. The columns of $\mathbf{U}_I$ are assumed to span the interference subspace, and may be taken to be orthogonal and of unit length. The interference parameter $\mathbf{c}_i$ is such that it differs from one observation to the next, and does not depend on the parameter $\mathbf{c}_j$ associated with any other observation $\mathbf{x}_j$. Therefore, if multiple samples $\mathbf{x}_1, \ldots, \mathbf{x}_M$ of noise are observed, then each sample corresponds to a *different* value of $\mathbf{c}_i$, and the elements of the set $\mathbf{c}_1, \ldots, \mathbf{c}_M$ do not depend on one another in any quantifiable manner.

The probability density of the combined noise $\mathbf{x}_i$ depends on the vector $\mathbf{c}_i$, which plays the role of an unknown deterministic parameter in the problem formulation. This parameter specifies the location of the distribution: each value of $\mathbf{c}_i$ corresponds to a probability density function whose location has been shifted by an amount $\mathbf{U}_I\mathbf{c}_i$. For example, if $\mathbf{z}_i$ is zero-mean MVN, then $\mathbf{x}_i$ will have the same covariance structure, but a mean value equal to $\mathbf{U}_I\mathbf{c}_i$.

Figure 4.2 depicts the case of a 1-D interference subspace with 2-D observations. The dashed contours represent the distribution of the random component of the noise. Through the action of the interference, the location of this distribution may be shifted by an arbitrary amount in the direction of the interference vector $\mathbf{i}$. The solid contours therefore represent one possible density function for the noise-plus-

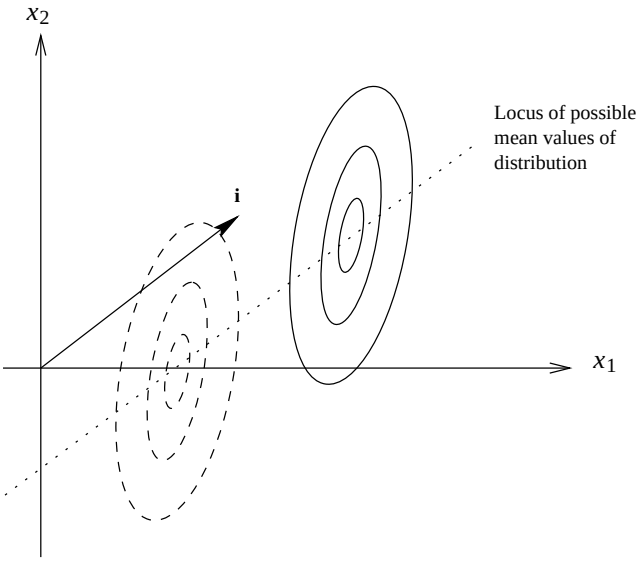



Figure 4.2: Possible probability densities of multivariate normal observations corrupted by subspace interference in the direction of **i**.

interference process. Other allowable densities may be centred anywhere along the broken line in the figure, parallel to the vector **i**, which traces out the total set of allowed locations of the distribution.

Since the noise distribution depends on a deterministic unknown parameter, the likelihood ratio statistic for simple hypotheses cannot be used to provide a test. The generalised likelihood ratio test formalism may be used, however. Maximum likelihood estimates of the unknown parameters under each hypothesis are then used to obtain approximate likelihood functions, which are used to form the generalised likelihood ratio. For example, in the event of a single vector sample being observed, the parameter $\mathbf{c}_1$ may be so estimated under each hypothesis, and used in the GLRT.

In general the process of estimating unknown parameters and using them in a detection statistic is suboptimal. This was clearly demonstrated in the previous chapter. In some instances the procedure may fail altogether, as discussed by Lehmann [87, p. 16]. A more powerful methodology, where applicable, is to require tests to be invariant to the unknown parameters, thereby restricting the class of possible tests. If the invariance condition is natural for the problem, and if a single test in this reduced class can be shown to be consistently better than the others, then some measure of optimality is achieved.

In the two sections which follow, the problem of detecting a target in nominally Gaussian noise is discussed. For completeness and in order to establish notation, the simple case of detection without interference is presented in Section 4.2.1. This is followed, in Section 4.2.2, by the more relevant subject of invariant detection, for the case of the noise containing an additional low-rank interference component. It is argued that a natural symmetry for the problem is obtained by considering all

observations which differ by an additive component in the interference subspace to be equivalent. This leads to the projection of the observation out of the interference subspace as a maximal invariant statistic, which permits the simple derivation of an optimal invariant test.

### 4.2.1 Detection without invariance

In the simplest form, the hypotheses are $\mathbf{z} : N[\mathbf{m}_0, \mathbf{C}]$ under $H_0$ (the target-absent hypothesis) and $\mathbf{z} : N[\mathbf{m}_1, \mathbf{C}]$ under $H_1$ (the target-present hypothesis). The target to be detected is $\mathbf{s} = \mathbf{m}_1 - \mathbf{m}_0$, and it is assumed that $\mathbf{m}_0$, $\mathbf{m}_1$, and $\mathbf{C}$ are known. For the case of real random vectors, the log-likelihood ratio for this problem is [119]

$$L(\mathbf{z}) = (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{C}^{-1}(\mathbf{z} - \mathbf{z}_0), \tag{4.7}$$

where $\mathbf{z}_0 = 1/2(\mathbf{m}_1 + \mathbf{m}_0)$. (For the complex case the log-likelihood ratio is $L(\mathbf{z}) = 2\mathrm{Re}\{(\mathbf{m}_1 - \mathbf{m}_0)^\dagger \mathbf{C}^{-1}(\mathbf{z} - \mathbf{z}_0)\}$, with $\mathbf{z}_0$ defined in the same way.) The optimal test is to compare this to a threshold $\eta$, and select $H_1$ when exceeded: that is, choose $H_1$ when

$$(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{C}^{-1}(\mathbf{z} - \mathbf{z}_0) > \eta. \tag{4.8}$$

This test is optimal in both the Bayes and the Neyman-Pearson sense, with $\eta$ chosen accordingly. Letting $\mathbf{w}^T = (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{C}^{-1}$, the decision inequality can be written as

$$\mathbf{w}^T \mathbf{z} \underset{H_0}{\overset{H_1}{\gtrless}} \eta + \mathbf{w}^T \mathbf{z}_0, \tag{4.9}$$

which shows that the inner product $\mathbf{w}^T \mathbf{z}$ is sufficient for the decision process.

It is simple to calculate $\mathbf{w}$ for this problem: since $\mathbf{w}^T = (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{C}^{-1}$ it can be seen that

$$\mathbf{C}\mathbf{w} = \mathbf{m}_1 - \mathbf{m}_0, \tag{4.10}$$

which is just a linear equation with the coefficient matrix being the covariance $\mathbf{C}$. Since $\mathbf{C}$ is positive semidefinite, appropriate algorithms can reduce the complexity of finding a solution [47]. If $\mathbf{C}$ is further known to have a more specific structure, it may be possible to utilise corresponding low-computational methods. For example, under a stationarity assumption $\mathbf{C}$ is symmetric, positive definite, and Toeplitz, all of which can simplify the calculation of $\mathbf{w}$.

### 4.2.2 Invariant detection for subspace interference

The optimal invariant test is now derived for the problem of detecting a known target $\mathbf{s} = \mathbf{m}_1 - \mathbf{m}_0$ in noise which is essentially MVN but has an additive subspace interference component. It is assumed that the nominal distribution of the random component of the noise is $\mathbf{z} : N[\mathbf{0}, \mathbf{C}]$, and that the interference lies in the subspace spanned by $\mathbf{U}_I$. The interference subspace is initially assumed known, as is the covariance matrix $\mathbf{C}$. The invariant detector for this problem has been developed by Scharf and Friedlander [120], but the derivation is repeated here using a formulation which is more suited to topics addressed in later sections.

The method used in developing the test is that of invariant testing, as presented by Lehmann [87, p. 282] and Scharf [119, p. 127]: a group of transformations is proposed which are deemed natural for the problem, under which any reasonable test should be invariant. It is shown that the hypothesis testing problem is invariant under this group, and a maximal invariant statistic is found. This maximal invariant statistic is invariant to the required group of transformations, while at the same time retaining as much valid information regarding the hypothesis in force as possible. Using this statistic, the most powerful invariant test for the problem is derived. Insofar as the equivalence class implied by the transformation group is reasonable, the resulting test is optimal.

Continuing with the development, it is assumed that the vector $\mathbf{x} = \mathbf{z} + \mathbf{U}_I\mathbf{c}$ is observed. The interference component $\mathbf{U}_I\mathbf{c}$ of this observation depends on the unknown deterministic vector $\mathbf{c}$, and on the known matrix $\mathbf{U}_I$. Under the noise-only hypothesis the distribution of the random component $\mathbf{z}$ is $N[\mathbf{0}, \mathbf{C}]$, and under the target-present hypothesis $\mathbf{z} : N[\mathbf{s}, \mathbf{C}]$. It is required to test between the two hypotheses.

Figure 4.3 demonstrates the structure of the problem for a two-dimensional observation space. Under $H_0$, the mean of the distribution lies somewhere along the line $\mathbf{x} = \mathbf{i}\alpha$, where $\mathbf{i}$ spans the interference subspace. Under $H_1$, the mean lies on the line $\mathbf{x} = \mathbf{s} + \mathbf{i}\alpha$. In both cases the covariance matrix of the observation is $\mathbf{C}$.

Since the observed data vectors have been corrupted by low-rank subspace interference, it is reasonable to require the detector to be invariant to additive contributions within the subspace. This follows since, in accordance with the assumptions made, the component of the observation that lies in $\langle \mathbf{U}_I \rangle$ can take on any value whatsoever. The invariance group is therefore $\{g_{\mathbf{a}} : g_{\mathbf{a}}(\mathbf{x}) = \mathbf{x} + \mathbf{U}_I\mathbf{a}, \mathbf{a} \in \mathbb{R}^p\}$. This places an equivalence on $\mathbf{x}$ and the observation $\mathbf{x} + \mathbf{U}_I\mathbf{a}$, for every choice of $\mathbf{a}$. Enforcing this requirement on the test is not unreasonable, and should in no way compromise the solution. Figure 4.4 demonstrates the equivalences dictated by this transformation group for a simple two-dimensional case.

The partition of the problem into relevant subspaces is best demonstrated by expressing the quantities in the transformed coordinate system spanned by the unitary matrix $(\mathbf{U}_H \;\; \mathbf{U}_I)$, where $\langle \mathbf{U}_I \rangle$ is the invariance subspace and $\mathbf{U}_H$ spans the complementary subspace. Under this representation the observed data are

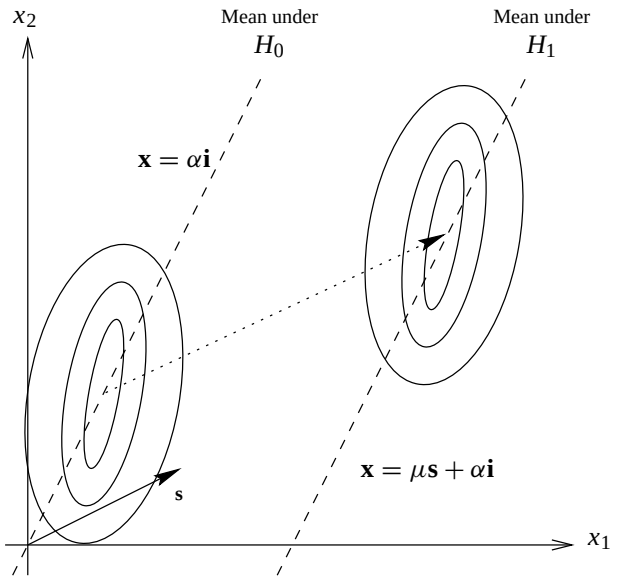



Figure 4.3: Structure of problem for detecting a known target **s** in MVN noise plus subspace interference. The interference is assumed to be in the direction of the vector **i**. The distributions shown are just one instance of many possible combinations.

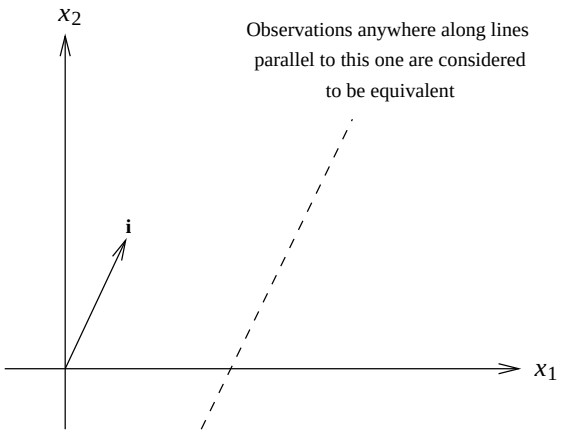



Figure 4.4: Set of equivalent observations under the assumption that subspace interference may be present in the direction of **i**.

$(\mathbf{U}_H \;\; \mathbf{U}_I)^T\mathbf{x}$, so

$$(\mathbf{U}_H \;\; \mathbf{U}_I)^T\mathbf{x} = \begin{pmatrix}\mathbf{U}_H^T\mathbf{x}\\ \mathbf{U}_I^T\mathbf{x}\end{pmatrix} = \begin{pmatrix}\mathbf{U}_H^T(\mathbf{z}+\mathbf{U}_I\mathbf{c})\\ \mathbf{U}_I^T(\mathbf{z}+\mathbf{U}_I\mathbf{c})\end{pmatrix} = \begin{pmatrix}\mathbf{U}_H^T\mathbf{z}\\ \mathbf{U}_I^T\mathbf{z}+\mathbf{c}\end{pmatrix}. \tag{4.11}$$

The first component of this vector describes the data in the subspace $\langle\mathbf{U}_H\rangle$, and the second the data in the complementary subspace $\langle\mathbf{U}_I\rangle$. Now, since $\mathbf{c}$ can take on any value, the final component of the transformed vector can take on any value. There is therefore no useful data contained in this component, and the subspace $\langle\mathbf{U}_I\rangle$ is useless for inference. It makes sense to require that a detector not make use of data contained in this subspace.

The invariant detection problem is readily solved in the original observation space, but more insight is provided if the problem is addressed using the alternative representation just provided. Without changing the decision problem, the unitary transformation $\mathbf{v} = (\mathbf{U}_H \;\; \mathbf{U}_I)^T\mathbf{x}$ is therefore made. In this transformed coordinate system the hypotheses are

$$\begin{aligned} H_0: \quad & \mathbf{v} : N\left[\begin{pmatrix}\mathbf{U}_H^T\mathbf{m}_0\\ \mathbf{U}_I^T\mathbf{m}_0+\mathbf{c}\end{pmatrix}, \begin{pmatrix}\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H & \mathbf{U}_H^T\mathbf{C}\mathbf{U}_I\\ \mathbf{U}_I^T\mathbf{C}\mathbf{U}_H & \mathbf{U}_I^T\mathbf{C}\mathbf{U}_I\end{pmatrix}\right], \text{ and}\\ H_1: \quad & \mathbf{v} : N\left[\begin{pmatrix}\mathbf{U}_H^T\mathbf{m}_1\\ \mathbf{U}_I^T\mathbf{m}_1+\mathbf{c}\end{pmatrix}, \begin{pmatrix}\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H & \mathbf{U}_H^T\mathbf{C}\mathbf{U}_I\\ \mathbf{U}_I^T\mathbf{C}\mathbf{U}_H & \mathbf{U}_I^T\mathbf{C}\mathbf{U}_I\end{pmatrix}\right]. \end{aligned} \tag{4.12}$$

Invariance is required to the unknown parameter $\mathbf{c}$, which can be attained using the group

$$\mathcal{G} = \left\{ g_{\mathbf{a}} : g_{\mathbf{a}}(\mathbf{v}) = \mathbf{v} + \begin{pmatrix}\mathbf{0}\\ \mathbf{a}\end{pmatrix}, \mathbf{a} \in \mathbb{R}^p \right\}. \tag{4.13}$$

This is a reasonable requirement for the testing problem, and corresponds exactly to the previously specified invariance group in the original coordinate system.

The transformed hypothesis testing problem is invariant-$\mathcal{G}$: under $H_0$, for example, the distribution of $g_{\mathbf{a}}(\mathbf{v})$ is

$$\mathbf{v} : N\left[\begin{pmatrix}\mathbf{U}_H^T\mathbf{m}_0\\ \mathbf{U}_I^T\mathbf{m}_0+\mathbf{c}+\mathbf{a}\end{pmatrix}, \begin{pmatrix}\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H & \mathbf{U}_H^T\mathbf{C}\mathbf{U}_I\\ \mathbf{U}_I^T\mathbf{C}\mathbf{U}_H & \mathbf{U}_I^T\mathbf{C}\mathbf{U}_I\end{pmatrix}\right], \tag{4.14}$$

which is still an element of the set of valid distributions of $\mathbf{v}$ under $H_0$. The same assertion can be made for the distribution of the transformed observation under $H_1$.

A maximal invariant statistic for a group $\mathcal{G}$ is a statistic which is invariant to each element of the group, and which has the added property that the mapping of each equivalence class (as defined by the group operation) to the statistic value is one-to-one. As described by Lehmann [87], for purposes of invariant hypothesis testing it is a sufficient statistic. For the problem described here, the maximal invariant is simply the component of the observation which is not associated with the interference subspace.

**Section 4.2: Detection in subspace interference**

That is, letting $\mathbf{v} = (\mathbf{p}^T \;\; \mathbf{q}^T)^T$ with the dimension of $\mathbf{q}$ the same as that of the interference subspace, $M(\mathbf{v}) = \mathbf{p}$ is a maximal invariant:

- **Invariant:**
$$M[g_{\mathbf{a}}(\mathbf{v})] = M\left[g_{\mathbf{a}}\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix}\right] = M\left[\begin{pmatrix}\mathbf{p}\\ \mathbf{q}+\mathbf{a}\end{pmatrix}\right] = \mathbf{p} = M[\mathbf{v}]. \tag{4.15}$$

- **Maximal:**
$$M[\mathbf{v}^{(1)}] = M[\mathbf{v}^{(2)}] \Longrightarrow \mathbf{v}^{(1)} = \mathbf{v}^{(2)}, \tag{4.16}$$

  so
$$\mathbf{v}^{(1)} = \mathbf{v}^{(2)} + \begin{pmatrix}\mathbf{0}\\ \mathbf{a}\end{pmatrix} = g_{\mathbf{a}}(\mathbf{v}^{(2)}) \quad \text{for} \quad g_{\mathbf{a}}(\mathbf{v}) = \mathbf{v} + \begin{pmatrix}\mathbf{0}\\ \mathbf{a}\end{pmatrix} \tag{4.17}$$

  and $\mathbf{a} = \mathbf{q}^{(1)} - \mathbf{q}^{(2)}$.

Since this statistic is sufficient for the invariant hypothesis testing problem, the most powerful invariant decision rule depends only on $M(\mathbf{v}) = \mathbf{U}_H^T\mathbf{x} = \mathbf{U}_H^T\mathbf{z}$. This quantity is $N[\mathbf{U}_H^T\mathbf{m}_0, \mathbf{U}_H^T\mathbf{C}\mathbf{U}_H]$ under $H_0$ and $N[\mathbf{U}_H^T\mathbf{m}_1, \mathbf{U}_H^T\mathbf{C}\mathbf{U}_H]$ under $H_1$. The decision rule resulting from just observing this maximal invariant statistic value is to choose $H_1$ when

$$(\mathbf{U}_H^T\mathbf{m}_1 - \mathbf{U}_H^T\mathbf{m}_0)^T(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}(\mathbf{U}_H^T\mathbf{z} - \mathbf{U}_H^T\mathbf{z}_0) > \eta, \tag{4.18}$$

or equivalently, when

$$(\mathbf{m}_1 - \mathbf{m}_0)^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T(\mathbf{z} - \mathbf{z}_0) > \eta, \tag{4.19}$$

where $\mathbf{z}_0 = 1/2(\mathbf{m}_1 + \mathbf{m}_0)$. This is the LRT for the maximal invariant, and is therefore most powerful in the class of all tests invariant to unknown additive interference in $\langle \mathbf{U}_I \rangle$.

By defining

$$\mathbf{w} = \mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T(\mathbf{m}_1 - \mathbf{m}_0), \tag{4.20}$$

the decision process becomes choosing $H_1$ when

$$\mathbf{w}^T\mathbf{z} > \eta + \mathbf{w}^T\mathbf{z}_0. \tag{4.21}$$

It is less obvious how to find the vector $\mathbf{w}$ for this case than for the noninvariant case. However, the following method provides a solution. Let $\mathbf{U}^T = (\mathbf{U}_H \;\; \mathbf{U}_I)^T$, and let

$$\mathbf{A} = \begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{pmatrix} \tag{4.22}$$

be the inverse of the partitioned matrix

$$\mathbf{C} = \mathbf{U}^T\mathbf{C}\mathbf{U} = \begin{pmatrix} \mathbf{U}_H^T\mathbf{C}\mathbf{U}_H & \mathbf{U}_H^T\mathbf{C}\mathbf{U}_I \\ \mathbf{U}_I^T\mathbf{C}\mathbf{U}_H & \mathbf{U}_I^T\mathbf{C}\mathbf{U}_I \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}. \tag{4.23}$$

Then, since $\mathbf{U}$ is unitary, $\mathbf{U}^{-1} = \mathbf{U}^T$ and $\mathbf{C}^{-1} = \mathbf{U}^T\mathbf{C}^{-1}\mathbf{U}$. Thus

$$\mathbf{A} = \begin{pmatrix} \mathbf{U}_H^T\mathbf{C}^{-1}\mathbf{U}_H & \mathbf{U}_H^T\mathbf{C}^{-1}\mathbf{U}_I \\ \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_H & \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I \end{pmatrix}. \tag{4.24}$$

Expanding the identity $\mathbf{C}\mathbf{A} = \mathbf{I}$ and performing some manipulation on the result, it can be shown that

$$\mathbf{C}_{11}^{-1} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}, \tag{4.25}$$

from which the following relation results:

$$(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1} = \mathbf{U}_H^T\mathbf{C}^{-1}\mathbf{U}_H - \mathbf{U}_H^T\mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I)^{-1}\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_H. \tag{4.26}$$

Noting that $\mathbf{U}_H\mathbf{U}_H^T = \mathbf{I} - \mathbf{U}_I\mathbf{U}_I^T$, we can form the product

$$\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T = (\mathbf{I} - \mathbf{U}_I\mathbf{U}_I^T)\left[\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I)^{-1}\mathbf{U}_I^T\mathbf{C}^{-1}\right](\mathbf{I} - \mathbf{U}_I\mathbf{U}_I^T), \tag{4.27}$$

which simplifies to

$$\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I)^{-1}\mathbf{U}_I^T\mathbf{C}^{-1}. \tag{4.28}$$

Substituting into equation 4.20 gives

$$\mathbf{w} = \mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_0) - \mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I)^{-1}\mathbf{U}_I^T\mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_0), \tag{4.29}$$

which is a function of the terms $\mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_0)$ and $\mathbf{C}^{-1}\mathbf{U}_I$. If $\mathbf{U}_I$ has a small number of columns with respect to $\mathbf{U}_H$, both of these terms can be calculated as solutions to a set of linear equations, eliminating the need to calculate $\mathbf{C}^{-1}$. Note that it is at no stage required to specify $\mathbf{U}_H$ explicitly.

### 4.2.3 Discussion

The hypothesis test in Equation 4.19 has been shown to be optimal in the class of all tests which are invariant to arbitrary additive contributions in the interference subspace. This optimality is absolute in the Neyman-Pearson sense.

**Section 4.2: Detection in subspace interference**

Using Equation 4.28 the test can be written as

$$(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{C}^{-1}(\mathbf{z} - \mathbf{z}_0) - (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T \mathbf{C}^{-1}\mathbf{U}_I)^{-1}\mathbf{U}_I^T \mathbf{C}^{-1}(\mathbf{z} - \mathbf{z}_0) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{4.30}$$

which is expressed instead in terms of the interference subspace matrix $\mathbf{U}_I$. This form is useful for interpreting the role of the interference subspace. Primarily, the matrix $\mathbf{U}_I$ can be replaced by $\mathbf{U}_I\mathbf{B}$ for any invertible $\mathbf{B}$ without changing the test. Since under these conditions $\langle \mathbf{U}_I \rangle = \langle \mathbf{U}_I\mathbf{B} \rangle$, the test depends only on the span of $\mathbf{U}_I$, not on the specific choice. This is expected, since the choice of representation for the interference subspace should not have an effect on the resulting test. Thus the interference component could just as reasonably be represented by $\mathbf{U}_I\mathbf{B}\mathbf{c}_i$, with $\mathbf{c}_i$ completely unknown. Since the maximal invariant simply projects the data out of the interference subspace, the details of the subspace representation have no effect on the ultimate detection.

With regard to invariant hypothesis tests, the equivalences imposed on the observations by the transformation group are paramount in any optimality discussion. If the transformation group is not *absolutely* reasonable for the problem, then even if the best invariant test can be found it cannot be claimed to be optimal: there may exist a test outside of the invariance class which exhibits better performance. For detection in subspace interference, the transformation group implies equivalence between observations which differ only by some component in the interference subspace. This equivalence is only reasonable if *absolutely* nothing is known about the interference component beyond the subspace constraint. This is often an unrealistic assumption — most physical systems are at the very least energy-constrained, and these constraints constitute information regarding the interference contribution. For example, interferences corresponding to excessively large values of $\mathbf{c}$ simply cannot occur in a voltage signal, since they will imply impossible energy outputs. Requiring invariance to subspace interference is then questionable, since the invariance condition is too restrictive to be entirely realistic.

For situations where the equivalence class *is* appropriate, the test given in Equation 4.19 cannot be improved upon. It is most powerful in the class of all tests which share the same set of equivalences. As such, it is the uniformly most powerful invariant (UMPI) test, where the specific invariance group is implied and is taken to be the set $\mathcal{G}$. The optimality is in terms of the Neyman-Pearson criterion. Therefore, for a fixed false alarm rate, the UMPI test has the highest detection probability out of all tests which are invariant-$\mathcal{G}$. Ferguson [34, p. 143] and Berger [10, p. 237] discuss the topic of invariance in a Bayesian context, where invariance of the loss function under the transformation group is also required, but that is not considered here.

For the tests discussed in this section, the only unknown parameters correspond to the contribution of the observation in the interference subspace. Since the maximal invariant projects the observed data out of this space, the distribution under each hypothesis is completely known and is independent of any unknown parameters. It is for this reason that the UMPI test is trivial to find. In general this will not be the case — the target to be detected may be parameterised by unknowns, and the noise contribution may

depend on parameters which are unknown and cannot be eliminated through a reasonable invariance constraint. If subspace interference is still present in the observation, however, the maximal invariant statistic continues to play a pivotal role in the formulation of any subsequent test: it contains more data than any other statistic which is independent of the subspace interference component.

Given a problem where there are unknown parameters in addition to the unknown interference, the procedure for invariant detection is then to find the maximal invariant statistic, calculate its parametric distribution under the hypotheses, and address the resulting problem in its own right. For example, once the maximal invariant has been found, the GLRT may be used on the reduced problem to find a reasonable test. If a test can be found which is optimal for *all* values of the unknown parameters, then once again a UMPI test results. The use of the maximal invariant completely eliminates the problems associated with the interference, and no useful information has been lost through the transformation.

The importance of the maximal invariant statistic applies to hypothesis testing paradigms outside the realm of parametric statistical classifiers. For example, ANN classifiers will also benefit from reductions brought about by invariance arguments, if indeed they are appropriate. The reduced variability in the data resulting from replacing the original data with the maximal invariant translates directly to better learning in the classifier: less of the available capacity of the network is wasted on trying to characterise corrupted data components.

Finally, when used to justify optimality, invariance can only be applied successfully to problems which exhibit a significant degree of symmetry [88, p. 85]. To place an equivalence class on sets of observations requires that the problem be entirely symmetric with regard to the associated transformations. This restricts the applicability of the concept somewhat — most problems do not have sufficient symmetry for invariance concepts to be applied to them without effecting a loss in possible performance. However, for those problems where symmetry does exist, the notion has considerable value.

## 4.3 Interference subspace estimation for invariant detection

In some applications it may be obvious which subspaces can be associated with interference, and are therefore candidates for inclusion in the invariance subspace of an invariant detector. For example, for detection of transients in a voltage signal which is subject to slow drift, it makes sense to require invariance to DC offset at the very least. That is, components of the observation spanned by the vector $(1\ 1\ \cdots\ 1)^T$ should be ignored in the detection.

The presence of interference may be more subtle, however, and a systematic method of identifying interference subspaces from samples of target-free data is desirable. In this section a maximum likelihood formulation is presented for this identification procedure. If this maximum likelihood estimate is plugged into the test statistic for the invariant detector, the resulting test is not longer

optimal. Nevertheless, its performance may still be better than a test which ignores the presence of interference when it is in fact present.

The results of this chapter make the assumption that the actual underlying distribution of the data is MVN, with known parameters. The dimension of the interference subspace is also assumed known. An invariant likelihood formulation is developed, which attempts to find an invariance subspace which has the property that the likelihood of observing the given data in the complementary subspace is maximised.

For all but the simplest formulations, the mathematical details are complicated and do not lead to tractable solutions. Some progress can nevertheless be made in general terms, and leads to insights into the characteristic properties of good invariance subspaces. Additionally, the necessary condition on the interference subspace estimate can be solved in some special instances, which can be used to develop approximate solutions for the general case.

The proposed identification method takes no account of the target to be detected. It is conceivable, therefore, that the estimated interference subspace completely contains a target which it is required to detect. Invariant detection then becomes impossible, since the valid components of the data are identical under both hypotheses. This cannot be considered a failing of the identification method, however: if a target does indeed lie in the interference subspace, then any detector *will* necessarily fail. Other advantages of decoupling the interference subspace estimate from the detection problem are that the noise modelling becomes independent of the target to be detected. This is useful if many different possible targets can be present in the data, and it is required to detect any single instance of one of them.

The assumption that the underlying data distribution is known is restrictive, and limits the utility of the proposed methods in practical problems. However, in Chapter 5 some methods are presented whereby this requirement is relaxed, and interference subspace estimates can be obtained for models which are not completely known.

Section 4.3.1 discusses the effect that subspace interference can have on the sample mean and covariance matrix of corrupted data samples. It is demonstrated that the interference can cause a shifting in the mean value of the sample distribution from its nominal position. Also, the interference can result in an increased variance in the directions associated with the interference subspace.

In Section 4.3.2, the invariant likelihood function is presented, which expresses the likelihood of observing the components of a data sample which lie in the subspace complementary to the interference subspace. It is this invariant likelihood which it is required to maximise, to obtain the maximum likelihood interference subspace estimate. The details of the maximisation are presented, and a necessary condition on the estimate is obtained.

Section 4.3.3 discusses the solutions to the estimation problem for two special cases. In particular, the

case of a white noise assumption on the covariance matrix of the underlying process is considered, and a closed-form solution is provided for the subspace estimate. The case of the covariance matrix being completely unknown is also presented: although this strictly belongs in the next chapter, it is discussed here on account of its similarity to the white noise case.

As mentioned, the necessary condition on the subspace estimate has not been solved for the general case. Section 4.3.4 therefore discusses an approximate procedure for the estimation, which makes use of the solution for the white noise case. In short, the observed data are prewhitened using the assumed known covariance, and the subspace identification performed in this whitened coordinate system. The estimate is then transformed back into the original coordinates, yielding an approximation to the required estimate. The accuracy of this procedure increases as the covariance of the data tends to some scaling of the identity matrix.

Finally, in Section 4.3.6 some results are presented which demonstrate the accuracy of the estimation procedures for simulated data. Some idea is given of the number of data samples required for the estimate to be reasonable.

### 4.3.1 Effect of interference on sample mean and covariance

If the nominal distribution of the data is known or can be accurately assumed, then samples of actual data can be used to find those subspaces which are most likely to contain interference. To this end, it is informative to characterise the manner in which interference manifests itself in the sample mean and sample covariance of the observations. In the remainder of the discussion, the sample covariance of the realisations $\mathbf{x}_1, \ldots, \mathbf{x}_M$ is defined to be

$$\widehat{\mathbf{S}}_{xx} = \frac{1}{M} \sum_{i=1}^{M} (\mathbf{x}_i - \widehat{\mathbf{m}}_x)(\mathbf{x}_i - \widehat{\mathbf{m}}_x)^T, \tag{4.31}$$

where $\widehat{\mathbf{m}}_x$ is the sample mean. In some texts the factor $1/(M-1)$ is used instead of $1/M$, which results in the sample covariance being an unbiased estimator of the distribution covariance, but this is not appropriate in a likelihood context.

The mean of a sample of data with interference is given by

$$\widehat{\mathbf{m}}_x = \frac{1}{M} \sum_{i=1}^{M} \mathbf{x}_i, \tag{4.32}$$

where $\mathbf{x}_i$ is assumed to be an observation of the form

$$\mathbf{x}_i = \mathbf{z}_i + \mathbf{U}_I \mathbf{c}_i, \tag{4.33}$$

and the noise-only hypothesis is in force. In terms of the underlying process, this sample mean is equal to

$$\begin{aligned}\widehat{\mathbf{m}}_x &= \frac{1}{M}\sum_{i=1}^{M}(\mathbf{z}_i + \mathbf{U}_I\mathbf{c}_i) \\ &= \frac{1}{M}\sum_{i=1}^{M}\mathbf{z}_i + \mathbf{U}_I\left(\frac{1}{M}\sum_{i=1}^{M}\mathbf{c}_i\right) \\ &= \widehat{\mathbf{m}}_z + \mathbf{U}_I\widehat{\mathbf{m}}_c,\end{aligned} \tag{4.34}$$

where $\widehat{\mathbf{m}}_z = 1/M\sum_{i=1}^{M}\mathbf{z}_i$ and $\widehat{\mathbf{m}}_c = 1/M\sum_{i=1}^{M}\mathbf{c}_i$. If the mean value of $\mathbf{z}_i$ is assumed to be $\mathbf{m}_0$ and the covariance matrix $\mathbf{C}$, then taking expectations yields

$$E\widehat{\mathbf{m}}_x = \mathbf{m}_0 + \mathbf{U}_I\widehat{\mathbf{m}}_c, \tag{4.35}$$

so the expected value of the sample mean is the distribution mean of the nominal data samples plus an unknown offset in the subspace $\langle\mathbf{U}_I\rangle$.

Using the corrupted data to calculate the sample covariance gives

$$\begin{aligned}\widehat{\mathbf{S}}_{xx} &= \frac{1}{M}\sum_{i=1}^{M}(\mathbf{x}_i - \widehat{\mathbf{m}}_x)(\mathbf{x}_i - \widehat{\mathbf{m}}_x)^T \\ &= \frac{1}{M}\sum_{i=1}^{M}[(\mathbf{z}_i + \mathbf{U}_I\mathbf{c}_i) - (\widehat{\mathbf{m}}_z + \mathbf{U}_I\widehat{\mathbf{m}}_c)][(\mathbf{z}_i + \mathbf{U}_I\mathbf{c}_i) - (\widehat{\mathbf{m}}_z + \mathbf{U}_I\widehat{\mathbf{m}}_c)]^T.\end{aligned} \tag{4.36}$$

Simplifying and noting that $E(\mathbf{z}_i - \widehat{\mathbf{m}}_z) = 0$, the expected value of the sample covariance becomes

$$E\widehat{\mathbf{S}}_{xx} = E\widehat{\mathbf{S}}_{zz} + \mathbf{U}_I\boldsymbol{\beta}\mathbf{U}_I^T. \tag{4.37}$$

Here, $\boldsymbol{\beta} = 1/M\sum_{i=1}^{M}(\mathbf{c}_i - \widehat{\mathbf{m}}_c)(\mathbf{c}_i - \widehat{\mathbf{m}}_c)^T$ is a symmetric positive semidefinite matrix, which apart from this condition can take on any values whatsoever.

Apart from the requirement that it be symmetric and positive semidefinite, the matrix $\boldsymbol{\beta}$ in Equation 4.37 has components which can take on any values. This is because $\boldsymbol{\beta}$ is the sample covariance of the unknown components $\mathbf{c}_1, \ldots, \mathbf{c}_M$. Since nothing is known about the distribution of $\mathbf{c}_i$, the matrix $\boldsymbol{\beta}$ is unknown.

The effect of the interference on the expected value of the sample covariance is therefore simply an

additive offset of the form $\mathbf{U}_I \boldsymbol{\beta} \mathbf{U}_I^T$. Letting

$$\mathbf{U}_I = \begin{pmatrix} \mathbf{i}_1 & \cdots & \mathbf{i}_p \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1p} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pp} \end{pmatrix}, \tag{4.38}$$

the product $\mathbf{U}_I \boldsymbol{\beta} \mathbf{U}_I^T$ expands to

$$\begin{aligned} \mathbf{U}_I \boldsymbol{\beta} \mathbf{U}_I^T = {} & \beta_{11}\mathbf{i}_1\mathbf{i}_1^T + \beta_{21}\mathbf{i}_2\mathbf{i}_1^T + \cdots + \beta_{p1}\mathbf{i}_p\mathbf{i}_1^T \\ & + \beta_{12}\mathbf{i}_1\mathbf{i}_2^T + \beta_{22}\mathbf{i}_2\mathbf{i}_2^T + \ldots + \beta_{p2}\mathbf{i}_p\mathbf{i}_2^T \\ & + \cdots + \beta_{1p}\mathbf{i}_1\mathbf{i}_p^T + \beta_{2p}\mathbf{i}_2\mathbf{i}_p^T + \cdots + \beta_{pp}\mathbf{i}_p\mathbf{i}_p^T, \end{aligned} \tag{4.39}$$

with $\beta_{jk} = \beta_{kj}$ since $\boldsymbol{\beta}$ is symmetric. Also, since $\boldsymbol{\beta}$ is positive semidefinite, the coefficients $\beta_{jj}$ are nonnegative for all $j$.

Consider pairs of terms in this sum with coefficients $\beta_{jk}$ and $\beta_{kj}$:

$$\beta_{jk}\mathbf{i}_j\mathbf{i}_k^T + \beta_{kj}\mathbf{i}_k\mathbf{i}_j^T = \beta_{jk}(\mathbf{i}_j\mathbf{i}_k^T + \mathbf{i}_k\mathbf{i}_j^T). \tag{4.40}$$

This sum (which is an $n \times n$ matrix) is symmetric since

$$(\mathbf{i}_j\mathbf{i}_k^T + \mathbf{i}_k\mathbf{i}_j^T)^T = \mathbf{i}_j\mathbf{i}_k^T + \mathbf{i}_k\mathbf{i}_j^T. \tag{4.41}$$

Rearranging and using symmetry, the expected value of the sample covariance of the corrupted observations is

$$\begin{aligned} E\widehat{\mathbf{S}}_{xx} = {} & E\widehat{\mathbf{S}}_{zz} + \beta_{11}\mathbf{i}_1\mathbf{i}_1^T + \beta_{12}(\mathbf{i}_1\mathbf{i}_2^T + \mathbf{i}_2\mathbf{i}_1^T) + \cdots + \beta_{1p}(\mathbf{i}_1\mathbf{i}_p^T + \mathbf{i}_p\mathbf{i}_1^T) \\ & + \beta_{22}\mathbf{i}_2\mathbf{i}_2^T + \cdots + \beta_{2p}(\mathbf{i}_2\mathbf{i}_p^T + \mathbf{i}_p\mathbf{i}_2^T) \\ & + \cdots + \beta_{pp}\mathbf{i}_p\mathbf{i}_p^T \end{aligned} \tag{4.42}$$

with all the coefficients $\beta_{jk}$ arbitrary within the constraint that $\boldsymbol{\beta}$ be nonnegative definite.

Thus the expected value of the sample covariance with interference is the expected value of the sample covariance without interference, plus a linear combination of terms from the set

$$\left\{ \mathbf{i}_j\mathbf{i}_k^T + \mathbf{i}_k\mathbf{i}_j;\ j,k = 1,2,\ldots,p \right\}, \tag{4.43}$$

where $p$ is the dimension of the interference subspace. Also, the components $\mathbf{i}_j\mathbf{i}_j^T$ must have nonnegative coefficients for all $j$.

By way of example, suppose the interference subspace contributes a constant additive offset with unknown amplitude to the observations, as well as an additive linear trend with unknown slope. The

basis vectors for this subspace can be taken to be the vector of ones and a zero-mean vector representing a linear ramp, both normalised to unit length. These basis vectors are shown in Figure 4.5. Denoting

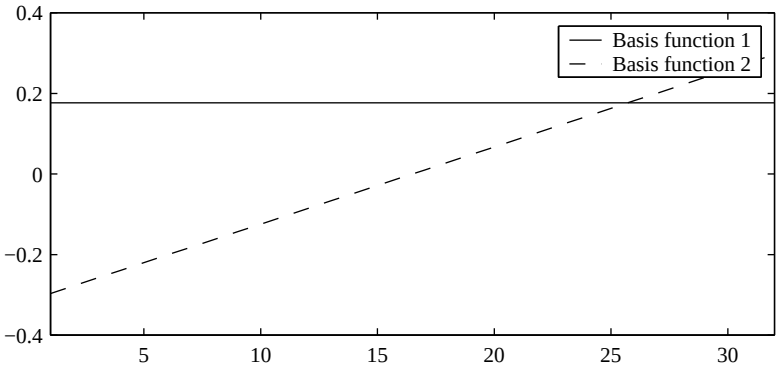



Figure 4.5: Basis functions for example 2-D interference subspace.

this first vector by $\mathbf{i}_1$ and the second by $\mathbf{i}_2$, the presence of interference will shift the location of the sample mean by $c_1\mathbf{i}_1 + c_2\mathbf{i}_2$, with $c_1$ and $c_2$ unknown. The sample covariance matrix will contain additive linear combinations of the form $\beta_{11}\mathbf{i}_1\mathbf{i}_1^T + \beta_{12}(\mathbf{i}_1\mathbf{i}_2^T + \mathbf{i}_2\mathbf{i}_1^T) + \beta_{22}\mathbf{i}_2\mathbf{i}_2^T$. These three components are shown in Figure 4.6. Note that the first and third components must have nonnegative coefficients

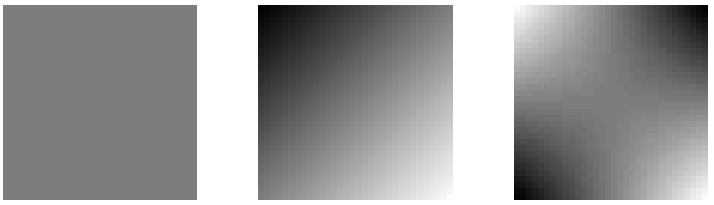


Figure 4.6: Components of interference contributions in the sample covariance matrix for corrupted data. The intensity in these plots is proportional to the value of the components.

in their contribution to the sample covariance. The second can take on any values within the constraint that the coefficient matrix be nonnegative. Since $\mathbf{i}_j^T\widehat{\mathbf{S}}_{xx}\mathbf{i}_j = \mathbf{i}_j^T\widehat{\mathbf{S}}_{zz}\mathbf{i}_j + \beta_{jj}$ is simply the variance of the samples in the direction of the vector $\mathbf{i}_j$, it is clear that the presence of interference can only *increase* the sample covariance in the directions of the basis vectors.

### 4.3.2 Likelihood function with interference

Suppose the nominal distribution of the noise-only data, before interference has been introduced, is $N[\mathbf{m}_0, \mathbf{C}]$ with $\mathbf{m}_0$ and $\mathbf{C}$ known. If the interference lies in the subspace $\langle \mathbf{U}_I \rangle$ with $\mathbf{U}_I^T\mathbf{U}_I = \mathbf{I}$, then the only useful data we have regarding the distribution are samples from the process $\mathbf{y} = \mathbf{U}_H^T\mathbf{z}$, where the columns of $\mathbf{U}_H = (\mathbf{h}_1 \cdots \mathbf{h}_{n-p})$ are orthonormal basis functions for the sample space of useful

data. These are effectively samples from the random process $\mathbf{y} : N[\mathbf{U}_H^T\mathbf{m}_0, \mathbf{U}_H^T\mathbf{C}\mathbf{U}_H]$. This is the distribution of the data that lies outside the interference subspace.

The estimation proceeds by finding the matrix $\widehat{\mathbf{U}}_H$ that maximises the likelihood of the observed data in the subspace $\langle\widehat{\mathbf{U}}_H\rangle$. The interference subspace estimate is then $\langle\widehat{\mathbf{U}}_I\rangle$, chosen so that $\langle\widehat{\mathbf{U}}_I\rangle \perp \langle\widehat{\mathbf{U}}_H\rangle$.

The probability density function for the data $\mathbf{y}_1, \ldots, \mathbf{y}_M$ is

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_M) = (2\pi)^{-M(n-p)/2}|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H|^{-M/2}e^{-1/2\sum_{i=1}^{M}(\mathbf{y}_i-\mathbf{U}_H^T\mathbf{m}_0)^T(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}(\mathbf{y}_i-\mathbf{U}_H^T\mathbf{m}_0)}. \tag{4.44}$$

Thus the log-likelihood is

$$L_y(\mathbf{U}_H) = K - \frac{M}{2}\ln|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H| - \frac{1}{2}\sum_{i=1}^{M}(\mathbf{y}_i - \mathbf{U}_H^T\mathbf{m}_0)^T(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}(\mathbf{y}_i - \mathbf{U}_H^T\mathbf{m}_0) \tag{4.45}$$

which, since $\mathbf{y}_i = \mathbf{U}_H^T\mathbf{z}_i = \mathbf{U}_H^T\mathbf{x}_i$, is equal to

$$\begin{aligned}
L_x(\mathbf{U}_H) &= K - \frac{M}{2}\ln|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H| - \frac{1}{2}\sum_{i=1}^{M}(\mathbf{x}_i - \mathbf{m}_0)^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T(\mathbf{x}_i - \mathbf{m}_0) \\
&= K - \frac{M}{2}\ln|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H| - \frac{M}{2}\frac{1}{M}\sum_{i=1}^{M}\mathrm{tr}\left\{\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H(\mathbf{x}_i - \mathbf{m}_0)(\mathbf{x}_i - \mathbf{m}_0)^T\right\} \\
&= K - \frac{M}{2}\ln|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H| - \frac{M}{2}\mathrm{tr}\left\{\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{S}_{xx}\right\} \\
&= K - \frac{M}{2}\ln|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H| - \frac{M}{2}\mathrm{tr}\left\{(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H\right\}.
\end{aligned} \tag{4.46}$$

The matrix $\mathbf{S}_{xx} = 1/M\sum_{i=1}^{M}(\mathbf{x}_i - \mathbf{m}_0)(\mathbf{x}_i - \mathbf{m}_0)^T$ is the sample covariance of the data under the assumption that the mean is known, and equal to $\mathbf{m}_0$. For maximum likelihood estimation, it is the quantity $L_x$ which has to be maximised with respect to the choice of $\mathbf{U}_H$, subject to the constraint that $\mathbf{U}_H^T\mathbf{U}_H = \mathbf{I}$. Note that this is a function of the corrupted data observations $\mathbf{x}_1, \ldots, \mathbf{x}_M$.

The requirement that $\mathbf{U}_H^T\mathbf{U}_H = \mathbf{I}$ leads to $(n-p) + [(n-p)-1] + \cdots + 1$ unique constraints, $n-p$ of which are (for $i = 1, \ldots, n-p$):

$$g_{ii}(\mathbf{h}_1, \ldots, \mathbf{h}_{n-p}) = \mathbf{h}_i^T\mathbf{h}_i - 1 = 0. \tag{4.47}$$

The remaining constraints can be written as

$$g_{jk}(\mathbf{h}_1, \ldots, \mathbf{h}_{n-p}) = 2\mathbf{h}_j^T\mathbf{h}_k = 0, \tag{4.48}$$

for $j = 1, \ldots, n-p$ and $k = j+1, \ldots, n-p$. The Lagrangian for the problem is given by

$$\mathcal{L} = -\frac{M}{2} \ln |\mathbf{U}_H^T \mathbf{C} \mathbf{U}_H| - \frac{1}{2} \sum_{i=1}^{M} (\mathbf{x}_i - \mathbf{m}_0)^T \mathbf{U}_H (\mathbf{U}_H^T \mathbf{C} \mathbf{U}_H)^{-1} \mathbf{U}_H^T (\mathbf{x}_i - \mathbf{m}_0) + \sum_{j=1}^{n-p} \sum_{k \geq j}^{n-p} \lambda_{jk} g_{jk}(\mathbf{h}_1, \ldots, \mathbf{h}_{n-p}). \quad (4.49)$$

Necessary conditions for an extremum are that $\frac{d\mathcal{L}}{d\mathbf{U}_H}\Big|_{\mathbf{U}_H = \widehat{\mathbf{U}}_H}$ equal the zero matrix. The matrix derivative of this expression can be found using the results presented in Section 4.6, yielding

$$\frac{d\mathcal{L}}{d\mathbf{U}_H} = -M\mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1} - M(\mathbf{I} - \mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T)\mathbf{S}_{xx}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1} + 2\mathbf{U}_H\mathbf{\Gamma}. \quad (4.50)$$

The necessary condition on the interference subspace estimate $\widehat{\mathbf{U}}_H$ is therefore

$$\frac{2}{M}\widehat{\mathbf{U}}_H\mathbf{\Gamma} = \mathbf{C}\widehat{\mathbf{U}}_H(\widehat{\mathbf{U}}_H^T\mathbf{C}\widehat{\mathbf{U}}_H)^{-1} + (\mathbf{I} - \mathbf{C}\widehat{\mathbf{U}}_H(\widehat{\mathbf{U}}_H^T\mathbf{C}\widehat{\mathbf{U}}_H)^{-1}\widehat{\mathbf{U}}_H^T)\mathbf{S}_{xx}\widehat{\mathbf{U}}_H(\widehat{\mathbf{U}}_H^T\mathbf{C}\widehat{\mathbf{U}}_H)^{-1}. \quad (4.51)$$

Evoking the constraint $\widehat{\mathbf{U}}_H^T\widehat{\mathbf{U}}_H = \mathbf{I}$ and solving for $\mathbf{\Gamma}$ gives

$$\mathbf{\Gamma} = \frac{M}{2}\left(\mathbf{I} + (\widehat{\mathbf{U}}_H^T - \widehat{\mathbf{U}}_H^T)\mathbf{S}_{xx}\widehat{\mathbf{U}}_H(\widehat{\mathbf{U}}_H^T\mathbf{C}\widehat{\mathbf{U}}_H)^{-1}\right) = \frac{M}{2}\mathbf{I}. \quad (4.52)$$

Substituting back and postmultiplying by $\widehat{\mathbf{U}}_H^T\mathbf{C}\widehat{\mathbf{U}}_H$ yields the final form of the necessary condition on the estimate $\widehat{\mathbf{U}}_H$ to be

$$\widehat{\mathbf{U}}_H(\widehat{\mathbf{U}}_H^T\mathbf{C}\widehat{\mathbf{U}}_H) = \mathbf{C}\widehat{\mathbf{U}}_H + \mathbf{S}_{xx}\widehat{\mathbf{U}}_H - \mathbf{C}\widehat{\mathbf{U}}_H(\widehat{\mathbf{U}}_H^T\mathbf{C}\widehat{\mathbf{U}}_H)^{-1}\widehat{\mathbf{U}}_H^T\mathbf{S}_{xx}\widehat{\mathbf{U}}_H. \quad (4.53)$$

This is a complicated matrix equation in $\widehat{\mathbf{U}}_H$. A few things can be said about the nature of the solutions:

- If $\widehat{\mathbf{U}}_H$ is a solution and $\mathbf{R}$ is an orthogonal matrix of the appropriate dimension, then $\widehat{\mathbf{U}}_H\mathbf{R}$ is also a solution. It is orthogonal, and its columns span the same space as $\widehat{\mathbf{U}}_H$. Thus all that can be ascertained is the space of the columns of $\widehat{\mathbf{U}}_H$.
- If $\widehat{\mathbf{U}}_H$ is multiplied on the left by an orthogonal matrix $\mathbf{L}$, and $\mathbf{C}$ is replaced by $\mathbf{L}\mathbf{C}\mathbf{L}^T$ and $\mathbf{S}_{xx}$ by $\mathbf{L}\mathbf{S}_{xx}\mathbf{L}^T$, the problem is unchanged. Since $\mathbf{C}$ and $\mathbf{S}_{xx}$ are symmetric, $\mathbf{L}$ can be chosen to diagonalise either of them. Thus either $\mathbf{C}$ or $\mathbf{S}_{xx}$ can be assumed diagonal without loss of generality.

Even with these insights, a closed-form solution to this equation has not been found. Nonetheless, it can be solved for two special cases, which provides some insight into the nature of the solutions.

### 4.3.3 Specific cases of interference subspace estimation

There are two cases where the estimate of $\mathbf{U}_H$ which maximises the likelihood has been found. The first is where the noise model is such that the samples are uncorrelated and identically distributed. The second occurs when no previous assumptions are made regarding the covariance $\mathbf{C}$, but it is instead estimated simultaneously from the observed data samples. It is shown in this section that both cases lead to the same necessary condition on the invariance subspace. A method of finding an orthogonal basis for the interference subspace estimate $\langle \widehat{\mathbf{U}}_I \rangle$ is then presented.

#### Interference subspace estimation for an uncorrelated noise model

The first special case of interest is when the components of the noise vector are uncorrelated and identically distributed. This corresponds to the condition $\mathbf{C} = \sigma^2 \mathbf{I}$, where $\sigma^2$ can be considered a known or unknown constant. The likelihood function to be maximised then becomes (from Equation 4.46)

$$L = K - \frac{M}{2}\left[n \ln \sigma^2 + \mathrm{tr}(\mathbf{U}_H^T \mathbf{S}_{xx} \mathbf{U}_H)\right], \tag{4.54}$$

so a constrained minimisation of the term $\mathrm{tr}\,(\mathbf{U}_H^T \mathbf{S}_{xx} \mathbf{U}_H)$ with respect to $\mathbf{U}_H$ yields the required estimate. Since the entire problem has been formulated in $\mathbb{R}^n$, it is reasonable to require that $\widehat{\mathbf{U}}_H$ be real.

The necessary condition on the maximum likelihood estimate of the retained subspace in the problem therefore becomes (from Equation 4.53)

$$\sigma^2 \widehat{\mathbf{U}}_H = \sigma^2 \widehat{\mathbf{U}}_H + \mathbf{S}_{xx} \widehat{\mathbf{U}}_H - \sigma^2 \widehat{\mathbf{U}}_H (\sigma^2)^{-1} \widehat{\mathbf{U}}_H^T \mathbf{S}_{xx} \widehat{\mathbf{U}}_H, \tag{4.55}$$

which upon simplification yields

$$\widehat{\mathbf{U}}_H \widehat{\mathbf{U}}_H^T \mathbf{S}_{xx} \widehat{\mathbf{U}}_H = \mathbf{S}_{xx} \widehat{\mathbf{U}}_H. \tag{4.56}$$

As with the general condition given in the previous section, the solutions to this equation can also be seen to be invariant to multiplication on the right by any orthogonal matrix $\mathbf{R}$. Thus all that may be ascertained is the space of the columns of $\widehat{\mathbf{U}}_H$.

The condition on $\widehat{\mathbf{U}}_H$ implies that

$$\widehat{\mathbf{U}}_H \widehat{\mathbf{U}}_H^T \mathbf{S}_{xx} \widehat{\mathbf{U}}_H \mathbf{w} = \mathbf{S}_{xx} \widehat{\mathbf{U}}_H \mathbf{w} \tag{4.57}$$

for all vectors $\mathbf{w}$, and the converse is also true. Letting $\mathbf{v} = \widehat{\mathbf{U}}_H \mathbf{w}$, as $\mathbf{w}$ varies, $\mathbf{v}$ takes on all vectors in $\langle \widehat{\mathbf{U}}_H \rangle$. Thus the condition can be expressed as follows: for all $\mathbf{v} \in \langle \widehat{\mathbf{U}}_H \rangle$,

$$\widehat{\mathbf{U}}_H \widehat{\mathbf{U}}_H^T \mathbf{S}_{xx} \mathbf{v} = \mathbf{S}_{xx} \mathbf{v}. \tag{4.58}$$

Additionally, the matrix $\widehat{\mathbf{U}}_H\widehat{\mathbf{U}}_H^T$ is an idempotent operator which projects orthogonally into the space $\langle\widehat{\mathbf{U}}_H\rangle$. Thus the implication is that the vector $\mathbf{S}_{xx}\mathbf{v}$ must lie in the span of $\widehat{\mathbf{U}}_H$.

Bearing in mind that only the space spanned by the columns is specified by the necessary condition, in the search for solutions $\widehat{\mathbf{U}}_H$ to Equation 4.56 one is therefore looking for a subspace $\langle\widehat{\mathbf{U}}_H\rangle$ which has the property that

$$\forall\mathbf{v}\in\langle\widehat{\mathbf{U}}_H\rangle,\quad \mathbf{S}_{xx}\mathbf{v}\in\langle\widehat{\mathbf{U}}_H\rangle \tag{4.59}$$

In other words, the vector subspace $\langle\widehat{\mathbf{U}}_H\rangle$ is mapped into itself by the transformation matrix $\mathbf{S}_{xx}$. Equivalently, $\langle\widehat{\mathbf{U}}_H\rangle$ must be an invariant subspace of the mapping.

Invariant subspaces are discussed by Golub and Van Loan [47, p. 190] and Watkins [129, p. 287], and from a more algebraic perspective by Gastinel [43, p. 183]. In short, a subspace $\mathcal{P}$ is invariant to transformation by the matrix $\mathbf{A}$ if for every $\mathbf{v}\in\mathcal{P}$ the transformed vector $\mathbf{Av}\in\mathcal{P}$. The notion has particular relevance to numerical computation of matrix eigenvalues and certain other decompositions. A useful theorem regarding invariant subspaces is provided and proved by Watkins [129, p. 288]:

> *Let $\mathcal{P}$ be a subspace of $\mathbb{F}^n$ with basis $\mathbf{x}_1,\ldots,\mathbf{x}_k$. Thus $\mathcal{P}=\langle\mathbf{x}_1,\ldots,\mathbf{x}_k\rangle$. Let $\hat{\mathbf{X}}=[\mathbf{x}_1\ \cdots\ \mathbf{x}_k]\in\mathbb{F}^{n\times k}$. Then $\mathcal{P}$ is invariant under $\mathbf{A}\in\mathbb{F}^{n\times n}$ if and only if there exists $\hat{\mathbf{B}}\in\mathbb{F}^{k\times k}$ such that $\mathbf{A}\hat{\mathbf{X}}=\hat{\mathbf{X}}\hat{\mathbf{B}}$.*

Here $\mathbb{F}$ can be any algebraic field, and in particular we may take $\mathbb{F}=\mathbb{R}$ or $\mathbb{F}=\mathbb{C}$.

Returning to the original minimisation problem, we are attempting to find a real matrix $\widehat{\mathbf{U}}_H$ which minimises $\mathrm{tr}(\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H)$ subject to the constraint that $\mathbf{U}_H^T\mathbf{U}_H=\mathbf{I}$. Furthermore, $\langle\widehat{\mathbf{U}}_H\rangle$ must be an invariant subspace of $\mathbf{S}_{xx}$. We proceed by finding the minimum achievable value of the likelihood term, and then find a suitable matrix $\widehat{\mathbf{U}}_H$ which attains this minimum.

Suppose $\mathbf{U}_H\in\mathbb{R}^{n\times(n-p)}$ spans an invariant subspace of $\mathbf{S}_{xx}$, and $\mathbf{U}_H^T\mathbf{U}_H=\mathbf{I}$. Then, by the theorem given earlier, there exists a matrix $\mathbf{B}$ such that $\mathbf{S}_{xx}\mathbf{U}_H=\mathbf{U}_H\mathbf{B}$. $\mathbf{B}$ can be written as $\mathbf{B}=\mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^{-1}$ for some diagonal $\boldsymbol{\Lambda}$ and some invertible $\mathbf{G}$ with unit-length columns. Thus $\mathbf{S}_{xx}\mathbf{U}_H\mathbf{G}=\mathbf{U}_H\mathbf{G}\boldsymbol{\Lambda}$. Letting $\mathbf{F}=\mathbf{U}_H\mathbf{G}$, we note that $\mathbf{F}$ also has unit-length columns. With this definition, the previous expression becomes $\mathbf{S}_{xx}\mathbf{F}=\mathbf{F}\boldsymbol{\Lambda}$, which is recognised as a standard eigenequation. Thus the diagonal elements of $\boldsymbol{\Lambda}$ are eigenvalues of $\mathbf{S}_{xx}$, with the columns of $\mathbf{F}$ being the corresponding eigenvectors. Furthermore, if $\mathbf{S}_{xx}$ is real and symmetric, the eigenvalues and eigenvectors are real.

With this insight, the term $\mathrm{tr}(\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H)$ is again considered. Since the columns of $\mathbf{U}_H$ are orthogonal, $\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H=\mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^{-1}$. Thus

$$\mathrm{tr}(\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H)=\mathrm{tr}(\mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^{-1})=\mathrm{tr}(\boldsymbol{\Lambda}). \tag{4.60}$$

Since the elements of $\boldsymbol{\Lambda}$ are eigenvalues of $\mathbf{S}_{xx}$, this trace can be minimised by choosing $\mathbf{F}$ such that its

columns are eigenvectors corresponding to the $n-p$ smallest eigenvalues. A solution for $\widehat{\mathbf{U}}_H$ can now be found as follows: since the columns of $\mathbf{F}$ can always be orthogonalised, there exists an orthogonal matrix $\mathbf{Q}$ and a matrix $\mathbf{R}$ such that $\mathbf{F} = \mathbf{QR}$ (the QR decomposition is one way of finding such an orthogonal $\mathbf{Q}$). Let $\widehat{\mathbf{U}}_H = \mathbf{QM}$ for $\mathbf{M}$ any $((n-p)\times(n-p))$-dimensional orthogonal matrix. Then

$$\mathrm{tr}(\widehat{\mathbf{U}}_H^T \mathbf{S}_{xx} \widehat{\mathbf{U}}_H) = \mathrm{tr}(\mathbf{M}^T \mathbf{Q}^T \mathbf{S}_{xx} \mathbf{QM}) = \mathrm{tr}(\mathbf{Q}^T \mathbf{S}_{xx} \mathbf{Q}). \tag{4.61}$$

Since $\mathbf{S}_{xx}\mathbf{QR} = \mathbf{QR}\boldsymbol{\Lambda}$, if $\mathbf{R}$ is invertible then $\mathbf{Q}^T\mathbf{S}_{xx}\mathbf{Q} = \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^{-1}$. Thus

$$\mathrm{tr}(\mathbf{Q}^T \mathbf{S}_{xx} \mathbf{Q}) = \mathrm{tr}(\mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^{-1}) = \mathrm{tr}(\boldsymbol{\Lambda}), \tag{4.62}$$

so $\mathrm{tr}(\mathbf{U}_H^T \mathbf{S}_{xx} \mathbf{U}_H)$ attains the minimum at $\mathbf{U}_H = \widehat{\mathbf{U}}_H$.

In summary, then, the required estimate of $\widehat{\mathbf{U}}_H$ may be found as follows:

- Find the eigenvectors of $\mathbf{S}_{xx}$ corresponding to the $(n-p)$ smallest eigenvalues. Arranging these eigenvectors into the columns of the matrix $\mathbf{F}$, we have $\mathbf{S}_{xx}\mathbf{F} = \mathbf{F}\boldsymbol{\Lambda}$.
- Find an orthogonal matrix $\mathbf{Q}$ and a matrix $\mathbf{R}$ such that $\mathbf{F} = \mathbf{QR}$.
- Let $\widehat{\mathbf{U}}_H = \mathbf{QM}$ for any orthogonal $\mathbf{M}$.

The procedure presented here for finding the solution is more general than is required, and characterises the entire class of possible solutions. It is also appropriate in the case where the matrix $\mathbf{S}_{xx}$ is not symmetric, where the eigenvectors are not orthogonal. If just one solution is required and the assumption of symmetry can be made, then in the previous steps we may always choose $\mathbf{R} = \mathbf{M} = \mathbf{I}$. Then the matrix $\widehat{\mathbf{U}}_H$ may simply be chosen to consist of the eigenvectors corresponding to the $(n-p)$ smallest eigenvalues of $\mathbf{S}_{xx}$. $\widehat{\mathbf{U}}_H$ is then orthogonal, as required, and minimises the likelihood term for the problem.

Clearly the solution is not unique — postmultiplication by any unitary matrix (or equivalently, rotation in the subspace spanned by the columns) leaves the likelihood unchanged. If the eigenvalues of $\mathbf{S}_{xx}$ are not unique then even the invariant subspace may not be uniquely specified. In that case there are multiple maximum likelihood estimates, and one may be selected arbitrarily.

#### Interference subspace estimation for unknown model covariance

The second special case occurs when the covariance matrix $\mathbf{C}$ of the model cannot be specified in advance. It is then possible to simultaneously estimate both the covariance and the interference

subspace from the same samples of data. This involves maximising the log-likelihood function

$$L_x = K - \frac{M}{2}\left[\ln|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H| + \mathrm{tr}\left\{(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}(\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H)\right\}\right] \tag{4.63}$$

with respect to both $\mathbf{C}$ and $\mathbf{U}_H$. To some extent this section preempts the discussion of Chapter 5, which deals with covariance matrix estimates in interference. However, the class of solutions is closely related to those presented in the previous section, so the discussion is not out of place at this point.

It is well-known that the conventional maximum likelihood estimate of an unknown covariance matrix is simply the sample covariance of the observations [98, p. 95]. Equivalently, the log-likelihood

$$L = K - \frac{M}{2}\left[\ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})\right] \tag{4.64}$$

is maximised with respect to $\mathbf{C}$ by taking $\mathbf{C} = \mathbf{S}$. Similarly, $L_x$ will be maximised with respect to $\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H$ by taking $\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H = \mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H$. There is more than one matrix $\widehat{\mathbf{C}}$ for which this is true. In fact, letting $\widehat{\mathbf{C}} = \mathbf{S}_{xx} + \mathbf{U}_H\boldsymbol{\beta}\mathbf{U}_H^T$ for any matrix $\boldsymbol{\beta}$ provides a solution. However, as far as the likelihood is concerned these different choices are equivalent, and for simplicity we may choose $\widehat{\mathbf{C}} = \mathbf{S}_{xx}$ as the maximum likelihood estimate.

Using this as the value of $\mathbf{C}$ in the noise model, the log-likelihood in Equation 4.63 becomes

$$L_x = K - \frac{M}{2}\left[\ln|\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H| + \mathrm{tr}\{\mathbf{I}\}\right], \tag{4.65}$$

which can be maximised by minimising the quantity $|\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H|$ subject to the constraint $\mathbf{U}_H^T\mathbf{U}_H = \mathbf{I}$. Using the results of Section 4.6 in a simple Lagrangian formulation, the necessary condition on the maximum likelihood estimate $\widehat{\mathbf{U}}_H$ of the parameter $\mathbf{U}_H$ becomes

$$\widehat{\mathbf{U}}_H\widehat{\mathbf{U}}_H^T\mathbf{S}_{xx}\widehat{\mathbf{U}}_H = \mathbf{S}_{xx}\widehat{\mathbf{U}}_H, \tag{4.66}$$

which is the same as for the uncorrelated noise case discussed previously.

Much of the analysis of the previous section carries over to this problem. In particular, since the condition given by Equation 4.66 is the same as that of Equation 4.56, we can again find an invertible matrix $\mathbf{G}$ with unit length columns and a diagonal matrix $\boldsymbol{\Lambda}$ of eigenvalues of $\mathbf{S}_{xx}$ such that $\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H = \mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^{-1}$. Therefore

$$|\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H| = |\mathbf{G}\boldsymbol{\Lambda}\mathbf{G}^{-1}| = |\mathbf{G}||\boldsymbol{\Lambda}||\mathbf{G}^{-1}| = |\boldsymbol{\Lambda}|. \tag{4.67}$$

Thus, in principle, the minimum value of $|\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H|$ which can be achieved is equal to the minimum product which can be obtained by multiplying $n - p$ of the eigenvalues of $\mathbf{S}_{xx}$ together. In practice, since $\mathbf{S}_{xx}$ is positive semidefinite, the eigenvalues are all nonnegative. Thus the minimum value which can be attained is simply the product of the $n - p$ smallest eigenvalues.

Once again the matrix $\mathbf{F}$, with columns equal to the eigenvectors associated with the smallest eigenvalues, has an orthogonal decomposition $\mathbf{F} = \mathbf{QR}$. Letting $\widehat{\mathbf{U}}_H = \mathbf{QM}$ for any orthogonal $\mathbf{M}$,

$$|\mathbf{U}_H^T \mathbf{S}_{xx} \mathbf{U}_H| = |\mathbf{M}^T \mathbf{Q}^T \mathbf{S}_{xx} \mathbf{QM}| = |\mathbf{Q}^T \mathbf{S}_{xx} \mathbf{Q}| = |\mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^{-1}| = |\boldsymbol{\Lambda}|. \tag{4.68}$$

Thus $|\mathbf{U}_H^T \mathbf{S}_{xx} \mathbf{U}_H|$ attains the minimum at $\mathbf{U}_H = \widehat{\mathbf{U}}_H$.

The maximum likelihood estimate of the interference subspace $\langle \widehat{\mathbf{U}}_I \rangle$ is simply the subspace complementary to $\langle \widehat{\mathbf{U}}_H \rangle$. Now, since the matrix $\mathbf{S}_{xx}$ is symmetric, it has orthogonal eigenvectors. Therefore, the basis functions of the invariance subspace may also be taken directly as the eigenvectors corresponding to the *p largest* eigenvalues of this sample covariance. Thus either $\mathbf{U}_I$ or $\mathbf{U}_H$ may be estimated from the eigenvectors of $\mathbf{S}_{xx}$, and both methods will yield the same result. Once again the choice of invariance subspace representation is only known to within an arbitrary rotation of the vectors comprising the basis functions.

### 4.3.4 Approximate estimate of interference subspace

It has been shown that if the noise samples are independent and identically distributed, then the maximum likelihood estimate of the invariance subspace is that subspace spanned by the eigenvectors of $\mathbf{S}_{xx}$ corresponding to the $p$ largest eigenvalues. A similar result applies when the the covariance matrix is assumed completely unknown, and a maximum likelihood criterion is used to estimate both the covariance and the interference subspace.

When the covariance matrix is known but is not some scaling of the identity matrix, the necessary condition on the interference subspace estimate given by Equation 4.53 is difficult to solve. However, an approximate method may be used to arrive at a solution which may be sufficiently accurate in some instances.

As demonstrated in Section 4.3.3, a closed-form solution to the estimation problem exists for $\mathbf{C} = \sigma^2\mathbf{I}$. The approach taken for the more general covariance case is therefore to prewhiten the observed data *before* finding the interference subspace. Once the subspace has been estimated in this whitened coordinate system, it is transformed back into the original data space. The estimate obtained using this procedure does not maximise the likelihood since the whitening transformation is in general not unitary. Nevertheless, as the covariance of the noise tends towards $\sigma^2\mathbf{I}$ the method will become more accurate.

Suppose the data $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are observed, and $\mathbf{x}_i = \mathbf{z}_i + \mathbf{U}_I \mathbf{c}_i$ for all $i$. In the noise-only case the

distribution of $\mathbf{z}_i$ is $N[\mathbf{m}_0, \mathbf{C}]$. The problem can be whitened by applying the transformation

$$\begin{aligned}\mathbf{w}_i &= \mathbf{C}^{-1/2}\mathbf{x}_i \\ &= \mathbf{C}^{-1/2}\mathbf{z}_i + \mathbf{C}^{-1/2}\mathbf{U}_I\mathbf{c}_i.\end{aligned} \tag{4.69}$$

The distribution of this transformed data is nominally $N[\mathbf{C}^{-1/2}\mathbf{m}_0, \mathbf{I}]$, with interference in the subspace $\langle \mathbf{C}^{-1/2}\mathbf{U}_I \rangle$.

Since the random components of the transformed data are nominally white, the results of Section 4.3.3 apply in this coordinate system. Thus the interference subspace matrix estimate $\widehat{\mathbf{U}}_{I_T}$ in these coordinates can be taken to be comprised of the eigenvectors corresponding to the $p$ largest eigenvectors of the sample covariance $\mathbf{S}_{ww}$. This sample covariance can be written as

$$\begin{aligned}\mathbf{S}_{ww} &= \frac{1}{M}\sum_{i=1}^{M}(\mathbf{w}_i - \mathbf{C}^{-1/2}\mathbf{m}_0)(\mathbf{w}_i - \mathbf{C}^{-1/2}\mathbf{m}_0)^T \\ &= \frac{1}{M}\sum_{i=1}^{M}\mathbf{C}^{-1/2}(\mathbf{x}_i - \mathbf{m}_0)(\mathbf{x}_i - \mathbf{m}_0)^T\mathbf{C}^{-1/2} \\ &= \mathbf{C}^{-1/2}\mathbf{S}_{xx}\mathbf{C}^{-1/2},\end{aligned} \tag{4.70}$$

which are all known quantities.

The interference and complementary subspace estimates $\widehat{\mathbf{U}}_{I_T}$ and $\widehat{\mathbf{U}}_{H_T}$ in the transformed coordinate system can therefore be found. All that remains is to transform the subspaces back into the original coordinate system. The inverse transformation matrix is $\mathbf{C}^{1/2}$, resulting in the invariance subspace estimate $\langle \mathbf{C}^{1/2}\widehat{\mathbf{U}}_{I_T} \rangle$ and the complementary subspace estimate $\langle \mathbf{C}^{1/2}\widehat{\mathbf{U}}_{H_T} \rangle$ in the coordinate system of the original observations. The columns of the matrices $\mathbf{C}^{1/2}\widehat{\mathbf{U}}_{I_T}$ and $\mathbf{C}^{1/2}\widehat{\mathbf{U}}_{H_T}$ are no longer orthogonal. However, they may be trivially orthogonalised, yielding the corresponding estimates $\widehat{\mathbf{U}}_I$ and $\widehat{\mathbf{U}}_H$.

There is a complication which arises when using this approximation technique, on account of orthogonality of vectors not being retained across general linear transformations. Thus, even though the matrices $\widehat{\mathbf{U}}_{I_T}$ and $\widehat{\mathbf{U}}_{H_T}$ in the whitened coordinate system are orthogonal and span disjoint subspaces, the subspace estimates $\langle \widehat{\mathbf{U}}_I \rangle$ and $\langle \widehat{\mathbf{U}}_H \rangle$ in the original coordinate system are not disjoint. Thus it makes a difference whether the estimates in this space are made based on estimates of the interference subspace or of the complementary subspace in the transformed coordinates.

In practice a choice has to be made regarding which of the two estimates to use. If the dimension of the interference subspace is small, it is simplest to just use the estimate of the interference subspace in the transformed problem. Alternatively, a more acceptable solution might be to calculate both possible solutions, and choose the one with the higher likelihood.

### 4.3.5 Discussion

The computational difficulties associated with the proposed method of interference subspace estimation are high. This is partially a result of the way that the estimation problem is formulated. Some justification is therefore required for posing the problem in a missing data framework, rather than just modelling the process directly with unknown parameters.

Specifically, in the literature the subspace interference problem is typically formulated by assuming a covariance matrix of the form $\mathbf{C} + \mathbf{U}_I \boldsymbol{\beta} \mathbf{U}_I$, with $\boldsymbol{\beta}$ an unknown parameter [83, 12, 44]. The assumption of multivariate normality is also made. The subspace interference estimation problem then becomes that of estimating the matrix $\mathbf{U}_I$.

There are a number of reasons why this explicit noise plus interference model is undesirable in this work:

- Firstly, the explicit model makes the assumption that the data are MVN, even in the interference subspace. This is an assumption which limits the utility of the subspace invariance formulation.
- Secondly, the matrix $\boldsymbol{\beta}$ is *not* an unknown deterministic quantity. It does not correspond to any characteristic of the process generating the data, only to the specific interference components in the set of observed data vectors. It is therefore inappropriate to attempt to estimate $\boldsymbol{\beta}$, and thereby obtain an estimate of $\mathbf{U}_I$. The viewpoints of Little and Rubin [89], discussed in more detail in the introduction to Chapter 5, confirm this observation.
- Thirdly, the invariant estimation approach can be used in cases where the interference interpretation is inappropriate. In particular, in Chapter 5 methods are presented for estimation in the presence of model mismatch, rather than interference. The invariant estimation methods can be applied to this alternative problem formulation without any modifications.

A point to note is that the interference subspace estimate presented in the previous section does not make use of the knowledge that the matrix $\boldsymbol{\beta}$ describing the interference components has to be positive semidefinite. There may therefore be a loss of performance with respect to an estimator which explicitly takes this factor into account. On the other hand, since the estimator provided does *not* make use of this knowledge, it is more general. Thus it could provide estimates which are appropriate for scenarios where a strict subspace interference interpretation is overly restrictive.

### 4.3.6 Example of interference subspace estimation

This section demonstrates the use of the proposed interference subspace estimation procedures for certain ideal cases. The results are based on statistical simulations, for which the correct solutions of

the interference subspace estimates are known.

In order to assess the accuracy of a subspace estimate, a figure of merit is required which expresses the extent to which the true subspace $\langle \mathbf{U}_I \rangle$ and the estimated subspace $\langle \widehat{\mathbf{U}}_I \rangle$ are similar. Since different matrices can represent the same subspace, this measure has to be independent of the specific details of the matrices $\mathbf{U}_I$ and $\widehat{\mathbf{U}}_I$ used to represent these subspaces. That is, only the subspaces spanned by the matrices are to be compared, not the matrices themselves.

If $\mathbf{U}_I$ and $\widehat{\mathbf{U}}_I$ are required to have orthogonal columns, one function which is suited to this task is

$$d(\mathbf{U}_I, \widehat{\mathbf{U}}_I) = |\mathbf{U}_I^T \widehat{\mathbf{U}}_I|. \tag{4.71}$$

Firstly, this measure has the property that if $\mathbf{U}_I$ and $\widehat{\mathbf{U}}_I$ are mutually orthogonal (so the subspaces $\langle \mathbf{U}_I \rangle$ and $\langle \widehat{\mathbf{U}}_I \rangle$ are disjoint), then it takes on a value of zero. In that case the two subspaces are completely disparate, and have no vectors in common. This quite naturally corresponds to a condition of complete mismatch of the subspaces. If on the other hand the subspaces are identical, then $\widehat{\mathbf{U}}_I = \mathbf{U}_I \mathbf{R}$ for some orthogonal matrix $\mathbf{R}$. In that case,

$$|\mathbf{U}_I^T \widehat{\mathbf{U}}_I| = |\mathbf{U}_I^T \mathbf{U}_I \mathbf{R}| = |\mathbf{U}_I^T \mathbf{U}_I||\mathbf{R}| = |\mathbf{U}_I^T \mathbf{U}_I|, \tag{4.72}$$

which for orthogonal $\mathbf{U}_I$ has an absolute value identically equal to unity. Thus the measure is zero when the subspaces are completely separate, and positive or negative unity when they are perfectly aligned. The quantity $|d(\mathbf{U}_I^T \widehat{\mathbf{U}}_I)|$ is therefore useful as a direct measure for subspace comparison.

A low value of $|d(\mathbf{U}_I^T \widehat{\mathbf{U}}_I)|$ does not necessarily imply that the subspaces are far apart, only that they do not coincide. For example, in the case of a 3-D invariance subspace, if two of the basis functions are correctly identified but the third is entirely incorrect, then the measure also returns zero. Thus all that is needed for a low subspace match measure is that *one* of the estimated basis functions be entirely incorrect. A high value of $|d(\mathbf{U}_I^T \widehat{\mathbf{U}}_I)|$, however, means that *all* the basis functions have been accurately identified.

Using this criterion to analyse the estimates, results were generated assessing the accuracy of the interference subspace estimation procedures discussed in the previous sections. Since the underlying covariance matrix has to be assumed known, two different cases were considered: one corresponding to a white noise assumption, and the other to a first-order Markov process with a fairly high correlation between samples. The autocorrelation functions for the two processes are shown in Figure 4.7. Each of the samples used were of length 32.

More specifically, the simulation results were generated as follows:

- For each autocorrelation function, $M$ samples $\mathbf{z}_1, \ldots, \mathbf{z}_M$ of zero-mean MVN noise were generated from the corresponding process. Two values of $M$, namely $M = 50$ and $M = 200$, were

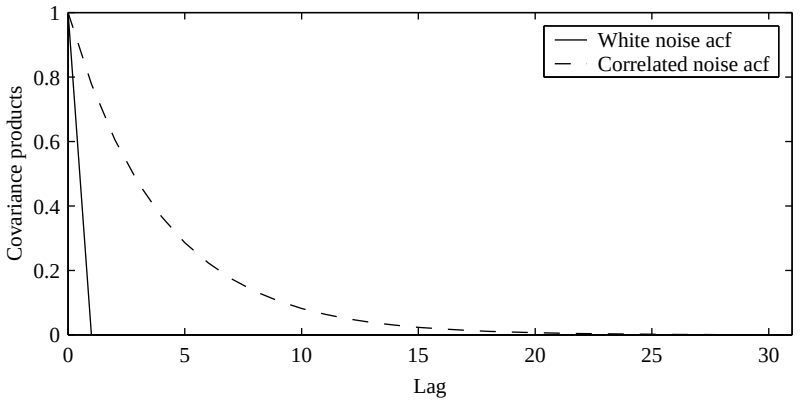


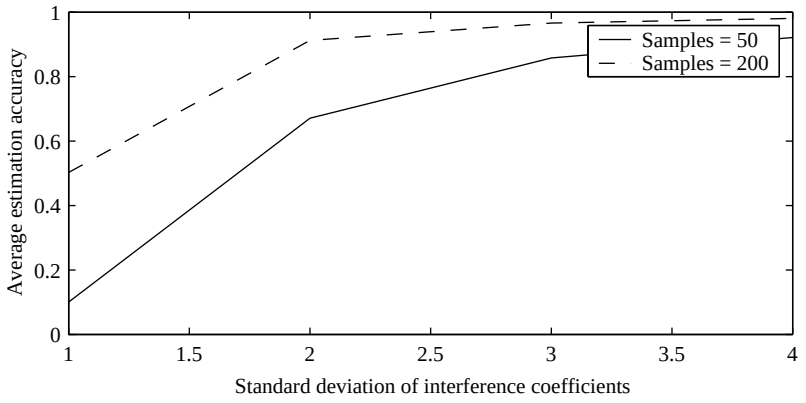

Figure 4.7: Autocorrelation functions for the samples generated in the simulations.

considered.

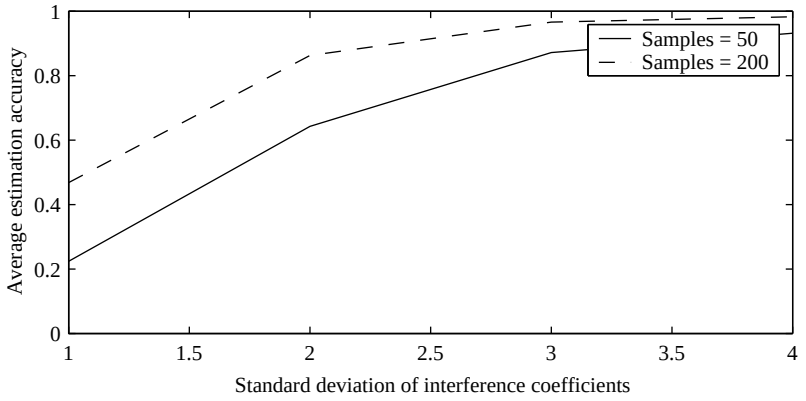
- A $p$-dimensional interference subspace $\langle \mathbf{U}_I \rangle$, with $p = 4$, was chosen at random, and represented by the matrix $\mathbf{U}_I$ with orthogonal columns.
- The generated samples were then corrupted by additive interference in the subspace $\langle \mathbf{U}_I \rangle$. This was done by randomly generating $M$ additional vectors $\mathbf{c}_1, \ldots, \mathbf{c}_M$ from a $N[\mathbf{0}, \sigma^2\mathbf{I}]$ process, and forming the corrupted data observations $\mathbf{x}_1, \ldots, \mathbf{x}_M$ with $\mathbf{x}_i = \mathbf{z}_i + \mathbf{U}_I\mathbf{c}_i$. Four different values of $\sigma$ were chosen for the interference coefficients $\mathbf{c}_i$, namely 1, 2, 3, and 4.
- The sample covariance matrix $\mathbf{S}_{xx}$ of these corrupted observations was then generated, and the methods of Sections 4.3.3 and 4.3.4 used to estimate the interference subspace. The known covariance matrix of the underlying process was used in this estimate. For the white noise case, the exact method of solution was used. For the correlated noise case, the approximate method presented in Section 4.3.4 was used. In this latter case the interference subspace estimate in the whitened coordinate system was transformed back and orthogonalised, yielding the required result.
- Finally, the estimated interference subspace was compared to the true interference subspace, using the figure of merit described earlier.

The procedure just outlined was performed 100 times for each combination of all the parameters, resulting in 100 samples of subspace match figures in each case. Note that each run uses a *different* random interference subspace matrix $\mathbf{U}_I$.

Figure 4.8 shows plots of the average value of these 100 numbers for each of the cases considered. The first plot corresponds to the white noise assumption on the underlying noise process, and the second to the correlated noise assumption. On the $x$-axis of these plots is the standard deviation of




(a) Results for white noise case.




(b) Results for correlated noise case.

Figure 4.8: Accuracy of interference subspace estimation using artificially corrupted ideal data.

the coefficients used for adding the interference to the samples. Thus, on average, large values of the standard deviation correspond to higher levels of added interference. On the $y$-axis is the subspace match measure discussed earlier, which reflects the degree to which the estimated subspace corresponds to the actual one.

A number of conclusions can be drawn from the results provided. Firstly, the interference subspace estimate improves as the number of noise samples increases. Secondly, if the added interference components are large in magnitude, then the estimation is facilitated. These results are precisely as would be expected. Additionally, the estimation procedure for the correlated noise case appears to work well, even though it is based on an approximate solution.

As discussed in Section 4.3.4, the approximate estimation technique can be performed in two ways: in the whitened coordinate system either the interference subspace or the complementary subspace can be estimated, and the result transformed back to the original observation space. At least for the value of $p$ considered in the results, if the experiment is repeated using the estimate of the complementary subspace in the whitened coordinates, then the performance is considerably worse. Thus, at least for small $p$, the direct use of the interference subspace estimate is indicated.

Although the approximate results for correlated noise samples appear to be good, in one regard they are not ideal. This relates to the consistency of the estimate: as the number of available samples from which to estimate the interference subspace increases to infinity, it is desirable that the estimate becomes perfect. Figure 4.9 shows a plot of how the estimates improve as the number of samples increases over the values 50, 500, 5000, and 50000. These results were obtained using $\sigma^2 = 1$ for the

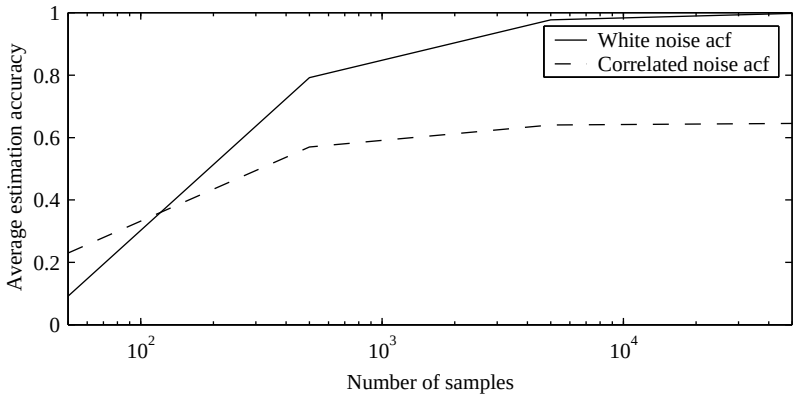



Figure 4.9: Performance of invariance subspace estimate as the number of sample observations increases.

added interference, again with a 4-dimensional interference subspace and 100 runs of the estimation. Apparently the subspace estimate for the white noise case *is* consistent, but this is not true for the correlated case. The reason for this is because the estimate is only approximate in the maximum likelihood sense.

## 4.4 Subspace invariance in general testing

In Section 4.2 a hypothesis test was presented which is optimal for detecting a known target in noise and subspace interference. The test statistic is invariant to additive contributions in the interference subspace, and ignores all components of the data which lie in this subspace. The test is optimal insofar as nothing is known about the interference components. In the previous section a method was then presented for estimating the interference subspace from samples of data, under the assumption that the underlying distribution of the random component of the noise is known.

**Section 4.4: Subspace invariance in general testing**

In this section the use of subspace invariant tests is presented for purposes other than detection in interference. In short, it is argued that such tests may be used in any circumstances where the inclusion of an invariance subspace either significantly improves the modelling accuracy, or substantially simplifies the resulting test. In such cases subspace invariance may not seem to be the most natural restriction, and the resulting tests have no reasonable claim to optimality. However, the formulation may nevertheless be desirable for other reasons:

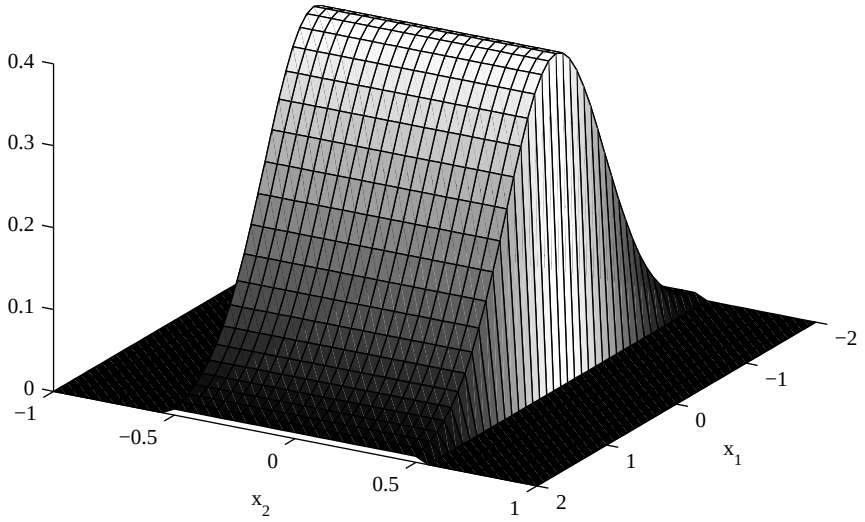
- The invariance condition may result in test statistics which are easier to calculate than when the condition is dropped. Under these circumstances it may be reasonable to find the most powerful test which has a desirable implementation.
- The data contained in some linear subspace may be in conflict with a convenient assumed distribution in the observation space. In this instance it may be simpler to just ignore the subspace where the mismatch is occurring, by requiring invariance to the subspace. The modelling and testing is then restricted to the conveniently-modelled complementary subspace.

In most cases both of these reasons usually play a role in the decision to use an invariant test.

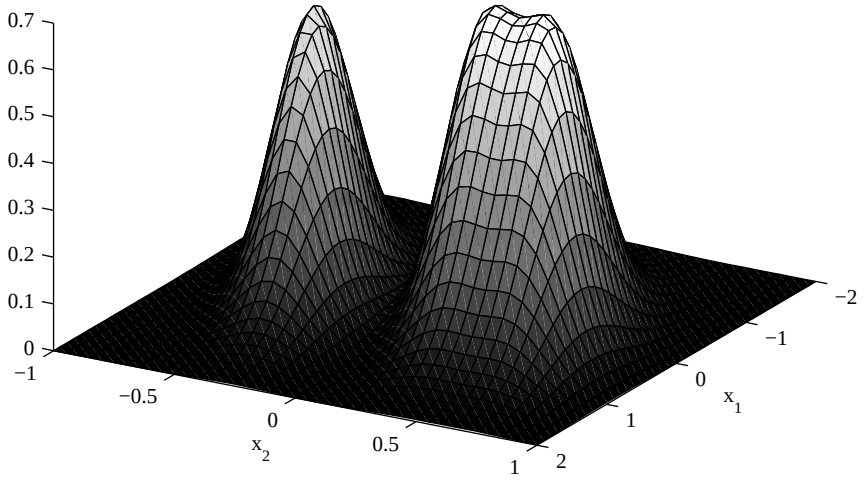
Figure 4.10 shows two hypothetical scenarios where the use of an invariance subspace may simplify detection. In both cases a noise probability density is demonstrated which is normally distributed in one dimension, but where the normal assumption is completely inappropriate in the orthogonal direction. Rather than attempting to characterise the densities accurately, it may be appropriate to simply ignore that component of each observation which does not fit the model. This can be achieved by performing the entire detection procedure on the data value corresponding to the $x_1$-direction, which has a normal distribution.

More generally, it can happen that the data contained in a $p$-dimensional subspace do not conform to an assumption which would otherwise be convenient in the entire $N$-dimensional space. Under the MVN assumption, for example, this could happen when the data contained in the subspace $\langle \mathbf{U}_I \rangle$ are either not normal, or when the disjoint subspaces $\langle \mathbf{U}_H \rangle$ and $\langle \mathbf{U}_I \rangle$ are not jointly normal. It may then be simpler to just ignore the components of the data in $\langle \mathbf{U}_I \rangle$, which would otherwise complicate the statistical description. As long as the target to be detected does not have a considerable component in the invariance subspace, the performance loss may not be too great. What is gained, however, is detection performance which is more predictable than if the modelling mismatch were simply ignored.

The notion of ignoring subspaces which do not conform to simple modelling has even more value when tight constraints are included in the allowable models. Often even the assumption of normality is not sufficient to simplify the parametric models to the point where they are useful or convenient in practice — further constraints may be required to yield an accurate parametric distribution which is effective for the detection. Under a normal assumption, for example, these constraints are brought about by incorporating dependencies into the mean and covariance parameters of the MVN distribution. The




(a) Example 1.




(b) Example 2.

Figure 4.10: Examples of probability density functions where one dimension does not fit a Gaussian assumption. A possible method of dealing with this is to ignore the component of the observation which lies in the $x_2$-direction.

constraints can enhance the convenience and tractability of the model, but also increase the possibility of mismatch between the assumed and the actual distributions. If the mismatches tend to be restricted to a linear subspace of the observation space, however, the use of subspace invariance can simplify the detection problem and improve performance.

In the previous section it was demonstrated that the invariant test in Equation 4.30 is optimal for the detection problem where interference in a known subspace is present in the observation. This is no longer true when invariance is applied in the context suggested here: because the invariance condition cannot be claimed to be natural for the problem, the requirement that a test be invariant involves a restriction which cannot by justified entirely on decision-theoretic grounds. Rather, the justification lies in the convenience of the representation, which in practice may ultimately result in better modelling accuracy. The test is still uniformly most powerful within the class of invariant tests, but this criterion is no longer of any use as an optimality property.

By incorporating invariance into the hypothesis test, the invariance dimensions are *entirely* ignored in the calculation of the statistic. If the target to be detected happens to lie in the invariance subspace, this will lead to a complete failure of the detection procedure. The only recourse then is to improve the accuracy of the statistical model used to characterise the noise. If *part* of the target lies in the ignored subspace, there will be a loss in detectability brought about by the effective reduction in the target energy. This will lead to a degradation in detection performance. However, if the invariance subspace significantly improves the modelling accuracy, an increase in the overall performance may still be effected.

Improved modelling accuracy will result in a test which exhibits performance closer to the design performance. Thus detection and false alarm probabilities will be closer to the values expected from the design stage. This is important in practical detectors. Also, it is easier to have confidence in a system which performs in accordance with predictions, since it reflects a deeper understanding of the processes being modelled.

## 4.5 Invariance subspace estimation for model mismatch

Section 4.3 discussed the problem of estimating the interference subspace from noise samples containing interference. The assumption was made that the distribution of the random component of the observations is multivariate normal, with known mean and covariance. It was demonstrated that interference causes an increase in the variance of the data in the corresponding dimensions. This increased variance can be identified, and estimated interference subspaces assigned accordingly.

However, as suggested in the previous section, the notion of subspace invariance in testing also has application in reducing the mismatch between assumed models and actual data realisations. Identifying

candidate invariance subspaces for this problem is different from that of identification of interference subspaces. The primary reason for this difference arises in that a mismatched model can exhibit variances in certain directions which are greater than *or smaller than* the variances predicted by the model. This cannot occur under the subspace interference interpretation, where only increased variances may be observed.

The likelihood criterion used for interference subspace estimation therefore cannot always be applied sensibly to the problem of model mismatch. A modified criterion is required which penalises both mismatch in terms of variances which are larger than those predicted by the model, as well as those which are smaller.

The solutions to this identification problem depend strongly on the criterion used to assess mismatch. Also, as with the subspace interference problem, most formulations are not tractable and are therefore not useful in practice. One criterion which is reasonable and useful is based on the Frobenius norm, which can be used to measure the distance between covariances matrices of the model and the observations. In this section a method of invariance subspace estimation is presented, based on this criterion. An example of invariance subspace estimation applied to a real-data problem is then presented in Section 4.5.2.

### 4.5.1 Formulation of invariance subspace estimate

Suppose that the assumed model covariance is $\mathbf{C}$, and that the sample covariance of the observations is $\mathbf{S}_{xx}$. As before, this sample covariance is calculated under the assumption that the noise mean under $H_0$ is known, and equal to $\mathbf{m}_0$. $\mathbf{C}$ is again assumed known, and may be chosen to have a form that is convenient from a computational perspective. The approach taken in the invariance subspace identification problem is to find a subspace $\langle \mathbf{U}_I \rangle$ where these two covariance matrices are maximally different. If such a subspace is ignored in subsequent estimation and detection, a reduction in model mismatch will be achieved.

Since the invariance subspace is spanned by $\mathbf{U}_I$, the model covariance in $\langle \mathbf{U}_I \rangle$ is $\mathbf{U}_I^T \mathbf{C} \mathbf{U}_I$. Similarly, the sample covariance in this subspace is $\mathbf{U}_I^T \mathbf{S}_{xx} \mathbf{U}_I$. Now, the Frobenius norm measures the mismatch between two matrices $\mathbf{F}$ and $\mathbf{G}$ as the sum of the squared differences between the corresponding elements. Thus, if $\mathbf{F} = (f_{ij})$ and $\mathbf{G} = (g_{ij})$, then the distance between the matrices in the specified metric is

$$\begin{aligned} d_F(\mathbf{F}, \mathbf{G}) &= \sum_{i=1}^{p} \sum_{j=1}^{p} (f_{ij} - g_{ij})^2 \\ &= \mathrm{tr}(\mathbf{F} - \mathbf{G})^T (\mathbf{F} - \mathbf{G}), \end{aligned} \tag{4.73}$$

where the dimension of the matrices is assumed to be $p \times p$. The distance between the model covariance

and the actual covariance is therefore given by

$$\begin{aligned} d_F(\mathbf{U}_I^T\mathbf{S}_{xx}\mathbf{U}_I, \mathbf{U}_I^T\mathbf{C}\mathbf{U}_I) &= \mathrm{tr}\left\{\mathbf{U}_I^T(\mathbf{S}_{xx}-\mathbf{C})\mathbf{U}_I\mathbf{U}_I^T(\mathbf{S}_{xx}-\mathbf{C})\mathbf{U}_I\right\} \\ &= \mathrm{tr}\left\{\mathbf{U}_I^T\mathbf{A}\mathbf{U}_I\mathbf{U}_I^T\mathbf{A}\mathbf{U}_I\right\}, \end{aligned} \tag{4.74}$$

where use has been made of the symmetry of $\mathbf{C}$ and $\mathbf{S}_{xx}$, and $\mathbf{A} = \mathbf{S}_{xx} - \mathbf{C}$ for convenience.

Under this proposed criterion, a good invariance subspace will be associated with a choice of $\mathbf{U}_I$ where the distance in Equation 4.74 is large. In the remainder of this section it is shown that the distance is maximised when the columns of $\mathbf{U}_I$ are chosen to be those eigenvectors of $\mathbf{A}$ corresponding to the $p$ eigenvalues with largest absolute magnitude. The distance measure chosen depends only on the subspace spanned by the matrix $\mathbf{U}_I$. That is, if $\mathbf{U}_I$ is replaced by any matrix $\mathbf{U}_I\mathbf{R}$ with $\mathbf{R}$ unitary, the distance remains unchanged:

$$\mathrm{tr}\left\{(\mathbf{U}_I\mathbf{R})^T\mathbf{A}(\mathbf{U}_I\mathbf{R})(\mathbf{U}_I\mathbf{R})^T\mathbf{A}(\mathbf{U}_I\mathbf{R})\right\} = \mathrm{tr}\left\{\mathbf{R}^T\mathbf{U}_I^T\mathbf{A}\mathbf{U}_I\mathbf{U}_I^T\mathbf{A}\mathbf{U}_I\mathbf{R}\right\} = \mathrm{tr}\left\{\mathbf{U}_I^T\mathbf{A}\mathbf{U}_I\mathbf{U}_I^T\mathbf{A}\mathbf{U}_I\right\}. \tag{4.75}$$

This is important since it again means that it is only the subspace $\langle\mathbf{U}_I\rangle$ that is important in the maximisation, not the specific details of $\mathbf{U}_I$.

As before, we assume that $\mathbf{U}_H$ is a matrix with orthogonal columns such that $\mathbf{U}_H^T\mathbf{U}_I = \mathbf{I}$. The distance $d_F$ can then be written as

$$\begin{aligned} d_F &= \mathrm{tr}\left\{\mathbf{U}_I^T\mathbf{A}(\mathbf{I}-\mathbf{U}_H\mathbf{U}_H^T)\mathbf{A}\mathbf{U}_I\right\} \\ &= \mathrm{tr}\left\{\mathbf{U}_I^T\mathbf{A}\mathbf{A}\mathbf{U}_I\right\} - \mathrm{tr}\left\{\mathbf{U}_I^T\mathbf{A}\mathbf{U}_H\mathbf{U}_H^T\mathbf{A}\mathbf{U}_I\right\}. \end{aligned} \tag{4.76}$$

Letting $\mathbf{U}_H = (\mathbf{h}_1 \cdots \mathbf{h}_{n-p})$ and $\mathbf{U}_I = (\mathbf{i}_1 \cdots \mathbf{i}_p)$, the second term in this expression is

$$\begin{aligned} \mathrm{tr}\left\{\mathbf{U}_I^T\mathbf{A}\mathbf{U}_H\mathbf{U}_H^T\mathbf{A}\mathbf{U}_I\right\} &= \mathbf{i}_1^T\mathbf{A}\left(\sum_{j=1}^{n-p}\mathbf{h}_j\mathbf{h}_j^T\right)\mathbf{A}\mathbf{i}_1 + \cdots + \mathbf{i}_p^T\mathbf{A}\left(\sum_{j=1}^{n-p}\mathbf{h}_j\mathbf{h}_j^T\right)\mathbf{A}\mathbf{i}_p \\ &= \sum_{k=1}^{p}\sum_{j=1}^{n-p}\mathbf{i}_k^T\mathbf{A}\mathbf{h}_j\mathbf{h}_j^T\mathbf{A}\mathbf{i}_k = \sum_{k=1}^{p}\sum_{j=1}^{n-p}(\mathbf{i}_k^T\mathbf{A}\mathbf{h}_j)^2, \end{aligned} \tag{4.77}$$

where the final result holds on account of the symmetry of $\mathbf{A}$. This result demonstrates that the second term in Equation 4.76 is a nonnegative quantity, irrespective of the choice of $\mathbf{U}_I$ and $\mathbf{U}_H$.

Using the previous observation, an upper bound on the value of $d_F$ under the required orthogonality constraints can be found. Specifically, since the second term in Equation 4.76 is nonnegative, the quantity $d_F$ must always be smaller than or equal to the maximum attainable value of $\mathrm{tr}\left\{\mathbf{U}_I^T\mathbf{A}\mathbf{A}\mathbf{U}_I\right\}$. Under the constraint that $\mathbf{U}_I^T\mathbf{U}_I = \mathbf{I}$, it is possible to calculate this maximum value. Forming the

Lagrangian

$$\mathcal{L} = \text{tr}\left\{\mathbf{U}_I^T \mathbf{A}\mathbf{A}\mathbf{U}_I\right\} + \sum_{j=1}^{p}\sum_{k\geq j}^{p} \lambda_{jk} g_{jk}(\mathbf{i}_1, \ldots, \mathbf{i}_p) \tag{4.78}$$

with the final term representing the orthogonality constraint, it is easy to show that

$$\frac{d\mathcal{L}}{d\mathbf{U}_I} = 2\mathbf{A}\mathbf{A}\mathbf{U}_I + 2\mathbf{U}_I\mathbf{\Gamma}. \tag{4.79}$$

To achieve a maximum, this derivative must be zero when evaluated at $\mathbf{U}_I = \widehat{\mathbf{U}}_I$:

$$2\mathbf{A}\mathbf{A}\widehat{\mathbf{U}}_I + 2\widehat{\mathbf{U}}_I\mathbf{\Gamma} = \mathbf{0}. \tag{4.80}$$

Premultiplying by $\widehat{\mathbf{U}}_I^T$,

$$2\mathbf{\Gamma} = -2\widehat{\mathbf{U}}_I^T\mathbf{A}\mathbf{A}\widehat{\mathbf{U}}_I, \tag{4.81}$$

so a necessary condition on $\widehat{\mathbf{U}}_I$ is that

$$\widehat{\mathbf{U}}_I\widehat{\mathbf{U}}_I^T\mathbf{A}\mathbf{A}\widehat{\mathbf{U}}_I = \mathbf{A}\mathbf{A}\widehat{\mathbf{U}}_I. \tag{4.82}$$

This is exactly the problem discussed in Section 4.3.3. There it was shown that to maximise the required trace, one solution is to choose as the columns of $\widehat{\mathbf{U}}_I$ the eigenvectors corresponding to the $p$ largest eigenvalues. The maximum value attained for $\text{tr}\left\{\mathbf{U}_I^T\mathbf{A}\mathbf{A}\mathbf{U}_I\right\}$ is in that case the sum of the corresponding eigenvalues. Now, since $\mathbf{A}\mathbf{A}$ is the Grammian of $\mathbf{A}$, it is positive semidefinite. Thus the eigenvalues involved in this sum are all nonnegative. Given $\mathbf{A}$, it is easy to calculate the maximum possible value that $d_F$ may have: it is simply the sum of the $p$ largest eigenvalues of $\mathbf{A}\mathbf{A}$.

It is now shown that there exists a choice of $\widehat{\mathbf{U}}_I$ and $\widehat{\mathbf{U}}_H$ for which the distance $d_F$ *attains* the maximum possible value derived in the previous paragraph. Clearly, then, this provides the required solution to the maximisation problem. In particular, we take $\widehat{\mathbf{U}}_I$ to be comprised of eigenvectors corresponding to the $p$ eigenvalues of $\mathbf{A}$ with maximum absolute magnitude. The matrix $\widehat{\mathbf{U}}_H$ is taken to be the $n - p$ remaining eigenvectors, which correspond to eigenvalues with smallest magnitude. Since $\mathbf{A}$ is symmetric, all these eigenvalues are real, although some may be negative. With the preceeding choice of $\widehat{\mathbf{U}}_I$ and $\widehat{\mathbf{U}}_H$, it is clearly true that $\widehat{\mathbf{U}}_I^T\mathbf{A}\widehat{\mathbf{U}}_I = \mathbf{\Lambda}_I$, $\widehat{\mathbf{U}}_H^T\mathbf{A}\widehat{\mathbf{U}}_H = \mathbf{\Lambda}_H$, and $\widehat{\mathbf{U}}_I^T\mathbf{A}\widehat{\mathbf{U}}_H = \mathbf{0}$, which are basic properties of eigenvectors. In these relationships $\mathbf{\Lambda}_I$ is a diagonal matrix containing the $p$ largest absolute eigenvalues of $\mathbf{A}$, and $\mathbf{\Lambda}_H$ is a diagonal matrix containing the smallest.

Returning to the expression in Equation 4.76, under the conditions described, the second term is identically zero since $\widehat{\mathbf{U}}_I^T\mathbf{A}\widehat{\mathbf{U}}_H = \mathbf{0}$. The achieved distance between the matrices is therefore

$$d_F = \text{tr}\left\{\widehat{\mathbf{U}}_I^T\mathbf{A}\mathbf{A}\widehat{\mathbf{U}}_I\right\}. \tag{4.83}$$

This distance can be written as

$$d_F = \mathrm{tr}\left\{(\mathbf{A}\widehat{\mathbf{U}}_I)^T(\mathbf{A}\widehat{\mathbf{U}}_I)\right\} = \mathrm{tr}\left\{(\widehat{\mathbf{U}}_I\mathbf{\Lambda}_I)^T(\widehat{\mathbf{U}}_I\mathbf{\Lambda}_I)\right\} = \mathrm{tr}\left(\mathbf{\Lambda}_I^2\right). \tag{4.84}$$

This is exactly the sum of the squared eigenvalues of $\mathbf{A}$ which have largest magnitude.

Finally, if $\lambda$ is an eigenvalue of $\mathbf{A}$, then $\lambda^2$ is an eigenvalue of $\mathbf{AA}$. Thus if $\{\lambda_1, \ldots, \lambda_p\}$ is the set of eigenvalues of $\mathbf{A}$ with largest magnitude, then $\{\lambda_1^2, \ldots, \lambda_p^2\}$ will be the set of largest eigenvalues of $\mathbf{AA}$. Now, since

$$\mathrm{tr}\left(\mathbf{\Lambda}_I^2\right) = \sum_{j=1}^{p}\lambda_j^2, \tag{4.85}$$

it is evident that $d_F$ *meets* the upper bound on the distance, since the sum of the squared values of the eigenvalues of $\mathbf{A}$ with maximum absolute value is identical to the sum of the maximum eigenvalues of $\mathbf{AA}$. Thus $\widehat{\mathbf{U}}_I$ maximises the distance between the covariance matrices, as required.

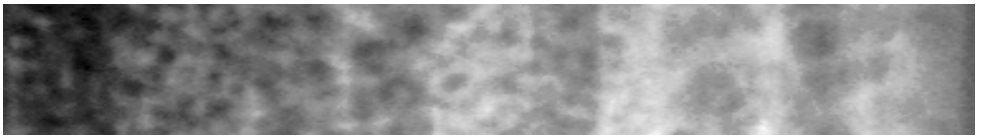
The method discussed can be used to find the invariance subspace which maximises the Frobenius distance between the model and sample covariances within that subspace. It seems, however, that the same method cannot be used to estimate the complementary subspace $\mathbf{U}_H$ through an analogous minimisation of the corresponding distances. For example, if it is required to find a single dimension $\mathbf{h}$ where the mismatch is small, any unit vector $\mathbf{h}$ such that $\mathbf{h}^T\mathbf{C}\mathbf{h} = \mathbf{h}^T\mathbf{S}_{xx}\mathbf{h}$ should be appropriate. Thus $\mathbf{h}$ should be chosen to be one of those directions where the variance of $\mathbf{C}$ and $\mathbf{S}_{xx}$ are equal. In general this will not occur for $\mathbf{h}$ an eigenvector of $\mathbf{S}_{xx} - \mathbf{C}$.

### 4.5.2 Example of invariance subspace estimation

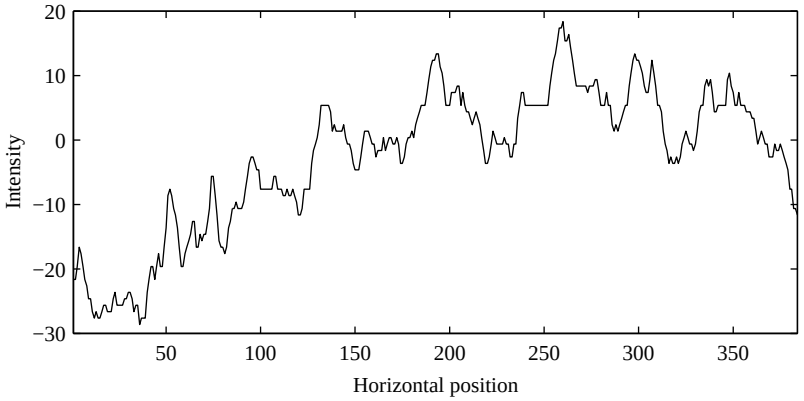
The invariance subspace estimate presented in the previous section can be used whenever the underlying data covariance matrix is assumed known. The invariance subspace then serves to reduce the degree of mismatch between the model and a set of data samples, by identifying those dimensions where the model and sample covariances differ most.

In most cases the underlying covariance matrix is not known. However, for *any* covariance assumption, the method presented will tend to produce an invariance subspace which will enhance the degree of match between the data and the model. For purposes of demonstrating the concept, therefore, the underlying covariance assumption is not critical. In this section a fairly arbitrary covariance assumption is made, and it is demonstrated how a subspace invariance restriction can result in a detector which exhibits both better and more predictable performance than a noninvariant detector.

The test data used for the demonstration is shown in Figure 4.11. It is a portion of a x-ray image of a scale model of a mill, of a type used in the mining industry. Also shown in the figure is a horizontal cut along the first line of the image, clearly demonstrating a nonstationary trend in the data. The problem



(a) Complete image.




(b) Horizontal cut through first line.

Figure 4.11: Test image used in analysis.

discussed here is that of detecting a known target in a horizontal 32 element sample of noise. For the demonstration, noise samples are obtained by simply extracting data from various locations in the image. Note that this procedure is in no direct way related to either the detection of a 2-D target in the image, or the problem of detecting a target with unknown location. The image is simply used to provide 1-D noise samples which are related to one another in a realistic but unknown manner.

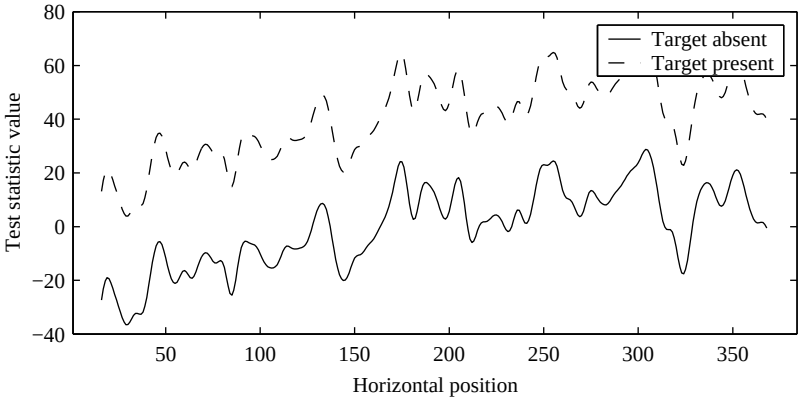
It is assumed that the target to be detected is a transient one-dimensional Gaussian signal with a standard deviation of 3 and an energy of $50 \times 50 = 2500$. The assumption is also made that the nominal distribution of the noise is stationary, with an autocorrelation function

$$r(j) = 15(0.8)^j. \tag{4.86}$$

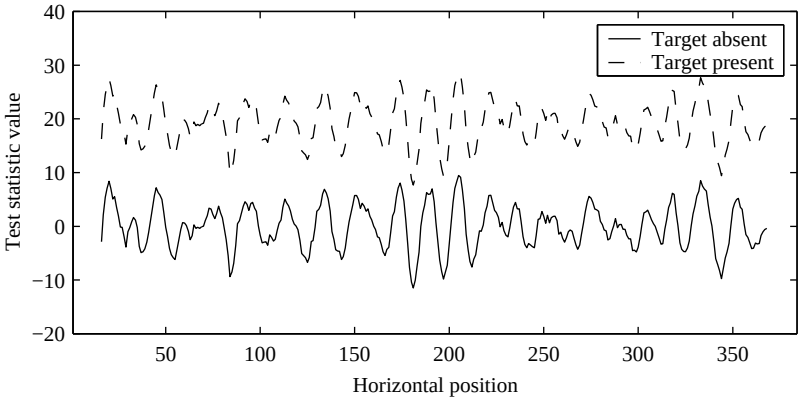
The image was made to have a mean of zero prior to the processing, so the noise in the model is assumed zero mean.

If a detector is developed using this arbitrary assumed model, it will clearly be inaccurate. Nevertheless, such a detector was designed, and applied to a set of overlapping noise samples extracted from the 20th to 30th lines in the test image. This led to a set of sample test statistic values under the $H_0$ assumption.

The solid plot in Figure 4.12(a) shows the values of each of these samples for noise data extracted from the 20th line. Also shown in Figure 4.12(a), and represented by the dashed plot, is the value that the




(a) No invariance subspace.




(b) Invariance subspace of dimension 3.

Figure 4.12: Example of test statistic applied to 20th line of test image.

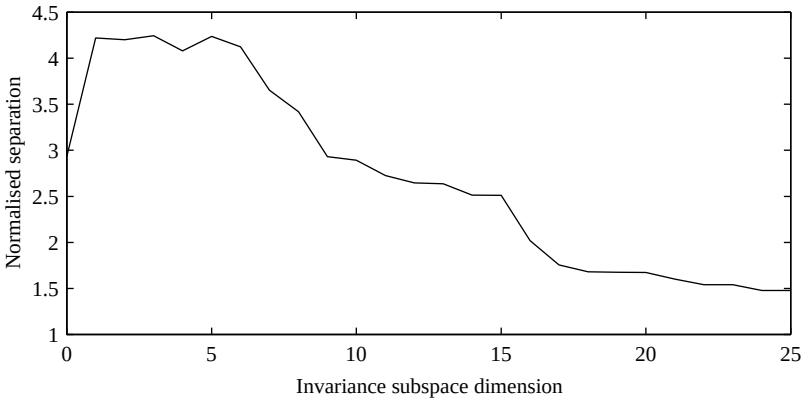
test statistic would take on in the event that there was a target present in each extracted sample of data. If the two sets of test statistic values were well-separated, then comparing the statistic to a threshold would provide a good test. However, in this case the distributions of the values overlap considerably, on account of the inaccuracy of the model used to derive the detector.

Figure 4.12(b) shows a similar result, this time using a detector designed with a 3-dimensional invariance subspace. This invariance subspace was estimated using noise samples obtained from the first 10 rows of the test data. In particular, for each of the 10 lines, a noise sample of length 32 was extracted from horizontal positions separated by a distance of 8 pixels, and a sample covariance matrix calculated. Using this sample covariance and the assumed noise covariance, a 3-dimensional

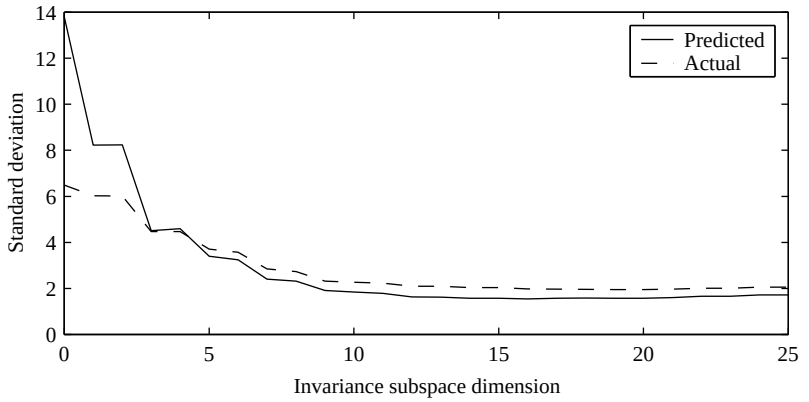
invariance subspace was obtained using the methods of the previous section. This invariance subspace was then used in conjunction with the assumed covariance to develop an invariant detection statistic, which was tested in the same manner described for the noninvariant detector. The invariant detector quite successfully ignores those components of the data which represent the most significant model mismatch. That is, the sets of values taken on under target-absent and target-present conditions are well separated and compact, and a threshold on the invariant test statistic will produce an effective test.

The invariance subspace has reduced the absolute separation between the values taken on between the two hypotheses. For the noninvariant detector this separation is approximately 40, while for the invariant detector it is about 20. However, on account of the better model match, the observed variance of the test statistic is considerably smaller for the invariant detector. If the separation in each case is considered as a fraction of the standard deviation of the noise in the statistic, then it is apparent that this normalised separation is greater for the invariant detector. Figure 4.13 shows this normalisation as a function of the invariance subspace dimension. As the invariance subspace increases, the normalised separation increases initially and then tails off as the dimension of the invariance subspace increases further. The initial rise is caused by the subspace invariance reducing the standard deviation of the noise in the test statistic; the subsequent dropoff is a result of progressively more of the target being nulled by the invariant detector.

Also shown in Figure 4.13 is the predicted and the observed standard deviations of the test statistic when applied to the noise data. The predicted value is obtained using the assumed model distribution, from which the original detectors were derived. It can be seen that as the dimension of the invariance subspace increases, the mismatch between the predicted standard deviation and the observed standard deviation decreases. This quite simply occurs because the invariant model is more accurate for representing the data than the noninvariant model.




(a) Normalised separation between hypotheses.




(b) Predicted and actual standard deviations of test statistic.

Figure 4.13: Detector performance as a function of the invariance subspace dimension.

## 4.6 Appendix: Some vector and matrix calculus results

In order to derive the necessary conditions on the maximum likelihood estimates in this chapter, some detailed vector and matrix calculus results are required. The relevant derivations are presented in this section.

### 4.6.1 $dy/d\boldsymbol{\theta}$ of $y = y(\mathbf{x}(\boldsymbol{\theta}))$:

$$\begin{aligned}\frac{dy}{d\boldsymbol{\theta}} &= \begin{pmatrix} \frac{d}{d\theta_1} y(x_1(\theta_1,\ldots,\theta_m),\ldots,x_n(\theta_1,\ldots,\theta_m)) \\ \vdots \\ \frac{d}{d\theta_m} y(x_1(\theta_1,\ldots,\theta_m),\ldots,x_n(\theta_1,\ldots,\theta_m)) \end{pmatrix} \\ &= \begin{pmatrix} \frac{dy}{dx_1}\frac{dx_1}{d\theta_1} + \cdots + \frac{dy}{dx_n}\frac{dx_n}{d\theta_1} \\ \vdots \\ \frac{dy}{dx_1}\frac{dx_1}{d\theta_m} + \cdots + \frac{dy}{dx_n}\frac{dx_n}{d\theta_m} \end{pmatrix} \\ &= \begin{pmatrix} \frac{dx_1}{d\theta_1} & \cdots & \frac{dx_n}{d\theta_1} \\ \vdots & \ddots & \vdots \\ \frac{dx_1}{d\theta_m} & \cdots & \frac{dx_n}{d\theta_m} \end{pmatrix} \begin{pmatrix} \frac{dy}{dx_1} \\ \vdots \\ \frac{dy}{dx_n} \end{pmatrix} \\ &= \frac{d\mathbf{x}}{d\theta}\frac{dy}{d\mathbf{x}}. \end{aligned} \tag{4.87}$$

### 4.6.2 $dy/d\boldsymbol{\theta}$ of $y = \mathbf{v}^T(\boldsymbol{\theta})\mathbf{w}(\boldsymbol{\theta})$:

Rewriting as $y = v_1(\boldsymbol{\theta})w_1(\boldsymbol{\theta}) + \cdots + v_n(\boldsymbol{\theta})w_n(\boldsymbol{\theta})$, the derivative with respect to $\boldsymbol{\theta}$ is

$$\begin{aligned}\frac{dy}{d\boldsymbol{\theta}} &= \begin{pmatrix} \frac{dy}{d\theta_1} \\ \vdots \\ \frac{dy}{d\theta_m} \end{pmatrix} \\ &= \begin{pmatrix} v_1\frac{dw_1}{d\theta_1} + w_1\frac{dv_1}{d\theta_1} + \cdots + v_n\frac{dw_n}{d\theta_1} + w_n\frac{dv_n}{d\theta_1} \\ \vdots \\ v_1\frac{dw_1}{d\theta_m} + w_1\frac{dv_1}{d\theta_m} + \cdots + v_n\frac{dw_n}{d\theta_m} + w_n\frac{dv_n}{d\theta_m} \end{pmatrix} \\ &= \frac{d\mathbf{w}}{d\boldsymbol{\theta}}\mathbf{v} + \frac{d\mathbf{v}}{d\boldsymbol{\theta}}\mathbf{w}. \end{aligned} \tag{4.88}$$

### 4.6.3 $d\mathbf{v}/d\boldsymbol{\theta}$ of $\mathbf{v} = \mathbf{A}(\boldsymbol{\theta})\mathbf{u}(\boldsymbol{\theta})$:

If $\mathbf{A}$ is the matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{pmatrix},$$

then the $i$th component of $\mathbf{v}$ is given by

$$v_i(\boldsymbol{\theta}) = \mathbf{a}_i^T(\boldsymbol{\theta})\mathbf{u}_i(\boldsymbol{\theta}).$$

By the result in Equation 4.88, the derivative of this quantity is

$$\frac{dv_i}{d\boldsymbol{\theta}} = \frac{d\mathbf{a}_i}{d\boldsymbol{\theta}}\mathbf{u} + \frac{d\mathbf{u}_i}{d\boldsymbol{\theta}}\mathbf{a}_i.$$

Thus

$$\begin{aligned}\frac{d\mathbf{v}}{d\boldsymbol{\theta}} &= \begin{pmatrix} \frac{dv_1}{d\boldsymbol{\theta}} & \cdots & \frac{dv_n}{d\boldsymbol{\theta}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{d\mathbf{a}_1}{d\boldsymbol{\theta}}\mathbf{u} & \cdots & \frac{d\mathbf{a}_n}{d\boldsymbol{\theta}}\mathbf{u} \end{pmatrix} + \frac{d\mathbf{u}}{d\boldsymbol{\theta}} \begin{pmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_n \end{pmatrix} \\ &= \begin{pmatrix} \frac{d\mathbf{a}_1}{d\boldsymbol{\theta}}\mathbf{u} & \cdots & \frac{d\mathbf{a}_n}{d\boldsymbol{\theta}}\mathbf{u} \end{pmatrix} + \frac{d\mathbf{u}}{d\boldsymbol{\theta}}\mathbf{A}^T. \end{aligned} \tag{4.89}$$

Assuming $\mathbf{A}$ is invertible, this in turn yields the result

$$\frac{d\mathbf{u}}{d\boldsymbol{\theta}} = \frac{d\mathbf{v}}{d\boldsymbol{\theta}}\mathbf{A}^{-T} - \begin{pmatrix} \frac{d\mathbf{a}_1}{d\boldsymbol{\theta}}\mathbf{u} & \cdots & \frac{d\mathbf{a}_n}{d\boldsymbol{\theta}}\mathbf{u} \end{pmatrix} \mathbf{A}^{-T}. \tag{4.90}$$

Thus the derivative of $\mathbf{u} = \mathbf{A}^{-1}(\boldsymbol{\theta})\mathbf{v}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is also known.

### 4.6.4 $dy/d\boldsymbol{\theta}$ of $y = \ln|\mathbf{A}(\boldsymbol{\theta})|$:

With $\mathbf{A}$ a matrix as in Section 4.6.3 and $a_{jk} = a_{jk}(\boldsymbol{\theta})$, the $i$th component of $dy/d\boldsymbol{\theta}$ is given by

$$\begin{aligned}\frac{dy}{d\theta_i} &= \frac{1}{|\mathbf{A}(\boldsymbol{\theta})|}\frac{d}{d\theta_i}|\mathbf{A}(\boldsymbol{\theta})| \\ &= \frac{1}{|\mathbf{A}(\boldsymbol{\theta})|}\sum_{j=1}^{n}\sum_{k=1}^{n}\left(\frac{d|\mathbf{A}|}{da_{jk}}\frac{da_{jk}}{d\theta_i}\right). \end{aligned} \tag{4.91}$$

By means of a cofactor expansion along the $j$th row [4, p. 88], $|\mathbf{A}| = \sum_{k=1}^{n} c_{jk}a_{jk}$ for all $j$, where $c_{jk}$ is the $(j,k)$th cofactor of $\mathbf{A}$. Letting

$$\mathbf{C} = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix} \tag{4.92}$$

be the matrix of cofactors from $\mathbf{A}$, the adjoint of $\mathbf{A}$ is given by $\text{adj}(\mathbf{A}) = \mathbf{C}^T$. The inverse of $\mathbf{A}$ can then be written as

$$\mathbf{A}^{-1} = \frac{\text{adj}(\mathbf{A})}{|\mathbf{A}|}, \tag{4.93}$$

and

$$\frac{d|\mathbf{A}|}{da_{jk}} = c_{jk}. \tag{4.94}$$

From Equation 4.91,

$$\begin{aligned}
\frac{dy}{d\theta_i} &= \frac{1}{|\mathbf{A}(\boldsymbol{\theta})|}\sum_{j=1}^{n}\sum_{k=1}^{n} c_{jk}\frac{da_{jk}}{d\theta_i} \\
&= \frac{1}{|\mathbf{A}(\boldsymbol{\theta})|}\mathrm{tr}\left[\begin{pmatrix} c_{11} & \dots & c_{n1} \\ \vdots & \ddots & \vdots \\ c_{1n} & \dots & c_{nn} \end{pmatrix}\begin{pmatrix} \frac{da_{11}}{d\theta_i} & \dots & \frac{da_{1n}}{d\theta_i} \\ \vdots & \ddots & \vdots \\ \frac{da_{n1}}{d\theta_i} & \dots & \frac{da_{nn}}{d\theta_i} \end{pmatrix}\right] \\
&= \frac{1}{|\mathbf{A}(\boldsymbol{\theta})|}\mathrm{tr}\left[\mathbf{C}^T\frac{d\mathbf{A}}{d\theta_i}\right] \\
&= \frac{1}{|\mathbf{A}(\boldsymbol{\theta})|}\mathrm{tr}\left[\mathrm{adj}(\mathbf{A})\frac{d\mathbf{A}}{d\theta_i}\right] \\
&= \mathrm{tr}\left[\frac{\mathrm{adj}(\mathbf{A})}{|\mathbf{A}|}\frac{d\mathbf{A}}{d\theta_i}\right] \\
&= \mathrm{tr}\left[\mathbf{A}^{-1}\frac{d\mathbf{A}}{d\theta_i}\right].
\end{aligned} \tag{4.95}$$

### 4.6.5 $dC/d\mathbf{h}_i$ of $C = \sum_{j=1}^{m}\sum_{k\geq j}^{m}\lambda_{jk}g_{jk}(\mathbf{h}_1,\dots,\mathbf{h}_m)$:

Suppose that

$$g_{jk}(\mathbf{h}_1,\dots,\mathbf{h}_m) = \begin{cases} 2\mathbf{h}_j^T\mathbf{h}_k & (j \neq k) \\ \mathbf{h}_j^T\mathbf{h}_k - 1 & (j = k). \end{cases} \tag{4.96}$$

In effect, $C$ then represents the constraint term in the Lagrangian formulation for maximisation subject to the condition that $\mathbf{U}_H^T\mathbf{U}_H = \mathbf{I}$. Written out explicitly with $\delta_{jk}$ the Kronecker delta,

$$\begin{aligned}
C(\mathbf{h}_1,\dots,\mathbf{h}_m) &= \sum_{j=1}^{m}\sum_{k>j}^{m} 2\lambda_{jk}\mathbf{h}_j^T\mathbf{h}_k + \sum_{j=1}^{m}\lambda_{jj}(\mathbf{h}_j^T\mathbf{h}_j - 1) \\
&= \sum_{j=1}^{m}\sum_{k=1,k\neq j}^{m}\lambda_{jk}\mathbf{h}_j^T\mathbf{h}_k + \sum_{j=1}^{m}\lambda_{jj}(\mathbf{h}_j^T\mathbf{h}_j - 1) \quad \text{with} \quad \lambda_{jk} = \lambda_{kj} \\
&= \sum_{j=1}^{m}\sum_{k=1}^{m}\lambda_{jk}(\mathbf{h}_j^T\mathbf{h}_k - \delta_{jk}) \quad \text{with} \quad \lambda_{jk} = \lambda_{kj}.
\end{aligned} \tag{4.97}$$

**Section 4.6: Appendix: Some vector and matrix calculus results**

The required derivatives are now given by

$$\begin{aligned}\frac{dC}{d\mathbf{h}_i} &= \frac{d}{d\mathbf{h}_i}\left[\sum_{k=1}^{m}\lambda_{ik}(\mathbf{h}_i^T\mathbf{h}_k-\delta_{ik})+\sum_{j=1}^{m}\lambda_{ji}(\mathbf{h}_j^T\mathbf{h}_i-\delta_{ji})-\lambda_{ii}(\mathbf{h}_i^T\mathbf{h}_i-\delta_{ii})\right]\\ &= \frac{d}{d\mathbf{h}_i}\left[\sum_{k=1,k\neq i}^{m}\lambda_{ik}\mathbf{h}_i^T\mathbf{h}_k+\sum_{j=1,j\neq i}^{m}\lambda_{ji}\mathbf{h}_j^T\mathbf{h}_i+\lambda_{ii}(\mathbf{h}_i^T\mathbf{h}_i-1)\right]\\ &= \sum_{k=1,k\neq i}^{m}\lambda_{ik}\mathbf{h}_k+\sum_{j=1,j\neq i}^{m}\lambda_{ji}\mathbf{h}_j+2\lambda_{ii}\mathbf{h}_i\\ &= \sum_{k=1}^{m}\lambda_{ik}\mathbf{h}_k+\sum_{j=1}^{m}\lambda_{ji}\mathbf{h}_j, \end{aligned} \tag{4.98}$$

which, since $\lambda_{jk}=\lambda_{kj}$, is equal to

$$\begin{aligned}\frac{dC}{d\mathbf{h}_i} &= \sum_{k=1}^{m}\lambda_{ik}\mathbf{h}_k+\sum_{j=1}^{m}\lambda_{ij}\mathbf{h}_j\\ &= 2\sum_{k=1}^{m}\lambda_{ik}\mathbf{h}_k\\ &= 2\begin{pmatrix}\mathbf{h}_1 & \cdots & \mathbf{h}_m\end{pmatrix}\begin{pmatrix}\lambda_{i1}\\ \vdots\\ \lambda_{im}\end{pmatrix}\\ &= 2\mathbf{U}_H\boldsymbol{\lambda}_i. \end{aligned} \tag{4.99}$$

Thus letting $\boldsymbol{\Gamma}$ be the symmetric matrix $\boldsymbol{\Gamma}=(\boldsymbol{\lambda}_1\ \cdots\ \boldsymbol{\lambda}_m)$ with $\boldsymbol{\lambda}_i^T=(\lambda_{i1}\ \cdots\ \lambda_{im})$,

$$\frac{dC}{d\mathbf{U}_H}=2\mathbf{U}_H(\lambda_{i1}\ \cdots\ \lambda_{im})=2\mathbf{U}_H\boldsymbol{\Gamma}. \tag{4.100}$$

### 4.6.6 $dy/d\mathbf{U}_H$ of $y=\ln|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H|$:

Suppose $\mathbf{U}_H=(\mathbf{h}_1\ \cdots\ \mathbf{h}_m)$ and

$$\mathbf{h}_j=\begin{pmatrix}h_{j1}\\ \vdots\\ h_{jn}\end{pmatrix}. \tag{4.101}$$

Letting $\mathbf{X}=\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H$, the result in Section 4.6.4 shows that

$$\frac{dy}{dh_{jk}}=\mathrm{tr}\left[\mathbf{X}^{-1}\frac{d\mathbf{X}}{dh_{jk}}\right]. \tag{4.102}$$

Initially, it is therefore required to calculate the derivative with respect to $h_{jk}$ of

$$\mathbf{X} = \begin{pmatrix} \mathbf{h}_1^T\mathbf{C}\mathbf{h}_1 & \cdots & \mathbf{h}_1^T\mathbf{C}\mathbf{h}_m \\ \vdots & \ddots & \vdots \\ \mathbf{h}_m^T\mathbf{C}\mathbf{h}_1 & \cdots & \mathbf{h}_m^T\mathbf{C}\mathbf{h}_m \end{pmatrix}. \tag{4.103}$$

Letting $\mathbf{C}^T = (\mathbf{c}_1 \ \cdots \ \mathbf{c}_m)$, there are four cases of interest which occur in the derivative of the elements $\mathbf{h}_r^T\mathbf{C}\mathbf{h}_s$ of $\mathbf{X}$:

- $j = r,\ j \neq s$ : Letting $\mathbf{y}_s = \mathbf{C}\mathbf{h}_s$, $\frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_r^T\mathbf{C}\mathbf{h}_s = \frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_j^T\mathbf{C}\mathbf{h}_s$. Therefore

$$\frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_r^T\mathbf{C}\mathbf{h}_s = \frac{d}{d\mathbf{h}_{jk}}\begin{pmatrix} h_{j1} & \cdots & h_{jn} \end{pmatrix}\begin{pmatrix} y_{s1} \\ \vdots \\ h_{sn} \end{pmatrix} = y_{nk} = \mathbf{c}_k^T\mathbf{h}_s. \tag{4.104}$$

- $j = s,\ j \neq r$ : $\frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_r^T\mathbf{C}\mathbf{h}_s = \frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_r^T\mathbf{C}\mathbf{h}_j = \mathbf{c}_k^T\mathbf{h}_r$.
- $j = r,\ j = s$ : $\frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_r^T\mathbf{C}\mathbf{h}_s = \frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_j^T\mathbf{C}\mathbf{h}_j$. However $\frac{d}{d\mathbf{h}_j}\mathbf{h}_j^T\mathbf{C}\mathbf{h}_j = 2\mathbf{C}\mathbf{h}_i = 2\mathbf{y}_i$, so

$$\frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_r^T\mathbf{C}\mathbf{h}_s = 2y_{jk} = 2\mathbf{c}_k^T\mathbf{h}_j. \tag{4.105}$$

- $j \neq r,\ j \neq s$ : $\frac{d}{d\mathbf{h}_{jk}}\mathbf{h}_r^T\mathbf{C}\mathbf{h}_s = 0$.

Using these results,

$$\frac{d}{dh_{jk}}\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H = \begin{pmatrix} 0 & \cdots & \mathbf{c}_k^T\mathbf{h}_1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ \mathbf{c}_k^T\mathbf{h}_1 & \cdots & 2\mathbf{c}_k^T\mathbf{h}_j & \cdots & \mathbf{c}_k^T\mathbf{h}_m \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & \mathbf{c}_k^T\mathbf{h}_m & \cdots & 0 \end{pmatrix},$$

where only the $j$th row and column have nonzero elements. This matrix can be written as

$$\begin{aligned} \frac{d}{dh_{jk}}\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H &= \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ \mathbf{c}_k^T\mathbf{h}_1 & \cdots & \mathbf{c}_k^T\mathbf{h}_j & \cdots & \mathbf{c}_k^T\mathbf{h}_m \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} 0 & \cdots & \mathbf{c}_k^T\mathbf{h}_1 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & \mathbf{c}_k^T\mathbf{h}_j & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & \mathbf{c}_k^T\mathbf{h}_m & \cdots & 0 \end{pmatrix} \\ &= \hat{\mathbf{e}}_j\mathbf{c}_k^T\mathbf{U}_H + \mathbf{U}_H^T\mathbf{c}_k\hat{\mathbf{e}}_j^T, \end{aligned} \tag{4.106}$$

where $\hat{e}_j$ is the $j$th unit vector.

Now, letting $\mathbf{A} = \mathbf{X}^{-1} = (\mathbf{a}_1 \ \cdots \ \mathbf{a}_m)$ and assuming $\mathbf{A}$ symmetric,

$$\begin{aligned}\mathbf{X}^{-1}\frac{d\mathbf{X}}{dh_{jk}} = \mathbf{A}\frac{d\mathbf{X}}{dh_{jk}} &= \mathbf{A}\hat{\mathbf{e}}_j\mathbf{c}_k^T\mathbf{U}_H + \mathbf{A}\mathbf{U}_H^T\mathbf{c}_k\hat{\mathbf{e}}_j^T = \mathbf{A}\hat{\mathbf{e}}_j\mathbf{c}_k^T\mathbf{U}_H + \mathbf{A}^T\mathbf{U}_H^T\mathbf{c}_k\hat{\mathbf{e}}_j^T \\ &= \mathbf{a}_j\begin{pmatrix}\mathbf{c}_k^T\mathbf{h}_1 & \cdots & \mathbf{c}_k^T\mathbf{h}_m\end{pmatrix} + \begin{pmatrix}\mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_m^T\end{pmatrix}\begin{pmatrix}\mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_m^T\end{pmatrix}\begin{pmatrix}\mathbf{0} & \cdots & \mathbf{c}_k & \cdots & 0\end{pmatrix} \\ &= \begin{pmatrix}a_{j1}\mathbf{c}_k^T\mathbf{h}_1 & \cdots & a_{j1}\mathbf{c}_k^T\mathbf{h}_m \\ \vdots & \ddots & \vdots \\ a_{jm}\mathbf{c}_k^T\mathbf{h}_1 & \cdots & a_{jm}\mathbf{c}_k^T\mathbf{h}_m\end{pmatrix} + \begin{pmatrix}\mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_m^T\end{pmatrix}\begin{pmatrix}\mathbf{0} & \cdots & \mathbf{h}_1^T\mathbf{c}_k & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{h}_1^T\mathbf{c}_k & \cdots & \mathbf{0}\end{pmatrix},\end{aligned} \tag{4.107}$$

where again the nonzero elements of the second term are in the $j$th column. The expression for the derivative can therefore be written as

$$\begin{aligned}\frac{dy}{dh_{jk}} &= \mathrm{tr}\left[\mathbf{X}^{-1}\frac{d\mathbf{X}}{dh_{jk}}\right] \\ &= a_{j1}\mathbf{c}_k^T\mathbf{h}_1 + \cdots + a_{jm}\mathbf{c}_k^T\mathbf{h}_m + \mathbf{a}_j^T\begin{pmatrix}\mathbf{h}_1^T\mathbf{c}_k \\ \vdots \\ \mathbf{h}_m^T\mathbf{c}_k\end{pmatrix} \\ &= 2\mathbf{a}_j^T\begin{pmatrix}\mathbf{h}_1^T\mathbf{c}_k \\ \vdots \\ \mathbf{h}_m^T\mathbf{c}_k\end{pmatrix} = 2\mathbf{a}_j^T\mathbf{U}_H^T\mathbf{c}_k.\end{aligned} \tag{4.108}$$

Combining this result over all $k$,

$$\begin{pmatrix}\frac{dy}{dh_{j1}} & \cdots & \frac{dy}{dh_{jn}}\end{pmatrix} = 2\mathbf{a}_j^T\mathbf{U}_H^T\begin{pmatrix}\mathbf{c}_1 & \cdots & \mathbf{c}_n)\end{pmatrix} = 2\mathbf{a}_j^T\mathbf{U}_H^T\mathbf{C}, \tag{4.109}$$

which transposed gives $\frac{dy}{d\mathbf{h}_i} = 2\mathbf{C}\mathbf{U}_H\mathbf{a}_j$. Similarly, by rearranging these into a matrix

$$\frac{dy}{d\mathbf{U}_H} = \begin{pmatrix}\frac{dy}{dh_1} & \cdots & \frac{dy}{dh_m}\end{pmatrix} = 2\mathbf{C}\mathbf{U}_H\begin{pmatrix}\mathbf{a}_1 & \cdots & \mathbf{a}_m\end{pmatrix} \tag{4.110}$$

it can be seen that $\frac{d}{d\mathbf{U}_H}\ln|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H| = 2\mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}$ for symmetric $\mathbf{C}$.

### 4.6.7 $d\mathbf{x}_j/d\mathbf{h}_i$ of $\mathbf{x} = \mathbf{U}_H^T\mathbf{C}\mathbf{U}_H$:

Assuming

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m \end{pmatrix}. \tag{4.111}$$

Then $\mathbf{x}_j^T = \mathbf{h}_j^T\mathbf{C}\mathbf{U}_H$, so

$$\mathbf{x}_j = \mathbf{U}_H^T\mathbf{C}\mathbf{h}_j = \begin{pmatrix} \mathbf{h}_1^T\mathbf{C}\mathbf{h}_j \\ \vdots \\ \mathbf{h}_m^T\mathbf{C}\mathbf{h}_j \end{pmatrix}. \tag{4.112}$$

Thus

$$\begin{aligned} \frac{d\mathbf{x}_j}{d\mathbf{h}_i} &= \begin{pmatrix} \frac{dx_{j1}}{d\mathbf{h}_i} & \cdots & \frac{dx_{jm}}{d\mathbf{h}_i} \end{pmatrix} \\ &= \begin{pmatrix} \frac{d}{d\mathbf{h}_i}\mathbf{h}_1^T\mathbf{C}\mathbf{h}_j & \cdots & \frac{d}{d\mathbf{h}_i}\mathbf{h}_m^T\mathbf{C}\mathbf{h}_j \end{pmatrix}. \end{aligned} \tag{4.113}$$

Two cases can now occur:

- **Case 1 (i $\neq$ j):** $\frac{d\mathbf{a}_j}{d\mathbf{h}_i} = (0 \cdots \mathbf{C}\mathbf{h}_j \cdots 0) = \mathbf{C}\mathbf{h}_j\hat{\mathbf{e}}_i^T$, where the nonzero element is in the $i$th position.
- **Case 2 (i = j):** $\frac{d\mathbf{a}_j}{d\mathbf{h}_i} = (\mathbf{C}\mathbf{h}_1 \cdots 2\mathbf{C}\mathbf{h}_i \cdots \mathbf{C}\mathbf{h}_m) = \mathbf{C}\mathbf{h}_j\hat{\mathbf{e}}_i^T + \mathbf{C}\mathbf{U}_H$, where the element with the preceeding factor of 2 is in the $j$th position.

### 4.6.8 $df/d\mathbf{U}_H$ of $f = \mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{y}$:

Let $\mathbf{v} = \mathbf{U}_H^T\mathbf{x}$, $\mathbf{A} = \mathbf{U}_H^T\mathbf{C}\mathbf{U}_H$, $\mathbf{w} = \mathbf{U}_H^T\mathbf{y}$, and $\mathbf{u} = \mathbf{A}^{-1}\mathbf{w}$. Then $f = \mathbf{v}^T\mathbf{u}$. Taking the derivative with respect to the $i$th column of $\mathbf{U}_H$ gives

$$\frac{df}{d\mathbf{h}_i} = \frac{d\mathbf{u}}{d\mathbf{h}_i}\mathbf{v} + \frac{d\mathbf{v}}{d\mathbf{h}_i}\mathbf{u}. \tag{4.114}$$

Now $\frac{d\mathbf{v}}{d\mathbf{h}_i}\mathbf{u} = \mathbf{x}\hat{\mathbf{e}}_i^T\mathbf{u}$ and

$$\begin{aligned} \frac{d\mathbf{u}}{d\mathbf{h}_i}\mathbf{v} &= \left[\frac{d\mathbf{w}}{d\mathbf{h}_i} - \begin{pmatrix} \frac{d\mathbf{a}_1}{d\mathbf{h}_i}\mathbf{u} & \cdots & \frac{d\mathbf{a}_m}{d\mathbf{h}_i}\mathbf{u} \end{pmatrix}\right]\mathbf{A}^{-T}\mathbf{v} \\ &= \mathbf{y}\hat{\mathbf{e}}_i^T\mathbf{A}^{-T}\mathbf{v} - \begin{pmatrix} \mathbf{C}\mathbf{h}_1\hat{\mathbf{e}}_i^T\mathbf{u} & \cdots & (\mathbf{C}\mathbf{h}_i\hat{\mathbf{e}}_i^T + \mathbf{C}\mathbf{U}_H)\mathbf{u} & \cdots & \mathbf{C}\mathbf{h}_m\hat{\mathbf{e}}_i^T\mathbf{u} \end{pmatrix}\mathbf{A}^{-T}\mathbf{v} \\ &= \mathbf{y}\hat{\mathbf{e}}_i^T\mathbf{A}^{-T}\mathbf{v} - \begin{pmatrix} \mathbf{C}\mathbf{h}_1\hat{\mathbf{e}}_i^T\mathbf{u} & \cdots & \mathbf{C}\mathbf{h}_m\hat{\mathbf{e}}_i^T\mathbf{u} \end{pmatrix}\mathbf{A}^{-T}\mathbf{v} - \mathbf{C}\mathbf{U}_H\mathbf{u}\hat{\mathbf{e}}_i^T\mathbf{A}^{-T}\mathbf{v}, \end{aligned} \tag{4.115}$$

so

$$\frac{df}{d\mathbf{h}_i} = \mathbf{x}\hat{\mathbf{e}}_i^T\mathbf{u} + \mathbf{y}\hat{\mathbf{e}}_i^T\mathbf{A}^{-T}\mathbf{v} - \mathbf{C}\mathbf{U}_H\mathbf{u}\hat{\mathbf{e}}_i^T\mathbf{A}^{-T}\mathbf{v} - \begin{pmatrix}\mathbf{C}\mathbf{h}_1\hat{\mathbf{e}}_i^T\mathbf{u} & \cdots & \mathbf{C}\mathbf{h}_m\hat{\mathbf{e}}_i^T\mathbf{u}\end{pmatrix}\mathbf{A}^{-T}\mathbf{v}. \tag{4.116}$$

But

$$\begin{aligned}\begin{pmatrix}\mathbf{C}\mathbf{h}_1\hat{\mathbf{e}}_i^T\mathbf{u} & \cdots & \mathbf{C}\mathbf{h}_m\hat{\mathbf{e}}_i^T\mathbf{u}\end{pmatrix}\mathbf{A}^{-T}\mathbf{v} &= (\hat{\mathbf{e}}_i^T\mathbf{u})\begin{pmatrix}\mathbf{C}\mathbf{h}_1 & \cdots & \mathbf{C}\mathbf{h}_m\end{pmatrix}\mathbf{A}^{-T}\mathbf{v} \\ &= (\hat{\mathbf{e}}_i^T\mathbf{u})\mathbf{C}\mathbf{U}_H\mathbf{A}^{-T}\mathbf{v} = \mathbf{C}\mathbf{U}_H\mathbf{A}^{-T}\mathbf{v}\hat{\mathbf{e}}_i^T\mathbf{u},\end{aligned} \tag{4.117}$$

which, letting $\mathbf{z} = \mathbf{A}^{-T}\mathbf{v}$, yields for the derivative

$$\begin{aligned}\frac{df}{d\mathbf{h}_i} &= (\mathbf{x} - \mathbf{C}\mathbf{U}_H\mathbf{A}^{-T}\mathbf{v})\hat{\mathbf{e}}_i^T\mathbf{u} + (\mathbf{y} - \mathbf{C}\mathbf{U}_H\mathbf{u})\hat{\mathbf{e}}_i^T\mathbf{A}^{-T}\mathbf{v} \\ &= (\mathbf{x} - \mathbf{C}\mathbf{U}_H\mathbf{A}^{-T}\mathbf{v})u_i + (\mathbf{y} - \mathbf{C}\mathbf{U}_H\mathbf{u})z_i.\end{aligned} \tag{4.118}$$

Arranging the individual results for each $i$ into a matrix and substituting the original assignments,

$$\begin{aligned}\frac{df}{d\mathbf{U}_H} &= \begin{pmatrix}\frac{df}{d\mathbf{h}_1} & \cdots & \frac{df}{d\mathbf{h}_m}\end{pmatrix} \\ &= (\mathbf{x} - \mathbf{C}\mathbf{U}_H\mathbf{A}^{-T}\mathbf{v})\mathbf{u}^T + (\mathbf{y} - \mathbf{C}\mathbf{U}_H\mathbf{u})\mathbf{z}^T \\ &= \mathbf{x}\mathbf{y}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-T} - \mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-T}\mathbf{U}_H^T\mathbf{x}\mathbf{y}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-T} \\ &\qquad + \mathbf{y}\mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1} - \mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{y}\mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1} \\ &= (\mathbf{I} - \mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-T}\mathbf{U}_H^T)\mathbf{x}\mathbf{y}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-T} + \\ &\qquad + (\mathbf{I} - \mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T)\mathbf{y}\mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}.\end{aligned} \tag{4.119}$$

If $\mathbf{C}$ is symmetric, this simplifies to

$$\frac{df}{d\mathbf{U}_H} = (\mathbf{I} - \mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T)(\mathbf{x}\mathbf{y}^T + \mathbf{y}\mathbf{x}^T)\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}, \tag{4.120}$$

and if additionally $\mathbf{y} = \mathbf{x}$, then

$$\frac{df}{d\mathbf{U}_H} = 2(\mathbf{I} - \mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T)\mathbf{x}\mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}. \tag{4.121}$$

Note that since $y = \mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}$ is equivalent to $y = \mathrm{tr}\left\{(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}\mathbf{x}^T\mathbf{U}_H\right\}$, this result can also be written as

$$\frac{d}{d\mathbf{U}_H}\mathrm{tr}\left\{(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}\mathbf{x}^T\mathbf{U}_H\right\} = 2(\mathbf{I} - \mathbf{C}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T)\mathbf{x}\mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}. \tag{4.122}$$

# Chapter 5

# Subspace invariant covariance estimation

In the previous chapter, the uniformly most powerful invariant test was presented for the problem of detecting a target in MVN noise with known parameters and interference in a known subspace. Methods were suggested for estimating the interference subspace from actual samples of data, still under the assumption that the noise mean and covariance parameters are known. The use of invariant detectors for more general problems was also proposed, where the invariance subspace is used to partially overcome the effects of model mismatch, specifically in the cases where such mismatch occurs in a low-dimensional subspace of the observation space.

All the results of Chapter 4 make the assumption that the parameters of the underlying noise process are known. Samples of data are then used to estimate a suitable invariance subspace. In this chapter the dual problem is addressed, where the invariance subspace is assumed known but the noise parameters are unknown and have to be estimated. Again the assumption is made of a nominally MVN distribution for the noise, at least within the subspace complementary to the invariance subspace.

In general, if not accounted for, the presence of subspace interference or low-rank model mismatch will have an effect on any noise parameter estimates. Take for example the simple 2-D situation depicted in Figure 5.1, where the actual noise is zero mean and uncorrelated, and interference is present in the $x_2$-direction. For purposes of estimation, the assumption is made that the noise mean is identically zero, and the covariance is some unknown multiple of the identity matrix, say $\sigma^2\mathbf{I}$. To specify the noise, it is required to estimate $\sigma^2$. Now, if $\sigma^2$ is estimated without suitable compensation for the interference, a reasonable estimator would be the average of the sample variance in the $x_1$ and $x_2$-directions. The resulting covariance estimate is demonstrated by the dotted contour in Figure 5.1, where it is clear that the value of $\sigma^2$ has been considerably overestimated.

For the case of the data being corrupted by interference in the subspace $\langle \mathbf{U}_I \rangle$, one remedy is to consider the interference subspace components unknown and estimate them along with the model parameters.

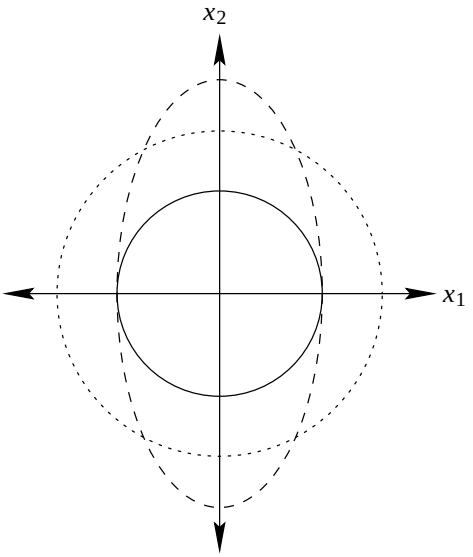



Figure 5.1: Contours of constant probability for the ideal underlying noise distribution (solid), the sample covariance in the presence of interference in the $x_2$-direction (dashed), and the estimated covariance (dotted).

For example, suppose that there are $M$ samples $\mathbf{x}_1, \ldots, \mathbf{x}_M$ on which to base the estimate, and that these samples are of the form $\mathbf{x}_i = \mathbf{z}_i + \mathbf{U}_I \mathbf{c}_i$. Here $\mathbf{z}_i$ is the assumed uncorrupted data observation, with distribution $N[\mathbf{m}(\boldsymbol{\theta}), \mathbf{C}(\boldsymbol{\theta})]$. This distribution is parameterised by the unknown $\boldsymbol{\theta}$, and the vectors $\mathbf{c}_1, \ldots, \mathbf{c}_M$ characterise the details of the interference. If it is required to estimate $\boldsymbol{\theta}$, then under a maximum likelihood formulation it may seem reasonable to maximise the likelihood of $\mathbf{x}_1, \ldots, \mathbf{x}_M$ over the entire set of parameters $\mathbf{c}_1, \ldots, \mathbf{c}_M$ and $\boldsymbol{\theta}$.

Little and Rubin [89] warn against such a formulation: since the number of unknowns increases with the sample size, the estimates obtained are neither asymptotically minimum variance nor unbiased. As in Section 3.4.1, the principle of invariant estimation can provide a better solution. Specifically, for the subspace invariance problem, candidate estimators of the noise parameters should be independent of the components of the data which lie in the invariance subspace. This completely eliminates the effect that the data components in the invariance dimensions have on the parameter estimates, which is desirable. Returning to the example in Figure 5.1, a reasonable invariant estimate would ignore the components of the data in the $x_2$-direction. That is, the sample variance of the data in the $x_1$-direction alone could provide a good estimate of $\sigma^2$, and has all the required invariance properties.

In principle, within the invariance requirement, any preferred statistical estimation criterion may be used to identify the parameter values. However, from the point of view of detection, maximum likelihood estimation is convenient and can be justified by its use in the GLRT. If this is the principle applied, then the required estimate will maximise the likelihood of the data in the subspace $\langle \mathbf{U}_H \rangle$ orthogonal to the invariance subspace $\langle \mathbf{U}_I \rangle$. Thus if $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are the observed data, then the only portion that is used

in the invariant estimation is $\mathbf{U}_H^T\mathbf{x}_1, \ldots, \mathbf{U}_H^T\mathbf{x}_M$, which are the components of the observations which lie outside of the invariance subspace. The invariant MLE will maximise the likelihood of observing these portions of the observations with respect to the unknown parameter values.

The presence of constraints on the noise parameters can cause some complications in the context of invariant estimation. This is because constraints on the assumed MVN model may conceptually be applied in two ways: either to the distribution of the data in the original observation space, or to the distribution of the data in the subspace that is considered valid for estimation. By placing the constraint on the valid components of the data, the invariance subspace is entirely removed from the problem. The constrained estimate then makes no reference whatsoever to any conditions or constraints in the original observation space, and the constrained maximisation can be performed directly. For example, an assumption might be made that the portion of an observation outside of the invariance subspace is subject to a white noise constraint. This constraint is phrased entirely within the subspace $\langle \mathbf{U}_H \rangle$, and no reference is made to conditions on the covariance of the data in the invariance subspace $\langle \mathbf{U}_I \rangle$.

More often, however, constraints are meaningful in the context of the original data space, rather than in the reduced subspace of valid data. Take for example the detection problem where the noise is assumed to be stationary except for the presence of low-rank additive interference or model mismatch. It then makes sense to consider a stationary constraint on the original full-rank data, which only indirectly imposes itself on the subspace of valid data — the invariance subspace cannot conceptually be removed from the problem formulation without abstracting the description of the constraint. Naturally it is always possible to translate a constraint on the complete data to a constraint on the valid data subspace, and perform the entire estimation problem in this subspace. However, from the point of view of computational tractability this approach is often not desirable.

In the previous chapter it was suggested that a missing data interpretation is appropriate for that portion of each observation which lies in the invariance subspace. This is useful since it provides a formalism in which the invariance aspects of the problem can be conveniently dealt with from a statistical viewpoint. In the context of this chapter, it permits the use of the expectation-maximisation (EM) algorithm of Dempster, Laird, and Rubin [28] for finding invariant maximum likelihood estimates. To use this formalism, it is assumed that there is an underlying process which generates data samples (henceforth referred to as the *complete* data). It is required to estimate the parameters of this process. The complete data may or may not be hypothetical, depending on the context in which invariance is being utilised. Either way, the complete data are not considered observable: only those components which lie outside of the invariance subspace (referred to as the *observable* or *incomplete* data components) may be considered valid and used in the estimation.

Subspace invariant estimates are interesting in their own right, but in the context of this work the emphasis is on their use in target detection. That is, invariant estimators and invariant detectors need to be combined in some manner, so that true subspace invariant detectors are obtained even in the

case where there are unknown noise parameters in the formulation. For many applications where invariant tests are desirable, invariant estimators are required of the noise parameters. This is because the training samples themselves exhibit modelling mismatch for the components which lie inside the invariance subspace, which can affect the estimates if not accounted for.

Some paradigms for including invariant estimates into the detection problem are presented in Section 5.1. In particular, both the plug-in classifier and the GLRT principles are applied to the detection problem where an invariance subspace is present. These formulations are not necessarily proposed as solutions to problems: they mainly serve to demonstrate the concepts involved in invariant detection and estimation.

In Section 5.2 the invariant estimation of the noise mean and covariance matrix is discussed for the simple case where these parameters are assumed completely unknown, and no assumptions are made regarding their structure. In this case, nothing can be said about the parameter values related to dimensions corresponding to the invariance subspace. For purposes of invariant detection, however, this lack of complete specification does not present a problem: the resulting test statistic is independent of the parameters which are not identifiable. In that case there is a rich background of literature which may be drawn on, and the performance of the resulting detector can be effectively quantified.

The case of unconstrained noise parameters is mainly of interest when large amounts of training data are available on which to base estimates. However, when data are scarce then constraints are required. Such constraints reduce the number of samples required to accurately estimate the parameters, and if appropriate can improve the detection performance.

In Sections 5.4, 5.5, and 5.6, the task of making invariant estimates of unknown noise parameters is discussed under certain constraints on the allowed values that these parameters may take on. In particular, the problems of subspace invariant covariance estimation under circulant, doubly block circulant, Toeplitz, doubly block Toeplitz, and ARMA constraints on the covariance matrix are considered. It is demonstrated that a missing data interpretation is appropriate in each case, and that the expectation-maximisation algorithm can sometimes be used to iteratively arrive at maximum likelihood estimates.

Section 5.7 briefly discusses the problem of simultaneous invariance subspace and covariance matrix estimation, in both the contexts of subspace interference and low-rank model mismatch. The methods are only appropriate in instances where direct maximisation of an objective function over the set of allowable covariances is possible. The ARMA assumption is one such case.

Finally, Section 5.8 provides some results demonstrating the simultaneous estimation of parameters, and the use of invariant detectors for real-data applications. It is shown that the invariant detectors can improve detectability, and exhibit better generalisation properties. The predictability of the invariant tests is also better than that of the noninvariant tests.

## 5.1 Parameter estimates for subspace invariant detectors

The detector presented in Section 4.2.1 is optimal for the problem of detecting a known target in MVN noise with known parameters. Similarly, the invariant detector of Section 4.2.2 is optimal for the case where observations are corrupted by interference in a known subspace, again in MVN noise with known mean and covariance.

If the noise parameters are unknown, the tests cannot be directly applied to the observed data: the test statistic will in general itself depend on these unknown parameters. A modified test formulation is therefore required which is not conditional on these unknowns.

In the literature, there are two common paradigms for dealing with this problem. Both make the assumption that there are additional observations available, with distributions which are related to the distribution of the noise in the data to be tested. Often these observations are required to contain only noise, in which case they share the distribution of the noise in the primary data.

The first paradigm uses the additional data samples to make a direct estimate of the parameters, leading to a plug-in formulation, where the estimates are substituted into the optimal test statistic derived for the known parameter problem. The second considers all the data at once, using a GLRT for the resulting problem. An estimate of the noise parameters is also implicit in this approach, in the form of the maximisation of density functions over the allowed space of parameters. Although less common, constraints can be included in both formulations, and can result in considerably more powerful tests [39].

If an invariance subspace is included in the problem, then the parameter estimation may become complicated. The reason for this is because the training data are themselves corrupted by interference, or exhibit model mismatch in the invariance subspace. As discussed, this can have an effect on the noise parameter estimates, and if not taken into account may have a degrading effect on the resulting test. Therefore, for use in an invariant test, the parameter estimates are required to be invariant to the components of the data in the invariance subspace.

The two sections which follow discuss detection formulations for the problem of detecting a known target in parametrically-known noise, when an invariance subspace is present and additional noise-only observations are available. In both cases it is demonstrated that subspace invariant noise parameter estimates are required. The remainder of the chapter discusses methods for obtaining these estimates.

### 5.1.1 Invariant estimation for plug-in detection

Let $\mathbf{x}$ be the data to be tested for target presence, with the additional samples $\mathbf{x}_1, \ldots, \mathbf{x}_M$ being available for estimation. Assuming for now the case of no invariance subspace, these additional observations are

assumed to be realisations of the same random process that generated the noise in $\mathbf{x}$. A plug-in classifier formulation [115] may be appropriate in this case: a good or optimal test statistic is derived for the data $\mathbf{x}$ under the assumption that the noise parameters are known, and estimates of these parameters based on $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are plugged into this statistic. In some contexts these additional data may be regarded as training samples.

If an invariance subspace is included, then the requirement that the governing distribution be the same for all observations can be relaxed slightly. All that is then required is that the samples share the same distribution in the subspace complementary to the invariance subspace. A test which is invariant to the data in $\langle \mathbf{U}_I \rangle$ is developed for $\mathbf{x}$, and a suitable invariant estimate of the noise parameters plugged into the expression. The estimate is obtained from the samples $\mathbf{x}_1, \ldots, \mathbf{x}_M$, and is again required to be invariant to components in the subspace $\langle \mathbf{U}_I \rangle$. The resulting test does not depend on any data components in the invariance subspace.

As with the case discussed in Chapter 3, finding a suitable invariant estimate requires the specification of an invariance group which places certain equivalences on sets of observations. This group is the same as the one applied to the detection portion of the test. The maximal invariant statistic again plays an important role in the resulting estimation problem. In particular, since all invariant functions of the data must be functions of the maximal invariant, any invariant estimator must also be a function of it. As with invariant detection, the maximal invariant contains as much information as possible which is independent of the component of the observations in the invariance subspace. The distribution of the maximal invariant statistic therefore provides all information regarding the invariant estimates of the unknown parameters.

Once the maximal invariant has been found, and its parametric distribution calculated, the choice of estimator is open to the same conditions as any other estimation problem. Therefore the invariant estimator may be chosen to optimise any of the standard criteria, such as being minimum variance unbiased or having minimum mean square error. However, one estimator which stands out from the others is the maximum likelihood estimator, which has certain useful invariance properties. In particular, if $\hat{\boldsymbol{\theta}}$ is a maximum likelihood estimate of a parameter $\boldsymbol{\theta}$, then $\mathbf{w}(\hat{\boldsymbol{\theta}})$ will be the maximum likelihood estimate of the transformed parameter $\mathbf{w}(\boldsymbol{\theta})$. This property justifies the use of the estimate as a plug-in value in a detection statistic: if the conditional detection statistic is $t(\mathbf{x}|\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then $t(\mathbf{x}|\hat{\boldsymbol{\theta}})$ is the MLE of the test statistic.

Since the estimates are used only in specifying the details of the statistic to be used in the detector, the only way to really assess the estimate is in terms of the detection performance which results from its use. The standard desirable properties of an estimator have no direct bearing to this problem. For example, there is no a priori reason to believe that an unbiased estimator or a minimum variance estimator will be better for detection than any other estimator.

Assuming that maximum likelihood is the estimation paradigm of choice, the invariant MLE of the

parameter $\boldsymbol{\theta}$ is that value which maximises the likelihood of the maximal invariant statistic. As in the case of invariant detection, if the observed additional data are $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$ then the maximal invariant statistic is $(\mathbf{y}_1, \ldots, \mathbf{y}_M)$, with $\mathbf{y}_i = \mathbf{U}_H^T \mathbf{x}_i$. This simply contains the component of each observation which lies outside of the invariance subspace. If the nominal distribution of the noise is $N[\mathbf{m}(\boldsymbol{\theta}), \mathbf{C}(\boldsymbol{\theta})]$, then each element of this maximal invariant statistic is $N[\mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}), \mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H]$. With the observations independent, the invariant probability density function is

$$\begin{aligned} p(\mathbf{y}_1, \ldots, \mathbf{y}_M) &= (2\pi)^{-M(n-p)/2} |\mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H|^{-M/2} e^{-\frac{1}{2}\sum_{j=1}^{M} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}))^T (\mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H)^{-1} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}))} \\ &= (2\pi)^{-M(n-p)/2} |\mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H|^{-M/2} e^{-\frac{1}{2}\sum_{j=1}^{M} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}))^T (\mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H)^{-1} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}))}, \end{aligned} \tag{5.1}$$

where $\mathbf{U}_H$ is a matrix with orthonormal columns which span the subspace orthogonal to $\langle \mathbf{U}_I \rangle$. As is usual for normal distributions, the calculation of this value can be simplified by considering instead the log-likelihood

$$L = K - \frac{M}{2} \ln |\mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H| - \frac{1}{2} \sum_{j=1}^{M} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}))^T (\mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H)^{-1} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta})), \tag{5.2}$$

and maximising it accordingly.

The estimate $\hat{\boldsymbol{\theta}}$ obtained from this estimation is substituted into the conditional invariant test statistic for the data observation $\mathbf{x}$, resulting in the final invariant plug-in test statistic (using Equation 4.29)

$$t(\mathbf{x}) = \mathbf{s}^T [\mathbf{C}(\hat{\boldsymbol{\theta}})^{-1} - \mathbf{C}(\hat{\boldsymbol{\theta}})^{-1} \mathbf{U}_I (\mathbf{U}_I^T \mathbf{C}(\hat{\boldsymbol{\theta}})^{-1} \mathbf{U}_I)^{-1} \mathbf{U}_I^T \mathbf{C}(\hat{\boldsymbol{\theta}})^{-1}] (\mathbf{x} - \mathbf{m}(\hat{\boldsymbol{\theta}})), \tag{5.3}$$

with $\mathbf{s}$ being the target to be detected. The test compares the value of this statistic to a threshold, and decides $H_1$ if exceeded. In general the threshold should be obtained by applying the test statistic to samples of actual data, rather than through analytical techniques which assume the distribution of the maximal invariant to be $N[\mathbf{U}_H^T \mathbf{m}(\hat{\boldsymbol{\theta}}), \mathbf{U}_H^T \mathbf{C}(\hat{\boldsymbol{\theta}}) \mathbf{U}_H]$. This is because in the latter case the uncertainty in the estimate of $\hat{\boldsymbol{\theta}}$ is not taken into account.

The test which results from this development is identical to what would be obtained if it were required to test for the presence of a target $\mathbf{U}_H^T \mathbf{s}$ in the observation $\mathbf{y} = \mathbf{U}_H^T \mathbf{x}$, with the additional observations $\mathbf{y}_1, \ldots, \mathbf{y}_M$ containing only noise. The noise distribution of these samples is $N[\mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}), \mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H]$, with $\boldsymbol{\theta}$ unknown. Thus the entire formulation is equivalent to calculating a maximal invariant for the observations, and proceeding with a conventional plug-in classifier development.

The plug-in formalism can in no way be claimed to be optimal unless the parameter estimates are perfect, which can never occur in practice. However, if the parameter estimates are good, then the resulting test statistic should exhibit good discrimination ability and the test performance will be high. More specifically, if the relevant parameter estimates are consistent, then the asymptotic performance

of the test (that is, as the number of training samples goes to infinity) will be the same as if the parameter values were known in advance [81, p. 246].

The reason for the suboptimality of the plug-in classifier lies in the fact that the sampling distribution of the estimator is not taken into account when developing the test statistic. Thus each estimator will result in a different test statistic. A good estimator is in general expected to yield a good test, but this cannot be guaranteed.

An alternative hypothesis testing paradigm is that of the generalised likelihood ratio test, which deals with this problem in a formal manner. This is discussed in the next section.

### 5.1.2 Invariant estimation for GLRT detection

The plug-in classifier is not the only way in which the detection problem can be approached when additional observations are available. Kelly [80], for example, argues that the GLRT formulation should be used simultaneously on the data to be tested as well as on the training data. This formulation addresses the problem in its entirety, and overcomes some of the arbitrariness of the plug-in formalism. To use the GLRT for this problem, it is assumed that the combined set of data $\mathbf{x}_1, \ldots, \mathbf{x}_M$ is observed. In contrast to the discussion in the previous section, however, here it is assumed that the vector $\mathbf{x}_1$ represents the data to be tested for the presence of target, and $\mathbf{x}_2, \ldots, \mathbf{x}_M$ are the additional noise-only data available for estimation. In Kelly's terminology, $\mathbf{x}_1$ is referred to as the *primary* data, and the additional observations constitute the *secondary* data.

Once again, if an invariance subspace is included in the problem, then a maximal invariant statistic is $(\mathbf{y}_1, \ldots, \mathbf{y}_M)$, with $\mathbf{y}_i = \mathbf{U}_H^T \mathbf{x}_i$. As in the previous section, the entire formulation can be applied to this maximal invariant. Under $H_0$, the distribution of each component $\mathbf{y}_i$ is $N[\mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}), \mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H]$. Under $H_1$, $\mathbf{y}_1 : N[\mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}) + \mathbf{U}_H^T \mathbf{s}, \mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H]$, with $\mathbf{y}_2, \ldots \mathbf{y}_M$ distributed as before. With the problem formulated in this way, the GLRT follows quite simply:

$$t_{\text{GLRT}} = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p_{H_1}(\mathbf{y}_1, \ldots, \mathbf{y}_M | \boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} p_{H_0}(\mathbf{y}_1, \ldots, \mathbf{y}_M | \boldsymbol{\theta})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{5.4}$$

where $p_{H_1}(\mathbf{y}_1, \ldots, \mathbf{y}_M | \boldsymbol{\theta})$ is the probability density of the combined set of observations under $H_1$, and $p_{H_0}(\mathbf{y}_1, \ldots, \mathbf{y}_M | \boldsymbol{\theta})$ the density under $H_0$. These densities are conditional on the unknown vector parameter $\boldsymbol{\theta}$. It is assumed that the set of values which this parameter can take on is the same under both hypotheses, namely $\boldsymbol{\Theta}$.

If the noise components in the observations are independent, then the only difference between the probability densities under the two hypotheses is that under $H_1$ the mean of $\mathbf{y}_1$ is offset by $\mathbf{U}_H^T \mathbf{s}$. The details of the test statistic are easy to derive, but in this general case are not very enlightening. The important point to note, however, is that implicit in the test is the maximisation of the likelihood function

under each assumed hypothesis. This once again corresponds to maximum likelihood estimation of the noise parameter which, since performed over the distribution of the maximal invariant statistic, is itself invariant to data contributions in the invariance subspace. For implementing the test, invariant maximum likelihood estimates of the parameters are therefore again required.

To demonstrate, consider the calculation of the denominator term in Equation 5.4. For independent observations, the probability density $p_{H_0}(\mathbf{y}_1, \ldots, \mathbf{y}_M)$ is identical to $p(\mathbf{y}_1, \ldots, \mathbf{y}_M)$ in Equation 5.1, so the log-likelihood is as given in Equation 5.2. Maximisation over $\boldsymbol{\theta}$ under the assumption of $H_0$ is therefore identical to the maximisation required for the plug-in detector formulation, and the invariant maximum likelihood estimate of $\boldsymbol{\theta}$ is exactly what is required to calculate the denominator term. Similarly, it is easy to show that estimation of $\boldsymbol{\theta}$ in the numerator term in Equation 5.4 can be performed by calculating the maximum likelihood estimate for the observation $(\mathbf{y}_1 - \mathbf{U}_H^T\mathbf{s}, \mathbf{y}_2, \ldots, \mathbf{y}_M)$.

Aside from the improvement brought about by incorporating the parameter estimation into the test formulation, the GLRT may be expected to perform better than the plug-in test because it also uses the potential target-carrying observation $\mathbf{y}_1$ in the noise parameter estimation. This observation contains information regarding the noise distribution, and it makes no sense not to use it in the estimate. The difference between the GLRT and the plug-in test becomes less significant as the number of secondary data vectors increases.

The GLRT formulation introduces the possibility of testing for target presence in the event that no additional noise-only samples are avaliable. In that case, any noise parameter estimate has to be based solely on the single sample of data on which testing is to be performed. This represents a situation where only a small amount of data is available on which to base the parameter estimates, and a corresponding decrease in test performance may be expected.

This shortage of data may be remedied by placing appropriate and sufficiently tight constraints on the set of allowable noise densities. By reducing the number of effective free parameters, more accurate noise density estimates can be obtained even from a small amount of data. This reduced variance comes at the risk of increased bias in the estimates, since the constrained model has less capacity to capture specific noise details. However, a biased estimate is in many instances superior to an unbiased estimate with a high variance.

Kay [74], for example, makes the assumption that the noise is a realisation of a white noise driven rational transfer function filter, and uses the GLRT to test for target presence in a single observation vector. This is equivalent to the assumption that the noise is governed by an ARMA(p,q) process, where the orders $p$ and $q$ are fixed in advance. The number of free parameters in this model is $p + q$, which can be chosen to be much smaller than the number of parameters required to specify a general covariance matrix. This approach presupposes that the noise can indeed be effectively modelled as a realisation of an ARMA process.

The section which follows discusses the problem of making a subspace invariant estimate of the noise parameters under the assumption that there are no constraints on the mean vector and the covariance matrix. The estimate maximises the likelihood of the data component which lies outside of the invariance subspace $\langle \mathbf{U}_I \rangle$, which is assumed known. The result can either be substituted directly into the invariant test statistic in a plug-in formulation, or can be used as the maximisation stage of a GLRT formulation. Since the estimates are unconstrained, they are only appropriate in instances where the number of secondary data vectors is large, and the resulting estimates can be considered good.

## 5.2 Simple subspace invariant estimation

If the noise distribution is unknown, it has to be estimated from actual data samples. In the context of parametric statistics, this involves postulating a probability density which depends on unknown parameters, and fixing these parameters using some estimator. As discussed in the previous section, for the invariant detection problem the data samples used to estimate the noise parameters are only valid in the subspace orthogonal to the invariance subspace. Thus the estimators themselves need to be invariant to the same subspaces that the detector needs to be.

Also discussed in the previous section, if the estimation paradigm is maximum likelihood then the required invariant estimate has to maximise the log-likelihood in Equation 5.2. Writing this in terms of the maximal invariant components $\mathbf{y}_1, \ldots, \mathbf{y}_M$, this likelihood is

$$L = K - \frac{M}{2} \ln |\mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H| - \frac{1}{2} \sum_{j=1}^{M} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta}))^T (\mathbf{U}_H^T \mathbf{C}(\boldsymbol{\theta}) \mathbf{U}_H)^{-1} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}(\boldsymbol{\theta})). \quad (5.5)$$

In the event that the mean and covariance matrix of the noise are unconstrained, the maximisation can be performed quite easily. In this case, the explicit dependence on the parameter vector $\boldsymbol{\theta}$ can be dropped, and the estimation problem is equivalent to finding the values of $\mathbf{m}$ and $\mathbf{C}$ that maximise the invariant log-likelihood

$$L = K - \frac{M}{2} \ln |\mathbf{U}_H^T \mathbf{C} \mathbf{U}_H| - \frac{1}{2} \sum_{j=1}^{M} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m})^T (\mathbf{U}_H^T \mathbf{C} \mathbf{U}_H)^{-1} (\mathbf{y}_j - \mathbf{U}_H^T \mathbf{m}). \quad (5.6)$$

For this unconstrained case, a necessary condition on the maximum likelihood estimate $\widehat{\mathbf{m}}$ of the mean

is that $\frac{dL}{d\mathbf{m}}\big|_{\mathbf{m}=\widehat{\mathbf{m}}} = 0$. Now,

$$\begin{aligned}\frac{dL}{d\mathbf{m}} &= -\frac{1}{2}\sum_{j=1}^{M}\left[\frac{d}{d\mathbf{m}}(\mathbf{y}_j - \mathbf{U}_H^T\mathbf{m})\right]2(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}(\mathbf{y}_j - \mathbf{U}_H^T\mathbf{m}) \\ &= \mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\sum_{j=1}^{M}(\mathbf{y}_j - \mathbf{U}_H^T\mathbf{m}) \\ &= M\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}(\widehat{\mathbf{m}}_y - \mathbf{U}_H^T\mathbf{m}), \end{aligned} \tag{5.7}$$

where $\widehat{\mathbf{m}}_y = 1/M\sum_{j=1}^{M}\mathbf{y}_j$.

Let $\mathbf{U} = (\mathbf{U}_H\ \mathbf{U}_I)$ be an orthogonal matrix, with $\mathbf{U}_H$ and $\mathbf{U}_I$ chosen in the usual manner. Premultiplying the necessary condition by the invertible matrix $\mathbf{U}^T$ yields

$$\begin{aligned}&\begin{pmatrix}\mathbf{U}_H^T\\ \mathbf{U}_I^T\end{pmatrix}M\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}(\widehat{\mathbf{m}}_y - \mathbf{U}_H^T\widehat{\mathbf{m}}) = \mathbf{0} \\ &\quad\Longrightarrow \begin{pmatrix}(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}(\widehat{\mathbf{m}}_y - \mathbf{U}_H^T\widehat{\mathbf{m}})\\ \mathbf{0}\end{pmatrix} = \mathbf{0}, \end{aligned} \tag{5.8}$$

since $\mathbf{U}_H^T\mathbf{U}_H = \mathbf{I}$ and $\mathbf{U}_I^T\mathbf{U}_H = \mathbf{0}$. With $\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H$ invertible, it must be true that $\mathbf{U}_H^T\widehat{\mathbf{m}} = \widehat{\mathbf{m}}_y$. Thus possible solutions are $\widehat{\mathbf{m}} = \mathbf{U}_H\widehat{\mathbf{m}}_y + \mathbf{U}_I\mathbf{c}$ for all vectors $\mathbf{c} \in \mathbb{R}^p$. Each of these solutions result in the same likelihood value. Thus they are all maximum likelihood estimators, and we may choose arbitrarily between them. The most obvious choice is that of $\mathbf{c} = \mathbf{0}$, resulting in $\widehat{\mathbf{m}} = \mathbf{U}_H\widehat{\mathbf{m}}_y$. Since $\widehat{\mathbf{m}}_y = \mathbf{U}_H^T\widehat{\mathbf{m}}_x$ with $\widehat{\mathbf{m}}_x$ defined accordingly, the estimate of $\mathbf{m}$ can be written in terms of the corrupted observations as $\widehat{\mathbf{m}} = \mathbf{U}_H\mathbf{U}_H^T\widehat{\mathbf{m}}_x$. This is simply the projection of the sample mean into the space spanned by the matrix $\mathbf{U}_H$.

The invariant covariance matrix estimate can be obtained similarly. Using the standard results for maximum likelihood covariance matrix estimation the required estimate of $\mathbf{C}$, namely $\widehat{\mathbf{C}}$, should satisfy

$$\begin{aligned}\mathbf{U}_H^T\widehat{\mathbf{C}}\mathbf{U}_H &= \frac{1}{M}\sum_{j=1}^{M}(\mathbf{y}_j - \mathbf{U}_H^T\widehat{\mathbf{m}})(\mathbf{y}_j - \mathbf{U}_H^T\widehat{\mathbf{m}})^T \\ &= \frac{1}{M}\sum_{j=1}^{M}(\mathbf{y}_j - \widehat{\mathbf{m}}_y)(\mathbf{y}_j - \widehat{\mathbf{m}}_y)^T = \widehat{\mathbf{S}}_{yy}, \end{aligned} \tag{5.9}$$

where $\widehat{\mathbf{S}}_{yy}$ is the sample covariance matrix of $\mathbf{y}_1, \ldots, \mathbf{y}_M$. Candidate values for $\widehat{\mathbf{C}}$ are

$$\widehat{\mathbf{C}} = \mathbf{U}_H\widehat{\mathbf{S}}_{yy}\mathbf{U}_H^T + \mathbf{U}_I\boldsymbol{\beta}\mathbf{U}_I^T \tag{5.10}$$

for all matrices $\boldsymbol{\beta}$, which again all result in the same likelihood value for the observations. Thus we

may choose $\boldsymbol{\beta} = \mathbf{0}$, so $\widehat{\mathbf{C}} = \mathbf{U}_H \widehat{\mathbf{S}}_{yy} \mathbf{U}_H^T$. Writing in terms of the sample covariance $\widehat{\mathbf{S}}_{xx}$ of the original corrupted data observations yields the result

$$\widehat{\mathbf{C}} = \mathbf{U}_H \mathbf{U}_H^T \widehat{\mathbf{S}}_{xx} \mathbf{U}_H \mathbf{U}_H^T \tag{5.11}$$

as the required invariant maximum likelihood covariance estimate.

The fact that the estimates are not unique arises because portions of data are missing. These missing data would be required to determine *anything* about some of the parameters. In the statistical parlance, these parameters are not identifiable; the invariant probability density function does not depend on their values, so they cannot be estimated. More specifically, the available valid data contain absolutely no information regarding the mean value of the original data distribution in the invariance subspace, nor of the variance and covariance components related to these directions. From the point of view of the log-likelihood in Equation 5.2, there are therefore many equivalent choices. The available choices all result in the same values for the quantities $\mathbf{U}_H^T \widehat{\mathbf{m}}$ and $\mathbf{U}_H^T \widehat{\mathbf{C}} \mathbf{U}_H$, which can therefore in general be estimated uniquely.

For the case of the plug-in detector, the proposed detection statistic in Equation 5.3 depends on the parameters $\mathbf{m}$ and $\mathbf{C}$ only through these quantities which may be uniquely estimated. The invariant plug-in test can therefore be specified completely and unambiguously. In fact, since $\mathbf{U}_H^T \widehat{\mathbf{m}} = \mathbf{U}_H^T \widehat{\mathbf{m}}_x$ and $\mathbf{U}_H^T \widehat{\mathbf{C}} \mathbf{U}_H = \mathbf{U}_H^T \widehat{\mathbf{S}}_{xx} \mathbf{U}_H$, the test can be correctly specified using the *corrupted* sample mean and covariance matrix based on the original observations $\mathbf{x}_1, \ldots, \mathbf{x}_M$. The test statistic is itself independent of the corrupted components of the estimate, so for the detector the estimation problem is trivial.

Similar conclusions apply to the GLRT formulation. Since the invariant likelihood does not depend on the interference or mismatch components, the required maximum likelihood estimates can simply be taken to be the sample mean and sample covariance of the corrupted data observations. Under each hypothesis, the maximised likelihood is the same as if an explicit invariant estimate is used.

The plug-in test statistic proposed for the unconstrained noise problem is identical to the sample matrix inversion technique discussed by Reed, Mallet, and Brennan [111], except that an invariance subspace has been included in this formulation. Since this invariance requirement can be eliminated entirely through the transformation to the maximal invariant, the analysis provided by Reed et al. them applies directly to this situation. Results are therefore readily available which characterise the performance of the plug-in formulation in this invariant testing problem.

The plug-in test produces a statistic which exhibits high discrimination ability between the two hypotheses. That is, the difference in value taken on under the condition of target absent and target present is large. However, a problem exists when the test threshold has to be set to yield the required performance characteristics: since the distribution of the statistic under $H_0$ depends on unknown covariance parameters, the threshold is difficult to set meaningfully. The same problem exists with the

subspace invariant GLRT. In Chapter 6 alternative test formulations are discussed which do not exhibit this problem for the unconstrained noise case. However, the test statistic formulations presented in this chapter are common and intuitive, and clearly demonstrate the need for invariant estimators in detection.

Thus, for the invariant test with no constraints on the mean and covariance matrix, invariant parameter estimates are *not* required for either of the formulations presented. In the more general case, however, where constraints are imposed on the parameters, this is not true. In that case, the presence of subspace mismatch between the data and the model can have an effect on the parameter estimates which is *not* cancelled out in the calculation of the test statistic. The remainder of this chapter discusses invariant estimation procedures for various constraints on the covariance matrix of the data.

## 5.3 EM algorithm for structured covariance matrix estimation

In this section a general formulation of the EM algorithm is presented which covers a large class of structured covariance matrix estimation problems. The development is fairly abstract, to accommodate as many cases of interest as possible. In particular, the solution is formulated in a manner such that the inclusion of an invariance subspace is facilitated. The sections following this one provide details of how the solution can be applied to real estimation problems, which may or may not incorporate an invariance subspace.

In the formulations, it is important to distinguish between results which apply to the field of complex numbers, and those which apply to real numbers. Where appropriate, the distinction will be clearly made. To some extent, the complex formulations are simpler, and are typically dominant in the radar literature. The real case formulations, at least in the problems discussed here, are conceptually similar, but additional restrictions are required to ensure that all results remain real. Recall also that the form of a complex MVN density differs from that of the real MVN density function: for the complex case if $\mathbf{x} : N[\mathbf{m}, \mathbf{C}]$ is an $n$-dimensional random vector, then

$$p(\mathbf{x}) = \pi^{-n}|\mathbf{C}|^{-1}e^{-(\mathbf{x}-\mathbf{m})^{\dagger}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})} \tag{5.12}$$

is the implied density function [75, p. 44].

To formulate the problem, we recall that subspace invariant estimation can be approached using a missing data interpretation. That is, the portions of the data which lie in the invariance subspace can be considered unobservable, and are therefore not available for estimation. Nonetheless, it is still required to make an estimate of the parameters from the data which *are* observed and considered valid, namely the portions of the observations orthogonal to the subspace to which invariance is required.

The expectation-maximisation algorithm is a statistical formulation for calculating maximum likeli-

hood estimates of unknown parameters in the presence of missing data. It was first formally presented in a general context by Dempster, Laird, and Rubin [28], although the principle had been used in an ad hoc capacity for some time before the publication of their paper. It is an iterative procedure, where each iteration is guaranteed to refine the current best estimate of the parameters in such a way that the likelihood is increased (or at least not decreased). As such, the algorithm provides an ideal computational mechanism for finding the constrained covariance matrix parameters needed for the invariant estimation problem.

### 5.3.1 Basic EM algorithm overview

Miller et al. [95] describe the EM algorithm very succinctly as follows (with some changes in notation):

> The EM algorithm is an iterative method for determining the MLE of some parameter vector, call it $\boldsymbol{\phi}$, from some observed data, call them $\mathbf{U}$. Using the terminology of Dempster and others, $\mathbf{U}$ is termed the "incomplete data." The incomplete data take values in a sample space $\overline{\mathbf{U}} = \{\text{all possible values of } \mathbf{U}\}$ and have a density $p(\boldsymbol{\phi}; \mathbf{U})$ over the sample space. It is often difficult to maximise $p(\boldsymbol{\phi}; \mathbf{U})$ with respect to $\boldsymbol{\phi}$. To circumvent this difficulty, $\overline{\mathbf{U}}$ is embedded in a large space $\overline{\mathbf{Z}}$ in which some hypothetical data $\mathbf{Z}$, termed the "complete data," take values. A many-to-one mapping from $\overline{\mathbf{Z}}$ to $\overline{\mathbf{U}}$, defined by some function $H$ such that $\mathbf{U} = H(\mathbf{Z})$ is assumed, as is a density $f(\boldsymbol{\phi}; \mathbf{Z})$ of $\mathbf{Z}$ over $\overline{\mathbf{Z}}$. The densities of the incomplete and complete data are related according to
>
> $$p(\boldsymbol{\phi}; \mathbf{U}) = \int_{\{\mathbf{Z}:\mathbf{U}=H(\mathbf{Z})\}} f(\boldsymbol{\phi}; \mathbf{Z}) d\mathbf{Z}.$$
>
> There are two steps for each iteration of the EM algorithm, an E step and M step. In the E (for expectation) step of the $g+1$ iteration, the conditional expectation of the complete-data likelihood function $E\{\ln f(\boldsymbol{\phi}; \mathbf{Z}) | \mathbf{U}, \boldsymbol{\phi}^{(g)}\}$ is determined, where $\boldsymbol{\phi}^{(g)}$ is the estimate of $\boldsymbol{\phi}$ determined at iteration $g$. In the M (for maximisation) step, the conditional expectation of the complete data log-likelihood is maximised with respect to $\boldsymbol{\phi}$, yielding the $\boldsymbol{\phi}^{(g+1)}$ iterate. As proven by Dempster and others, the sequence of log-likelihood functions of the incomplete-data $\ln p(\boldsymbol{\phi}^{(g)}; \mathbf{U})$, $\ln p(\boldsymbol{\phi}^{(g+1)}; \mathbf{U})$, ..., will be nondecreasing.

A requirement for the EM iteration to be useful is that the maximisation of the complete data likelihood function with respect to the unknown parameters be simple and efficient. If this is not the case, then computational reduction is not achieved. Specifically, if the complete data MLE is difficult to find, then by using the EM algorithm one has effectively substituted a difficult maximisation problem (the incomplete data problem) with a sequence of iterations, each of which involves a difficult maximisation problem (over the complete data). Clearly there is then little point in adopting the EM framework.

To reiterate, at each stage the EM algorithm formulation embeds the estimation problem into a larger one, which has the desirable property of computational tractability. The expected likelihood function for this complete data problem is then calculated, conditional on the observed data and the current best estimates of the unknown parameters. This constitutes the expectation stage. The maximisation portion of each iteration involves finding the value of the parameters which maximises this conditional likelihood. For practical purposes this maximisation has to be simple, preferably with a closed-form solution. The values of the parameters which maximises the conditional likelihood are then taken to be the current best estimates, and the iteration is repeated.

The EM algorithm has also been used for finding maximum likelihood estimates of covariance matrices under the constraint that these estimates be Toeplitz [95]. Although a less obvious application of the EM algorithm, in this context each observed data vector is assumed to constitute the first $n$ samples of an $s$-dimensional process with an unknown positive semidefinite circulant covariance matrix. As long as the order of the extension to circulant is large enough (that is, $s$ is sufficiently large), the maximisation of the observed data covariance is effectively performed over the set of positive definite Toeplitz matrices.

Although each EM iteration is guaranteed to not reduce the likelihood, the problem of whether it converges to a global maximum is difficult to answer in general [27]. Furthermore, the fact that the set of possible parameter values is constrained can in some instances impede the efficiency of the search: when restricted to the constraint set, the path between two solutions which are close in real terms may be very long. Also, it has been reported that the algorithm sometimes converges very slowly — this may be considered to be the price paid for guaranteed convergence. Finally, as with all iterative search techniques, the initial estimates of the unknown parameters can have an effect on the final solution.

### 5.3.2 Constrained covariance matrix estimation with missing data

The EM algorithm is well-suited to the task of estimating noise parameters in the presence of subspace interference or low-rank model mismatch: the valid portions of the data (in other words the portions which do not lie in the invariance subspace) constitute the observed or incomplete data, while the complete data are chosen to facilitate the maximisation step of the EM iteration. As mentioned earlier in this chapter, it is convenient to place any constraints on the complete data rather than on the observed or incomplete data. Therefore, as long as the constraint is tractable in terms of maximum likelihood estimation of the parameters from the complete data, the EM algorithm is an appropriate method of solution.

In the remainder of this chapter, the models are all assumed to have a zero mean. In the context of subspace invariant detection and estimation, this restriction is not severe: as long as the actual mean vector lies in the invariance subspace, the assumption is appropriate. In most cases this condition can

be either implicitly or explicitly enforced. Alternatively, an estimate of the mean can be obtained from the data, and the samples modified accordingly.

Suppose that one is given $M$ independent $q$-dimensional observations $\mathbf{v}_1, \ldots, \mathbf{v}_M$, distributed according to $\mathbf{v}_i : N[\mathbf{0}, \mathbf{C}_{vv}]$. These constitute the incomplete or observed data. It is required to find the maximum likelihood estimate of $\mathbf{C}_{vv}$ subject to certain constraints. For the EM algorithm these constraints are formulated indirectly, as conditions on the assumed complete data.

To continue, the relationship between the incomplete data and the complete data has to be specified. In this development the relationship is assumed to be through a deterministic and known linear transformation. Furthermore, the complete data are assumed to consist of $s$-dimensional multivariate normal random vectors $\mathbf{u}_1, \ldots, \mathbf{u}_M$, each with a $N[\mathbf{0}, \mathbf{C}_{uu}]$ distribution. Therefore, each observed vector $\mathbf{v}_i$ is related to the corresponding complete data vector $\mathbf{u}_i$ through the linear relationship

$$\mathbf{v}_i = \mathbf{A}\mathbf{u}_i. \tag{5.14}$$

This ensures that the distribution of $\mathbf{v}_i$ is still normal.

The choice of the complete data is not unique. There is not even a requirement that the relationship between the complete and the incomplete data be linear. The formulation given here is therefore mainly for convenience, since MVN distributions and linear transformations are easy to work with. However, even within the constraints provided, the choice is not unique: the only requirement is that $\mathbf{C}_{uu}$ and $\mathbf{A}$ be chosen in such a way that $\mathbf{u}_i$ has the required distribution. The specification of these parameters therefore has to be based on additional criteria. Primarily, the properties desired for the complete data distribution are:

- The constraint on $\mathbf{C}_{vv}$ must be easily expressed in terms of the complete data covariance $\mathbf{C}_{uu}$, and
- the unknown covariance matrix parameters must be more easily calculated for the constrained complete data problem than for the original problem.

For some problems there may be one formulation which stands out as more natural than the others. In that case, as long as the previous conditions are met, then that formulation should be used. In the absence of an obvious candidate, however, any formulation which matches the requirements may be used.

As explained, the constraint on $\mathbf{C}_{vv}$ is expressed in terms of a constraint on $\mathbf{C}_{uu}$. This is possible due to the strict equality $\mathbf{C}_{vv} = \mathbf{A}\mathbf{C}_{uu}\mathbf{A}^\dagger$: a constraint on $\mathbf{C}_{uu}$ enforces a corresponding requirement upon $\mathbf{C}_{vv}$. For purposes of this development, it is convenient to require that only matrices $\mathbf{C}_{uu}$ satisfying

$$\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^\dagger = \mathbf{C}_{uu} \tag{5.15}$$

for some diagonal matrix $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_s^2)$ be considered candidates for the maximum likelihood covariance estimate, where $\mathbf{V}$ is a known and fixed unitary matrix. This is a restriction which proves to be useful for the problems discussed in later sections, but is still moderately general.

Since $\mathbf{C}_{uu}$ is a covariance matrix, it has to be Hermitian and positive semidefinite. Recognising the previous equations as an eigenequation, the diagonal elements of $\mathbf{\Sigma}$ (which are the eigenvalues of $\mathbf{C}_{uu}$) may be restricted to be real and nonnegative (thus justifying the notation used for the elements of $\mathbf{\Sigma}$). If necessary further restrictions on these values may be applied, and will correspond to tighter constraints on the possible values of $\mathbf{C}_{uu}$ (and consequently on $\mathbf{C}_{vv}$). Combining these results and letting $\mathcal{D}$ be a subset of all real diagonal matrices with positive elements, a constraint set

$$\mathcal{M}_{\mathbf{V},\mathcal{D}} = \{\mathbf{V}\mathbf{\Sigma}\mathbf{V}^\dagger, \mathbf{\Sigma} \in \mathcal{D}\} \tag{5.16}$$

is defined. The criterion placed on $\mathbf{C}_{uu}$ is that it belongs to this set. $\mathcal{D}$ must be chosen so that *only* valid candidates for the maximum likelihood covariance estimate are included in this set.

The form of the constraint in Equation 5.16 is such that the complete data are conveniently parameterised: since the complete data observations are decorrelated by the matrix $\mathbf{V}^\dagger$, the diagonal elements of $\mathbf{\Sigma}$ completely describe the covariance matrix. For maximum likelihood estimation, the constrained complete data covariance estimate can be obtained by maximising the likelihood over the allowed values of $\sigma_1^2, \ldots, \sigma_s^2$. To this end, it is useful to define the transformed vector $\mathbf{d}_i = \mathbf{V}^\dagger \mathbf{u}_i$, which is distributed according to $\mathbf{d}_i : N[\mathbf{0}, \mathbf{\Sigma}]$. In this transformed coordinate system, the maximum likelihood estimate $\widehat{\mathbf{\Sigma}}$ of the complete data covariance matrix can be calculated by maximising the likelihood of the observations $\mathbf{d}_1, \ldots, \mathbf{d}_M$ over $\mathbf{\Sigma}$, subject to the condition $\mathbf{\Sigma} \in \mathcal{D}$. If required, the constrained MLE of the data in the original coordinate system is then $\widehat{\mathbf{C}}_{uu} = \mathbf{V}\widehat{\mathbf{\Sigma}}\mathbf{V}^\dagger$.

From a statistical point of view the decorrelated data vectors are *entirely* equivalent to the original complete data — they are just represented in a different coordinate system which has been rotated around the origin. However, because the decorrelated data are easier to work with, the derivation of the EM algorithm is simplified by considering $\mathbf{d}_1, \ldots, \mathbf{d}_M$ to be the complete data, rather than $\mathbf{u}_1, \ldots, \mathbf{u}_M$. For purposes of the algorithm, the relationship between the incomplete and complete data is therefore taken to be

$$\mathbf{v}_i = \mathbf{A}\mathbf{V}\mathbf{d}_i, \tag{5.17}$$

with $\mathbf{d}_i$ assumed to have a $N[\mathbf{0}, \mathbf{\Sigma}]$ distribution. The constraint on $\mathbf{C}_{vv}$ is thereby expressed in terms of the constraint on $\mathbf{\Sigma}$, namely that $\mathbf{\Sigma} \in \mathcal{D}$. Note that this reformulation by no means contradicts the previous discussion regarding the selection and utility of the hypothetical complete data observations $\mathbf{u}_1, \ldots, \mathbf{u}_M$; the conditions imposed on $\mathbf{\Sigma}$ are entirely equivalent to those imposed on $\mathbf{C}_{uu}$.

Having outlined the optimisation criteria and the relationship between the complete and incomplete data, the expectation and maximisation steps of the EM algorithm can be described. The development

is fairly straightforward, but leads to complicated expressions for the relevant quantities involved in the calculation. From an implementation viewpoint, the expectation and maximisation steps are iterated until convergence occurs, at which point the solution is obtained by transforming back into the original parameter space.

### 5.3.3 Expectation step

Recall that the estimation step of the EM algorithm involves finding the expected value of the complete data log-likelihood function, conditional on the current best estimate of the unknown parameters and the observed data $\mathbf{v}_1, \ldots, \mathbf{v}_M$. The log-likelihood function of the complete data therefore first has to be specified.

The log-likelihood for the data is different under conditions of real and complex numbers. For real numbers, it is given by

$$L(\boldsymbol{\Sigma}; \mathbf{d}_1, \ldots, \mathbf{d}_M) = -\frac{Ms}{2} \ln 2\pi - \frac{M}{2} \left[ \ln |\boldsymbol{\Sigma}| + \frac{1}{M} \sum_{j=1}^{M} \mathbf{d}_j^T \boldsymbol{\Sigma}^{-1} \mathbf{d}_j \right], \tag{5.18}$$

while for the complex case it is

$$L(\boldsymbol{\Sigma}; \mathbf{d}_1, \ldots, \mathbf{d}_M) = -Ms \ln \pi - M \left[ \ln |\boldsymbol{\Sigma}| + \frac{1}{M} \sum_{j=1}^{M} \mathbf{d}_j^\dagger \boldsymbol{\Sigma}^{-1} \mathbf{d}_j \right]. \tag{5.19}$$

These expressions are very similar, and can both be written in the form

$$L(\boldsymbol{\Sigma}; \mathbf{d}_1, \ldots, \mathbf{d}_M) = K - J \left[ \ln |\boldsymbol{\Sigma}| + \frac{1}{M} \sum_{j=1}^{M} \mathbf{d}_j^\dagger \boldsymbol{\Sigma}^{-1} \mathbf{d}_j \right], \tag{5.20}$$

with the constants $K$ and $J$ chosen accordingly. By considering this latter form, the complex and real cases may be considered at once. To maximise the likelihood, it is required to minimise the quantity in square brackets. Note that the log-likelihood in these expressions would be exactly the same if

calculated for the alternative complete data observations $\mathbf{u}_1, \ldots, \mathbf{u}_M$, with $\mathbf{C}_{uu} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^\dagger$:

$$\begin{aligned} L(\mathbf{C}_{uu}; \mathbf{u}_1, \ldots, \mathbf{u}_M) &= K - J\left[\ln|\mathbf{C}_{uu}| + \frac{1}{M}\sum_{j=1}^{M}\mathbf{u}_j^\dagger\mathbf{C}_{uu}^{-1}\mathbf{u}_j\right] \\ &= K - J\left[\ln|\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^\dagger| + \frac{1}{M}\sum_{j=1}^{M}\mathbf{d}_j^\dagger\mathbf{V}^\dagger(\mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^\dagger)^{-1}\mathbf{V}\mathbf{d}_j\right] \\ &= K - J\left[\ln|\boldsymbol{\Sigma}| + \frac{1}{M}\sum_{j=1}^{M}\mathbf{d}_j^\dagger\boldsymbol{\Sigma}^{-1}\mathbf{d}_j\right]. \end{aligned} \tag{5.21}$$

The equivalence of the complete data representations is therefore evident.

With $\boldsymbol{\Sigma}$ diagonal, the general log-likelihood in Equation 5.20 can be rewritten as

$$L(\boldsymbol{\Sigma}; \mathbf{d}_1, \ldots, \mathbf{d}_M) = K - J\left[\sum_{k=1}^{s}\ln\sigma_k^2 + \frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{s}\frac{|d_j(k)|^2}{\sigma_k^2}\right], \tag{5.22}$$

where $d_j(k)$ is the $k$th element of the vector $\mathbf{d}_j$. The required conditional expectation of the complete data log-likelihood at the $g$th iteration is therefore

$$\begin{aligned} E\{L|\boldsymbol{\Sigma}^{(g)}; \mathbf{v}_1, \ldots, \mathbf{v}_M\} &= K - J\left[\sum_{k=1}^{s}\ln\sigma_k^2 + \frac{1}{M}\sum_{k=1}^{s}\sum_{j=1}^{M}\frac{E\{|d_j(k)|^2|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_1, \ldots, \mathbf{v}_M\}}{\sigma_k^2}\right] \\ &= K - J\left[\sum_{k=1}^{s}\ln\sigma_k^2 + \sum_{k=1}^{s}\frac{1}{\sigma_k^2}\left(\frac{1}{M}\sum_{j=1}^{M}E\{|d_j(k)|^2|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_1, \ldots, \mathbf{v}_M\}\right)\right]. \end{aligned} \tag{5.23}$$

Taking a similar approach to Miller et al. [95], the quantity in round brackets can be seen to be the $k$th diagonal element of the matrix

$$\mathbf{C}_{dd}^{(g)} = \frac{1}{M}\sum_{j=1}^{M}E\{\mathbf{d}_j\mathbf{d}_j^\dagger|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_1, \ldots, \mathbf{v}_M\}. \tag{5.24}$$

Since the observations are independent,

$$E\{\mathbf{d}_j\mathbf{d}_j^\dagger|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_1, \ldots, \mathbf{v}_M\} = E\{\mathbf{d}_j\mathbf{d}_j^\dagger|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_j\}. \tag{5.25}$$

To calculate this expectation, the joint distribution of $\mathbf{d}_j$ and $\mathbf{v}_j$ is required: with $\mathbf{K}_{vv} = \mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)}\mathbf{V}^\dagger\mathbf{A}^\dagger$ and $\mathbf{K}_{dd} = \boldsymbol{\Sigma}^{(g)}$ we have

$$\begin{pmatrix}\mathbf{v}_j \\ \mathbf{d}_j\end{pmatrix} : N\left[\begin{pmatrix}\mathbf{0} \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{K}_{vv} & \mathbf{K}_{vd} \\ \mathbf{K}_{dv} & \mathbf{K}_{dd}\end{pmatrix}\right]. \tag{5.26}$$

The quantity $\mathbf{K}_{dv} = \mathbf{K}_{vd}^{\dagger}$ can be calculated as

$$\begin{aligned}\mathbf{K}_{dv} &= E\{\mathbf{d}_j\mathbf{v}_j^{\dagger}\} = E\{\mathbf{d}_j\mathbf{d}_j^{\dagger}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger}\} = E\{\mathbf{d}_j\mathbf{d}_j^{\dagger}\}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger} \\ &= \boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger}. \end{aligned} \tag{5.27}$$

The conditional distribution of $\mathbf{d}_j$ given $\mathbf{v}_j$ and $\boldsymbol{\Sigma}^{(g)}$ is therefore [113]

$$\mathbf{d}_j|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_j : N[\mathbf{K}_{dv}\mathbf{K}_{vv}^{-1}\mathbf{v}_j, \boldsymbol{\Sigma}^{(g)} - \mathbf{K}_{dv}\mathbf{K}_{vv}^{-1}\mathbf{K}_{vd}], \tag{5.28}$$

from which it can be shown that

$$E\{\mathbf{d}_j\mathbf{d}_j^{\dagger}|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_j\} = \mathbf{K}_{dv}\mathbf{K}_{vv}^{-1}\mathbf{v}_j\mathbf{v}_j^{\dagger}\mathbf{K}_{vv}^{-1}\mathbf{K}_{vd} + \boldsymbol{\Sigma}^{(g)} - \mathbf{K}_{dv}\mathbf{K}_{vv}^{-1}\mathbf{K}_{vd}. \tag{5.29}$$

This latter expression may be obtained by considering the conditional variance of $\mathbf{d}_j$, and using the Gauss-Markov theorem. Using this result in Equation 5.24 we have that

$$\mathbf{C}_{dd}^{(g)} = \mathbf{K}_{dv}\mathbf{K}_{vv}^{-1}\mathbf{S}_{vv}\mathbf{K}_{vv}^{-1}\mathbf{K}_{vd} + \boldsymbol{\Sigma}^{(g)} - \mathbf{K}_{dv}\mathbf{K}_{vv}^{-1}\mathbf{K}_{vd}, \tag{5.30}$$

where $\mathbf{S}_{vv} = 1/M\sum_{j=1}^{M}\mathbf{v}_j\mathbf{v}_j^{\dagger}$ is the sample covariance of the incomplete data observations. Writing explicitly, this gives

$$\begin{aligned}\mathbf{C}_{dd}^{(g)} &= \boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger}(\mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger})^{-1}\mathbf{S}_{vv}(\mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger})^{-1}\mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)} \\ &\qquad + \boldsymbol{\Sigma}^{(g)} - \boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger}(\mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger})^{-1}\mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)}. \\ &= \boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger}(\mathbf{C}_{vv}^{(g)})^{-1}\mathbf{S}_{vv}(\mathbf{C}_{vv}^{(g)})^{-1}\mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)} + \boldsymbol{\Sigma}^{(g)} - \boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger}(\mathbf{C}_{vv}^{(g)})^{-1}\mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)}, \end{aligned} \tag{5.31}$$

where use has been made of the fact that the current best estimate of the required covariance matrix on the $g$th iteration is

$$\mathbf{C}_{vv}^{(g)} = \mathbf{A}\mathbf{V}\boldsymbol{\Sigma}^{(g)}\mathbf{V}^{\dagger}\mathbf{A}^{\dagger}. \tag{5.32}$$

Once this matrix has been calculated, the diagonal elements may be substituted into the expression for the expected conditional log-likelihood in Equation 5.23. It is this resulting expression which is required for the maximisation stage of the iteration.

### 5.3.4 Maximisation step

This step involves maximising the conditional likelihood $E\{L|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_1, \ldots, \mathbf{v}_M\}$ in Equation 5.23 over the allowed space $\boldsymbol{\Sigma} \in \mathcal{D}$. Any constraints on $\boldsymbol{\Sigma}$ relate directly to conditions on the allowable set of values of the parameters $\sigma_1^2, \ldots, \sigma_s^2$ in Equation 5.23. In some instances no constraints are required on these values, beyond the need for them to be real and nonnegative. However, sometimes this does not

sufficiently restrict the values of the complete data covariance $\mathbf{C}_{uu}$, in which case additional conditions are required.

To be as general as possible, the maximisation problem could be formulated with a general linear constraint on the parameters, for example with $L$ constraints of the form

$$\mathbf{c}_i^T \boldsymbol{\sigma} = \mathbf{b}_i, \qquad i = 1, \ldots, L, \tag{5.33}$$

with $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_s^2)^T$. However, this does not lead to a closed-form solution due to the nonlinearity of the objective function $E\{L|\boldsymbol{\Sigma}^{(g)}, \mathbf{v}_1, \ldots, \mathbf{v}_M\}$. It is possible to formulate an effective search for an extremum given such linear constraints [128, p. 155]. Such a formalism will however not be presented: for our purposes the constraints are generally simple, and specific solutions will be provided for each individual case of interest.

Having determined the values of the parameters $(\sigma_1^2)^{(g+1)}, \ldots, (\sigma_s^2)^{(g+1)}$ which maximises the conditional expected log-likelihood, the next estimate of the parameter matrix to be used in the iteration is

$$\boldsymbol{\Sigma}^{(g+1)} = \mathrm{diag}((\sigma_1^2)^{(g+1)}, \ldots, (\sigma_s^2)^{(g+1)}). \tag{5.34}$$

This constitutes the new best estimate of the complete data covariance, and can be used in the succeeding expectation step. Alternatively, if the likelihood is considered high enough, relevant estimates of $\mathbf{C}_{uu}$ or $\mathbf{C}_{vv}$ can be calculated directly from this quantity, and the iterations terminated.

## 5.4 Subspace invariant estimation for circulant covariance matrices

In this section the results of the previous discussion are applied to the problem of finding the maximum likelihood covariance matrix estimate under a positive semidefinite circulant constraint. It is assumed that the estimate is required to be invariant to data components lying in a known linear subspace of the original observation space. The presentation of a solution for the circulant case is useful on two counts: firstly, it highlights the differences in formulations for the cases of complex and real data; secondly, it provides the basic framework for a more interesting constraint, namely that of the covariance matrix being Toeplitz. The case of complex data is dealt with first — it permits a simpler specification of the constraint set, and results in easier calculations. This is followed by a discussion of the real case.

Two different interpretations of the invariant estimation problem are possible, depending on whether the data are considered to be corrupted by subspace interference or low-rank model mismatch. In both cases it is assumed that the corrupted full-rank $n$-dimensional observations $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are given, and that the invariance subspace is $\langle \mathbf{U}_I \rangle$:

- In the first case, the data $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are assumed to be corrupted by additive subspace interference

in the subspace $\langle \mathbf{U}_I \rangle$. That is, each observation $\mathbf{x}_i$ is of the form $\mathbf{x}_i = \mathbf{z}_i + \mathbf{U}_I \mathbf{c}_i$, with $\mathbf{z}_i$ a realisation of an underlying $N[\mathbf{0}, \mathbf{C}]$ process and $\mathbf{c}_i$ completely unknown. It is required to estimate the covariance matrix $\mathbf{C}$, subject to the constraint that it be circulant, Hermitian, and positive semidefinite.

- In the second case, the data $\mathbf{x}_1, \ldots, \mathbf{x}_M$ are assumed to be drawn from a random process which only partially conforms to the assumption of zero-mean MVN with a circulant covariance. Specifically, the model is only considered appropriate in the subspace $\langle \mathbf{U}_H \rangle$, orthogonal to $\langle \mathbf{U}_I \rangle$. The problem is to estimate the underlying circulant covariance matrix, under the additional requirement that it be Hermitian and positive semidefinite.

In both instances the only portions of the original data which may be considered valid for estimation are $\mathbf{y}_1, \ldots, \mathbf{y}_M$, with $\mathbf{y}_i = \mathbf{U}_H^T \mathbf{x}_i$. In a natural way, these vectors constitute the incomplete data for the estimation problem. The interpretation for the complete data is different for the two cases, however. The subspace interference problem is the simplest: the complete data can be considered to be the underlying observations $\mathbf{z}_1, \ldots, \mathbf{z}_M$, which have not yet been corrupted by interference. For the subspace mismatch problem, on the other hand, it is possible to *postulate* similar (but hypothetical) complete data observations $\mathbf{z}_1, \ldots, \mathbf{z}_M$, which do not suffer from the mismatch present in the original observations, but nevertheless share the same conditional distribution in the subspace $\langle \mathbf{U}_H \rangle$.

Irrespective of whether the process for which the covariance is to be estimated actually exists or not, the formulation of the EM algorithm is identical. The complete data are taken to be $\mathbf{z}_1, \ldots, \mathbf{z}_M$, each with a $N[\mathbf{0}, \mathbf{C}]$ distribution. The incomplete data are taken to be $\mathbf{y}_1, \ldots, \mathbf{y}_M$, with $\mathbf{y}_i = \mathbf{U}_H^\dagger \mathbf{x}_i$ for each $i$. By assumption $\mathbf{U}_H^\dagger \mathbf{x}_i = \mathbf{U}_H^\dagger \mathbf{z}_i$, so the relationship between the complete and incomplete data is $\mathbf{y}_i = \mathbf{U}_H^\dagger \mathbf{z}_i$.

The EM formulation given in the previous section can therefore be applied by taking $(\mathbf{v}_1, \ldots, \mathbf{v}_M) = (\mathbf{y}_1, \ldots, \mathbf{y}_M)$, each with dimension $q = n - p$. The complete data observations $(\mathbf{u}_1, \ldots, \mathbf{u}_M)$ in that section are taken to be $(\mathbf{z}_1, \ldots, \mathbf{z}_M)$ as described above, each with dimension $s = n$. The transformation matrix between the complete and incomplete data is $\mathbf{A} = \mathbf{U}_H^\dagger$. All that remains is to specify the parameters dictating the constraint set, namely $\mathbf{V}$ and $\mathcal{D}$.

It is at this point that the development between the real and complex cases differ. In particular, for the real case the set of candidate covariance matrices is required to only have real elements. The two formulations therefore differ quite considerably.

### 5.4.1 Mathematical preliminaries

For the problem under discussion in this section, it is required to maximise the likelihood of the complete data over the set of positive definite Hermitian circulant matrices. In the first instance, these

matrices are permitted to be complex, but in the second an additional requirement is made that the covariance matrix be real. In both cases the set $\mathcal{M}_{\mathbf{V},\mathcal{D}}$ in Equation 5.16 needs to be specified in such a way that it contains exactly the matrices of interest. To this end, some results pertaining to circulant matrices are required. In particular, a useful property of general circulant matrices is that they are diagonalised by the discrete Fourier transform (DFT).

Before continuing, it is useful to clearly define the $n$-dimensional unitary DFT matrix $\mathbf{W}_n$. Using the notation of [102], we let $w_n = e^{-i\frac{2\pi}{n}}$ and define

$$\mathbf{W}_n = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & w_n^{1(1)} & \cdots & w_n^{1(n-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & w_n^{(n-1)1} & \cdots & w_n^{(n-1)(n-1)} \end{pmatrix}. \tag{5.35}$$

Now, $1/\sqrt{n}\mathbf{W}_n\mathbf{z}$ is the Fourier transform of $\mathbf{z}$ as usually defined. Since $\mathbf{W}_n$ is unitary, it is useful to let

$$\mathbf{W}_n^\dagger = \begin{pmatrix} \mathbf{w}_n^0 & \mathbf{w}_n^1 & \cdots & \mathbf{w}_n^{n-1} \end{pmatrix}, \tag{5.36}$$

where

$$\mathbf{w}_n^k = \frac{1}{\sqrt{n}} \begin{pmatrix} w_n^{-k(0)} \\ w_n^{-k(1)} \\ \vdots \\ w_n^{-k(n-1)} \end{pmatrix} \tag{5.37}$$

constitutes the $k$th orthonormal DFT basis vector.

Now, let $\mathbf{C}$ be the circulant matrix

$$\mathbf{C} = \begin{pmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \cdots & c_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & c_3 & \cdots & c_0 \end{pmatrix}. \tag{5.38}$$

The following two theorems are proved by Davis [24, p. 72]:

- $\mathbf{C}$ is diagonalised by $\mathbf{W}_n$, so $\mathbf{W}_n\mathbf{C}\mathbf{W}_n^\dagger = \mathbf{\Lambda}$ for some $\mathbf{\Lambda} = \operatorname{diag}(\lambda_0, \ldots, \lambda_{n-1})$, and
- $\mathbf{\Lambda} = \operatorname{diag}(\lambda_0, \ldots, \lambda_{n-1})$ implies that $\mathbf{W}_n^\dagger\mathbf{\Lambda}\mathbf{W}_n$ is circulant.

In both cases, $\lambda_k \in \mathbb{C}$. These provide necessary and sufficient conditions for a matrix to be circulant. However, the results are not specific enough, even for the complex case, since *all* circulant matrices are included in the characterisation — the requirement is for positive semidefinite Hermitian matrices.

**Section 5.4: Subspace invariant estimation for circulant covariance matrices**

When $\mathbf{C}$ is Hermitian, then it has real eigenvalues [85, p. 75]. Noting that the expression $\mathbf{W}_n\mathbf{C}\mathbf{W}_n^\dagger = \mathbf{\Lambda}$ in the previous paragraph is an eigenequation, the diagonal elements of $\mathbf{\Lambda}$ must therefore be real, since they are identically the eigenvalues of $\mathbf{C}$. Furthermore, if $\mathbf{C}$ is positive semidefinite, then these elements must additionally be nonnegative. Therefore it may be asserted that if $\mathbf{C}$ is circulant and positive semidefinite, then $\mathbf{W}_n\mathbf{C}\mathbf{W}_n^\dagger$ is diagonal, with real nonnegative elements. The converse of this result is also true: firstly, if $\lambda_k \in \mathbb{R}$ then $\mathbf{W}_n^\dagger\mathbf{\Lambda}\mathbf{W}_n$ is Hermitian, since

$$(\mathbf{W}_n^\dagger\mathbf{\Lambda}\mathbf{W}_n)^\dagger = \mathbf{W}_n^\dagger\mathbf{\Lambda}^\dagger\mathbf{W}_n = \mathbf{W}_n^\dagger\mathbf{\Lambda}\mathbf{W}_n. \tag{5.39}$$

Secondly, if $\lambda_k$ is nonnegative for each $k$, then the matrix $\mathbf{\Lambda}$ is nonnegative. Since $\mathbf{W}_n$ is invertible, $\mathbf{W}_n^\dagger\mathbf{\Lambda}\mathbf{W}_n$ is nonnegative definite.

Summarising these results, if $\mathbf{C} = \mathbf{W}_n^\dagger\mathbf{\Lambda}\mathbf{W}_n$ then:

- If $\mathbf{\Lambda}$ is real and diagonal with nonnegative elements, then $\mathbf{C}$ is circulant, Hermitian, and nonnegative definite, and
- If $\mathbf{C}$ is circulant, Hermitian, and nonnegative definite, then $\mathbf{\Lambda}$ is real and diagonal with nonnegative elements.

Thus the set

$$\{\mathbf{W}_n^\dagger\mathbf{\Lambda}\mathbf{W}_n | \lambda_i \in \mathbb{R}, \lambda_i \geq 0\} \tag{5.40}$$

is *precisely* the set of all positive semidefinite Hermitian matrices with a circulant structure. There are no extra or missing elements in this set. Additionally, if a characterisation is required for the set of all positive definite matrices of this form, then it is sufficient to restrict the elements $\lambda_0, \ldots, \lambda_{n-1}$ to be positive.

A similar characterisation can be developed for the set of positive semidefinite symmetric circulant matrices, which are required to be real. Since these matrices are all included in the set described previously, further conditions are required on the diagonal elements of $\mathbf{\Lambda}$ to ensure that the matrix $\mathbf{W}_n^\dagger\mathbf{\Sigma}\mathbf{W}_n$ also be real.

In this context it is useful to characterise the nature of the diagonal elements of the product $\mathbf{W}_n\mathbf{S}\mathbf{W}_n^\dagger$, where $\mathbf{S}$ is an arbitrary real symmetric matrix: denoting the diagonal elements by $\delta_0, \ldots, \delta_{n-1}$, we

have

$$\begin{aligned}
\delta_k &= \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} s_{jl} w_n^{k(j-l)} \\
&= \frac{1}{n}\sum_{j=1}^{n} s_{jj} + \frac{1}{n}\sum_{j=1}^{n}\sum_{l<j} s_{jl} w_n^{k(j-l)} + \frac{1}{n}\sum_{j=1}^{n}\sum_{l>j} s_{jl} w_n^{k(j-l)} \\
&= \frac{1}{n}\sum_{j=1}^{n} s_{jj} + \frac{1}{n}\sum_{j=1}^{n}\sum_{l<j} \left( s_{jl} w_n^{k(j-l)} + s_{lj} w_n^{k(l-j)} \right) \\
&= \frac{1}{n}\sum_{j=1}^{n} s_{jj} + \frac{1}{n}\sum_{j=1}^{n}\sum_{l<j} 2 s_{jl} \left( \frac{w_n^{k(j-l)} + w_n^{k(l-j)}}{2} \right) \\
&= \frac{1}{n}\sum_{j=1}^{n} s_{jj} + \frac{1}{n}\sum_{j=1}^{n}\sum_{l<j} 2 s_{jl} \cos\left( \frac{2\pi}{n} k(j-l) \right).
\end{aligned} \tag{5.41}$$

On account of the cosine function being even and periodic,

$$\cos\left( \frac{2\pi}{n} k(j-1) \right) = \cos\left( \frac{2\pi}{n} (n-k)(j-1) \right), \tag{5.42}$$

from which it can be seen that $\delta_k = \delta_{n-k}$. Thus under the conditions described here, the diagonal elements of $\mathbf{W}_n \mathbf{S} \mathbf{W}_n^\dagger$ form an even sequence.

Combining this result with those presented for the complex case, it is evident then that if $\mathbf{C}$ is real, symmetric, circulant, and positive definite, then it can be written in the form $\mathbf{C} = \mathbf{W}_n^\dagger \mathbf{\Lambda} \mathbf{W}_n$ for some real and diagonal matrix $\mathbf{\Lambda}$, which has nonnegative diagonal elements which form an even sequence. (In the context of eigenequations, this means that $\mathbf{C}$ has repeated eigenvalues, as it must if it is symmetric.) All that remains then, for purposes of characterising the required set of real matrices, is to demonstrate that the converse is also true.

Once again drawing on previous results in this section, if $\mathbf{\Lambda}$ is real and diagonal with nonnegative elements, then $\mathbf{G} = \mathbf{W}_n^\dagger \mathbf{\Lambda} \mathbf{W}_n$ is nonnegative definite, Hermitian, and circulant. Therefore, it is required to show that the additional condition $\lambda_k = \lambda_{n-k}$ on the diagonal elements of $\mathbf{\Lambda}$ imply that $\mathbf{G}$ is real. The Hermitian property in that case then ensures the required symmetry. Now, the nature of the elements of $\mathbf{G}$ can be determined by examining the first column of $\mathbf{G}$ in the relation given. Denoting this column by the vector $\mathbf{g}_1$, we have

$$\mathbf{g}_1 = 1/\sqrt{n} \mathbf{W}_n^\dagger \mathbf{\Lambda} \mathbf{1} = 1/\sqrt{n} \mathbf{W}_n^\dagger \boldsymbol{\lambda}, \tag{5.43}$$

where $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_{n-1})^T$. The conjugate of $\mathbf{g}_1$ is

$$\mathbf{g}_1^* = 1/\sqrt{n} \mathbf{W}_n^T \mathbf{\Lambda}^* = 1/\sqrt{n} \mathbf{W}_n \boldsymbol{\lambda}, \tag{5.44}$$

since $\boldsymbol{\lambda}$ is real. Noting that $\lambda_k = \lambda_{n-k}$ simply implies that the sequence $\lambda_0, \ldots, \lambda_{n-1}$ is even (see for example [45, p. 33]), we see that $\mathbf{g}_1^*$ is proportional to the DFT of an even sequence. From the properties of the DFT, $\mathbf{g}_1^*$ is therefore real [45, p. 61]. Thus $\mathbf{g}_1$ is real, and so is $\mathbf{G}$.

The set

$$\{\mathbf{W}_n^\dagger \boldsymbol{\Lambda} \mathbf{W}_n | \lambda_i \in \mathbb{R}, \lambda_i \geq 0, \lambda_i = \lambda_{n-i}^2\} \tag{5.45}$$

is therefore identically the set of all real, symmetric, and positive semidefinite matrices which have a circulant structure. The additional constraint that the diagonal elements of $\boldsymbol{\Lambda}$ form an even sequence is all that is required to restrict the set of matrices to be real.

It may seem disturbing that the set of real matrices with the required characteristics is formulated in terms of the complex matrix $\mathbf{W}_n$. However, a complex formulation is not essential if it is not desired. Specifically, on account of the constraint on the diagonal elements of $\boldsymbol{\Lambda}$, the choice of $\mathbf{W}_n$ to diagonalise the circulant matrices is not unique and may always be chosen to be real. This is because some eigenvalues of $\mathbf{C}$ are necessarily repeated, on account of the matrix being symmetric. In particular, writing $\mathbf{C} = \mathbf{W}_n^\dagger \boldsymbol{\Lambda} \mathbf{W}_n$, the even sequence constraint implies that the eigenvalues corresponding to the eigenvectors $\mathbf{w}_n^j$ and $\mathbf{w}_n^{n-j}$ must be equal. This being the case, only the subspace spanned by $a\mathbf{w}_n^j + b\mathbf{w}_n^{n-j}$ for $a = b \in \mathbb{R}$ may be uniquely specified. Noting that $\mathbf{w}_n^j$ is identically the transpose of $\mathbf{w}_n^{n-j}$, this condition means that the eigenspace corresponding to each set of repeated eigenvalues lies in the real space $\mathbb{R}^n$. Therefore, each pair of eigenvectors with the same eigenvalue may be replaced by two real orthonormal eigenvectors which span the same space. Thus $\mathbf{W}_n^\dagger \boldsymbol{\Lambda} \mathbf{W}_n$ may always be written as $\mathbf{P}^T \boldsymbol{\Lambda} \mathbf{P}$ for some real (and fixed) matrix $\mathbf{P}$. Basilevsky [6, p. 224] discusses the real eigendecomposition of real symmtric circulant matrices in more detail: such a matrix can in general be decomposed using a set of sampled real sine and cosine functions.

For purposes of implementation, however, the complex formulation is preferred because it is more easily implemented using the FFT.

### 5.4.2 Case of complex circulant covariance matrices

With the preliminaries completed, the remainder of the EM formulation can be completed. From the results of the previous section, the set of all complex Hermitian positive semidefinite circulant matrices of dimension $n \times n$ is given by

$$\{\mathbf{W}_n^\dagger \boldsymbol{\Sigma} \mathbf{W}_n | \sigma_i^2 \in \mathbb{R}, \sigma_i^2 \geq 0\}, \tag{5.46}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Thus we may take this to be the covariance matrix constraint set by choosing $\mathbf{V} = \mathbf{W}_n^\dagger$ and $\mathcal{D}$ to be the set of all real diagonal matrices with nonnegative elements.

With $\mathbf{A}$, $\mathbf{V}$, and $\mathcal{D}$ suitably defined, the iterations described in Section 5.3 can be used directly. The expectation stage proceeds trivially, although certain computational savings can be made by exploit-

ing the relationships between circulant matrices, diagonal matrices, and the DFT. The conditional expectation of the complete data log-likelihood at the $g$th iteration is

$$E\{L|\mathbf{\Sigma}^{(g)};\mathbf{y}_1,\ldots,\mathbf{y}_M\} = K - J\left[\sum_{k=1}^{n}\ln\sigma_k^2 + \sum_{k=1}^{n}\frac{1}{\sigma_k^2}\left(\frac{1}{M}\sum_{j=1}^{M}E\{|d_j(k)|^2|\mathbf{\Sigma}^{(g)},\mathbf{y}_1,\ldots,\mathbf{y}_M\}\right)\right], \tag{5.47}$$

with the quantity in round brackets being the $k$th diagonal element of

$$\mathbf{C}_{dd}^{(g)} = \mathbf{\Sigma}^{(g)}\mathbf{W}_n\mathbf{U}_H(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{S}_{yy}(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{U}_H^\dagger\mathbf{W}_n^\dagger\mathbf{\Sigma}^{(g)} + \mathbf{\Sigma}^{(g)} - \mathbf{\Sigma}^{(g)}\mathbf{W}_n\mathbf{U}_H(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{U}_H^\dagger\mathbf{W}_n^\dagger\mathbf{\Sigma}^{(g)}. \tag{5.48}$$

In this expression, $\mathbf{C}_{yy}^{(g)}$ is the current best estimate of the incomplete data covariance at the $g$th iteration, namely

$$\mathbf{C}_{yy}^{(g)} = \mathbf{U}_H^\dagger\mathbf{W}_n^\dagger\mathbf{\Sigma}^{(g)}\mathbf{W}_n\mathbf{U}_H. \tag{5.49}$$

This completes the specification of the expectation step.

The maximisation stage is quite simple, with the conditional expectation in Equation 5.47 being easy to maximise over the choice of parameters $\sigma_1^2,\ldots,\sigma_n^2$. Taking the derivative with respect to $\sigma_i^2$ and equating to zero yields the following necessary condition on the new estimate $\hat{\sigma}_i^2$ of $\sigma_i^2$:

$$-J\left[\frac{1}{\hat{\sigma}_i^2} - \frac{1}{(\hat{\sigma}_i^2)^2}\left(\frac{1}{M}\sum_{j=1}^{M}E\{|d_j(i)|^2|\mathbf{\Sigma}^{(g)},\mathbf{y}_1,\ldots,\mathbf{y}_M\}\right)\right] = 0. \tag{5.50}$$

Thus the required estimates are given by

$$\hat{\sigma}_i^2 = \frac{1}{M}\sum_{j=1}^{M}E\{|d_j(i)|^2|\mathbf{\Sigma}^{(g)},\mathbf{y}_1,\ldots,\mathbf{y}_M\} \tag{5.51}$$

for each value of $i$. Clearly these estimates are nonnegative, as required, so no further work is needed to enforce the constraint.

These estimated values, which maximise the conditional log-likelihood of the complete data, are then used to form the matrix $\mathbf{\Sigma}^{(g+1)}$, which describes the new best estimate of the covariance matrix at the $(g+1)$th iteration. The corresponding covariance matrix estimate for the complete data is given by $\mathbf{C}^{(g+1)} = \mathbf{W}_n^\dagger\mathbf{\Sigma}^{(g+1)}\mathbf{W}_n$.

The iterations described in this section are fairly computationally expensive, but certain savings can be made in the calculation of the conditional likelihood. Miller et al. [95] provide some possible methods for reducing this burden.

### 5.4.3 Case of real circulant covariance matrices

The formulation for the EM iterations in the real case is similar to that of the complex case, except that certain restrictions have to be made on the set of allowable covariances. Specifically, these allowable covariances are

$$\{\mathbf{W}_n^\dagger \mathbf{\Sigma} \mathbf{W}_n | \sigma_i^2 \in \mathbb{R}, \sigma_i^2 \geq 0, \sigma_i^2 = \sigma_{n-i+2}^2\}, \tag{5.52}$$

with $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. The condition $\sigma_i^2 = \sigma_{n-i+2}^2$ implies that the diagonal elements of $\mathbf{\Sigma}$ form an even sequence, with the top left element labelled $\boldsymbol{\sigma}_1^2$. (Note that this numbering is different from that used in Section 5.4.1, where the numbering started at zero to better demonstrate the required relationships.) As for the complex case, we may take $\mathbf{V} = \mathbf{W}_n^\dagger$. The set $\mathcal{D}$ is now the set of all real diagonal matrices whose diagonal elements form an even sequence.

The expectation stage of the algorithm is identical to that discussed in the previous section, since it does not depend on the constraint set $\mathcal{D}$. The formulation can therefore be applied unmodified.

The maximisation stage is complicated by the requirement that the diagonal elements of $\mathbf{\Sigma}$ form an even sequence. This represents a constraint which has to be enforced on these elements. One approach could be to use a Lagrangian multiplier formulation to effect these constraints. However, they are sufficiently simple to be worked directly into the maximisation.

The condition $\sigma_i^2 = \sigma_{n-i+2}^2$ implies that the only unique parameters are $\sigma_1^2, \ldots, \sigma_{\lfloor n/2 \rfloor + 1}^2$. Restricting the discussion to the case of $n$ even (the case of $n$ odd is similar, but leads to slightly different expressions), the conditional expectation in Equation 5.23 can be written, in terms of the unique parameters $\sigma_1^2, \ldots, \sigma_{n/2+1}^2$, as

$$\begin{aligned} E\{L|\mathbf{\Sigma}^{(g)}, \mathbf{y}_1, \ldots, \mathbf{y}_M\} = K - \frac{M}{2}\left\{\ln \sigma_1^2 + 2\sum_{k=2}^{n/2} \ln \sigma_k^2 + \sigma_{n/2+1}^2\right\} \\ - \frac{M}{2}\left\{\frac{c_1^2}{\sigma_1^2} + \sum_{k=2}^{n/2} \frac{c_k^2 + c_{n-k+2}^2}{\sigma_k^2} + \frac{c_{n/2+1}^2}{\sigma_{n/2+1}^2}\right\}, \end{aligned} \tag{5.53}$$

where $c_k^2 = 1/M \sum_{j=1}^M E\{|d_j(k)|^2 | \mathbf{\Sigma}^{(g)}, \mathbf{y}_1, \ldots, \mathbf{y}_M\}$. The derivative with respect to $\sigma_i^2$ of this quantity is

$$\frac{dE\{L|\mathbf{\Sigma}^{(g)}, \mathbf{y}_1, \ldots, \mathbf{y}_M\}}{d\sigma_i^2} = \begin{cases} -\frac{M}{2}\frac{1}{\sigma_i^2} + \frac{M}{2}\frac{c_i^2}{(\sigma_i^2)^2} & i = 1, n/2+1 \\ -M\frac{1}{\sigma_i^2} + \frac{M}{2}\frac{c_i^2 + c_{n-i+2}^2}{(\sigma_i^2)^2} & i = 2\ldots, n/2, \end{cases} \tag{5.54}$$

which, equating to zero and solving, gives

$$\hat{\sigma}_i^2 = \begin{cases} c_i^2 & i = 1, n/2+1 \\ 1/2(c_i^2 + c_{n+2-i}^2) & i = 2\ldots, n/2. \end{cases} \tag{5.55}$$

As with the previous case, these parameters can easily be calculated from the matrix $\mathbf{C}_{dd}^{(g)}$. Again these estimates are clearly nonnegative, as required. They can be used directly as the elements of the best estimate $\mathbf{\Sigma}^{(g+1)}$ of $\mathbf{\Sigma}$ at the $(g+1)$th iteration.

Less apparent is the fact that the exact same algorithm can be used for this real case as for the complex case, as long as the initial covariance estimate is real and symmetric. In order to see this, consider the quantity $\mathbf{C}_{dd}^{(g)}$ in Equation 5.48: using the relation $\mathbf{C}^{(g)} = \mathbf{W}_n^{\dagger}\mathbf{\Sigma}^{(g)}\mathbf{W}_n$, this can be written as

$$\mathbf{C}_{dd}^{(g)} = \mathbf{W}_n[\mathbf{C}^{(g)}\mathbf{U}_H(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{S}_{yy}(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{U}_H^T\mathbf{C}^{(g)}]\mathbf{W}_n^{\dagger} + \mathbf{\Sigma}^{(g)} - \mathbf{W}_n[\mathbf{C}^{(g)}\mathbf{U}_H(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{U}_H^T\mathbf{C}^{(g)}]\mathbf{W}_n^{\dagger}. \quad (5.56)$$

Noting that the quantities in square brackets are both symmetric, the argument presented in Section 5.4.1 demonstrates that the first and last terms in this expression have diagonal elements which form an even sequence. Therefore, if we have $\sigma_i^2 = \sigma_{n+i+2}^2$, then the elements of $\mathbf{C}_{dd}^{(g)}$ are such that $\mathbf{C}_{dd}^{(g)}(k,k) = \mathbf{C}_{dd}^{(g)}(n-k+2, n-k+2)$. Thus the form of the new parameter estimate $\mathbf{\Sigma}^{(g+1)}$ is such that $\mathbf{C}^{(g+1)}$ is real and symmetric, as required.

### 5.4.4 Example of invariant circulant covariance estimation

Some results of applying the invariant constrained covariance matrix estimation procedure to three different real-valued observations are provided in this section. The results assume that only a single noise observation is available for the estimation, so $M = 1$. The samples were generated from ideal zero-mean normal random processes. In each case the actual covariance matrix of the process was chosen to be circulant, but with different correlation properties. The relevant autocorrelation functions for the processes are demonstrated in Figure 5.2. Since the covariance matrices of the processes

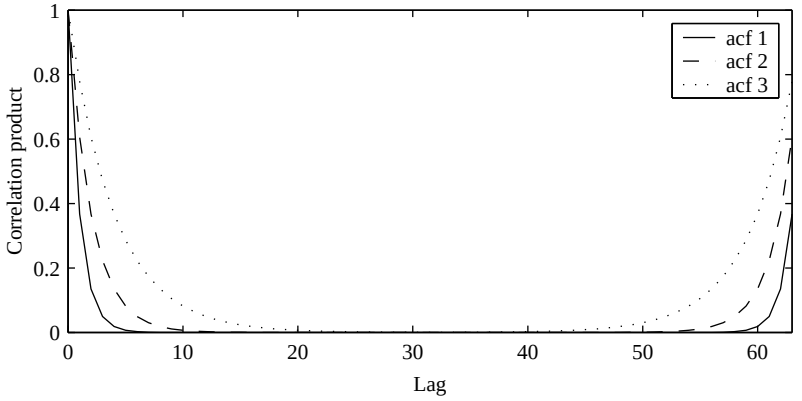



Figure 5.2: Autocorrelation functions for the three different processes used to generate the samples for testing the circulant constrained invariant estimation.

are circulant, these autocorrelation functions completely describe the associated covariances. The autocorrelation functions are each of length 64, as was the sample generated for each process. Since the process is zero mean, the autocorrelation function is the same as the autocovariance function.

Figure 5.3 shows the invariance subspace which was used in the estimation problem in all cases. The

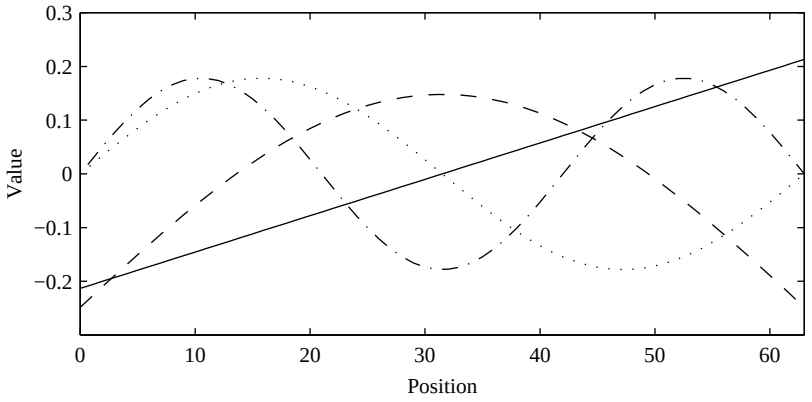
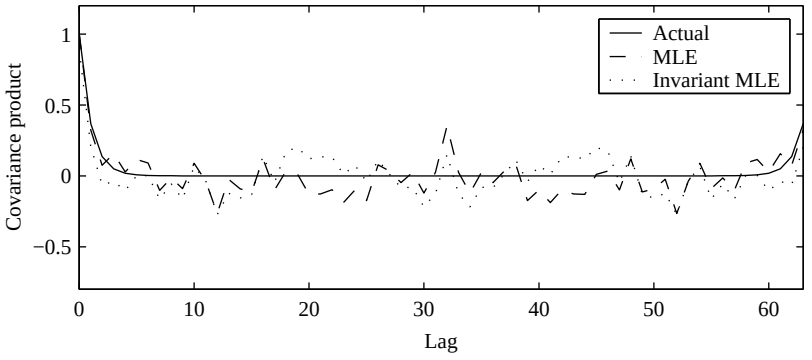



Figure 5.3: Nonorthogonal basis functions for invariance subspace.

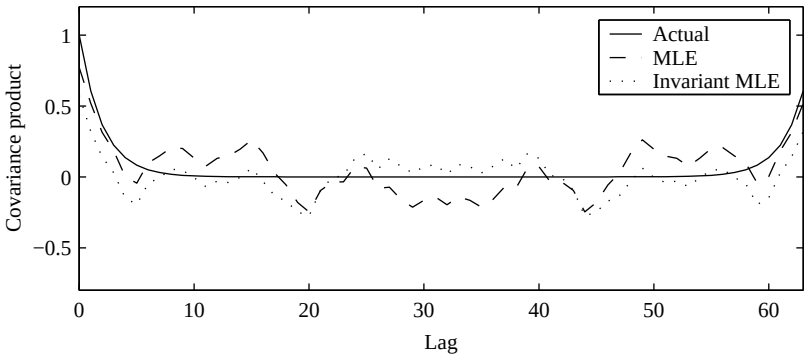
first basis function is a zero-mean ramp, and the remaining three functions are portions of sampled sinusoids with different periods. Thus the invariance subspace is 4-dimensional. This invariance subspace was chosen arbitrarily, simply for purposes of demonstration. Note that although the basis functions have been scaled to have unit energy, they are not orthogonal. This has no effect on the result, and the invariance subspace representation may be orthogonalised if desired.

The results of the estimation for the three cases are shown in the three plots of Figure 5.4. Note that each plot corresponds to estimates based on just a *single* sample signal generated according to the required conditions. The solid line corresponds to the actual autocorrelation function of the process used to generate the sample. Since the data do in fact conform to the circulant assumption, the dashed line in each plot shows the maximum likelihood estimate of the covariance matrix, under a circulant, symmetric, and positive semidefinite constraint, using *all* of the available data. That is, even the data components in the invariance subspace have been used. This represents the best possible estimate that can be made. The dotted plots, however, have only made use of the data in the subspace orthogonal to the invariance subspace in the estimation, still under the constraint that the estimate be circulant, positive definite, and symmetric.

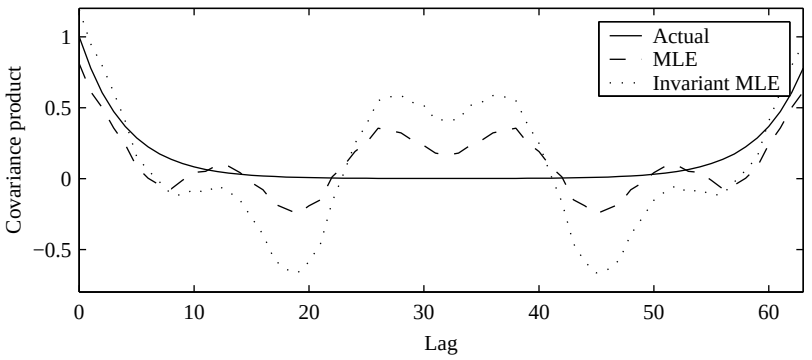
The results are not intended to quantify the extent to which the presence of the invariance subspace degrades the estimation. Such an analysis is difficult to perform in general, since the accuracy of the estimates depends very strongly on the relationship between the invariance subspace and the actual data covariance. However, it may be inferred even from these limited plots that the invariant estimates are not as good as the estimates which have access to all the data. Additionally, as the correlation between




(a) Autocorrelation function 1.




(b) Autocorrelation function 2.




(c) Autocorrelation function 3.

Figure 5.4: Examples of ML and invariant ML covariance estimates under a circulant constraint, for the three different correlation conditions. The solid line in each plot demonstrates the actual autocorrelation function of the process used to generate the samples.

the samples increases, the quality of *both* estimates is reduced. This is common to all estimation problems: high correlations effectively mean less independent data on which to base the estimates, resulting in poorer performance.

One has to be careful not to read too much into the results presented in Figure 5.4, particularly when the estimates are ultimately to be used in an invariant detector. Under this condition, it is not the estimate of $\mathbf{C}(\boldsymbol{\theta})$ which is of interest, but rather the estimate of $\mathbf{U}_H^T\mathbf{C}(\boldsymbol{\theta})\mathbf{U}_H$. The reason for this is quite simple: since the invariant estimator uses no information contained in the subspace $\mathbf{U}_I$, one would expect that the quality of the portion of the estimate relating to $\langle\mathbf{U}_I\rangle$ would be compromised. In fact, for the circulant estimation problem it may happen, as in the case of the unconstrained invariant estimation problem, that certain components of the covariance are not identifiable with subspace invariant estimation. The simplest case where this happens is when the invariance subspace contains the constant vector $\mathbf{1} = (1\ 1\ \ldots\ 1)^T$. In that case the candidate covariances $\mathbf{C}(\boldsymbol{\theta})$ and $\mathbf{C}(\boldsymbol{\theta}) + \alpha\mathbf{1}\mathbf{1}^T$ are both circulant, and both result in the same value for the invariant log-likelihood. Thus the mean value of the overall covariance is not identifiable, and cannot be estimated. As for the unconstrained invariant estimation problem, however, this has no effect on the corresponding invariant detector, which ignores this component of the estimate in any case.

Finally, Figure 5.5 shows the invariant log-likelihood of the covariance estimate at each iteration, for each of the samples discussed. The likelihood initially rises very fast, but the rate of increase drops

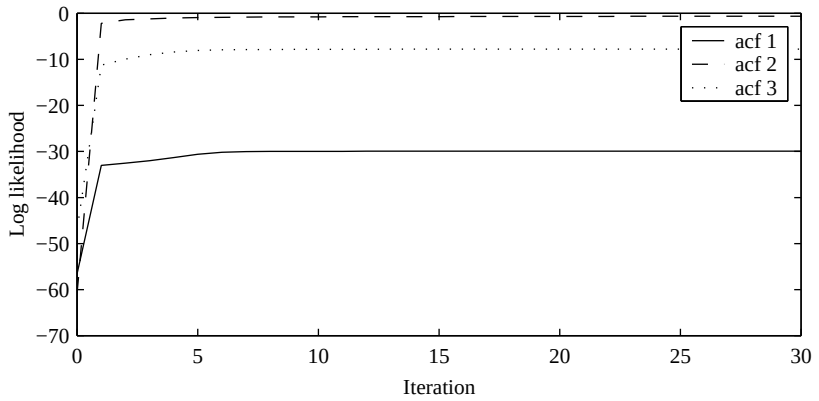



Figure 5.5: Examples of invariant log-likelihood as iterations progress, for the three different autocorrelation functions.

significantly after this climb. The likelihood continues to rise as the iterations commence and the estimate becomes more refined. It has been observed that sometimes a second rise, occasionally quite significant, occurs after quite a large number of iterations. One therefore has to be careful not to terminate the iterations prematurely. While the algorithm has been in use, it has always been seen to converge. Furthermore, it has never been observed that any iteration has decreased the invariant log-likelihood value.

## 5.5 Subspace invariant estimation for Toeplitz covariance matrices

In this section, the previous results are extended to the problem of covariance matrix estimation under a Toeplitz constraint, both with and without an invariance subspace. Initially, the case of no invariance is considered, since it is slightly simpler to explain. This case is fairly extensively discussed in the literature, usually as applied to complex observations and covariances. Once the simpler case has been discussed, extending the results to include an invariance subspace is easy.

### 5.5.1 Estimation without subspace invariance

The standard approach to estimating covariance matrices under a Toeplitz constraint is to once again use a missing data interpretation. The relationship between the complete and incomplete data is of a very specific form: each observed data vector $\mathbf{x}_i$ is assumed to consist of the first $n$ elements of an hypothetical $s$-dimensional complete data vector $\mathbf{u}_i$, which is MVN with an unknown positive definite circulant covariance matrix. The vectors $\mathbf{x}_1, \ldots, \mathbf{x}_M$ play the role of the incomplete data in the EM formulation. The relationship between $\mathbf{x}_i$ and $\mathbf{u}_i$ is therefore

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{u}_i, \tag{5.57}$$

where $\mathbf{I}_{n\times n}$ is the $n \times n$ identity matrix, and $\mathbf{0}$ is a matrix of zeros of the appropriate order. This transformation simply extract the first $n$ elements of $\mathbf{u}_i$.

The approach used for the estimation makes use of the observation that the top-left block of a circulant matrix has a Toeplitz structure. Furthermore, if the matrix is positive semidefinite, then the corresponding block is also positive semidefinite. Explicitly, if $\mathbf{C}$ is circulant and positive definite, then

$$\begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{C} \begin{pmatrix} \mathbf{I}_{n\times n} \\ \mathbf{0} \end{pmatrix} \tag{5.58}$$

is positive semidefinite and Toeplitz. Noting that this matrix is precisely the covariance of $\mathbf{x}_i$ if $\mathbf{u}_i : N[\mathbf{0}, \mathbf{C}]$, it is apparent that maximising the likelihood subject to a circulant constraint on the complete data will result in a Toeplitz estimate for the covariance matrix of the incomplete data.

A circulant positive definite Hermitian constraint on the complete data is therefore equivalent to a Toeplitz positive definite Hermitian constraint on the incomplete data. What is not apparent, however, is to what extent the converse is true: what proportion of $n \times n$ Toeplitz matrices can be embedded into the top-left block of a $s \times s$ circulant matrix? This is an important question, since it is over precisely *these* matrices that the effective maximisation will be performed. Specifically, if the complete data circulant covariance matrix is required to be Hermitian and positive definite, then the implied set of

**Section 5.5: Subspace invariant estimation for Toeplitz covariance matrices**

Toeplitz covariance matrices over which the incomplete data likelihood is maximised is precisely

$$\left\{ \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{C} \begin{pmatrix} \mathbf{I}_{n\times n} \\ \mathbf{0} \end{pmatrix} \middle| \, \mathbf{C} \text{ is Hermitian, positive definite, and circulant} \right\}. \tag{5.59}$$

Evidently this is not exactly equivalent to the set of all positive definite Hermitian Toeplitz matrices — the latter set contains elements not included in the former.

The extent to which the formulation is appropriate for the problem depends on the dimension chosen for the complete data, namely the parameter $s$. This parameter is entirely free in the formulation, and the only effect that it has is on the extent to which the implied constraint approximates the required constraint. For example, if $s$ is chosen to equal $n$, the effective constraint set is just the set of circulant matrices with the required properties. The elements of this set satisfy the requirement that they be Toeplitz (since a circulant matrix is always Toeplitz), but at the same time the set is far smaller than the complete set of Toeplitz matrices. Thus a more stringent constraint has been placed on the candidate covariances than was actually desired. As $s$ is chosen successively larger than $n$, this disparity becomes smaller. Dembo, Mallows, and Shepp [27] demonstrate that *every* finite-dimensional positive definite Toeplitz matrix can be embedded into a circulant matrix of sufficiently large order. They also provide bounds on $s$ required for such an embedding, based on the value of the minimum eigenvalue of the Toeplitz matrix.

In practice the difference between the actual constraint in force and the desired constraint appears not to be critical. As long as $s$ is fairly large relative to $n$ (a factor of two seems acceptible), the resulting estimates are reasonable and useful. However, it should be remembered that the estimate obtained may in fact only be an approximation to what is required.

Returning to the formal development of the algorithm for implementing the estimation, as in the previous section two cases have to be considered: real and of complex covariances. Each requires a different specification of the constraint set. Once described, however, the development of the algorithm details is straightforward.

For the case of complex covariances, the constraint on the complete data covariance $\mathbf{C}$ is that it be Hermitian, positive semidefinite, and circulant. This is identical to the constraint used in Section 5.4.2, so the choice of $\mathbf{V} = \mathbf{W}_n^\dagger$ and $\mathcal{D}$ to be the set of all real diagonal matrices with nonnegative elements is appropriate. The only difference is in the definition of the transformation matrix $\mathbf{A}$. With the incomplete data $(\mathbf{v}_1, \ldots, \mathbf{v}_M)$ taken to be the actual observations $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$, and the complete data $\mathbf{u}_1, \ldots, \mathbf{u}_M$ as described, the relationship provided in Equation 5.57 implies that

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix}. \tag{5.60}$$

This completely specifies the required quantities in the EM formulation.

The implementation details are fundamentally identical to those provided in the section on invariant estimation for circulant matrices, although on account of the simple nature of the relationship between the complete and incomplete data some computational simplifications may be included [95]. The only difference is that now the incomplete data are taken to be the actual observations, rather than the data contained in the valid subspace.

Similarly, the real case is essentially solved using the results of Section 5.4.3, with a similar set of candidates for the complete data covariance $\mathbf{C}$. Once again it may be observed that as long as the initial estimate of the covariance is real and the observations are real, the algorithm for the solution to the complex problem always results in real estimates for the covariance at the subsequent iteration. Thus the two cases may be solved using the same implementation, by simply using an appropriate starting condition.

Finally, some necessary conditions on the convergence of the EM algorithm are known [132], but it is difficult to ascertain whether they are met for this Toeplitz covariance estimation problem [27]. In practice it has been observed that it always seems to provide a good solution. Nevertheless, convergence is sometimes quite slow, and the rate of increase in likelihood as the iterations progress is by no means consistent. This can make it difficult to specify a stopping criterion for the iterations, especially if accuracy is the dominant factor. For signal processing applications it may be reasonable to execute a predetermined number of iterations to improve an estimate, particularly when execution time is a constraint. Since the likelihood is nondecreasing, a fixed number of iterations will *always* provide a better solution (in the maximum likelihood sense) than the original estimate.

### 5.5.2 Estimation with subspace invariance

With the generalised EM formulation presented in Section 5.3, the extension of the Toeplitz covariance estimation problem to include subspace invariance is easily achieved. Suppose the observed data are $\mathbf{x}_1, \ldots, \mathbf{x}_M$, each assumed to nominally be normally distributed, with a zero mean and a Toeplitz covariance matrix. These observations are corrupted by mismatch components in the invariance subspace $\langle \mathbf{U}_I \rangle$. Thus the components $\mathbf{y}_i, \ldots, \mathbf{y}_M$ are all that can be considered valid for estimation, where $\mathbf{y}_i = \mathbf{U}_H^T \mathbf{x}_i$.

Conceptually, it is quite useful to postulate the hypothetical data $\mathbf{z}_1, \ldots, \mathbf{z}_M$, which are identical to the observed data $\mathbf{x}_1, \ldots, \mathbf{x}_M$ except that they have not been corrupted in the invariance subspace. Thus $\mathbf{z}_i$ actually *does* conform to the Toeplitz assumption. The sets of observations however may differ in the invariance subspace, so all that can be guaranteed is that $\mathbf{U}_H^T \mathbf{z}_i = \mathbf{U}_H^T \mathbf{x}_i$.

If $\mathbf{z}_1, \ldots, \mathbf{z}_M$ could actually be observed, the estimation of the Toeplitz covariance matrix could be achieved using the results of the previous section. In particular, the $s$-dimensional complete data vectors $\mathbf{u}_1, \ldots, \mathbf{u}_M$ are proposed, which have a circulant covariance matrix $\mathbf{C}$. The relationship between these

complete data and the hypothetical ideal uncorrupted observations $\mathbf{z}_1, \ldots, \mathbf{z}_M$ is

$$\mathbf{z}_i = \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{u}_i. \tag{5.61}$$

These uncorrupted observations are not observed: all that can be considered observable are the components in the subspace orthogonal to the invariance subspace, namely $\mathbf{U}_H^T\mathbf{z}_1, \ldots, \mathbf{U}_H^T\mathbf{z}_M$. These quite naturally take the role of the incomplete data observations in the EM formulation. It is required to maximise the log-likelihood of these incomplete data observations over the complete data covariance matrix $\mathbf{C}$. Once this estimate is obtained, the required Toeplitz covariance matrix is

$$\begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{C} \begin{pmatrix} \mathbf{I}_{n\times n} \\ \mathbf{0} \end{pmatrix}. \tag{5.62}$$

Now, the observable components $\mathbf{U}_H^T\mathbf{z}_1, \ldots, \mathbf{U}_H^T\mathbf{z}_M$ are precisely $\mathbf{y}_1, \ldots, \mathbf{y}_M$, which can therefore be considered the incomplete data.

With these quantities all specified, it is easy to establish the relationship between the complete data $\mathbf{u}_1, \ldots, \mathbf{u}_M$ and the incomplete data $\mathbf{y}_1, \ldots, \mathbf{y}_M$:

$$\mathbf{y}_i = \mathbf{U}_H^T\mathbf{x}_i = \mathbf{U}_H^T\mathbf{z}_i = \mathbf{U}_H^T \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{u}_i. \tag{5.63}$$

Thus for the EM algorithm, it is appropriate to take

$$\mathbf{A} = \mathbf{U}_H^T \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix}. \tag{5.64}$$

The constraint sets for the complex and real cases are chosen exactly as in the previous discussion, and the form of the quantities to be calculated in the iterations is identical. That is, the conditional expectation of the complete data log-likelihood is

$$L(\mathbf{\Sigma}; \mathbf{d}_1, \ldots, \mathbf{d}_M) = K - J \left[ \sum_{k=1}^{s} \ln \sigma_k^2 + \frac{1}{M} \sum_{j=1}^{M} \sum_{k=1}^{s} \frac{|d_j(k)|^2}{\sigma_k^2} \right], \tag{5.65}$$

which can be calculated using the diagonal elements of

$$\begin{aligned} \mathbf{C}_{dd}^{(g)} &= \mathbf{\Sigma}^{(g)}\mathbf{V}^\dagger\mathbf{A}^\dagger(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{S}_{yy}(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{A}\mathbf{V}\mathbf{\Sigma}^{(g)} + \mathbf{\Sigma}^{(g)} - \mathbf{\Sigma}^{(g)}\mathbf{V}^\dagger\mathbf{A}^\dagger(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{A}\mathbf{V}\mathbf{\Sigma}^{(g)} \\ &= \mathbf{\Sigma}^{(g)}\mathbf{W}_n \begin{pmatrix} \mathbf{I}_{n\times n} \\ \mathbf{0} \end{pmatrix} \mathbf{U}_H(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{S}_{yy}(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{U}_H^\dagger \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{W}_n^\dagger\mathbf{\Sigma}^{(g)} + \mathbf{\Sigma}^{(g)} \qquad (5.66) \\ &\quad - \mathbf{\Sigma}^{(g)}\mathbf{W}_n \begin{pmatrix} \mathbf{I}_{n\times n} \\ \mathbf{0} \end{pmatrix} \mathbf{U}_H(\mathbf{C}_{yy}^{(g)})^{-1}\mathbf{U}_H^\dagger \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{W}_n^\dagger\mathbf{\Sigma}^{(g)}, \qquad (5.67) \end{aligned}$$

with

$$(\mathbf{C}_{yy}^{(g)}) = \mathbf{U}_H^{\dagger} \begin{pmatrix} \mathbf{I}_{n\times n} & \mathbf{0} \end{pmatrix} \mathbf{W}_n^{\dagger} \boldsymbol{\Sigma}^{(g)} \mathbf{W}_n \begin{pmatrix} \mathbf{I}_{n\times n} \\ \mathbf{0} \end{pmatrix} \mathbf{U}_H. \tag{5.68}$$

This completes the specification of the expectation stage. The maximisation step is exactly as described in Sections 5.4.2 and 5.4.3 for the complex and real cases respectively.

### 5.5.3 Example of invariant Toeplitz covariance estimation

Results analogous to those presented in Section 5.4.4 are now given, except for a Toeplitz constraint on the components of the observed data which lie outside the invariance subspace. For the demonstration, the basis functions of this subspace are again taken to be those shown in Figure 5.3.

Again three real-valued noise samples were generated, each of length 64 and from normal random processes with zero mean. This time, however, the actual covariance matrix of each of these processes is taken to be Toeplitz. The autocorrelation functions of the three samples are given in Figure 5.6. Along with the assumption that the covariance matrix be symmetric, these autocorrelation functions

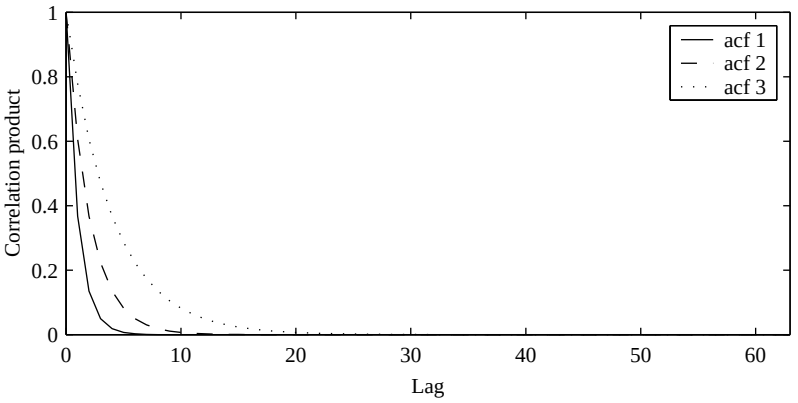
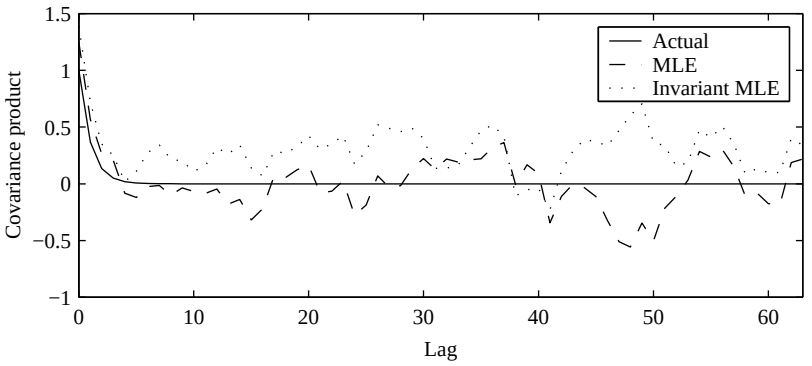



Figure 5.6: Autocorrelation functions for the three different processes used to generate the samples for testing the Toeplitz-constrained invariant estimation.

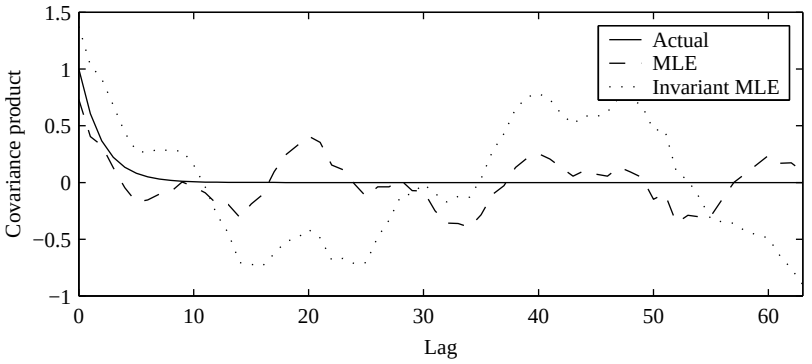
completely determine the associated Toeplitz covariance matrices.

Results for the estimation in each case are shown in Figure 5.7. Again the plots show the true autocorrelation function of the sample, the maximum likelihood estimate using all of the data, and the maximum likelihood estimate using only the data orthogonal to the invariance subspace. In all cases the constraint that the estimated covariance matrix be real, symmetric, positive semidefinite, and Toeplitz is applied.

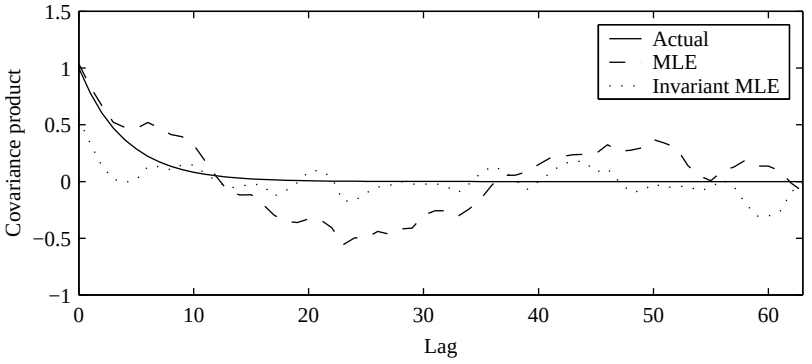
As with the case of the circulant constraint, the estimation appears to improve when the samples are less correlated. Again this can be attributed to there being more effective data available for the estimation.




(a) Autocorrelation function 1.




(b) Autocorrelation function 2.




(c) Autocorrelation function 3.

Figure 5.7: Examples of ML and invariant ML covariance estimates under a Toeplitz constraint, for three different correlation conditions.

Also, as with the circulant case, the estimates may in fact appear to be worse than they are when used in a detector, on account of the reduced accuracy associated with making estimates in the directions of the invariance subspace. The extent to which which such estimates *can* be made depends on the relationships enforced by the Toeplitz constraint on the invariance and complementary subspaces.

Plots of the log-likelihood as the iterations commence are shown in Figure 5.8. It can be seen that the

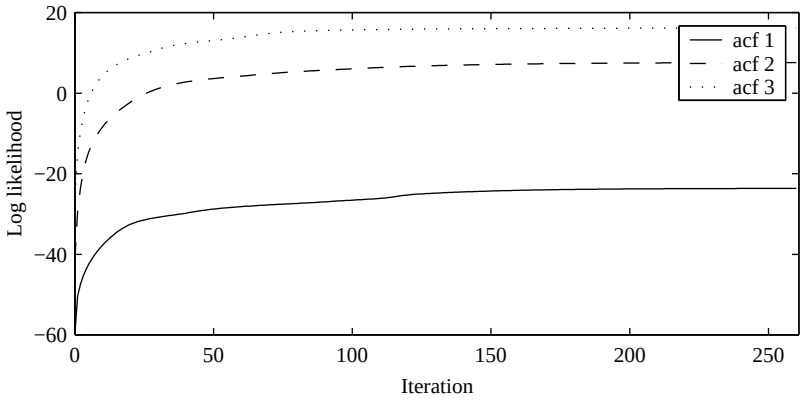



Figure 5.8: Examples of invariant log-likelihood as iterations progress, for three different autocorrelation functions.

convergence is considerably slower than for estimation under the circulant constraint. Additionally, the rate of increase in log-likelihood is quite erratic, with small jumps occurring even after a large number of iterations. Both of these characteristics can be attributed to the fact that for the Toeplitz case there are many more parameters to be optimised over. For example, in the real case the number of free parameters in the circulant estimation is approximately 32, while for the Toeplitz estimation there are approximately 64. The increased number of parameters increases the difficulty of the search, reflected in the slower convergence.

## 5.6 Subspace invariant estimation under alternative constraints

The results presented thus far may be extended to many other cases of potential interest. In particular, for application in two-dimensional imaging environments, maximum likelihood estimation of covariances under doubly block circulant and doubly block Toeplitz constraints is of interest. In the section which follows, the necessary background for these situations is presented, although details of the final iterations are not provided.

Other constraints which are of interest are those of the processes in question being AR, MA, or ARMA. Depending on the order chosen for these processes, the implied constraint on the covariance matrix can be more restrictive than a general stationarity assumption. This is often desirable in instances

where the amount of data available for estimation is small. As with the Toeplitz case, however, as far as invariant detection is concerned it is useful to be able to make invariant estimates of the associated parameter values. This case is discussed briefly in Section 5.6.2.

### 5.6.1 Doubly block circulant and doubly block Toeplitz constraints

A doubly block circulant matrix is a matrix which has a circulant block struture, each block of which is circulant. A doubly block Toeplitz matrix is defined similarly. These matrices arise in statistical characterisation of covariances of cyclic or stationary data, when 2-D observations are reordered into a column vector through column or row ordering.

The constraints can also be cast into the form required by the general algorithm developed in Section 5.3. The most difficult part of the formulation involves the diagonalisation of general doubly block circulant matrices. This section begins with a description of how such matrices can be diagonalised in the complex case. The required quantities for specifying the EM iterations under doubly circulant and doubly Toeplitz constraints are then provided. In the interests of simplicity only complex matrix formulations are considered, and estimation with an invariance subspace is not presented. The extensions are however quite simple, and conceptually follow the same procedures as for the simpler cases of Toeplitz and circulant covariances.

Let $\mathbf{B}$ be a doubly block circulant matrix, with $r \times r$ blocks each of dimension $s \times s$:

$$\mathbf{B} = \begin{pmatrix} \mathbf{C}_0 & \mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_{r-1} \\ \mathbf{C}_{r-1} & \mathbf{C}_0 & \mathbf{C}_1 & \cdots & \mathbf{C}_{r-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{C}_3 & \cdots & \mathbf{C}_0 \end{pmatrix}, \tag{5.69}$$

Note that $\mathbf{B}$ is in general itself not circulant.

Davis [24, p. 185] demonstrates the following points:

- $\mathbf{B}$ is diagonalisable by the unitary matrix $\mathbf{W}_r \otimes \mathbf{W}_s$, where $\otimes$ is the Kronecker product operator [49]. Furthermore, if the matrices of eigenvalues of the circulant blocks are given by $\mathbf{\Lambda}_k,\ k = 0, \ldots, r-1$, then the diagonal matrix of the eigenvalues of $\mathbf{B}$ is given by $\sum_{k=0}^{r-1} \mathbf{\Omega}_r^k \otimes \mathbf{\Lambda}_k$. In this expression, $\mathbf{\Omega}_r = \mathrm{diag}(\mathbf{w}_r^{-0}, \mathbf{w}_r^{-1}, \ldots, \mathbf{w}_r^{-(r-1)})$.
- Any matrix $\mathbf{A} = (\mathbf{W}_r \otimes \mathbf{W}_s)\mathbf{\Lambda}(\mathbf{W}_r \otimes \mathbf{W}_s)$, where $\mathbf{\Lambda}$ is diagonal, is doubly block circulant, with $r \times r$ blocks each of dimension $s \times s$. Since $(\mathbf{W}_r \otimes \mathbf{W}_s) = (\mathbf{W}_r \otimes \mathbf{W}_s)^\dagger$, this can be written in the more conventional form $\mathbf{A} = (\mathbf{W}_r \otimes \mathbf{W}_s)^\dagger \mathbf{\Lambda}(\mathbf{W}_r \otimes \mathbf{W}_s)$.

These are analogous to the results provided for circulant matrices, and give necessary and sufficient

conditions for a matrix to be doubly block circulant. As before, if **B** is Hermitian then its eigenvalues are necessarily real. The converse is also true, although it is not explicitly demonstrated here.

We assume that the observed $(m \times n)$-dimensional data constitute the top-left block of the $(r \times s)$-dimensional complete data, as demonstrated in Figure 5.9. In both cases the data are raster reordered

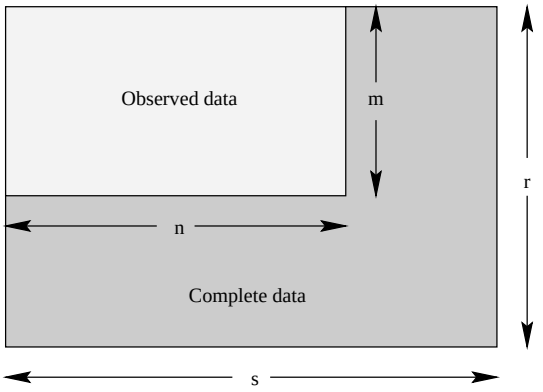



Figure 5.9: Embedding the $(m \times n)$-dimensional observation (incomplete data) into the $(r \times s)$-dimensional complete data.

into a column vector for statistical purposes. (This means that the data are scanned from left-to-right and top-to-bottom, filling the vector sequentially.) Therefore the observed data vector $\mathbf{x}_i$ can be extracted from the complete data vector $\mathbf{u}_i$ by means of the relation

$$\mathbf{x}_i = (\mathbf{I}_{m\times r} \otimes \mathbf{I}_{n\times s})\mathbf{u}_i. \tag{5.70}$$

$\mathbf{I}_{j\times k}$ is a $j \times k$ identity matrix of zeros with ones along the main diagonal. This implies that $\mathbf{A} = (\mathbf{I}_{m\times r} \otimes \mathbf{I}_{n\times s})$ in Equation 5.14.

The constraint placed on the complete data covariance is that it be doubly block circulant of the required dimension. It is easy to see that if the vector $\mathbf{u}_i$ is normally distributed with such a covariance, then the covariance of $\mathbf{x}_i$ will be doubly block Toeplitz. Additionally, the matrix must be nonnegative definite so that the complete data probability density function is defined. The results just presented demonstrate how this class of matrices can be usefully represented: we take $\mathcal{M}$ to be the set

$$\mathcal{M} = \{(\mathbf{W}_r \otimes \mathbf{W}_s)^\dagger \mathbf{\Lambda} (\mathbf{W}_r \otimes \mathbf{W}_s),\, \mathbf{\Lambda} \in \mathcal{D}\}, \tag{5.71}$$

with $\mathcal{D}$ the set of all diagonal matrices with real nonnegative elements. Then $\mathcal{M}$ contains all the elements which we wish to optimise over. Thus $\mathbf{V} = (\mathbf{W}_r \otimes \mathbf{W}_s)^\dagger$ in Equation 5.16, and $\mathbf{\Sigma} = \mathbf{\Lambda}$.

As for the case of optimisation over matrices with circulant extensions, a simple closed form can be obtained for the parameter values for the situation discussed in this section. In a parallel result to

Equation 5.51,

$$\hat{\sigma}_i^2 = \frac{1}{M}\sum_{j=1}^{M} E\{|\mathbf{d}_j(i)|^2|\boldsymbol{\Sigma}^{(g)}, \mathbf{x}_1, \ldots, \mathbf{x}_M\}, \tag{5.72}$$

where all the relevant quantities have been derived from the raster reordered observations. For each $i$, this value is simply the $i$th diagonal element of $\mathbf{C}_{dd}^{(g)}$ in Equation 5.31.

A final simplification arises by considering the value of the product $\mathbf{V}^\dagger\mathbf{A}$ in the calculation. This product is given by

$$\begin{aligned}\mathbf{V}^\dagger\mathbf{A}^\dagger &= (\mathbf{W}_r \otimes \mathbf{W}_s)(\mathbf{I}_{m\times r} \otimes \mathbf{I}_{n\times s})^\dagger \\ &= (\mathbf{W}_r \otimes \mathbf{W}_s)(\mathbf{I}_{r\times m} \otimes \mathbf{I}_{s\times n}) \\ &= \mathbf{W}_r\mathbf{I}_{r\times m} \otimes \mathbf{W}_s\mathbf{I}_{s\times n} \\ &= \mathbf{W}_r^{(m)} \otimes \mathbf{W}_s^{(n)}\end{aligned} \tag{5.73}$$

where $\mathbf{W}_r^{(m)}$ contains the first $m$ columns of $\mathbf{W}_r$, and $\mathbf{W}_s^{(n)}$ contains the first $n$ columns of $\mathbf{W}_s$.

### 5.6.2 ARMA constraints

The final case of interest is that of subspace invariant estimation of covariance matrices under AR, MA, and ARMA constraints. These are especially useful when the amount of data available for estimation is small.

Even without the invariance subspace, maximum likelihood estimation for these models is difficult. In most applications, approximations are used to simplify the objective likelihood function, which in turn leads to an approximate maximum likelihood estimate. To some extent the estimation under an autoregressive assumption is simpler than the cases of MA and ARMA processes, although it still does not permit a convenient solution.

There are some fairly efficient methods available for calculating the likelihood of a general ARMA process given a seqence of successive scalar observations. For example, an algorithm which finds this likelihood using an innovations decomposition is provided by Mélard [94].

Unfortunately, it is difficult to use this type of approach within the EM framework presented in this chapter. This is because the maximisation stage of the EM iteration, at least in the form presented, requires a way of estimating the ARMA parameters from the conditional expectation of the sample covariance matrix, rather than from a single sequence of observed values. Exact likelihood methods for this problem are difficult to obtain, and Porat [104, p. 200] recommends simplifying the problem through appropriate approximate methods. However, this casts the EM procedure in doubt, and since a nondecreasing likelihood can no longer be guaranteed, convergence may not occur with a modified

estimate. Unfortunately the EM method does not appear to permit an alternative formulation, where this problem is overcome.

For this reason, a simpler approach is to maximise the invariant log-likelihood function directly, over the space of allowed parameter values for the general ARMA process. Computationally this is an expensive procedure, but is sufficient as far as investigating ideas is concerned. Furthermore, since the number of parameters in the models is generally low, search methods are not unreasonable for the estimation. Thus, where required, the invariant log-likelihood is simply maximised using the Nelder-Mead simplex algorithm [108]. The matrices over which this maximisation is performed is the set of all possible valid covariance matrices of the assumed process. The set of parameter values is chosen such that the implied ARMA process is stable and invertible. Kay [75, p. 131] provides a method of calculating the autocorrelation function of an ARMA process, which is all that is required in the likelihood calculation.

## 5.7 Simultaneous estimation of interference subspace and covariance matrix

Chapter 4 discussed the problem of estimating the invariance subspace from actual data, when the true covariance matrix is known. Chapter 5 discussed the dual problem of estimating the covariance matrix when the invariance subspace is known. A problem which has not yet been addressed is how to make simultaneous estimates of these two quantities when neither are known.

This is a difficult problem, for which no ideal solution has been found. For the interference subspace estimate, the primary cause of this difficulty is the fact that the proposed estimation method is only approximate, with no maximum likelihood solution having been obtained. For the invariance subspace in the model mismatch problem, the difficulty arises in that an estimate is obtained of the invariance subspace, and not of the complementary subspace.

One method which presents itself is to iterate between the invariance subspace estimation and the covariance matrix estimation, in the hope that the procedure will converge to a reasonable solution for both quantities. There is no guarantee that this method will provide a good solution set, however. In fact, it is difficult to determine whether the procedure will even converge. In some cases it has been seen to provide a reasonable solution, but this is difficult to validate in general.

At least for the subspace interference problem, a better solution is to attempt to maximise the likelihood over both the unknown covariance matrix and the unknown invariance subspace. Thus the invariant likelihood should be maximised over the joint parameters $\mathbf{U}_I$ (or equivalently $\mathbf{U}_H$) and $\mathbf{C}$. Since the value of $\mathbf{U}_H$ that approximately maximises this likelihood is known with respect to $\mathbf{C}$, the explicit dependence on $\mathbf{U}_H$ can be eliminated: we know $\widehat{\mathbf{U}}_H(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ parameterises the covariance matrix.

An approximate modified invariant likelihood can therefore be formulated as

$$L_a = (2\pi)^{-M(n-p)/2}|\widehat{\mathbf{U}}_H^T(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\widehat{\mathbf{U}}_H(\boldsymbol{\theta})|^{-M/2}e^{-\frac{1}{2}\sum_{j=1}^{M}\mathbf{y}_j^T(\widehat{\mathbf{U}}_H^T(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\widehat{\mathbf{U}}_H(\boldsymbol{\theta}))^{-1}\mathbf{y}_j}. \tag{5.74}$$

Maximising over $\boldsymbol{\theta}$ then yields the approximate covariance estimate, from which the corresponding invariance subspace estimate can be obtained.

Unfortunately this procedure can only be applied using direct maximisation over the parameter $\boldsymbol{\theta}$. The invariant covariance estimates for circulant and Toeplitz matrices cannot be used in this context, since they require that the invariance subspace be held constant over the maximisation. Nevertheless, estimates using ARMA processes can be obtained, by adopting the search methods described in the previous section.

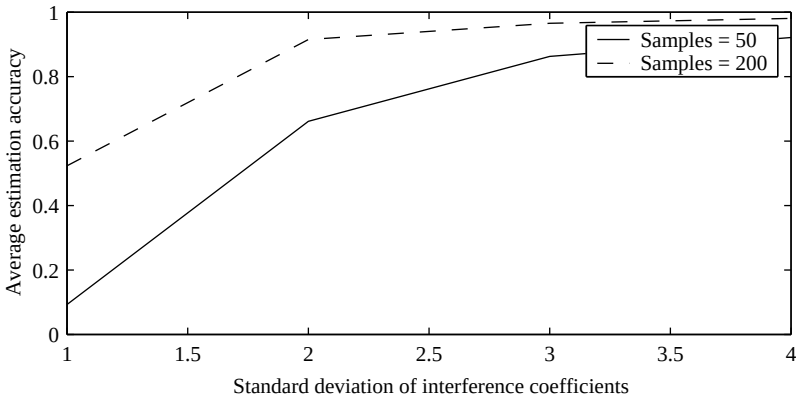
To demonstrate the performance of this proposed procedure for the subspace interference case, the experiment presented in Section 4.3.6 was repeated. Data samples were again generated from first-order AR processes with autocorrelation functions shown in Figure 4.7, and subspace interference added as before. This time the underlying covariance matrix, which generated the samples prior to interference being added, was *not* assumed known. Instead, under just the assumption of a first-order autoregressive model, the procedure outlined in this section was used.

As for the earlier case, this interference subspace estimation procedure was repeated 100 times for each set of parameters of interest. Figure 5.10 shows the results of the simulations, and is analogous to the results in Figure 4.8 for the known covariance case. The performance of the proposed procedure is good, and indicates that the fact that the covariance matrix is unknown has a negligible effect on the accuracy of the interference subspace estimates.

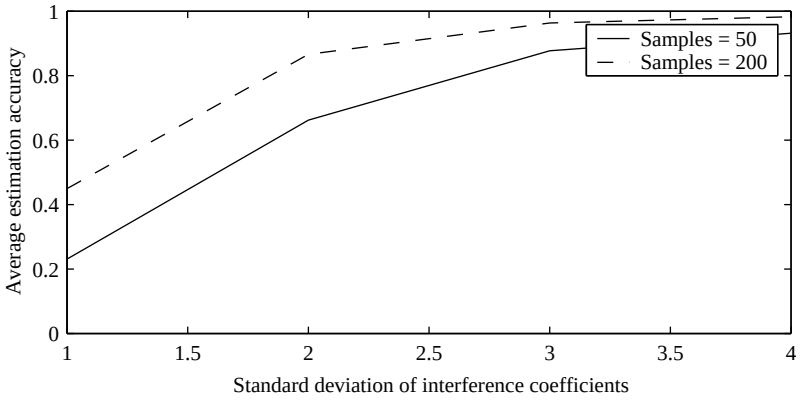
It must be noted that the assumption of a first-order autoregressive process was completely appropriate and valid for the results presented. In general this may not be the case, and it is difficult to predict the effect on the algorithm if the assumed model is invalid. Also, the model used has only two degrees of freedom. For estimating the parameters of this model, the 50 or 200 noise samples represent a huge amount of data on which to base these estimates. It is therefore not surprising that the covariance matrix and interference subspace estimates are good.

If assumptions of more general covariance structures are required, a reasonable approach would be to estimate the interference subspace using the method just described, using an ARMA process of fairly high order. Once estimated, the invariance subspace can be held fixed while an invariant Toeplitz or circulant constrained covariance matrix estimate is made. Although not ideal, if the ARMA process is sufficiently flexible, the interference subspace estimate will be fairly accurate. At the very least, the main causes of nonstationarity will have been eliminated.

For the case of subspace invariance for model mismatch, the same procedure can be used. This time, however, the relationship between the invariance subspace estimate and the covariance matrix is




(a) Results for white noise case.




(b) Results for correlated noise case.

Figure 5.10: Accuracy of interference subspace estimation using artificially corrupted ideal data, when the covariance matrix is not known.

obtained using the result provided in Section 4.5. Although not meaningful in a maximum likelihood context, the procedure still leads to sensible and useful results.

An alternative is to attempt to minimise the Frobenius distance between the covariance and sample covariance in the retained subspaces. That is, the distance $d_F(\mathbf{U}_H^T\mathbf{S}_{xx}\mathbf{U}_H, \mathbf{U}_H^T\mathbf{C}(\boldsymbol{\theta})\mathbf{U}_H)$ can be minimised simultaneously over $\mathbf{U}_H$ and $\boldsymbol{\theta}$. Since this minimisation is not trivial, a similar procedure to that used in the subspace interference problem may be applied: a good invariance subspace is known as a function of $\boldsymbol{\theta}$, so $\widehat{\mathbf{U}}_H(\boldsymbol{\theta})$ can be calculated. Thus an approximate solution is obtained by directly

minimising the modified distance measure

$$d_F(\widehat{\mathbf{U}}_H^T(\boldsymbol{\theta})\mathbf{S}_{xx}\widehat{\mathbf{U}}_H(\boldsymbol{\theta}), \widehat{\mathbf{U}}_H^T(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\widehat{\mathbf{U}}_H(\boldsymbol{\theta})) = \\ \text{tr}\left\{\widehat{\mathbf{U}}_H^T(\boldsymbol{\theta})(\mathbf{S}_{xx} - \mathbf{C}(\boldsymbol{\theta}))\widehat{\mathbf{U}}_H(\boldsymbol{\theta})\widehat{\mathbf{U}}_H^T(\boldsymbol{\theta})(\mathbf{S}_{xx} - \mathbf{C}(\boldsymbol{\theta}))\widehat{\mathbf{U}}_H(\boldsymbol{\theta})\right\} \quad (5.75)$$

with respect to $\boldsymbol{\theta}$.

## 5.8 Practical examples of invariant detectors

The primary justification for invariant detectors is to eliminate the effect of subspace interference or low-rank model mismatch from the detection problem. The invariant detector ignores these components, thereby gaining a potential performance advantage via improved modelling.

A useful characteristic of the invariant detector is that it achieves this improved accuracy without the need for adaptivity. This facilitates the calculation of the detection statistic, and is useful in environments where speed is critical. Ideally, if the invariance subspace is correctly identified, the resulting detector will be robust to the types of interference or model mismatch which commonly occur.

In this section, it is shown how an invariant detector can improve overall performance in the case of certain real-data applications. To make a fair assessment, it is necessary to clarify the characteristics that a good detector should have. Firstly, the detection statistic should be highly discriminatory: that is, it should differ maximally under the conditions of target absence and target presence. However, in most cases this is not sufficient to define a good detector. Just as important, is the requirement that the values which the statistic take on under the various hypotheses be consistent and well characterised. At some point a decision has to be made based on the value of the detection statistic; if this decision cannot be made meaningfully, the detector is useless in practice.

### 5.8.1 Invariance subspace for interference

The first example uses real samples of data, and attempts to assess whether the noise plus subspace interference assumption can be successfully used to model the data for purposes of detection. In particular, subspace interference of various dimensionalities is assumed to be present in the observations, and simultaneous estimates are obtained of the interference subspaces and covariance matrix parameters. A highly-constrained covariance model is initially assumed, namely that of a first-order autoregressive process, so the procedure presented in Section 5.7 for the estimation is appropriate.

Two samples of data were used, extracted from disjoint regions in an x-ray image of a human torso.

The first was assigned the role of training data, and the second of testing data. The samples were specifically chosen to be fairly homogeneous, so that the assumption of spatial stationarity is largely appropriate. Also, they were obtained from regions in the vicinity of the spine, resulting in a high degree of texture. The samples, each of dimension $64 \times 384$, are shown in Figure 5.11. Thay are

(a) Training data.

(b) Testing data.

Figure 5.11: Training and test data for subspace interference demonstration.

sufficiently similar that ordinarily one would feel justified in using the same model to describe local regions of each.

The procedure adopted for the design and assessment of the detectors is similar to that used in Section 4.5.2. The training and testing data are therefore used only for obtaining noise samples, from which the corresponding detectors can be assessed. Thus the implication is not that it is required to detect targets in the images, only that it is required to perform detection in noise which resembles the noise present in the images.

For each assumed invariance subspace dimension, interference subspace and covariance matrix estimates were obtained using only data from the training image. These were based on a sample covariance matrix obtained using data sequences extracted from various locations. In particular, horizontal samples of length 32 were used, the centres of which were separated by a distance of 16 pixels in the horizontal direction and 1 pixel in the vertical.

To test the resulting detectors, a set of artificial targets was used. These targets are shown in Figure 5.12. Specifically, the targets are transient Gaussian signals with standard deviations taking integer values in the range 1 to 5. They were scaled to have equal energy.

Firstly, assuming no invariance subspace, the noninvariant detection statistic was developed by finding the MLE of the noise, and substituting this estimate into the noninvariant test in Section 4.2.1. Then,

**Section 5.8: Practical examples of invariant detectors**

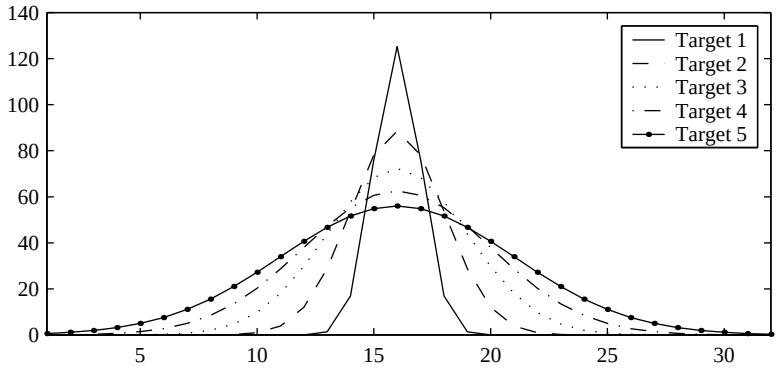



Figure 5.12: Targets used for testing the invariant detection.

for assumed interference subspace dimensions from 1 to 10, the covariance matrix and interference subspace estimates described earlier were obtained, and simply plugged into the subspace invariant test statistic of Equation 4.2.2. Figure 5.13 shows an example of the effective detection templates corresponding to noninvariant and invariant tests, for one specific target and interference subspace dimension.

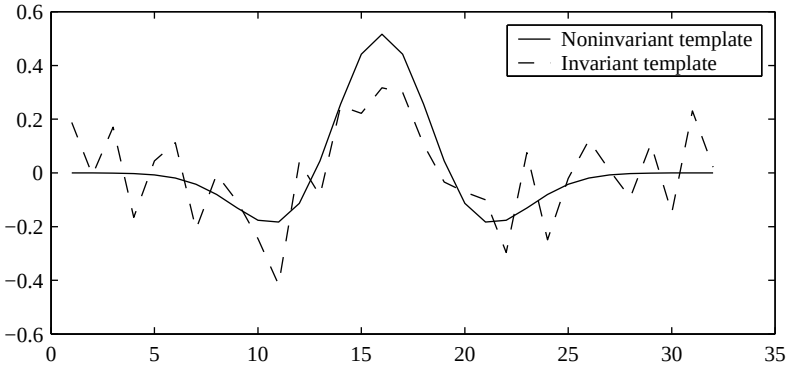



Figure 5.13: Effective templates for noninvariant and invariant detection of Target 3. This result is for an invariance subspace of dimension 4.

Using the methods described, 11 detection statistics, 11 covariance matrix estimates, and 10 invariance subspace estimates (since there is no estimate for the noninvariant test) were obtained. Each of these detection statistics was then applied to noise samples taken from both the training and the testing data, and independent sets of results were formed. Additionally, targets were added to the noise samples, simulating conditions of target presence, and the detection statistics were again calculated.

Table 5.1 shows the average normalised separation between the sets of values taken on by the test statistics under the conditions of target absence and target presence. That is, the absolute difference

| Invariance dimension | Training data | | | | | Testing data | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Targ1 | Targ2 | Targ3 | Targ4 | Targ5 | Targ1 | Targ2 | Targ3 | Targ4 | Targ5 |
| 0 | 3.40 | 2.46 | 1.81 | 1.34 | 1.05 | 2.88 | 1.82 | 1.07 | 0.69 | 0.49 |
| 1 | 3.57 | 2.44 | 1.81 | 1.46 | 1.23 | 3.04 | 2.04 | 1.49 | 1.19 | 1.00 |
| 2 | 3.43 | 2.23 | 1.73 | 1.43 | 1.22 | 2.88 | 1.84 | 1.41 | 1.17 | 0.99 |
| 3 | 3.60 | 2.37 | 1.64 | 1.09 | 0.70 | 3.06 | 1.93 | 1.30 | 0.86 | 0.55 |
| 4 | 3.58 | 2.34 | 1.59 | 1.04 | 0.64 | 3.04 | 1.92 | 1.28 | 0.83 | 0.51 |
| 5 | 3.43 | 1.86 | 1.00 | 0.53 | 0.32 | 2.98 | 1.63 | 0.86 | 0.45 | 0.27 |
| 6 | 3.24 | 1.68 | 0.87 | 0.48 | 0.30 | 2.82 | 1.47 | 0.75 | 0.41 | 0.26 |
| 7 | 2.33 | 0.95 | 0.63 | 0.43 | 0.29 | 2.00 | 0.82 | 0.54 | 0.36 | 0.24 |
| 8 | 2.28 | 0.81 | 0.57 | 0.41 | 0.28 | 1.96 | 0.69 | 0.49 | 0.35 | 0.24 |
| 9 | 2.24 | 0.78 | 0.53 | 0.37 | 0.26 | 1.93 | 0.66 | 0.45 | 0.31 | 0.22 |
| 10 | 2.21 | 0.69 | 0.50 | 0.36 | 0.25 | 1.90 | 0.59 | 0.43 | 0.30 | 0.21 |

Table 5.1: Normalised separation between hypotheses for different invariance subspace dimensions and different targets.

between the two cases is calculated (which is constant for each detector), and this is normalised by the average standard deviation observed in the test statistic when only noise is present. This figure represents the change in value of the detection statistic when a target is added, expressed as a ratio of the average standard deviation of the noise in the statistic. As such, it is a direct indication of detectability.

For the training data, on which the detector was designed, the normalised separation tends to decrease for each target as the dimension of the invariance subspace is increased. Some exceptions may be noted where the average detectability increases when an invariance subspace is added. For these targets, the presence of the invariance subspace improves the match between the model and the data *more* than it decreases performance by nulling the target. In other instances, however, this is not the case — although the invariance subspace may improve the modelling, this effect is not sufficient to improve the overall performance of the test statistic. If the data were highly nonstationary or exhibited strong trends, this would probably not be the case. An invariance subspace would then improve the modelling, perhaps to the extent that the overall detectability improves.

Also shown in Table 5.1 are detectability figures for the detection statistics applied to the testing data. Here the invariant detectors are better than the noninvariant detector for all the targets. In some cases, the detectability improves by more than a factor of two. This is an interesting result, in that it demonstrates that the invariance subspace estimate obtained from the training data to some extent captures model deviations or interference which is consistent over *both* images. The invariant testing paradigm therefore results in a detector which generalises better than the noninvariant detector, at least for the case shown here.

The improved discrimination ability of the test statistic is only one advantage of the invariant detectors. They have the additional property of reducing the mismatch between the assumed model and the actual

data. Recall that when the invariance subspace was estimated, an invariant estimate of the covariance matrix was obtained. This covariance estimate can be used to predict the variance of the test statistic under the null hypothesis.

The presence of the invariance subspace tends to reduce the degree of mismatch between this predicted variance and the variance actually observed when the test statistic is applied to the data. Table 5.2 shows results comparing these values for the cases of the five different targets. Again a breakdown

| Inv. dim. | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 |
|---|---|---|---|---|---|
| 0 | 31.17 (37.00) | 43.06 (67.76) | 62.31 (100.7) | 90.66 (133.5) | 124.5 (164.0) |
| 1 | 33.66 (40.03) | 53.28 (65.86) | 70.59 (78.99) | 80.75 (83.45) | 84.66 (83.38) |
| 2 | 40.90 (40.40) | 63.87 (52.56) | 77.02 (55.67) | 84.11 (56.34) | 87.25 (56.07) |
| 3 | 35.99 (37.57) | 44.77 (44.22) | 47.44 (44.84) | 47.42 (43.74) | 45.54 (41.56) |
| 4 | 36.76 (36.01) | 44.58 (40.46) | 47.18 (40.95) | 47.39 (40.37) | 45.80 (39.04) |
| 5 | 34.08 (33.68) | 37.22 (35.60) | 35.53 (33.79) | 32.00 (30.73) | 29.48 (28.69) |
| 6 | 34.18 (32.84) | 37.38 (34.49) | 35.60 (32.95) | 31.84 (30.26) | 29.66 (28.70) |
| 7 | 30.55 (29.71) | 30.02 (28.59) | 30.66 (28.98) | 30.84 (29.12) | 30.13 (28.70) |
| 8 | 30.91 (29.54) | 29.35 (28.07) | 30.77 (28.68) | 31.26 (28.95) | 30.48 (28.56) |
| 9 | 30.85 (29.12) | 29.78 (28.01) | 28.83 (27.42) | 28.54 (27.26) | 28.41 (27.23) |
| 10 | 30.10 (28.42) | 27.90 (26.88) | 28.81 (27.17) | 28.62 (27.10) | 28.54 (27.09) |

(a) Results on training data.

| Inv. dim. | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 |
|---|---|---|---|---|---|
| 0 | 36.80 (37.00) | 58.17 (67.76) | 105.3 (100.7) | 177.5 (133.5) | 267.6 (164.0) |
| 1 | 39.46 (40.03) | 63.75 (65.86) | 85.93 (78.99) | 98.72 (83.45) | 103.5 (83.38) |
| 2 | 48.60 (40.40) | 77.50 (52.56) | 94.24 (55.67) | 103.2 (56.34) | 107.3 (56.07) |
| 3 | 42.36 (37.57) | 54.97 (44.22) | 59.66 (44.84) | 60.20 (43.74) | 57.94 (41.56) |
| 4 | 43.24 (36.01) | 54.21 (40.46) | 58.48 (40.95) | 59.18 (40.37) | 57.23 (39.04) |
| 5 | 39.17 (33.68) | 42.60 (35.60) | 41.23 (33.79) | 37.77 (30.73) | 35.02 (28.69) |
| 6 | 39.25 (32.84) | 42.70 (34.49) | 41.19 (32.95) | 37.48 (30.26) | 35.11 (28.70) |
| 7 | 35.54 (29.71) | 34.82 (28.59) | 35.79 (28.98) | 36.27 (29.12) | 35.58 (28.70) |
| 8 | 35.96 (29.54) | 34.41 (28.07) | 36.07 (28.68) | 36.73 (28.95) | 35.95 (28.56) |
| 9 | 35.81 (29.12) | 34.89 (28.01) | 33.85 (27.42) | 33.65 (27.26) | 33.60 (27.23) |
| 10 | 35.00 (28.42) | 32.79 (26.88) | 33.80 (27.17) | 33.72 (27.10) | 33.72 (27.09) |

(b) Results on testing data.

Table 5.2: Actual standard deviations of test statistics for different invariance subspace dimensions and different targets. Shown in brackets in each case is the predicted value based on the model.

is provided of the results as applied to the training data, and those as applied to the testing data. For

the training data, the discrepancy between the predicted value and the observed value decreases as the invariance subspace dimension increases. For subspace dimensions of above 5, the difference is almost negligible. This indicates that the presence of the invariance subspace reduces the mismatch between the data and the model.

Also shown in Table 5.2 are similar results obtained for the testing data. Again the discrepancy between the predicted and the observed values tends to decrease as the invariance subspace dimension is increased. The extent of the improvement is considerably less than for the training data. However, since the predicted values in both cases were obtained from the covariance matrix estimated from the training data, this reduced improvement is expected. Figure 5.14 shows the predicted standard deviation and the actual standard deviations for the testing and training data, for the problem of detecting Target 3.

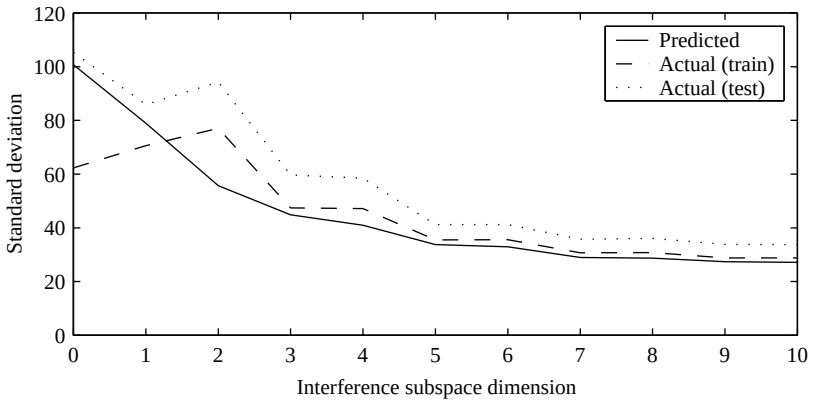



Figure 5.14: Predicted and actual standard deviations of noninvariant and invariant test statistics for Target 3. The predicted values are based on the invariant covariance matrix estimates.

To assess the use of an interference subspace with less restrictive covariance constraints, the procedure just described was repeated using a Toeplitz constraint instead of the first-order autoregressive assumption.

For estimating the interference subspace, a different method was adopted: instead of just using the training image, the estimate was made using a sample covariance matrix which was comprised of data samples from both images. The assumption of an ARMA(4,1) process was made, and interference subspace and covariance matrix estimates obtained. Subspace dimensionalities of 1 to 10 were considered, as before. For each dimension, the subspace estimates were then fixed, and assumed known in the subsequent detector formulations. The corresponding covariance matrix estimates were ignored in the subsequent development.

Plug-in detectors were then designed using only samples from the training data set. That is, for each

assumed interference subspace, an invariant covariance matrix estimate was made from data in this set. This estimate was constrained to be Toeplitz. The final detector was then obtained by plugging the resulting covariance matrix estimate into the subspace invariant detection statistic.

Table 5.3 shows the average target detectability figures as a function of the interference subspace dimension. In this case the presence of the invariance subspace at no stage increases the detectability.

| Invariance | Training data | | | | | Testing data | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| dimension | Targ1 | Targ2 | Targ3 | Targ4 | Targ5 | Targ1 | Targ2 | Targ3 | Targ4 | Targ5 |
| 0 | 3.76 | 2.65 | 2.02 | 1.59 | 1.31 | 3.23 | 2.23 | 1.66 | 1.29 | 1.05 |
| 1 | 3.76 | 2.65 | 2.01 | 1.58 | 1.29 | 3.23 | 2.24 | 1.66 | 1.30 | 1.06 |
| 2 | 3.71 | 2.56 | 1.92 | 1.50 | 1.22 | 3.16 | 2.13 | 1.55 | 1.20 | 0.98 |
| 3 | 3.65 | 2.38 | 1.57 | 1.01 | 0.64 | 3.14 | 2.03 | 1.32 | 0.83 | 0.53 |
| 4 | 3.63 | 2.30 | 1.42 | 0.81 | 0.42 | 3.13 | 1.97 | 1.20 | 0.67 | 0.35 |
| 5 | 3.36 | 1.78 | 0.83 | 0.36 | 0.20 | 2.91 | 1.55 | 0.72 | 0.30 | 0.17 |
| 6 | 3.31 | 1.68 | 0.74 | 0.30 | 0.18 | 2.87 | 1.48 | 0.65 | 0.26 | 0.16 |
| 7 | 2.82 | 1.06 | 0.42 | 0.24 | 0.19 | 2.42 | 0.91 | 0.35 | 0.20 | 0.16 |
| 8 | 2.81 | 1.00 | 0.37 | 0.21 | 0.17 | 2.42 | 0.86 | 0.31 | 0.18 | 0.15 |
| 9 | 2.63 | 0.73 | 0.29 | 0.20 | 0.16 | 2.25 | 0.62 | 0.25 | 0.17 | 0.14 |
| 10 | 2.24 | 0.51 | 0.26 | 0.19 | 0.16 | 1.92 | 0.43 | 0.22 | 0.16 | 0.13 |

Table 5.3: Normalised separation between hypotheses for different invariance subspace dimensions and different targets.

The reason can be attributed to the high capacity of the general Toeplitz constrained model, and the relative homogeneity of the data. That is, under a Toeplitz assumption the model is still flexible enough to accurately model the data even when interference is apparently present. This may not be the case for data which exhibit greater variability.

The test statistic also appears to generalise well, as can be determined from results when applied to the testing data. The presence of the invariance subspace only reduces the detectability. The average detectability in each case is smaller than that for the training data, as expected, but the difference is quite small.

Table 5.4 shows actual standard deviations of the test statistics versus the standard deviations predicted by the estimated model. Again the results are separated into training and testing data. For the training data, the actual observed values are almost identical to the predicted values. This can again be attributed to the relative freedom of the model under just a Toeplitz constraint. The model mismatch is therefore almost identically zero in all cases, and the introduction of an invariance subspace then only serves to reduce performance.

The presence of the interference subspace does reduce the mismatch between the predicted and the observed test statistic variance for the test image. The difference is small, however, and hardly warrants

| Inv. dim. | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 |
|---|---|---|---|---|---|
| 0 | 35.94 (35.84) | 45.96 (45.79) | 55.05 (54.66) | 65.66 (65.07) | 77.54 (77.06) |
| 1 | 35.89 (35.83) | 45.90 (45.84) | 55.00 (54.73) | 65.48 (65.00) | 76.93 (76.48) |
| 2 | 36.12 (35.80) | 46.49 (45.59) | 55.94 (54.60) | 67.34 (65.63) | 79.62 (77.49) |
| 3 | 35.03 (34.94) | 42.19 (42.36) | 45.29 (45.76) | 45.22 (46.02) | 41.81 (42.50) |
| 4 | 34.89 (34.62) | 41.43 (41.14) | 44.07 (43.82) | 43.72 (43.56) | 38.49 (38.46) |
| 5 | 33.49 (33.37) | 37.23 (37.24) | 35.87 (35.96) | 31.24 (31.11) | 28.54 (28.22) |
| 6 | 33.37 (33.28) | 37.35 (37.58) | 36.54 (36.94) | 31.90 (31.97) | 28.91 (28.70) |
| 7 | 31.35 (30.94) | 32.20 (31.68) | 28.89 (28.54) | 27.68 (27.30) | 28.11 (27.78) |
| 8 | 31.42 (31.03) | 32.82 (32.49) | 29.24 (29.02) | 27.78 (27.48) | 28.07 (27.74) |
| 9 | 30.22 (29.81) | 29.79 (28.90) | 26.90 (26.25) | 27.04 (26.56) | 27.60 (27.16) |
| 10 | 28.87 (28.49) | 27.03 (26.34) | 26.53 (25.82) | 27.02 (26.53) | 27.55 (27.12) |

(a) Results on training data.

| Inv. dim. | Target 1 | Target 2 | Target 3 | Target 4 | Target 5 |
|---|---|---|---|---|---|
| 0 | 41.84 (35.84) | 54.55 (45.79) | 66.93 (54.66) | 81.05 (65.07) | 96.93 (77.06) |
| 1 | 41.77 (35.83) | 54.35 (45.84) | 66.52 (54.73) | 79.95 (65.00) | 94.04 (76.48) |
| 2 | 42.35 (35.80) | 55.87 (45.59) | 68.97 (54.60) | 84.18 (65.63) | 99.90 (77.49) |
| 3 | 40.70 (34.94) | 49.54 (42.36) | 54.09 (45.76) | 54.65 (46.02) | 50.77 (42.50) |
| 4 | 40.50 (34.62) | 48.43 (41.14) | 52.31 (43.82) | 52.48 (43.56) | 46.23 (38.46) |
| 5 | 38.63 (33.37) | 42.58 (37.24) | 41.28 (35.96) | 36.55 (31.11) | 33.68 (28.22) |
| 6 | 38.44 (33.28) | 42.53 (37.58) | 41.69 (36.94) | 36.97 (31.97) | 33.97 (28.70) |
| 7 | 36.51 (30.94) | 37.60 (31.68) | 34.10 (28.54) | 32.76 (27.30) | 33.19 (27.78) |
| 8 | 36.59 (31.03) | 38.31 (32.49) | 34.49 (29.02) | 32.85 (27.48) | 33.14 (27.74) |
| 9 | 35.28 (29.81) | 35.14 (28.90) | 32.14 (26.25) | 32.09 (26.56) | 32.62 (27.16) |
| 10 | 33.72 (28.49) | 32.00 (26.34) | 31.62 (25.82) | 32.01 (26.53) | 32.54 (27.12) |

(b) Results on testing data.

Table 5.4: Actual standard deviations of test statistics for different invariance subspace dimensions and different targets. Shown in brackets in each case is the predicted value based on the model.

the use of the invariant tests. A plot of the predicted standard deviations versus actual standard deviations for Target 3 is shown in Figure 5.15

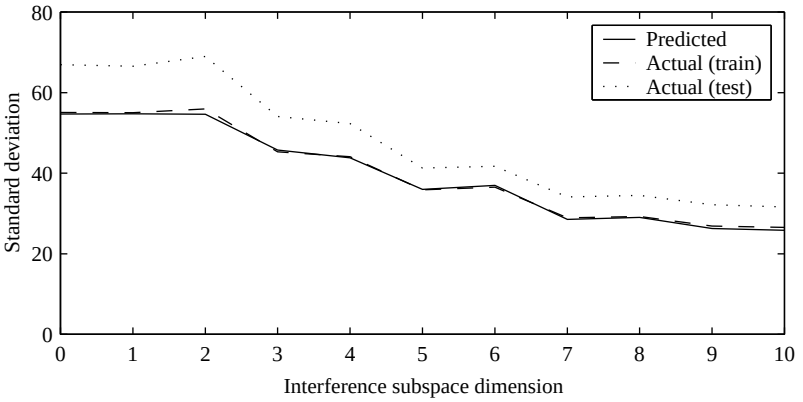



Figure 5.15: Predicted and actual standard deviations of noninvariant and invariant test statistics for Target 3. The predicted values are based on the invariant covariance matrix estimates.

The conclusion that may be drawn from these results is that the notion of the invariance subspace has merit primarily in the context of highly-constrained models or highly complex data. If the models have the capacity to accurately represent the data, then there is little need for invariance. To some extent the inclusion of an invariance subspace can be used to generalise models which are otherwise quite data-specific, but the effect is quite limited. It must be noted that the data samples used in this case *are* very amenable to the assumption of stationarity. Thus it is hardly surprising that the Toeplitz constraint leads to an accurate model. For more complex data, however, even this flexible model could benefit from the introduction of an invariance subspace.

One paradigm where highly constrained models are essential is when detectors are made adaptive. This is usually required when data are not easily statistically characterisable, but are predictable on the short term. In Chapter 6 invariance subspaces are applied to these types of problem.

### 5.8.2 Invariance subspace for reduced model mismatch

In Section 4.5 the role of the invariance subspace in reducing model mismatch was discussed. The invariance subspace is then chosen to include those components of the observation which are most in conflict with a convenient assumed noise model. By ignoring these components, a test is obtained which is more predictible from the point of view of actual attained false alarm rates and detection probabilities. This comes at the cost of reduced overall detectability.

The results generated in the previous section have been repeated under the assumption of low-rank

model mismatch, instead of a subspace interference. The invariance subspace is then estimated using the method presented in Section 5.7, where simultaneous minimisation of the Frobenius norm over the invariance subspace and the covariance matrix is performed. The results are very similar to those using the assumption of subspace interference. They will therefore not be repeated here. It has been observed that the invariance subspace estimates are quite similar in both cases, and the advantages of including invariance are apparent for both formulations. This is a promising result, suggesting that the precise details of the invariance formulation are not critical. It also raises the possibility that good results may be obtained without the need for rigorous maximum likelihood procedures.

To explicitly demonstrate the concept of invariance subspace estimation for model mismatch, results have therefore been generated for a different problem. In particular, the same test image used in Section 4.5.2 and shown in Figure 4.11 was used, and invariant detectors designed for it. Again the image was simply used to obtain noise samples, with no unknown target location problem implied. The assumption was made that data samples extracted along horizontal lines in this image can be treated as a realisation of the same stationary random process, at least in the subspace orthogonal to the invariance subspace.

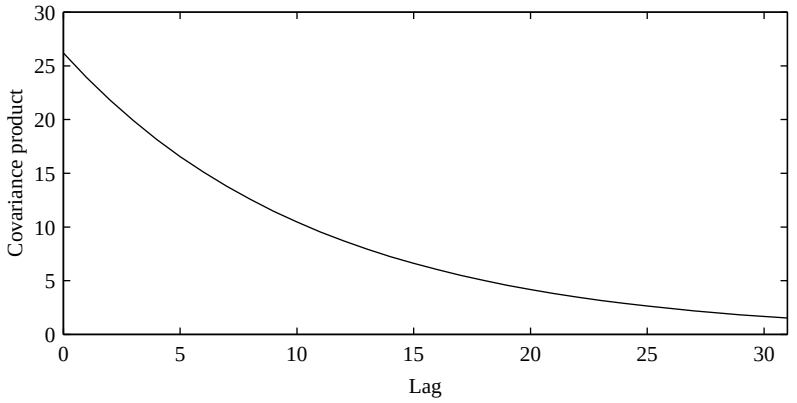
In anticipation of one of the primary topics of the next chapter, namely predictability in detection, the data analysis was performed differently from the analysis in the previous section. In particular, the following steps were performed:

- Initially the assumption of a first-order autoregressive process was made for the noise samples. These samples were each of length 32, and obtained by extracting horizontal windows of data from the image.
- A sample covariance matrix was calculated, using nonoverlapping data extracted from the entire image. For the noninvariant detector, a simple constrained estimate of the covariance was made. For the invariant detector, a three-dimensional invariance subspace was assumed. The simultaneous estimation procedure which minimises the Frobenius distance between the relevant covariance matrices, presented in Section 5.7, was used to obtain a covariance matrix estimate and an invariance subspace estimate.
- Assuming accuracy all the estimates, a noninvariant and an invariant detector were developed for two different targets. These targets were taken to be Target 2 and Target 3 from the previous section. The estimate of the noise covariance matrix in each case was used to select the test threshold, for 8 required false alarm rates ($P_{FA}$).
- Each detector was then applied to a number of samples of data extracted from the image. Since the samples contain only noise, the actual false alarm rate for each detector and each threshold value could be found.

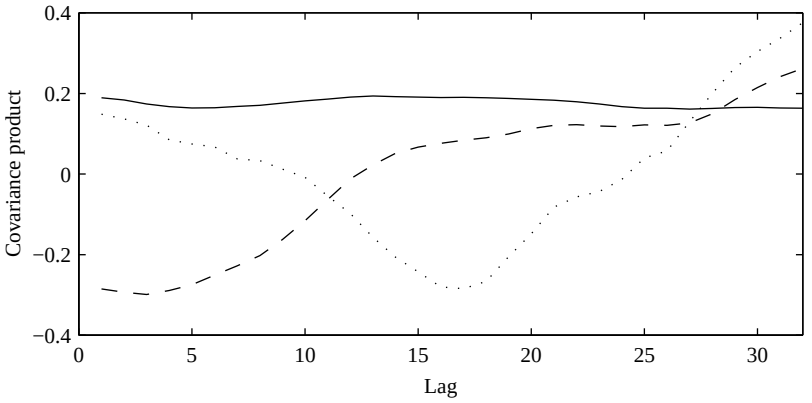
**Section 5.8: Practical examples of invariant detectors**

For both the noninvariant and invariant detectors, and for each of the two targets, the method described was used to obtain plots of actual false alarm rates versus desired false alarm rates. If the models are perfectly accurate for the data, then the difference between the actual and desired rates will be small.

Figure 5.16 shows the autocorrelation function and invariance subspace estimates obtained using this method. These estimates do not depend on the targets to be detected. It may be observed that the




(a) Autocorrelation function estimate.

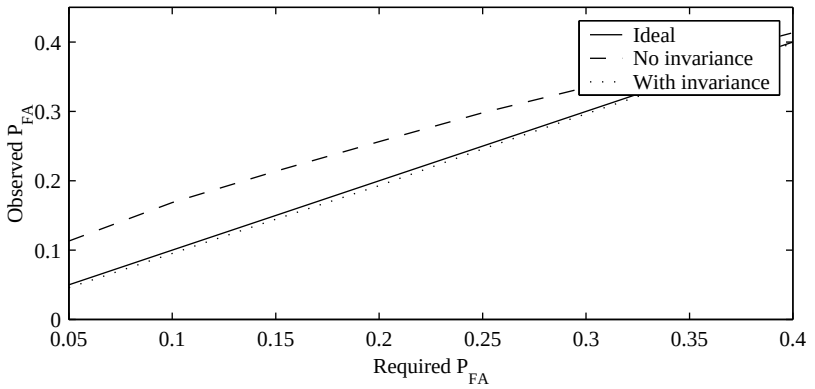



(b) Basis functions of three-dimensional invariance subspace estimate.

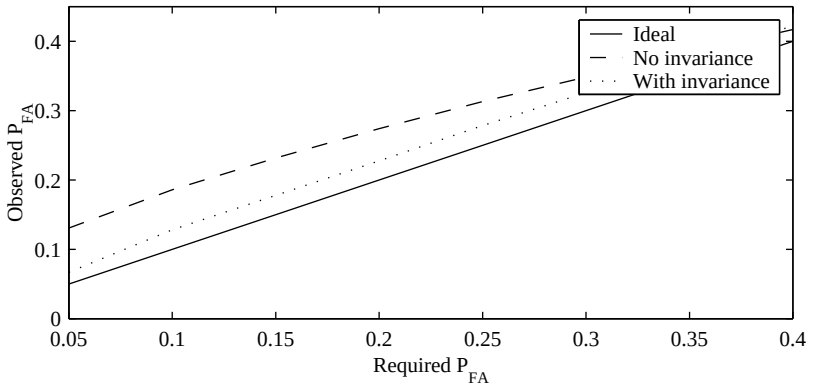
Figure 5.16: Invariant autocorrelation function and invariance subspace basis estimates under first-order autoregressive assumption.

dotted basis function in Figure 5.16(b) quite closely resembles target 3. It may be expected that the detectability for the invariant test will suffer for this target.

Figure 5.17 shows results for the subspace invariant detectors using the autocorrelation function and invariance subspace estimates shown in Figure 5.16. The first plot corresponds to the predictability of




(a) Target 2.




(b) Target 3.

Figure 5.17: Actual false alarm rates versus ideal rates for noninvariant and invariant detection using AR(1) model. A 3-D invariance subspace was assumed.

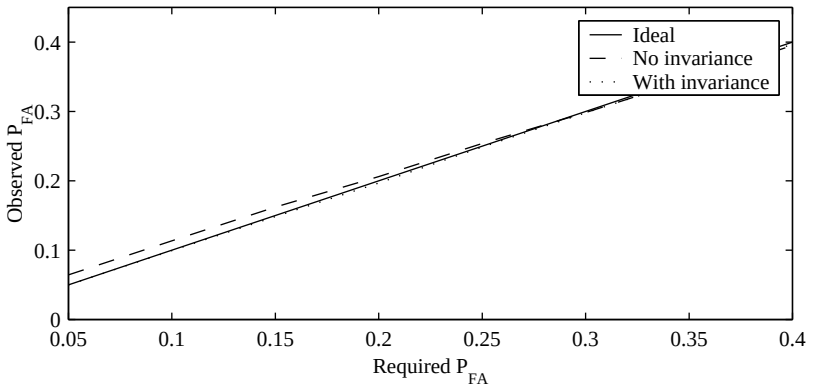
the test for Target 2, and the second for Target 3. Also shown are results for the noninvariant detectors, which at no time use data in the invariance subspace in their calculation. It is clear that the invariant detector produces a test that exhibits performance closer to the design conditions. In particular, for Target 2, the invariance subspace has improved the modelling considerably. The improvement for Target 3 is smaller, but still significant.

What is not demonstrated in these plots is that the overall detectability has decreased with the introduction of the invariance subspace: the ratio of the normalised separation between the invariant and noninvariant test is about 0.7 for Target 2, and 0.45 for Target 3. There is therefore a significant reduction in detectability associated with the inclusion of the invariance subspace.

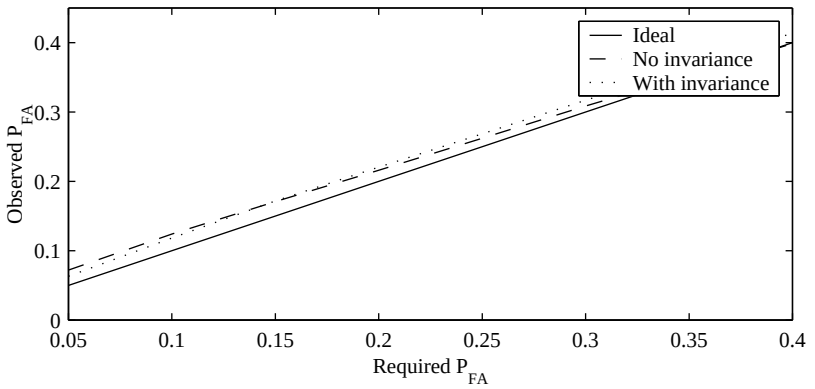
**Section 5.8: Practical examples of invariant detectors**

As discussed, if the performance of the invariant test is not as desired, the options are to change the assumed model, or to alter the assumed dimension of the invariance subspace. In general, increasing the dimensionality will decrease detectability, while improving the predictability. If satisfactory performance is not obtained for any assumed invariance subspace dimension, it indicates that the assumed model is inappropriate even when an invariance subspace is included. The only option then is to modify the assumed class of noise distributions, in this case perhaps by considering ARMA models of higher order.

Figure 5.18 shows similar plots to those of Figure 5.17, but this time using an ARMA(4,1) model for the noise. This improves the predictability of both the invariant and noninvariant tests, since the




(a) Target 2.




(b) Target 3.

Figure 5.18: Actual false alarm rates versus ideal rates for noninvariant and invariant detection using ARMA(4,1) model. A 3-D invariance subspace was assumed.

model has the capacity to characterise the data better. The invariance subspace continues to reduce the

mismatch, however. With the better modelling the detectability for the invariant detectors improves in both cases, with the ratio of the normalised separation increasing to 0.79 and 0.51, for Target 2 and Target 3 respectively.

It may be noted that, even with the invariance subspace, the predictability of the test statistic for Target 3 is fairly low. This may be attributed directly to the fact that the target lies substantially in the invariance subspace. The target therefore resembles components of the noise which are difficult to characterise effectively. Any detector will suffer under these conditions.

Predictability in detection is especially important when detectors are made adaptive. For these detectors to be effective, however, constraints are essential to improve the parameter estimation. If the constrained models are too restrictive, an invariance subspace can be introduced. This subspace can improve the validity of the constrained models, thereby improving overall performance. This topic is addressed in the next chapter.

## Section 5.8: Practical examples of invariant detectors

# Chapter 6

# Adaptive detection with invariance

In this chapter the role of subspace invariant detectors in adaptive detection is presented. Two cases are considered in detail, namely where secondary noise-only observations are available for estimation, and where simultaneous estimation and detection has to be performed on the same data sample. In this latter case constraints on the set of allowed noise densities are essential.

As discussed in the literature review chapter, in an adaptive context with secondary observations available, the work by Burgess and Van Veen [20] is of considerable interest. They propose subspace projections as a general method of dimensionality reduction when covariance matrices are completely unknown. That is, by reducing the dimensionality, the ratio of the amount of data available for estimation to the number of noise parameters is improved. The subspace projections can therefore lead to better performance even though data are discarded.

Because covariance constraints are used in the majority of this thesis, the subspace projection method for dimensionality reduction is not appropriate. Instead, the data-to-parameter ratio is improved by limiting the set of possible values for the unknown parameters. Thus in this work certain dimensions of the observation space are ignored for a different reason: because they do not conform to the modelling assumptions.

The chapter begins with a decomposition of the detection problem into disjoint subspaces. This is useful for demonstrating the way in which the different components of an observation contribute to the overall detection statistic. In particular, for the known noise parameter case, it is shown that the noise-only components serve only to reduce the variance in the conditional distribution of the target-carrying components. When unknown noise parameters are included, a separate mechanism arises by which the noise-only subspace can improve detection: the observed data in this subspace contain additional information regarding the distribution of the noise in the target subspace. Thus further improvements may be obtained from using the noise-only component.

**Section 6.1: Subspace decomposition of the detection problem**

Section 6.2 discusses in detail the problem of adaptive detection when secondary noise-only observations are available which share the distribution of the primary data. Two tests are considered, namely the AMF and the GLRT discussed in previous chapters. It is argued that the formulations of these tests presented in the literature are inappropriate in some instances, and modified derivations are presented which more accurately represent common problems. The resulting tests are shown to have a CFAR property in the case where there are no constraints on the covariance matrix estimate. In both these cases it is demonstrated that the inclusion of an invariance subspace for improving modelling is trivial, and the tests continue to be CFAR. Finally, a simplified AMF formulation is presented which has some computational savings.

Section 6.3 discusses issues relating to covariance constraints in the testing problem. When constrained estimates are used the tests presented no longer have a CFAR property, but they are nevertheless asymptotically CFAR whenever the noise estimates are consistent. The tests are then applied to the problem of simultaneous estimation and detection from a single sample of data, where covariance constraints are essential. It is shown that the AMF and GLRT formulations can be used for this problem. The use of estimates which are invariant to possible target presence are also proposed in these instances. These estimates overcome certain difficulties which otherwise arise.

In Section 6.4 results are presented showing the use of invariance subspaces in the detectors discussed. In all cases it is assumed that there is only a single sample of data available for the training and testing. It is demonstrated that the invariance subspace can result in a detector which is considerably more predictable than the corresponding noninvariant detector. It also overcomes the need for a preprocessing stage to make the data better suit the model.

## 6.1 Subspace decomposition of the detection problem

In time-series analysis, it is commonly required to predict the value of a future sample based on previously observed samples. This prediction is achieved through the use of a statistical model for the process. In the event that the process generating the samples is MVN and the model parameters are known, a linear predictor formulation is appropriate. If all assumptions are valid, then the predicted value minimises the expected mean square error between itself and the actual value obtained [113].

A similar interpretation may be applied to the general problem of predicting one set of components of a MVN observation from a different set of components. In particular, it is possible to reformulate the optimal detector in coloured noise into an estimation stage, followed by detection. The estimation stage makes use of the data components in the noise-only subspace to predict the nominal value of the noise component in the target-carrying subspace. The difference between the observed data in the target subspace and this predicted value then constitutes a sufficient statistic for the optimal detection. The interpretation may also be extended to the invariant detector discussed in Chapter 4, by appropriately

ignoring the invariance subspace in the prediction.

### 6.1.1 Two-stage decomposition of the detection problem

The development is as follows: suppose the observation space is decomposed into three mutually disjoint subspaces $\langle \mathbf{U}_I \rangle$, $\langle \mathbf{U}_T \rangle$, and $\langle \mathbf{U}_N \rangle$, which are $p$, $t$, and $(n-p-t)$-dimensional respectively. These spaces are spanned by the matrices $\mathbf{U}_I$, $\mathbf{U}_T$, and $\mathbf{U}_N$, each with orthonormal columns. $\mathbf{U}_I$ is chosen to span the invariance subspace, which is ignored in the detection. $\langle \mathbf{U}_T \rangle$ is the projection of the target-carrying subspace out of the invariance subspace — in other words, all targets that may be present lie in $\langle (\mathbf{U}_I \ \mathbf{U}_T) \rangle$, with $\mathbf{U}_I \perp \mathbf{U}_T$. $\langle \mathbf{U}_N \rangle$ is then the remaining subspace after $\langle \mathbf{U}_T \rangle$ and $\langle \mathbf{U}_I \rangle$ have been removed. It is a noise-only subspace, containing data with the same distribution under each possible hypothesis.

Suppose that it is required to detect any instance of the parametric target $\mathbf{s}_{\boldsymbol{\theta}}$ in additive noise $\mathbf{n} : N[\mathbf{0}, \mathbf{C}]$. By expanding on the alternative basis, the observed data $\mathbf{x}$ may be represented by the components $\mathbf{x}_t = \mathbf{U}_T^T \mathbf{x}$, $\mathbf{x}_i = \mathbf{U}_I^T \mathbf{x}$, and $\mathbf{x}_n = \mathbf{U}_N^T \mathbf{x}$. Useful interpretations may be applied to these components:

- $\mathbf{x}_t$ is the potential target-carrying portion of the observation, which is not affected by the invariance subspace. It is also subject to noise.
- $\mathbf{x}_i$ is the component to which the test is required to be invariant, and should therefore be ignored in any subsequent development.
- $\mathbf{x}_n$ is the component of the observation that is completely unaffected by target presence, and contains only valid noise data. It therefore *only* provides information regarding the additive noise in the observation.

The reformulation of the detection problem is obtained by means of the following interpretation: since target presence only affects the component $\mathbf{x}_t$ in a useful manner, it should be possible to perform the entire detection process in the subspace $\langle \mathbf{U}_T \rangle$. However, since the noise-only component $\mathbf{x}_n$ contains information regarding the noise in this target subspace, it is still useful from the point of view of detection. The approach is therefore to use the data $\mathbf{x}_n$ to *predict* the value of the noise in the target subspace. It is then shown that the *difference* between the actual observed data in this subspace and the predicted value is sufficient for the detection. Thus it may be concluded that the *only* role of the noise-only subspace (at least for the case of a known noise distribution) is to provide information regarding the data in the target subspace.

The assertions made in the previous paragraph may be proved using the Gauss-Markov theorem and some inversion formulae for partitioned matrices. Firstly, with $\mathbf{x}_t$ and $\mathbf{x}_n$ as defined, and with $\mathbf{x}$ :

**Section 6.1: Subspace decomposition of the detection problem**

$N[\mathbf{0}, \mathbf{C}]$, under $H_0$

$$\begin{pmatrix}\mathbf{x}_t\\ \mathbf{x}_n\end{pmatrix} : N\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{U}_T^T\mathbf{C}\mathbf{U}_T & \mathbf{U}_T^T\mathbf{C}\mathbf{U}_N\\ \mathbf{U}_N^T\mathbf{C}\mathbf{U}_T & \mathbf{U}_N^T\mathbf{C}\mathbf{U}_N\end{pmatrix}\right]. \tag{6.1}$$

Similarly, under $H_1$,

$$\begin{pmatrix}\mathbf{x}_t\\ \mathbf{x}_n\end{pmatrix} : N\left[\begin{pmatrix}\mathbf{U}_T^T\mathbf{s}_{\boldsymbol{\theta}}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{U}_T^T\mathbf{C}\mathbf{U}_T & \mathbf{U}_T^T\mathbf{C}\mathbf{U}_N\\ \mathbf{U}_N^T\mathbf{C}\mathbf{U}_T & \mathbf{U}_N^T\mathbf{C}\mathbf{U}_N\end{pmatrix}\right]. \tag{6.2}$$

The target parameter $\boldsymbol{\theta}$ is assumed known for purposes of the detection problem.

As described, the noise-only subspace is used to predict the value of the noise component in the subspace $\langle \mathbf{U}_T \rangle$. Under the MVN assumption, the natural estimate for this purpose is the conditional expectation $E\{\mathbf{x}_t|\mathbf{x}_n\}$. By applying the Gauss-Markov theorem, the appropriate conditional distribution is given by

$$\mathbf{x}_t|\mathbf{x}_n : N[\mathbf{U}_T^T\mathbf{C}\mathbf{U}_N(\mathbf{U}_N^T\mathbf{C}\mathbf{U}_N)^{-1}\mathbf{x}_n, \mathbf{Q}], \tag{6.3}$$

where $\mathbf{Q} = \mathbf{U}_T^T\mathbf{C}\mathbf{U}_T - \mathbf{U}_T^T\mathbf{C}\mathbf{U}_N(\mathbf{U}_N^T\mathbf{C}\mathbf{U}_N)^{-1}\mathbf{U}_N^T\mathbf{C}\mathbf{U}_T$. Note that this distribution applies under $H_0$, since it is assumed to describe only the noise component of the observation $\mathbf{x} = \mathbf{s}_{\boldsymbol{\theta}} + \mathbf{n}$. It is now shown that with $\hat{\mathbf{x}}_t = E\{\mathbf{x}_t|\mathbf{x}_n\}$, the statistic $\mathbf{x}_t - \hat{\mathbf{x}}_t$ is sufficient for the invariant detection problem, where invariance is required to data components in the subspace $\langle \mathbf{U}_I \rangle$.

The distribution of the statistic $\mathbf{x}_t - \hat{\mathbf{x}}_t$ is $N[\mathbf{0}, \mathbf{Q}]$ under $H_0$, and $N[\mathbf{U}_T^T\mathbf{s}_{\boldsymbol{\theta}}, \mathbf{Q}]$ under $H_1$. For known $\boldsymbol{\theta}$, the optimal test statistic of $H_0$ versus $H_1$ given the data $\mathbf{x}_t - \hat{\mathbf{x}}_t$ is therefore

$$\begin{aligned} t &= \mathbf{s}_{\boldsymbol{\theta}}^T\mathbf{U}_T\mathbf{Q}^{-1}(\mathbf{x}_t - \hat{\mathbf{x}}_t)\\ &= \mathbf{s}_{\boldsymbol{\theta}}^T\mathbf{U}_T\mathbf{Q}^{-1}(\mathbf{x}_t - \mathbf{U}_T^T\mathbf{C}\mathbf{U}_N(\mathbf{U}_N^T\mathbf{C}\mathbf{U}_N)^{-1}\mathbf{x}_n. \end{aligned} \tag{6.4}$$

A fairly lengthy simplification process now follows. Noting that the conditional covariance matrix $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \mathbf{U}_T^T\left[\mathbf{C} - \mathbf{C}\mathbf{U}_N(\mathbf{U}_N^T\mathbf{C}\mathbf{U}_N)^{-1}\mathbf{U}_N^T\mathbf{C}\right]\mathbf{U}_T, \tag{6.5}$$

Equation 4.28 can be used to show that

$$\mathbf{Q} = \mathbf{U}_T^T\left[\begin{pmatrix}\mathbf{U}_T & \mathbf{U}_I\end{pmatrix}\left(\begin{pmatrix}\mathbf{U}_T^T\\ \mathbf{U}_I^T\end{pmatrix}\mathbf{C}^{-1}\begin{pmatrix}\mathbf{U}_T & \mathbf{U}_I\end{pmatrix}\right)^{-1}\begin{pmatrix}\mathbf{U}_T^T\\ \mathbf{U}_I^T\end{pmatrix}\right]\mathbf{U}_T. \tag{6.6}$$

Note that Equation 4.28 requires orthogonality between $\mathbf{U}_H$ and $\mathbf{U}_I$ to hold, which is why $\mathbf{U}_H$ in that expression is taken to be $(\mathbf{U}_T \;\; \mathbf{U}_I)$ when applied here. This last expression for $\mathbf{Q}$ may be simplified to

$$\mathbf{Q} = \begin{pmatrix}\mathbf{I}_{t\times t} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_T & \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_I\\ \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_T & \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{I}_{t\times t}\\ \mathbf{0}\end{pmatrix}. \tag{6.7}$$

Therefore, $\mathbf{Q}$ is simply the top left $t \times t$ elements of the inverse of the block matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_T & \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_I \\ \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_T & \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I \end{pmatrix}. \tag{6.8}$$

The partitioned matrix inversion formula [6, p. 179] may then be used to derive a final useful form for the conditional covariance:

$$\mathbf{Q}^{-1} = \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_T - \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I)^{-1}\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_T. \tag{6.9}$$

The second portion of the simplification relates to the quantity $\mathbf{x}_t - \hat{\mathbf{x}}_t$. In this case, noting that

$$\begin{aligned} \mathbf{x}_t - \hat{\mathbf{x}}_t &= \mathbf{U}_T^T\mathbf{x} - \mathbf{U}_T^T\mathbf{C}\mathbf{U}_N(\mathbf{U}_N^T\mathbf{C}\mathbf{U}_N)^{-1}\mathbf{U}_N^T\mathbf{x} \\ &= \mathbf{U}_T^T[\mathbf{C} - \mathbf{U}_T^T\mathbf{C}\mathbf{U}_N(\mathbf{U}_N^T\mathbf{C}\mathbf{U}_N)^{-1}\mathbf{U}_N^T\mathbf{C}]\mathbf{C}^{-1}\mathbf{x}, \end{aligned} \tag{6.10}$$

the same results as before may be used to show that

$$\mathbf{x}_t - \hat{\mathbf{x}}_t = \begin{pmatrix} \mathbf{I}_{t\times t} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_T & \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_I \\ \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_T & \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{x} \\ \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{x} \end{pmatrix}. \tag{6.11}$$

Once again the partitioned matrix inversion formula may be used (in conjunction with the result obtained in Equation 6.9) to show that

$$\mathbf{x}_t - \hat{\mathbf{x}}_t = \begin{pmatrix} \mathbf{Q} & -\mathbf{Q}\mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{x} \\ \mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{x} \end{pmatrix}, \tag{6.12}$$

so that

$$\mathbf{x}_t - \hat{\mathbf{x}}_t = \mathbf{Q}\mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{x} - \mathbf{Q}\mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I)^{-1}\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{x}. \tag{6.13}$$

Finally, substituting this result into the test statistic in Equation 6.4 yields the following expression:

$$t = \mathbf{s}_{\boldsymbol{\theta}}^T\mathbf{U}_T\mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{x} - \mathbf{s}_{\boldsymbol{\theta}}^T\mathbf{U}_T\mathbf{U}_T^T\mathbf{C}^{-1}\mathbf{U}_I(\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{U}_I)^{-1}\mathbf{U}_I^T\mathbf{C}^{-1}\mathbf{x}. \tag{6.14}$$

Again using the relationship in Equation 4.28 can be written as

$$t = \mathbf{s}_{\boldsymbol{\theta}}^T\mathbf{U}_T\mathbf{U}_T^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}, \tag{6.15}$$

where $\mathbf{U}_H$ is a matrix spanning the subspace complementary to $\langle\mathbf{U}_I\rangle$. Now $\mathbf{U}_T\mathbf{U}_T^T\mathbf{s}_{\boldsymbol{\theta}}$ is the orthogonal projection of $\mathbf{s}_{\boldsymbol{\theta}}$ into the space $\langle\mathbf{U}_T\rangle$, which is contained in the space $\langle\mathbf{U}_H\rangle$. Thus $\mathbf{U}_T\mathbf{U}_T^T\mathbf{s}_{\boldsymbol{\theta}} = \mathbf{U}_H\mathbf{U}_H^T\mathbf{s}_{\boldsymbol{\theta}}$, and

$$t = \mathbf{s}_{\boldsymbol{\theta}}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}. \tag{6.16}$$

Comparing this last expression for $t$ to the form of the optimal invariant hypothesis test given in Equation 4.19, it may be observed that $t$ is sufficient for the detection problem. That is, knowledge of the value of $t$ is all that is required to implement the test in Equation 4.30. Therefore, as far as the detection problem is concerned, it may be concluded that nothing has been lost by replacing the original observation $\mathbf{x}$ with the calculated quantity $\mathbf{x}_t - \hat{\mathbf{x}}_t$.

The decomposition presented here therefore results in a one-dimensional test statistic which is sufficient for the detection. At this point it may be observed that the formal development of the matched filter for invariant detection in coloured noise results in the same one-dimensional sufficient statistic. However, the two interpretations are very different: the results of this section demonstrate that the noise-alone subspace is *only* useful to the detection insofar as it can provide information regarding the noise component in the target-carrying subspace. Figure 6.1 demonstrates the principle: knowledge of the noise-only component of the data *changes* the known distribution of the data in the target subspace, from a marginal distribution to a conditional distribution. Once the conditional distribution of the data

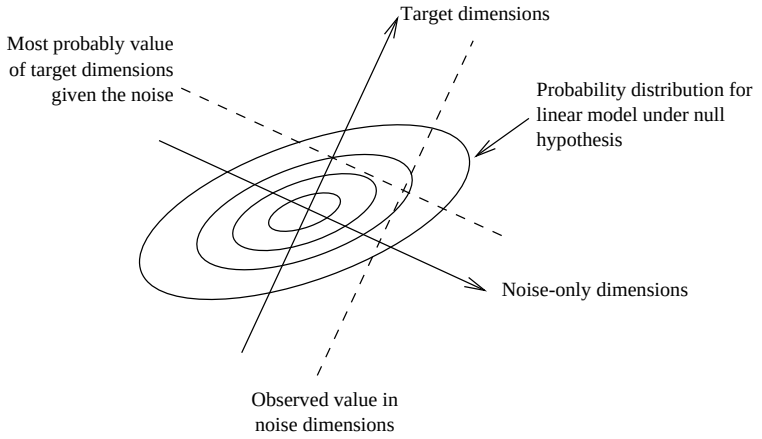



Figure 6.1: Estimating the value in the target dimension from the noise-only component, under the target-absent hypothesis.

in the target subspace has been found, the detection may be performed using only data in $\langle \mathbf{U}_T \rangle$. The remaining components are of no further use.

### 6.1.2 The role of the noise-only subspace in detection

The primary result of the previous section is in interpreting the role of the noise-alone subspace in the detection problem. However, it must be noted that the insights provided in this section are only appropriate in the event that the noise distribution is known. In that case, the *only* useful information that may be obtained from the noise-only subspace is the predicted value of the noise in the target subspace.

The decomposition presented also provides a reasonable solution to the problem of assessing whether an assumed noise model is appropriate for detecting a known target, given a sample of actual noise data. In general, it is difficult to ascertain whether model mismatch or inaccuracy will have a detrimental effect on detector performance: it is quite conceivable that a partially-incorrect model is still accurate for detection. However, the subspace decomposition presented can provide a means of obtaining such information: if the best predicted value of the noise in the target-carrying subspace is similar to the value that is actually observed under the null hypothesis, then model accuracy for detection is indicated. Note that this validation process has the advantage that it specifically takes the target subspace into account.

The results presented suggest that the use of the noise-alone subspace is only useful insofar as observing $\mathbf{x}_n$ reduces the uncertainty of the data in the target-carrying subspace. Although true when the noise distribution is assumed known, if the noise contains unknown parameters this statement is no longer true. When unknown parameters are included in the formulation, the noise-only subspace also provides information regarding the *values* of these parameters. This provides further information regarding the *distribution* of the data in the target subspace, in addition to the actual noise component in this subspace.

To explain these results in more detail, suppose that the nominal noise distribution is $N[\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})]$. In that case,

$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_n \end{pmatrix} : N\left[ \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{C}_{tt}(\boldsymbol{\theta}) & \mathbf{C}_{tn}(\boldsymbol{\theta}) \\ \mathbf{C}_{nt}(\boldsymbol{\theta}) & \mathbf{C}_{nn}(\boldsymbol{\theta}) \end{pmatrix} \right]. \tag{6.17}$$

Suppose for now that $\boldsymbol{\theta}$ is fixed and known. The noise covariance matrix is then known, and the results presented previously apply. That is, the component $\mathbf{x}_n$ can be used to make an estimate of the noise component in $\mathbf{x}_t$, which is useful in the subsequent detection. This, however, is only possible when $\mathbf{C}_{tn}(\boldsymbol{\theta})$ and $\mathbf{C}_{nt}(\boldsymbol{\theta})$ are nonzero — in the event that the coupling between the subspaces *is* zero, knowledge regarding the component of the data in the subspace $\langle \mathbf{U}_N \rangle$ provides *no* information regarding the noise in $\langle \mathbf{U}_T \rangle$. In that case, observing $\mathbf{x}_n$ contributes nothing to the final detector, which may effectively be performed entirely in the subspace $\langle \mathbf{U}_T \rangle$. The subspaces are effectively independent.

When $\boldsymbol{\theta}$ is unknown, this last conclusion no longer applies. Then, even if $\mathbf{C}_{tn}(\boldsymbol{\theta})$ and $\mathbf{C}_{nt}(\boldsymbol{\theta})$ are zero for all values of $\boldsymbol{\theta}$, the data in the noise-only subspace still contain information that is useful to the decision process in the subspace $\langle \mathbf{U}_T \rangle$. This is because the distribution of $\mathbf{x}_n$ depends on $\boldsymbol{\theta}$. Thus, by observing $\mathbf{x}_n$, information is obtained regarding the details of the distribution of the data component $\mathbf{x}_t$.

## 6.2 Adaptive detection with secondary noise-only observations

Section 5.1 presented an outline of solutions to the problem of detecting a known target in the presence of partially-known noise, when an invariance subspace is included in the formulation. It was demonstrated

that the resulting tests require subspace invariant estimates of the noise parameters.

The case of completely known targets does not generally lead to tests which exhibit a constant false alarm rate. For example, Reed, Mallet and Brennan [111] discuss a plug-in formulation for this known target detection problem when the covariance matrix is completely unspecified, and the distribution of the test statistic under $H_0$ depends on the unknown covariance. This makes it difficult to select a test threshold meaningfully, since the properties of the test will in general depend on the unknown parameter values.

In [116], Robey et al. demonstrate that the plug-in principle, when applied to the problem of detecting the presence of the target $\alpha\mathbf{s}$ for unknown $\alpha$, leads to a CFAR test. That is, the introduction of the parameter $\alpha$ results in the test having a normalisation component which improves its statistical properties under the noise-only hypothesis. They call this test the adaptive matched filter (AMF). Similarly, Kelly [80] shows that the introduction of an unknown target scaling parameter results in the GLRT for the same problem having a CFAR property.

Both the AMF test and the GLRT in these cases assume the presence of secondary target-free observations, which share the distribution of the noise component of the data to be tested for target presence. The observations are all assumed to be mutually independent and zero mean, and the only unknown parameter in their distribution is the covariance matrix. No constraints are placed on the set of allowable covariances.

The tests presented by Robey et al. and Kelly both deal with the problem of complex signals and targets. The unknown target scaling parameter $\alpha$ is also complex, and is permitted to take on any value whatsoever. For real observations, this formulation is inappropriate for many problems: since the parameter $\alpha$ is unconstrained, it can take on both positive and negative values. Thus the tests will produce the same result when presented with the data observation $\mathbf{x}$ as they will when presented with $-\mathbf{x}$. They will therefore detect instances of both the target $\alpha\mathbf{s}$ and the target $-\alpha\mathbf{s}$, without discrimination.

When a noise model is used to represent an environment comprised substantially of clutter, this two-sided property of the tests is highly undesirable if only positive scale coefficients are relevant. In x-ray images, for example, tumours *always* exhibit greater attenuation than surrounding tissue. It makes no sense then to allow false alarms based on regions which resemble tumours in shape, but have inverted attenuation profiles. Thus it is desirable to reformulate the AMF and GLRT to better represent this class of problems.

In the two sections which follow, the AMF and GLRT are modified for the problem of one-sided hypothesis testing, where the target coefficient is still considered unknown, but is constrained to be nonnegative. It is demonstrated that the resulting tests are once again CFAR. Initially the inclusion of an invariance subspace is not considered: this case is discussed in Section 6.2.3. Also, constraints on

the allowed covariance matrix are not considered, but are presented in Section 6.3. In Section 6.2.4 a simplified AMF formulation is presented which is less adaptive, but avoids the need for a matrix inversion in its calculation.

### 6.2.1 AMF reformulated for positive target coefficients

The derivation in this section is similar to that presented in [116], with the exception that the target amplitude parameter is required to be nonnegative. Thus it is required to detect the target $\alpha\mathbf{s}$ in coloured noise, with $\alpha > 0$. With $\mathbf{x}$ the data to be tested for target presence, the null hypothesis is $\mathbf{x} : N[\mathbf{0}, \mathbf{C}]$. The composite alternative is that $\mathbf{x} : N[\alpha\mathbf{s}, \mathbf{C}]$, for some $\alpha > 0$.

In practice, the covariance matrix $\mathbf{C}$ is unknown, and is estimated from the secondary noise-only observations $\mathbf{x}_1, \ldots, \mathbf{x}_M$. However, for purposes of the AMF test development, it is assumed that $\mathbf{C}$ is known, and the GLRT is derived for detecting the presence of the target $\mathbf{x}$ under these conditions. The GLRT statistic for this known covariance matrix problem is

$$\begin{aligned} t_{\text{GLRT}}(\mathbf{x}) &= \ln \frac{\max_{\alpha\geq 0}(2\pi)^{-n/2}|\mathbf{C}|^{-1/2}e^{-1/2(\mathbf{x}-\alpha\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{x}-\alpha\mathbf{s})}}{(2\pi)^{-n/2}|\mathbf{C}|^{-1/2}e^{-1/2\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}}} \\ &= \max_{\alpha\geq 0}\left\{\alpha\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x} - 1/2\alpha^2\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}\right\}. \end{aligned} \tag{6.18}$$

It is easy to show that the value $\hat{\alpha}$ of $\alpha$ that maximises the previous expression is given by

$$\hat{\alpha} = \begin{cases} \frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} & \mathbf{s}^T\mathbf{C}^{-1}\mathbf{x} \geq 0 \\ 0 & \mathbf{s}^T\mathbf{C}^{-1}\mathbf{x} < 0, \end{cases} \tag{6.19}$$

which assumes the constraint $\alpha > 0$. Substituting back into 6.18, the GLRT statistic becomes

$$t_{\text{GLRT}}(\mathbf{x}) = \begin{cases} \frac{1}{2}\frac{(\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x})^2}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}} & \mathbf{s}^T\mathbf{C}^{-1}\mathbf{x} \geq 0 \\ 0 & \mathbf{s}^T\mathbf{C}^{-1}\mathbf{x} < 0. \end{cases} \tag{6.20}$$

The resulting test is to compare this to a threshold and decide $H_1$ if exceeded.

If only positive values of the threshold are permitted, then the test $t_{\text{GLRT}}(\mathbf{x}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta'$ can be cast into the simplified form

$$\frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x}}{\sqrt{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{6.21}$$

This follows since $H_1$ is never decided if $\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x} < 0$. To form the final AMF test statistic, the estimate $\widehat{\mathbf{C}}$ of $\mathbf{C}$ obtained from the secondary data vectors is substituted into this expression. Thus as long as

the threshold $\eta$ is always chosen to be positive, the AMF is

$$t_{\text{AMF}}(\mathbf{x}) = \frac{\mathbf{s}^T \widehat{\mathbf{C}}^{-1}\mathbf{x}}{\sqrt{\mathbf{s}^T \widehat{\mathbf{C}}^{-1}\mathbf{s}}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{6.22}$$

Note that the square of this test statistic is precisely the test developed by Robey et al. for the corresponding two-sided problem.

Using the same methods presented in [116], it can be demonstrated that the distribution of the test statistic $t_{\text{AMF}}(\mathbf{x})$ is independent of the actual noise covariance $\mathbf{C}$. Firstly, consider the term $\mathbf{s}^T \widehat{\mathbf{C}}^{-1}\mathbf{x}$ in the numerator. Under the null hypothesis, $\mathbf{x} : N[\mathbf{0}, \mathbf{C}]$ and $M\widehat{\mathbf{C}} : W[\mathbf{C}; n, M]$, where $W[\mathbf{C}; n, M]$ denotes an $n$-dimensional Wishart distribution with $M$ degrees of freedom, and matrix scale parameter $\mathbf{C}$ [30, p. 302]. It is assumed that $\widehat{\mathbf{C}}$ has been obtained using $M$ additional independent observations. Now

$$\mathbf{s}^T \widehat{\mathbf{C}}^{-1}\mathbf{x} = \mathbf{s}^T \mathbf{C}^{-1/2}\mathbf{C}^{1/2}\widehat{\mathbf{C}}^{-1}\mathbf{C}^{1/2}\mathbf{C}^{-1/2}\mathbf{x} = \mathbf{u}^T \tilde{\mathbf{C}}^{-1}\mathbf{y}, \tag{6.23}$$

where $\mathbf{u} = \mathbf{C}^{-1/2}\mathbf{s}$, $\mathbf{y} = \mathbf{C}^{-1/2}\mathbf{x}$, and $\tilde{\mathbf{C}} = \mathbf{C}^{-1/2}\widehat{\mathbf{C}}\mathbf{C}^{-1/2}$. The transformed vector $\mathbf{y}$ has a $N[\mathbf{0}, \mathbf{I}]$ distribution. Also, from the properties of the Wishart distribution [30, p. 302], the distribution of $M\tilde{\mathbf{C}}$ is $W[\mathbf{I}; n, M]$. The distribution of both of these quantities is therefore independent of the unknown covariance $\mathbf{C}$. The only term in this alternative form of the numerator which depends on $\mathbf{C}$ is the vector $\mathbf{u}$.

It can be shown that this term depends on $\mathbf{C}$ only through a scalar coefficient $d$. Let $\mathbf{U}$ be a unitary transformation chosen such that

$$d\hat{\mathbf{e}}_1 = \mathbf{U}^T\mathbf{u}, \tag{6.24}$$

with $\hat{\mathbf{e}}_1$ the first unit vector $(1\ 0\ \cdots\ 0)^T$. Then

$$\mathbf{u}^T \tilde{\mathbf{C}}^{-1}\mathbf{y} = \mathbf{u}^T\mathbf{U}\mathbf{U}^T \tilde{\mathbf{C}}^{-1}\mathbf{U}\mathbf{U}^T\mathbf{y} = d\hat{\mathbf{e}}_1^T \tilde{\mathbf{G}}^{-1}\mathbf{z}, \tag{6.25}$$

where now $\tilde{\mathbf{G}} = \mathbf{U}^T\tilde{\mathbf{C}}\mathbf{U}$ and $\mathbf{z} = \mathbf{U}^T\mathbf{y}$. Since the distribution of $\mathbf{z} : N[\mathbf{0}, \mathbf{I}]$ and $M\tilde{\mathbf{G}} : W[\mathbf{I}; n, M]$, it is evident that the unknown covariance matrix $\mathbf{C}$ enters the distribution of $\mathbf{s}^T \widehat{\mathbf{C}}^{-1}\mathbf{x}$ only through $d$.

Using a similar analysis on the denominator term $\mathbf{s}^T \widehat{\mathbf{C}}^{-1}\mathbf{s}$,

$$\begin{aligned}\mathbf{s}^T \widehat{\mathbf{C}}^{-1}\mathbf{s} &= \mathbf{s}^T \mathbf{C}^{-1/2}\mathbf{C}^{1/2}\widehat{\mathbf{C}}^{-1}\mathbf{C}^{1/2}\mathbf{C}^{-1/2}\mathbf{s} = \mathbf{u}^T \tilde{\mathbf{C}}^{-1}\mathbf{u} \\ &= \mathbf{u}^T\mathbf{U}\mathbf{U}^T \tilde{\mathbf{C}}^{-1}\mathbf{U}\mathbf{U}^T\mathbf{u} = d^2\hat{\mathbf{e}}_1^T \tilde{\mathbf{G}}^{-1}\hat{\mathbf{e}}_1,\end{aligned} \tag{6.26}$$

where once again the distribution of $M\tilde{\mathbf{G}}$ is $W[\mathbf{I}; n, M]$. Thus the distribution of the denominator also only depends on $\mathbf{C}$ through the value of $d$.

Returning to the AMF, the test statistic can therefore be written as

$$t_{\text{AMF}} = \frac{d\hat{\mathbf{e}}_1^T \tilde{\mathbf{G}}^{-1}\mathbf{z}}{\sqrt{d^2\hat{\mathbf{e}}_1^T \tilde{\mathbf{G}}^{-1}\hat{\mathbf{e}}_1}} = \frac{\hat{\mathbf{e}}_1^T \tilde{\mathbf{M}}^{-1}\mathbf{z}}{\sqrt{\hat{\mathbf{e}}_1^T \tilde{\mathbf{G}}^{-1}\hat{\mathbf{e}}_1}}. \tag{6.27}$$

Since $d$ can be eliminated from the statistic, the remaining quantity in no way depends on $\mathbf{C}$. Thus under $H_0$ the distribution of the test statistic is completely determined, and does not depend on the underlying covariance parameters. A fixed test threshold can therefore be used to yield a desired false alarm rate.

In practice, the restriction that the threshold be positive means that this result only holds for tests of sufficiently small size. Since the distribution under $H_0$ is symmetric about the origin, the test presented is applicable whenever the required false alarm rate is less than 0.5. In most cases this is hardly a restriction, unless high detection rates are required at the cost of a large number of false alarms.

It is easy to see how this modified formulation, which permits only positive target coefficients, can improve detection performance. As mentioned, the AMF test statistic presented in the literature makes use of the square of the statistic obtained here. Thus, when compared to a threshold, the unconstrained test makes a decision of target presence whenever the quantity $t_{\text{AMF}}$ above is either greater than or smaller than a set threshold. The modified formulation for positive target coefficients, however, only decides $H_1$ when the value of $t_{\text{AMF}}$ is greater than a threshold. Thus for the same threshold value, the two tests have the same detection probability, but the false alarm rate in the latter case is halved. This translates directly to better detection.

No attempt has been made to develop a closed-form solution for the density of this test statistic under either hypothesis. However, the limiting distribution as the number of secondary data vectors increases can be found, and is useful for subsequent developments. Suppose that the distribution of the secondary vectors and of $\mathbf{x}$ under $H_0$ is $N[\mathbf{0}, \mathbf{C}]$ under $H_0$. Since the sample covariance matrix is a consistent estimator of the actual covariance, as $M$ increases, $\widehat{\mathbf{C}}$ tends towards $\mathbf{C}$. The variance in this estimate also goes to zero. Thus the AMF test statistic

$$\frac{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}}{\sqrt{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s}}} \longrightarrow \frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x}}{\sqrt{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}}, \qquad \text{as} \qquad M \longrightarrow \infty. \tag{6.28}$$

With $\mathbf{x} : N[\mathbf{0}, \mathbf{C}]$, this quantity has a $N[0, 1]$ distribution under $H_0$.

Through a similar set of arguments it can be seen that this test is also asymptotically optimal for detecting the target $\alpha\mathbf{s}$ in noise, for any $\alpha > 0$. It approaches the optimal matched filter for the known covariance matrix case.

Thus it has been shown that the AMF of Robey et al. can be reformulated to detect only instances of targets which have a positive coefficient. In the next section it is shown that Kelly's GLRT can be

similarly modified, again resulting in a CFAR test.

A final point to note with regard to this test, and those which follow, is that when written in the form

$$\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} \underset{H_0}{\overset{H_1}{\gtrless}} \sqrt{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s}}, \tag{6.29}$$

an adaptive thresholding interpretation is appropriate. Thus the value of the test component $\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}$ may be calculated and compared to the threshold value $\sqrt{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s}}$, which is itself a function of the estimated noise characteristics. The same test is achieved using this interpretation.

### 6.2.2 GLRT reformulated for positive target coefficients

As in the previous section, it is assumed that it is required to detect the presence of some positive scaling of the target $\mathbf{s}$ in zero-mean noise with an unknown and unconstrained covariance matrix. The data to be tested for target presence are $\mathbf{x}$, and it is assumed that $M$ additional noise-only observations are available for estimation. (Note that this is in contrast to the multiple observation GLRT formulation in Section 5.1.2, where the primary data vector was taken to be $\mathbf{x}_1$.) Thus the null hypothesis is that $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_M : N[\mathbf{0}, \mathbf{C}]$, and the alternative is that $\mathbf{x} : N[\alpha\mathbf{s}, \mathbf{C}]$ for some $\alpha > 0$, with $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M : N[\mathbf{0}, \mathbf{C}]$. The covariance matrix $\mathbf{C}$ is completely unknown.

The derivation of the test proceeds in the same way as in [80], up to the point where it is necessary to maximise the likelihood ratio over the unknown factor $\alpha$. That is, the GLRT statistic is given by

$$t_{\text{GLRT}}(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_M) = \frac{\max_{\alpha,\mathbf{C}}(2\pi)^{-(M+1)n/2}|\mathbf{C}|^{-(M+1)/2}e^{-1/2(\mathbf{x}-\alpha\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{x}-\alpha\mathbf{s})-1/2\sum_{i=1}^M\mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i}}{\max_{\alpha,\mathbf{C}}(2\pi)^{-(M+1)n/2}|\mathbf{C}|^{-(M+1)/2}e^{-1/2\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}-1/2\sum_{i=1}^M\mathbf{x}_i^T\mathbf{C}^{-1}\mathbf{x}_i}}. \tag{6.30}$$

Using standard results of MLEs of unknown covariance matrices, it is easy to show that the numerator is maximised over $\mathbf{C}$ by the choice $\widehat{\mathbf{C}}_1 = ((\mathbf{x}-\alpha\mathbf{s})(\mathbf{x}-\alpha\mathbf{s})^T + M\widehat{\mathbf{C}})/(M+1)$, where

$$\widehat{\mathbf{C}} = \frac{1}{M}\sum_{i=1}^M \mathbf{x}_i\mathbf{x}_i^T. \tag{6.31}$$

Similarly, the denominator takes on a maximum for $\widehat{\mathbf{C}}_0 = (\mathbf{x}\mathbf{x}^T + M\widehat{\mathbf{C}})/(M+1)$. Substituting these values into the test statistic and simplifying yields the result

$$t_{\text{GLRT}}(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_M) = \frac{\max_{\alpha\geq 0}|\widehat{\mathbf{C}}_1|^{-(M+1)/2}}{|\widehat{\mathbf{C}}_0|^{-(M+1)/2}} = \frac{\max_{\alpha\geq 0}|(\mathbf{x}-\alpha\mathbf{s})(\mathbf{x}-\alpha\mathbf{s})^T + M\widehat{\mathbf{C}}|^{-(M+1)/2}}{|\mathbf{x}\mathbf{x}^T + M\widehat{\mathbf{C}}|^{-(M+1)/2}}. \tag{6.32}$$

Without changing the test, the monotonically increasing transformation $t \longrightarrow t^{2/(M+1)}$ can be made,

resulting in the modified test statistic

$$t_{\text{GLRT}} = \frac{|\mathbf{x}\mathbf{x}^T + M\widehat{\mathbf{C}}|}{\min_{\alpha\geq 0}|(\mathbf{x}-\alpha\mathbf{s})(\mathbf{x}-\alpha\mathbf{s})^T + M\widehat{\mathbf{C}}|}. \tag{6.33}$$

Using the same identities presented by Kelly, namely that $|\mathbf{z}\mathbf{z}^T + \mathbf{S}| = |\mathbf{S}|(1+\mathbf{z}^T\mathbf{S}^{-1}\mathbf{z})$, this statistic can be written as

$$t_{\text{GLRT}} = \frac{(1+1/M\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x})}{\min_{\alpha\geq 0}(1+1/M(\mathbf{x}-\alpha\mathbf{s})^T\widehat{\mathbf{C}}^{-1}(\mathbf{x}-\alpha\mathbf{s}))}. \tag{6.34}$$

At this point the minimisation over $\alpha$ can be performed.

As for the case of the AMF, the constrained minimum of the denominator over $\alpha \geq 0$ occurs for

$$\hat{\alpha} = \begin{cases} \frac{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}}{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s}} & \mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} \geq 0 \\ 0 & \mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} < 0. \end{cases} \tag{6.35}$$

The resulting likelihood ratio test statistic is therefore

$$\begin{aligned} t_{\text{GLRT}} &= \begin{cases} \frac{1+1/M\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}}{1+1/M\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}-1/M\frac{(\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x})^2}{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s}}} & \mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} \geq 0 \\ 1 & \mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} < 0 \end{cases} \\ &= \begin{cases} \frac{1}{1-\gamma} & \mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} \geq 0 \\ 1 & \mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} < 0, \end{cases} \end{aligned} \tag{6.36}$$

with

$$\gamma = \frac{(\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x})^2}{(\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s})(M+\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x})}. \tag{6.37}$$

Now, using the Cauchy-Schwarz inequality with the inner product $\langle\mathbf{u},\mathbf{v}\rangle = \mathbf{u}^T\widehat{\mathbf{C}}^{-1}\mathbf{v}$ [4, p. 227], it can be seen that

$$(\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x})^2 \leq (\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s})(\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}) < (\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s})(M+\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}), \tag{6.38}$$

where the additional result has been used that $\widehat{\mathbf{C}}$ is nonsingular and positive definite. Thus it must always be the case that $0 \leq \gamma < 1$. On this interval $t_{\text{GLRT}} \geq 1$, and the monotonically increasing transformation $t \longrightarrow M(1-1/t)$ can be applied to the value of the test statistic. For testing purposes, this yields the equivalent statistic

$$t_{\text{GLRT}} = \begin{cases} \frac{M(\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x})^2}{(\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s})(M+\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x})} & \mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} \geq 0 \\ 0 & \mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} < 0. \end{cases} \tag{6.39}$$

The GLRT compares this value to a threshold, and decides $H_1$ when exceeded.

Once again for sufficiently large threshold values this statistic can be modified, by taking square roots,

to produce the test

$$t_{\text{GLRT}} = \frac{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}}{\sqrt{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s}}\sqrt{1+1/M\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{6.40}$$

For this to hold requires that $\eta > 0$, which in account of the symmetry of the distribution corresponds to a false alarm rate of less than 0.5. Thus as long as the required size is smaller than 0.5, the above test represents the GLRT for the problem.

The distribution of this test statistic under $H_0$ is independent of the unknown parameters. From the discussion of the AMF, it is apparent that the distribution of the ratio $\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}/\sqrt{\mathbf{s}^T\widehat{\mathbf{C}}^{-1}\mathbf{s}}$ is independent of both $\alpha$ and $\mathbf{C}$. Additionally, the term $\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}$ can be shown to have a distribution independent of $\mathbf{C}$ under $H_0$:

$$\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x} = \mathbf{x}^T\mathbf{C}^{-1/2}\mathbf{C}^{1/2}\widehat{\mathbf{C}}^{-1}\mathbf{C}^{1/2}\mathbf{C}^{-1/2}\mathbf{x} = \mathbf{y}^T\tilde{\mathbf{C}}^{-1}\mathbf{y}, \tag{6.41}$$

with $\mathbf{y} : N[\mathbf{0}, \mathbf{I}]$ and $M\tilde{\mathbf{C}} : W[\mathbf{I}; n, M]$ as before. Thus the density of the term $\sqrt{1+1/M\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}}$ does not depend on the unknown covariance $\mathbf{C}$, and the distribution of the test statistic $t_{\text{GLRT}}$ in Equation 6.40 is similarly independent.

Kelly [80] provides analytical solutions for the distribution of the test statistic, its size, and its power, for the case where the target amplitude parameter is unconstrained. The functions are complex, however, and provide little insight. Thus no attempt has been made to calculate these quantities for the modified test presented here. What can quite simply be observed, however, is the limiting distribution of the test statistic when large numbers of secondary noise-only observations are available. For large $M$, $\sqrt{1+1/M\mathbf{x}^T\widehat{\mathbf{C}}^{-1}\mathbf{x}} \approx 1$ and $\widehat{\mathbf{C}} \approx \mathbf{C}$. Thus the test statistic is approximately

$$t_{\text{GLRT}} \approx \frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{x}}{\sqrt{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}}. \tag{6.42}$$

Under $H_0$, $\mathbf{x} : N[\mathbf{0}, \mathbf{C}]$, so the limiting distribution of $t_{\text{GLRT}}$ is also $N[0, 1]$.

### 6.2.3 Subspace invariant AMF and GLRT formulations

When secondary noise-only observations are available, and it is required to detect a known target in noise with unknown covariance, the presence of an invariance subspace causes no complications if the covariance is unconstrained. That is, for observed secondary observations $\mathbf{x}_1, \ldots, \mathbf{x}_M$ and primary data $\mathbf{x}$, the invariant detection problem can simply be formulated in terms of the projected components $\mathbf{y}_1, \ldots, \mathbf{y}_M$ and $\mathbf{y}$, with $\mathbf{y}_i = \mathbf{U}_H^T\mathbf{x}_i$ and $\mathbf{y} = \mathbf{U}_H^T\mathbf{x}$. The presence of the target in the vector $\mathbf{y}$ will be of the form $\alpha\mathbf{U}_H^T\mathbf{s}$, if it appeared in the original as $\alpha\mathbf{s}$.

The subspace invariant detection problem can therefore be reformulated as follows: test for the presence of the target $\alpha\mathbf{U}_H^T\mathbf{s}$ in the observation $\mathbf{y}$, where $\mathbf{y}_1, \ldots, \mathbf{y}_M$ are secondary noise-only data, assumed to

be zero mean, with completely unknown covariance.

An advantage of the invariance subspace formulation is that the distributions of the secondary data no longer have to be considered equal: only the marginal distributions in the subspace orthogonal to the invariance subspace must be the same. This can extend the applicability of the tests to situations where they could otherwise not be applied.

The AMF presented in the previous section can therefore be trivially applied. Written in terms of the original corrupted observations, the invariant test is

$$t_{\text{AMF}} = \frac{\mathbf{s}^T\mathbf{U}_H(\mathbf{U}_H^T\widehat{\mathbf{C}}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}}{\sqrt{\mathbf{s}^T\mathbf{U}_H(\mathbf{U}_H^T\widehat{\mathbf{C}}^{-1}\mathbf{U}_H)\mathbf{U}_H^T\mathbf{s}}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{6.43}$$

This test again has a constant false alarm rate, since it is precisely the AMF for the transformed observations. The limiting distribution of this test statistic is $N[0, 1]$.

The subspace invariant GLRT for the unconstrained case can similarly be seen to be

$$t_{\text{GLRT}} = \frac{\mathbf{s}^T\mathbf{U}_H(\mathbf{U}_H^T\widehat{\mathbf{C}}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}}{\sqrt{\mathbf{s}^T\mathbf{U}_H(\mathbf{U}_H^T\widehat{\mathbf{C}}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{s}}\sqrt{1+1/M\mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\widehat{\mathbf{C}}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}}} \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{6.44}$$

The limiting distribution, as the number of additional observations becomes large, is the same as that of the invariant AMF, namely $N[0, 1]$.

The energy of the modified target $\alpha\mathbf{U}_H^T\mathbf{s}$ is less than or equal to the energy of the original target, which will have the effect of reducing its detectability. However, an improvement in overall detectability can still be brought about in two different ways. The first is through the reduced dimensionality of the observations, as discussed in the introduction to this chapter, which can result in more stable noise covariance parameter estimates. The second is through the potential for more accurate modelling by means of the invariance subspace.

### 6.2.4 Simplified AMF detector formulation

The AMF test requires the inversion of the possibly high-dimensional sample covariance matrix $\widehat{\mathbf{C}}$. If this is undesirable, it may be reasonable to use an approximate value of $\widehat{\mathbf{C}}$ in the numerator of the AMF test, which is then held constant. The numerator is then given by the product $\mathbf{s}^T\mathbf{C}_{\text{ave}}^{-1}\mathbf{x}$, with $\mathbf{C}_{\text{ave}}$ chosen from the data to result in high differentiation ability of the test statistic on average. It is still desirable to have a test which is adaptive, however, and the denominator term has to be adjusted accordingly.

Under the null hypothesis, if $\mathbf{x} : N[\mathbf{0}, \mathbf{C}]$ then the distribution of the numerator is $N[0, \mathbf{s}^T\mathbf{C}_{\text{ave}}^{-1}\mathbf{C}\mathbf{C}_{\text{ave}}^{-1}\mathbf{s}]$.

An appropriate normalisation is therefore to use the modified AMF test statistic

$$\frac{\mathbf{s}^T\mathbf{C}_{\text{ave}}^{-1}\mathbf{x}}{\sqrt{\mathbf{s}^T\mathbf{C}_{\text{ave}}^{-1}\mathbf{C}\mathbf{C}_{\text{ave}}^{-1}\mathbf{s}}}, \tag{6.45}$$

which has a $N[0,1]$ distribution under $H_0$ if $\mathbf{C}$ is known. In practice $\mathbf{C}$ is unknown and is estimated either from secondary data or from the data which are to be tested. Using the estimate $\widehat{\mathbf{C}}$ in the previous expression yields the modified AMF test statistic

$$t_{\text{MAMF}} = \frac{\mathbf{s}^T\mathbf{C}_{\text{ave}}^{-1}\mathbf{x}}{\sqrt{\mathbf{s}^T\mathbf{C}_{\text{ave}}^{-1}\widehat{\mathbf{C}}\mathbf{C}_{\text{ave}}^{-1}\mathbf{s}}}. \tag{6.46}$$

Once again the procedure is to calculate this statistic, compare it to a threshold, and decide $H_1$ when exceeded.

Under the assumption that the noise covariance matrix is unconstrained and is estimated from additional independent samples, this yields a CFAR test. Additionally, the statistic has a closed-form solution under $H_0$, so the threshold can be selected analytically.

To see this, let $\mathbf{w} = \mathbf{C}_{\text{ave}}^{-1}\mathbf{s}$. The modified AMF statistic is then

$$t_{\text{MAMF}} = \frac{\mathbf{w}^T\mathbf{x}}{\sqrt{\mathbf{w}^T\widehat{\mathbf{C}}\mathbf{w}}} = \frac{\mathbf{w}^T\mathbf{C}^{1/2}\mathbf{C}^{-1/2}\mathbf{x}}{\sqrt{\mathbf{w}^T\mathbf{C}^{1/2}\mathbf{C}^{-1/2}\widehat{\mathbf{C}}\mathbf{C}^{-1/2}\mathbf{C}^{1/2}\mathbf{w}}}. \tag{6.47}$$

Again letting $\tilde{\mathbf{C}} = \mathbf{C}^{-1/2}\widehat{\mathbf{C}}\mathbf{C}^{-1/2}$, and letting $\mathbf{w}_m = \mathbf{C}^{1/2}\mathbf{w}$ and $\mathbf{y} = \mathbf{C}^{-1/2}\mathbf{x}$, this becomes

$$\begin{aligned} t_{\text{MAMF}} &= \frac{\mathbf{w}_m^T\mathbf{y}}{\sqrt{\mathbf{w}_m^T\tilde{\mathbf{C}}\mathbf{w}_m}} \\ &= \frac{\hat{\mathbf{w}}_m^T\mathbf{y}}{\sqrt{\hat{\mathbf{w}}_m^T\tilde{\mathbf{C}}\hat{\mathbf{w}}_m}} \\ &= \frac{\hat{\mathbf{w}}_m^T\mathbf{y}}{\sqrt{(\hat{\mathbf{w}}_m^T M\tilde{\mathbf{C}}\hat{\mathbf{w}}_m)/M}} \end{aligned} \tag{6.48}$$

where $\hat{\mathbf{w}}_m$ is the unit vector in the direction of $\mathbf{w}_m$. Since $\mathbf{y} : N[\mathbf{0}, \mathbf{I}]$ under $H_0$, the numerator term $\hat{\mathbf{w}}_m^T\mathbf{y}$ has a $N[0,1]$ distribution. Also, since $M\tilde{\mathbf{C}} : W[\mathbf{I}; n, M]$, the denominator term $\mathbf{w}_m^T M\tilde{\mathbf{C}}\mathbf{w}_m$ has a $W[1; 1, M]$ distribution. This is identically a $\chi^2$ distribution with $M$ degrees of freedom. Furthermore, since the numerator and denominator depend on different sets of observations, they are independent. Thus the numerator of $t_{\text{MAMF}}$ has a central normal distribution with unit variance, and the denominator is the scaled square root of an independent $\chi^2$ distributed quantity with M degrees of freedom. The test statistic therefore has a $t$ distribution under $H_0$, with $M$ degrees of freedom [86, p. 313].

As the number of secondary data vectors increases, the covariance matrix can be better estimated. Thus the covariance estimate used in the MAMF test statistic progressively becomes error free and unbiased. In the limit, the distribution of the modified test statistic becomes normal, with unit variance. In terms of the previous discussion, this can be observed by noting that as the value of $M$ increases, the standard $t$ distribution approaches the $N[0, 1]$ distribution. Thus the distribution of the test statistic becomes approximately $N[0, 1]$ as the covariance parameter estimate improves.

This simplified detector also permits the incorporation of an invariance subspace. Following the same method used for the AMF and GLRT, the corresponding invariant test statistic is

$$t_{\text{MAMF}} = \frac{\mathbf{s}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}_{\text{ave}}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}}{\sqrt{\mathbf{s}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}_{\text{ave}}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\widehat{\mathbf{C}}\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}_{\text{ave}}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{s}}}. \tag{6.49}$$

The quantity $\mathbf{s}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}_{\text{ave}}\mathbf{U}_H)^{-1}\mathbf{U}_H^T$ can be calculated in advance, eliminating the need for matrix inversion at the adaptation stage.

## 6.3 Adaptive detection with covariance constraints

In this section, the use of covariance constraints in the AMF and GLRT formulations presented in the previous sections are discussed. Section 6.3.1 deals with the case where secondary noise-only observations are available. Section 6.3.2 then discusses the case where estimation and detection have to be performed using the same sample of data. In all cases the imposed constraint means that the tests no longer have a constant false alarm rate, although this property does hold asymptotically.

### 6.3.1 Testing with multiple observations

When the additional observations are low in number, constrained covariance matrix estimates are essential. Without these constraints, the variance in the estimates is so large that the test statistic, in both the AMF and GLRT instances, is subject to high uncertainty. This reduces the effectiveness of the resulting tests, often to the point that they are useless.

Even when the number of observations is high, constraints can regularise the estimates, yielding better detectors. As long as the constraints are valid, overall performance will increase. As discussed in the previous chapter, an invariance subspace can extend the validity of these assumptions if they prove overly restrictive. By ignoring those components of the observations which contradict the assumptions, better model validity can be attained. This is particularly important when stationary models are used to model nonstationary data.

Because the invariance subspace estimation procedures require large amounts of data, it is seldom

feasible to adapt these estimates to the particular set of observations being considered. A more viable approach is to make just a single invariance subspace estimate, which aims to eliminate the mismatch for as wide a variety of situations as possible. The invariance subspace is then held constant, while adaptive covariance matrix estimates continue to be made.

An "average" invariance subspace estimate is one approach to the subspace estimation problem which proves useful in practice. A large number of training data observations are used to estimate a candidate invariance subspace and covariance matrix estimate, using the results of the previous section. The estimates can be made in the context of invariance for subspace interference or for model mismatch reduction. The invariance subspace estimate obtained is then assumed to characterise the mismatch components, and corresponding data components are ignored in subsequent testing.

This is by no means a perfect solution, since the notion of an "average" covariance matrix estimate is not appropriate in highly-variable data. In a different context and for a different problem, Burgess and Van Veen [20] use an estimate which is based purely on the class of allowable noise distributions under the assumed model. A similar approach could perhaps be adopted for this problem, where the invariance subspace minimises the average or maximum mismatch over the entire class of valid noise parameters. Nevertheless, such an approach has not been formulated, and the average estimate has proven to be useful in applications. There are a rich set of possibilities for this estimation problem.

Thus, using an average estimate, the invariance subspace corresponding to a set of data can be specified. This estimate is then assumed valid and known, and held fixed when the detector is applied to data which are subsequently presented to it. In this context, the adaptive subspace invariant detection statistics presented in the previous section can simply be applied, with the covariance matrix estimates replaced with their invariant constrained counterparts.

The use of constrained estimates in the AMF proceeds trivially, with the subspace invariant constrained estimate $\widehat{\mathbf{C}}$ used directly in the expression for the test. The GLRT does not proceed so simply, however, since the constraint has to be worked directly into the maximisation of the numerator and denominator. This requirement arises purely as a result of the potential target-carrying observation $\mathbf{x}$, which is not included in the estimate $\widehat{\mathbf{C}}$. Nevertheless, a reasonable approach is to ignore this factor, and use the constrained covariance matrix estimate as $\widehat{\mathbf{C}}$ in the GLRT expression.

Because constrained estimates have been used, neither of the invariant tests can be expected to have a constant false alarm rate. That is, the covariance matrix is no longer just the sample covariance, and a Wishart distribution is not appropriate for representing its distribution. As long as the invariant covariance matrix estimates are consistent, however, the limiting distributions are still appropriate in each case. Thus if the amount of data available for estimation is large, the distributions of the test statistics in each case can be assumed to be $N[0, 1]$. An approximate CFAR property is therefore appropriate.

Again, because data have been eliminated through the invariance requirement, the tests cannot be claimed asymptotically optimal. Their performance may however be better than tests which *are* optimal, but are applied in contexts which make invalid assumptions.

### 6.3.2 Adaptive detection from a single sample

Constraints become even more important when adaptive detection has to be performed on just a single sample of data. Without constraints, it is not even possible to obtain a useful covariance matrix estimate, since it will be singular. The adaptive detectors usually require the inversion of this matrix, so the procedures fail entirely.

For stationary constraints, a single sample of data is sufficient for estimates to be made. The longer the sample, the better the estimates of the coefficients corresponding to low-order lag products. In some respects this is a more suitable formulation for many problems, where if a multiple-observation method is used, then the additional data have to be taken from different locations in the same set of data.

With constraints and only a single observation, the AMF test from the previous section can be applied unmodified. Thus the test statistic for the case of no invariance subspace is simply

$$t_{\text{AMF}} = \frac{\mathbf{s}^T \widehat{\mathbf{C}}^{-1} \mathbf{x}}{\sqrt{\mathbf{s}^T \widehat{\mathbf{C}}^{-1} \mathbf{s}}}, \tag{6.50}$$

where $\widehat{\mathbf{C}}$ is a suitable constrained covariance matrix estimate.

The choice of $\widehat{\mathbf{C}}$ in this case is less obvious, however, since additional observations containing only noise are not available. For CFAR detection, the assumption that only noise is present in the data is however appropriate. Then, if the observation does indeed only contain noise, the estimate will be valid and consistent. For long data samples the covariance matrix estimate will be good, and the distribution of the test statistic will be approximately $N[0, 1]$. The test threshold can therefore be set to yield an approximately constant false alarm rate.

The detection performance may however suffer using this approach, since the covariance matrix estimate will be invalid when the target is in fact present. Then the covariance matrix estimate becomes biased, and is no longer consistent. In some applications this reduced performance may be justified by the approximate CFAR property of the test.

To see how the performance can suffer, consider the term $\mathbf{s}^T \widehat{\mathbf{C}}^{-1} \mathbf{s}$. Ideally, if the estimate is perfect, then this term will equal $\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}$, where $\mathbf{C}$ is the true covariance matrix. Suppose, however, that there is a target present in the data from which the covariance is estimated, and the presence of this target is simply ignored. Ignoring for now the covariance constraint, the presence of the target $\alpha \mathbf{s}$ will tend

to change the value of the term to $\mathbf{s}^T(\mathbf{C}+\alpha^2\mathbf{s}\mathbf{s}^T)^{-1}\mathbf{s}$. Using Woodbury's identity [77, p. 571], this expression can be written as

$$\begin{aligned}\mathbf{s}^T(\mathbf{C}+\alpha^2\mathbf{s}\mathbf{s}^T)^{-1}\mathbf{s} &= \mathbf{s}^T\left[\mathbf{C}^{-1}-\alpha^2\mathbf{C}^{-1}\mathbf{s}(1+\alpha^2\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s})^{-1}\mathbf{s}^T\mathbf{C}^{-1}\right]\mathbf{s} \\ &= \mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}\left[1-\alpha^2\frac{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}{1+\alpha^2\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}\right]. \end{aligned} \tag{6.51}$$

The second term in square brackets is identically zero for $\alpha = 0$, but tends to unity as $\alpha$ grows without bound. Thus for finite $\alpha > 0$, it is always true that

$$\mathbf{s}^T(\mathbf{C}+\alpha^2\mathbf{s}\mathbf{s}^T)^{-1}\mathbf{s} < \mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}. \tag{6.52}$$

Furthermore, as $\alpha \longrightarrow \infty$, $\mathbf{s}^T(\mathbf{C}+\alpha^2\mathbf{s}\mathbf{s}^T)^{-1}\mathbf{s} \longrightarrow 0$.

Assuming zero uncertainty in the covariance estimate, and still ignoring the constraint on this estimate, under $H_1$ the value of the AMF test statistic will be approximately

$$\frac{\mathbf{s}^T(\mathbf{C}+\alpha^2\mathbf{s}\mathbf{s}^T)^{-1}\mathbf{s}}{\sqrt{\mathbf{s}^T(\mathbf{C}+\alpha^2\mathbf{s}\mathbf{s}^T)^{-1}\mathbf{s}}} = \sqrt{\mathbf{s}^T(\mathbf{C}+\alpha^2\mathbf{s}\mathbf{s}^T)^{-1}\mathbf{s}}. \tag{6.53}$$

In this form it can be seen that as the target amplitude $\alpha$ increases, the value tends towards zero. Thus, the performance of the AMF test proposed will become worse as the target amplitude increases. Now, if a constraint were imposed on the covariance matrix estimate, the presence of the target in the signal would still tend to shift the value of $\widehat{\mathbf{C}}$ towards $\mathbf{C}+\alpha^2\mathbf{s}\mathbf{s}^T$, although the effect may be reduced if this latter quantity does not conform to the constraint. Nonetheless, the same conclusions may be expected to apply, and the presence of the target will tend to reduce detection performance.

With the development of invariant detectors, another option exists: the estimate of the covariance matrix can be made invariant to the presence of the target. The estimator is then required to yield the same result irrespective of whether there is a target present in the data sample or not. For the simple problem of detecting a target with unknown amplitude in noise, such an invariant estimate can be obtained by including the vector spanning the target subspace into the invariance subspace.

As long as the invariant estimate is consistent, the test obtained will not suffer from the problems just outlined. That is, the presence of the target will not corrupt the covariance estimate, so the test will no longer tend to fail for targets with large amplitudes. In fact, the performance should improve for larger amplitude targets, as indeed one would expect from a test. Furthermore, the test will still have an asymptotic CFAR property.

If a fixed invariance subspace is included into the detection formulation, the same conclusions apply. That is, the test statistic in Equation 6.43 can be used, with an invariant estimate $\widehat{\mathbf{C}}$ substituted for the sample covariance matrix. This estimate is required to be invariant to both the invariance subspace

and the target subspace.

This proposed solution has interpretation in terms of the noise-alone reference of El Ayadi [32, 31] and El Ayadi and Picinbono [33]. In particular, the subspace orthogonal to both the invariance subspace and the target subspace can be considered to contain only noise under both hypotheses. Since the invariant estimator only makes use of data in this subspace, it can be considered a NAR estimate. The subspace decomposition presented at the start of this chapter can be used to interpret the resulting test in terms of these different components. It must be noted that in this application the components of the estimates related to data in the invariance subspace are *not* ignored by the final detection stage.

As an aside, the target-invariant covariance matrix estimates can be extended to the case where one of many different targets may be present in the data. The invariance subspace is then simply augmented with the additional basis functions spanning the space of all possible targets. In practice a principal components analysis can be used to find an approximate target subspace in this instance.

An invariance subspace has additional advantages in the context of adaptive detection. In particular, by enhancing the validity of a stationary model, it permits the use of longer data samples in the estimation procedures. That is, because the primary causes of nonstationarity or mismatch are ignored, the stationarity assumption is appropriate over longer distances in space or time than would otherwise be true. Thus more data samples can be used in the estimate, and the conditions where asymptotic results are obtained is more easily reached. This is particularly important when covariances are constrained, and the test statistic distributions under $H_0$ are not known. In these instances the asymptotic results are all that are available for test threshold selection. If estimates can be enhanced to encourage the accuracy of these approximations, it can aid considerably in detection.

The GLRT does not suffer from the same disadvantages as the AMF, since separate estimates of the covariance matrix are made under the assumed hypotheses. However, a difficulty with the GLRT regards its implementation, particularly when the covariance matrix is structured. The NAR estimate can again provide a means of overcoming this problem.

For the case where the GLRT is used on a single observation, and constraints are included, the derivation presented in Section 6.2.2 is inappropriate. In particular, the test requires a *constrained* maximisation of the likelihood over the allowed covariances. The GLRT is therefore

$$t_{\text{GLRT}}(\mathbf{x}) = \frac{\max_{\alpha,\mathbf{C}}(2\pi)^{-n/2}|\mathbf{C}|^{-1/2}e^{-1/2(\mathbf{x}-\alpha\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{x}-\alpha\mathbf{s})}}{\max_{\mathbf{C}}(2\pi)^{-n/2}|\mathbf{C}|^{-1/2}e^{-1/2\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}}}. \tag{6.54}$$

The maximisation in both cases is performed over all valid covariances, and in the numerator over all appropriate $\alpha$.

The denominator in this expression can be calculated easily. It simply involves a constrained estimation of $\mathbf{C}$ from the data $\mathbf{x}$, under the assumption that $\mathbf{C}$ is of the required form. Any maximum likelihood

constrained estimation methods are appropriate for this term. A difficulty arises in the calculation of the numerator, however. In general the maximisation over the parameter $\alpha$ cannot be conveniently included into the constrained covariance matrix estimation. Therefore, in practice it may be necessary to consider a number of possible values for $\alpha$, make the associated constrained covariance matrix estimates, and find the combination which results in the highest likelihood. The number of constrained covariance matrix estimates required is equal to the number of values considered for $\alpha$.

The computational requirements involved in this calculation are prohibitive, and it is usually not practical to use this method. Once again, however, the invariant covariance matrix estimate can overcome the problem. Instead of making multiple covariance estimates for all possible values of $\alpha$, an estimate is simply used which is invariant to $\alpha$. In other words, it is invariant to the presence of the target. Under this approximation, the numerator term can be found using just a single covariance matrix estimate.

There is no reason that different covariance matrix estimates cannot still be used in both the numerator and the denominator, but if both estimates are made invariant then only one covariance matrix estimation calculation is required. That is, a target-invariant estimate of the noise can be made, and used simultaneously as the approximate maximising value for $\mathbf{C}$ in the numerator and the denominator. The numerator can then easily be maximised over the remaining parameter $\alpha$, yielding the final test. If this procedure is followed, then the exact same test is obtained as for the AMF using the invariant covariance matrix estimate, discussed earlier in this section. Thus the two approaches are reconciled.

Irrespective of the precise formulation used, the same conclusions as for the AMF hold when an invariance subspace is included for purposes of enhanced model match. In that case the GLRT test statistic would ordinarily be

$$t_{\text{GLRT}}(\mathbf{x}) = \frac{\max_{\alpha,\mathbf{C}}(2\pi)^{-(n-p)/2}|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H|^{-1/2}e^{-1/2(\mathbf{x}-\alpha\mathbf{s})^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T(\mathbf{x}-\alpha\mathbf{s})}}{\max_{\mathbf{C}}(2\pi)^{-(n-p)/2}|\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H|^{-1/2}e^{-1/2\mathbf{x}^T\mathbf{U}_H(\mathbf{U}_H^T\mathbf{C}\mathbf{U}_H)^{-1}\mathbf{U}_H^T\mathbf{x}}}, \tag{6.55}$$

which implies invariant maximum likelihood estimates of $\mathbf{C}$. If the maximisation in the numerator is taken to be invariant to $\langle\mathbf{s}\rangle$ in addition to $\langle\mathbf{U}_I\rangle$, then the implementation is facilitated.

The simplified AMF detector presented in Section 6.2.4 also extends easily to the case of adaptive detection using just a single data sample, and the same methods discussed for the AMF detector apply. The resulting test remains asymptotically CFAR, with a $N[0, 1]$ distribution under the null hypothesis.

## 6.4 Examples of detection performance using single samples

In this section the adaptive tests presented in Section 6.3.2 are applied to certain real-data applications. Results are obtained which demonstrate both the detectability and the predictability of the tests. It is

shown that in all cases the incorporation of an invariance subspace increases the predictability of the detectors. In most cases this comes at the cost of reduced detectability.

The data used to assess performance are the same as used in Section 5.8.1, and shown in Figure 5.11. Again the images were simply used to provide noise samples which come from a realistic source, and have a definite but unknown underlying relationship. The assumption was made that it was required to detect the presence of Target 2 in Figure 5.12 in the noise samples.

Results are presented for three different detectors, namely the AMF, the modified AMF, and the AMF using target-invariant covariance matrix estimates. The target amplitude is considered unknown, but constrained to be positive. In each case the effect of including an interference subspace is assessed.

As in the example presented in Section 5.8.1, it does not appear to be critical whether the interference subspace formulation or the general invariance subspace formulation is used for the subspace estimate. Similar invariance subspace results are obtained in each case, and conclusions which apply to the one case are appropriate in the other. Results are therefore not presented for the use of an invariance subspace for explicitly reducing model mismatch.

### 6.4.1 AMF results using single samples

To provide a reference against which to compare the performance of the invariant detectors, a conventional noninvariant adaptive detector was developed for samples obtained from the test image. The assumption was made that, over short distances, the data contained in horizontal lines across the image could be modelled as a first-order autoregressive process with zero mean.

Because the data are quite variable, the resulting detector performed very poorly. To enhance the match between the model and the data an unsharp-masking preprocessing stage was therefore employed, which attempted to remove the underlying trends in the data. This was done by calculating a moving average across the image data, on a line-by-line basis, and subtracting the resulting low-pass version of the signal from the original. The length of the window used to form the average is a variable that has to be specified.

With the preprocessing, the resulting data were more suited to the first-order AR model. The following procedure was then applied to the test data:

- A sample of data was extracted from a particular location along a horizontal line in the image, which is known to contain only noise. The sample was used to make an estimate of the local covariance properties, using maximum likelihood estimation under an AR(1) assumption. Two lengths of data samples were considered for the estimation window, namely 32 and 64. Using the sample an estimate $\widehat{\mathbf{C}}$ was obtained.

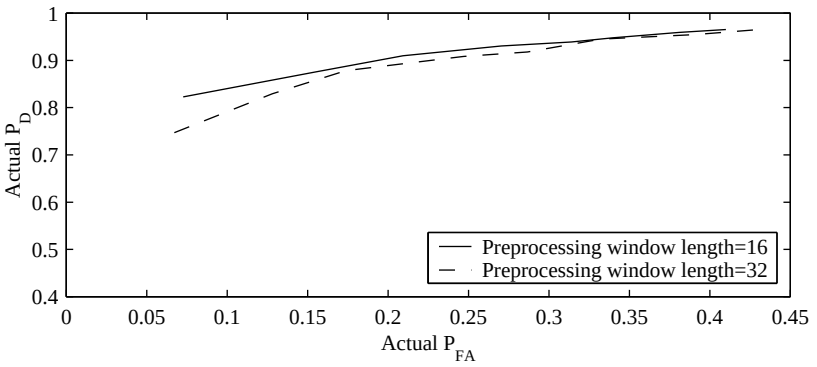
**Section 6.4: Examples of detection performance using single samples**

- The estimate $\widehat{\mathbf{C}}$ is assumed to accurately characterise the covariance properties at the location from which the estimation sample was extracted. Data on which to perform detection were then extracted from the same location, although the length of this data sample was fixed at 32. This constitutes the testing data $\mathbf{x}$.
- Using the two quantities obtained, the noninvariant AMF statistic value was calculated using Equation 6.50. This represents a sample value of the test statistic when $H_0$ is in force.
- The procedure was repeated on different data extracted from other locations in the test image. The centres of the locations considered were separated by a distance of 8 pixels in both the vertical and horizontal directions. A set of values was thus obtained which is representative of the values taken on by the test statistic $t_{\text{AMF}}$ in the absence of a target.
- The procedure was then repeated, but this time a target was inserted into every data sample extracted. The profile of the target was modified to take into account the preprocessing stage. For the covariance estimation, however, the assumption was still made that the data contained only noise. Using the same methods as before, a set of representative values of the test statistic $t_{\text{AMF}}$ were therefore obtained for the case of target presence.

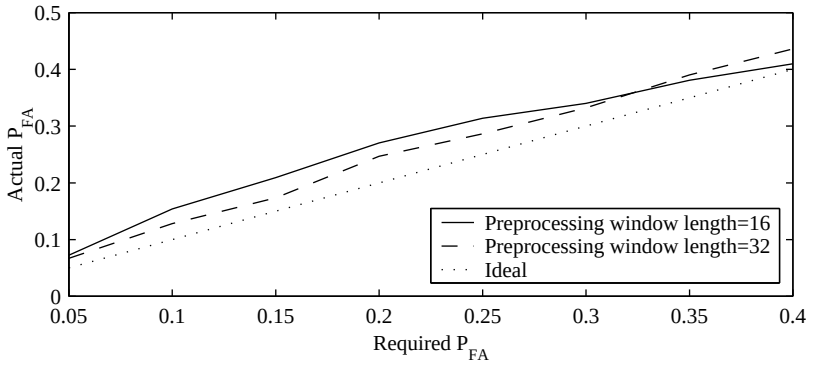
Once these sets of values under the two different hypotheses were obtained, an assessment of the test performance could be made. In each case considered, three plots were generated:

- Using the sample test statistic values, the performance of the detector can be assessed. Firstly, by considering different test statistic values, the ROC curve was obtained. This is a plot of the detection probability $P_{\text{D}}$ versus the false alarm rate $P_{\text{FA}}$, as the threshold value used in the test changes. It is a direct reflection of the discrimination ability of the test, also referred to as the detectability.
- To set the test threshold in practice, the assumption has to be made that the asymptotic results apply. That is, the distribution of the test statistic $t_{\text{AMF}}$ under $H_0$ is assumed to be $N[0, 1]$. For a required $P_{\text{FA}}$ it is therefore possible to find a threshold value to be used in the test. A second plot was produced which shows the actual observed $P_{\text{FA}}$ versus the required $P_{\text{FA}}$, indicating the extent to which the $N[0, 1]$ distribution assumption is valid. The degree to which the desired performance matches the actual performance is referred to as the predictability of the test.
- Finally, a third plot of the actual observed detection probability versus the required $P_{\text{FA}}$ was generated, which again uses the assumption that the null distribution of $t_{\text{AMF}}$ is $N[0, 1]$.

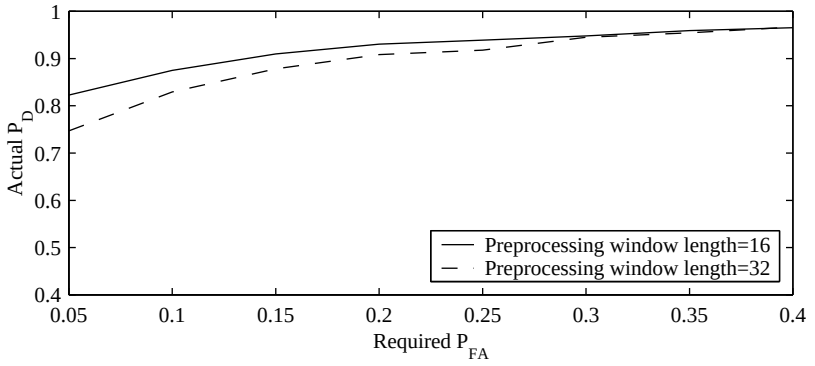
Figure 6.2 shows the results for an estimation window of length 32. Two different lengths of preprocessing windows were considered, namely that of 16 and 32 samples. The window of length 32 modifies the original data less than the window of length 16. Figure 6.3 shows similar results, but this




(a) ROC curves for detection.



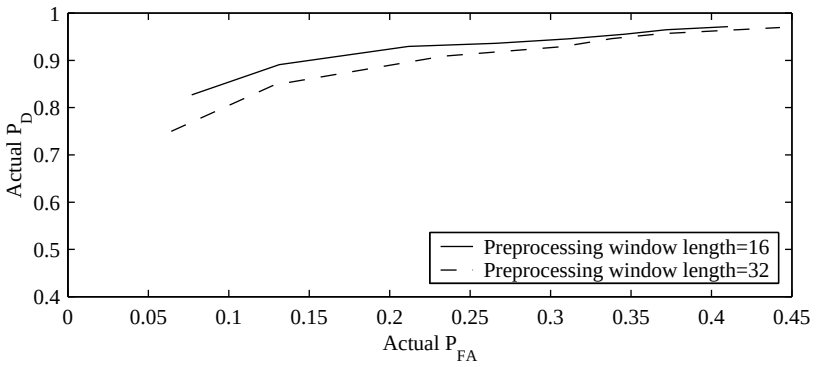

(b) Actual false alarm rate versus desired false alarm rate.



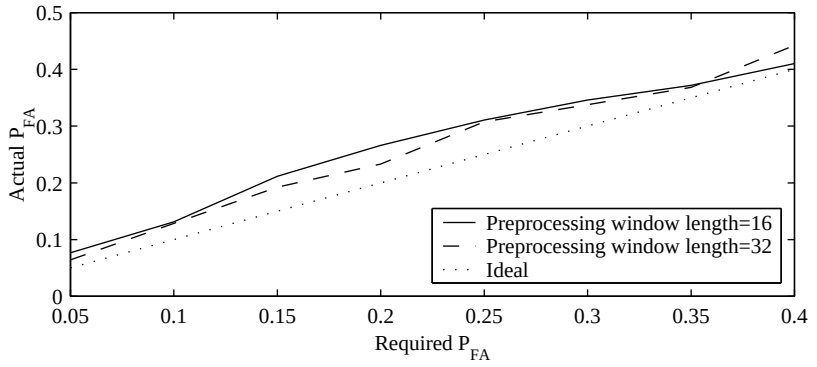

(c) Actual detection probability versus desired false alarm rate.

Figure 6.2: AMF detection results (without an invariance subspace), using an estimation window of length 32.

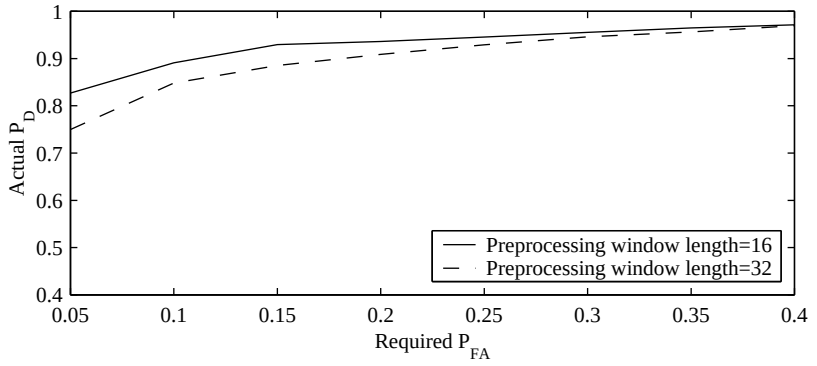
**Section 6.4: Examples of detection performance using single samples**




(a) ROC curves for detection.




(b) Actual false alarm rate versus desired false alarm rate.




(c) Actual detection probability versus desired false alarm rate.

Figure 6.3: AMF detection results (without an invariance subspace), using an estimation window of length 64.

time using an estimation window of length 64.

A number of conclusions can be drawn from the results presented. Firstly, longer estimation windows do not improve the performance of the test, either in terms of detectability (as shown by the ROC curves) or predictability (as shown by the plots of actual versus required $P_{\text{FA}}$). This can be attributed to the fact that there is model mismatch, so the covariance estimates are biased. Increasing the estimation window length does not decrease this bias.

Using more samples for estimation should however reduce the variance in the estimate, which could bring about improved performance. However, this is not observed. A possible reason for this lack of improvement is that the assumed model is not appropriate over the full length of the longer estimation window. Thus the decreased variance is offset by inappropriateness of the model over longer sample lengths.

Secondly, the difference between the required false alarm rate and the achieved false alarm rate is high. This indicates that the assumed $N[0, 1]$ distribution of the test statistic under $H_0$ is not appropriate. That is, even with long sequences of estimation data the expected asymptotic properties of the test are not coming into play. Again this can be attributed to model mismatch, and the fact that the data are not stationary across the length of the estimation window.

The conclusion may be drawn that even with the preprocessing, the assumption of an AR(1) process along the length of the data windows being considered is inappropriate. The primary effect of the mismatch is that the asymptotic properties of the test are not attained, and the detector is not predictable.

The invariant test formulation can be applied to the same problem. Because the model explicitly incorporates mismatch, the preprocessing stage is not required. That is, the invariant detector simply ignores those components which are in conflict with the model. An analysis of the invariant test was performed as follows:

- Candidate invariance subspaces were estimated from the training image in Figure 5.11. Using a subspace interference assumption, invariance subspaces of dimension 1, 3, and 5 were estimated from the training data. The method used was the simultaneous approximate maximum likelihood method presented in Section 5.7, using an assumed AR(1) model. The length of the invariance subspace basis functions was chosen to be the same as the length of the estimation window. The invariance subspaces were then assumed fixed, while the invariant detector was applied to the test data.
- A similar procedure was then adopted as for the noninvariant adaptive detector discussed earlier. In this case, however, the covariance matrix estimate was made to be invariant to the assumed invariance subspace. For the calculation of both test statistic values, the covariance estimates ignored the possible presence of a target. The test statistic $t_{\text{AMF}}$ from Equation 6.43 was used.

As with the noninvariant detector, this procedure results in two sets of sample values of the test statistic, for the conditions of target presence and target absence. The sets of values were analysed in the same manner as before. Figure 6.4 shows results for three different invariance subspace dimensions, and an estimation window of length 32. Figure 6.5 shows similar results, but for an estimation window of length 64.

The first point to note from these results is that the detectability of the test tends to decrease as the invariance subspace is increased. This can be seen from the ROC curves, which move downwards for higher invariance dimensions. This is a result of portions of the target lying in the invariance subspace, and not being available for detection. However, the penalty incurred is far smaller for the case of a longer estimation window length. This indicates that the longer estimation window has produced a covariance estimate which is better than that obtained for shorter estimation windows. The only reason that this improved estimation can be achieved is because the invariance subspace has improved the validity of the model, even over long data lengths.

The second point of interest is that as the invariance subspace dimension is increased, the actual false alarm rate tends towards the required rate. This makes for a very predictable detector, which performs as designed. The presence of the invariance subspace has eliminated those components of the problem where the data conflict with the AR(1) assumption. The bias in the covariance estimates is therefore reduced, making the asymptotic properties of the test more appropriate.

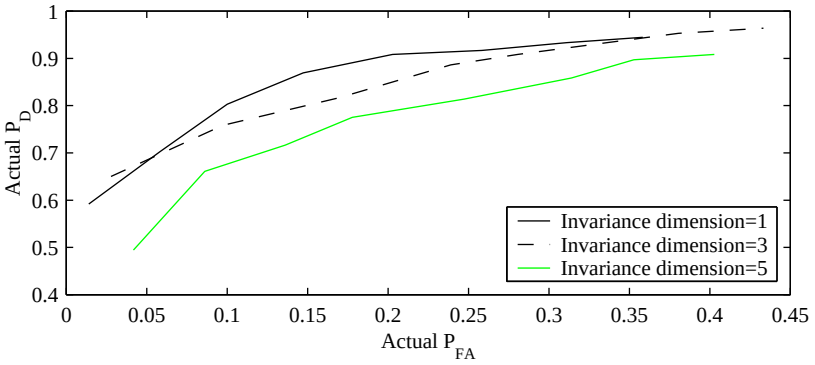
### 6.4.2 Modified AMF results using single samples

In this section, similar results are provided to those of the previous section. The difference is that in this case the modified AMF is used, which uses an average estimate of the covariance in the numerator. Thus the adaptivity only comes in through the denominator term, which changes to suit the noise properties of the particular sample of data being analysed.

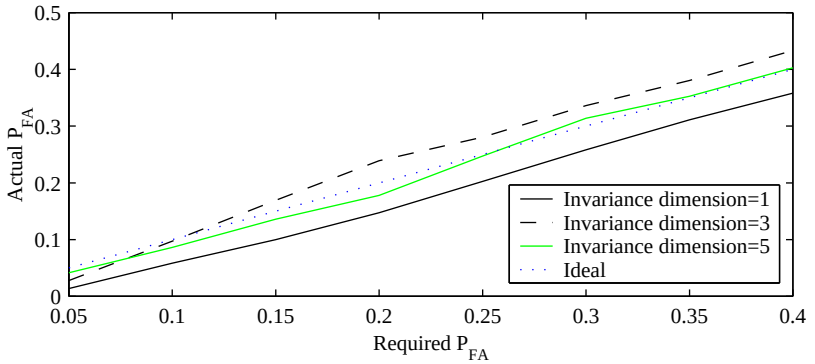
For assessing the noninvariant detector for this problem, the same procedure used in the previous section was adopted. The results for estimation windows of length 32 and 64 are shown in Figures 6.6 and 6.7. Shown in Figures 6.8 and 6.9 are analogous results for the modified AMF test when invariance subspaces of different orders are included.

An interesting point in these plots is that in the nonadaptive case the longer preprocessing window leads to better detectability, which is in contrast to the results presented in the previous section. The predictability for this case is also improved, as demonstrated in the plots of required versus actual false alarm rates. This result may well be specific to the target being detected. However, the reduced uncertainty brought about by fixing the numerator of the test statistic has clearly improved the detection.

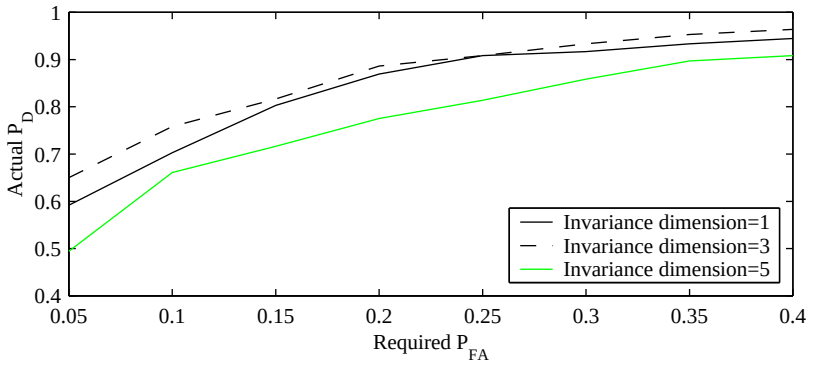
By fixing the detection statistic, the capacity of the test to adapt has been restricted. The test therefore




(a) ROC curves for detection.



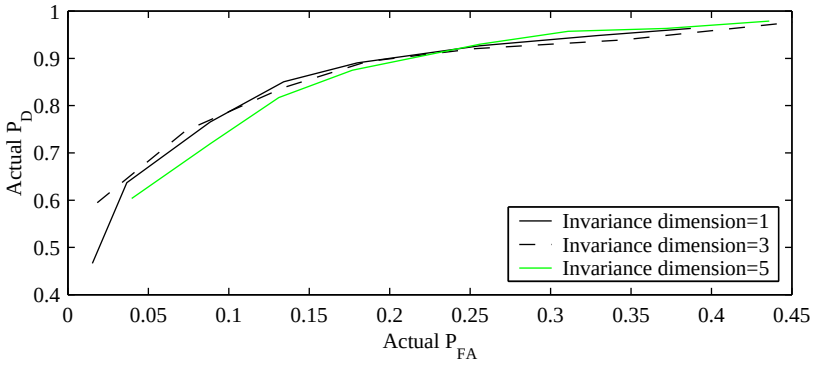

(b) Actual false alarm rate versus desired false alarm rate.



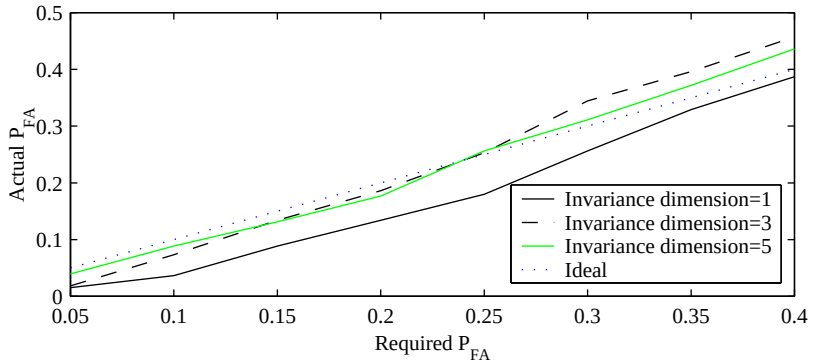

(c) Actual detection probability versus desired false alarm rate.

Figure 6.4: Invariant AMF detection results, using an estimation window of length 32.

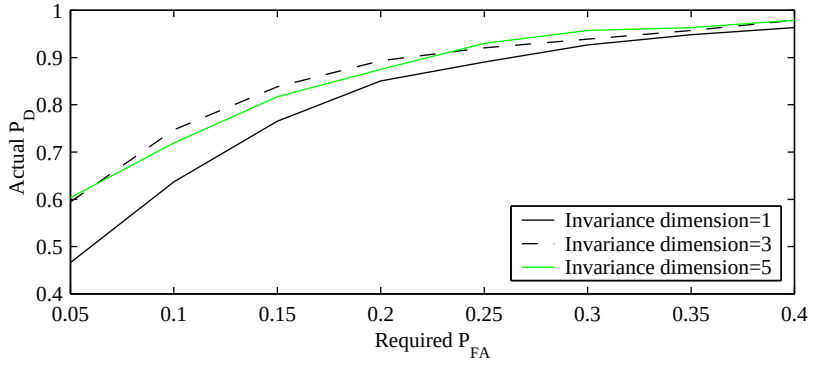
## Section 6.4: Examples of detection performance using single samples




(a) ROC curves for detection.

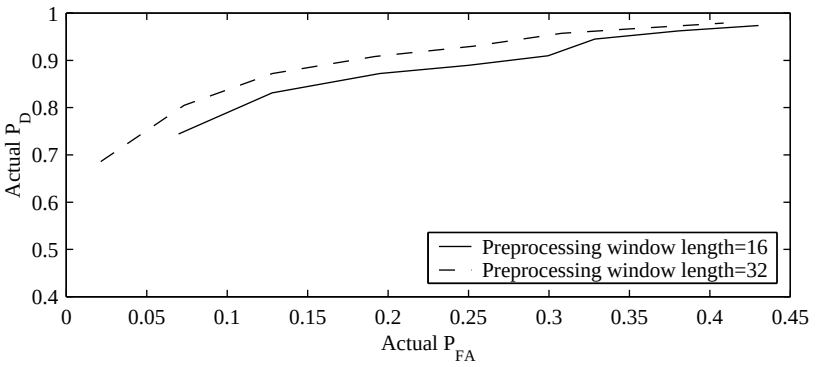



(b) Actual false alarm rate versus desired false alarm rate.

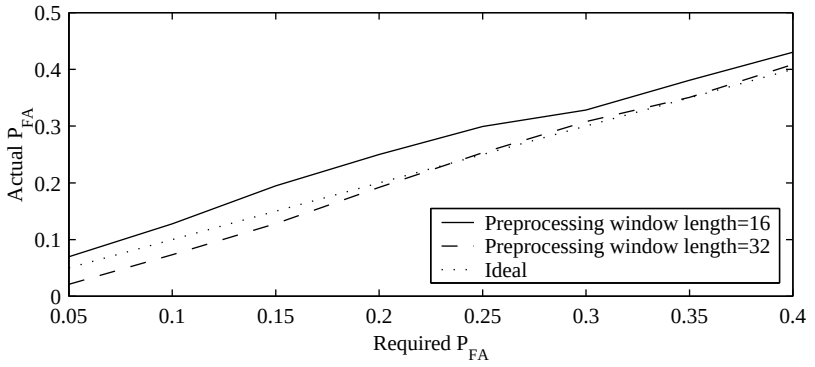



(c) Actual detection probability versus desired false alarm rate.

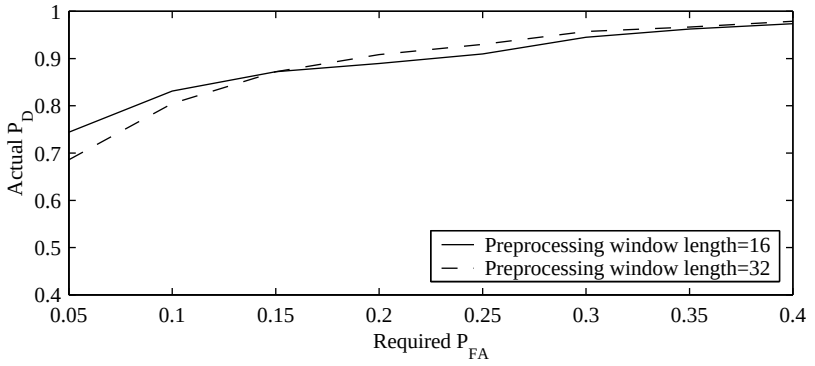
Figure 6.5: Invariant AMF detection results, using an estimation window of length 64.




(a) ROC curves for detection.



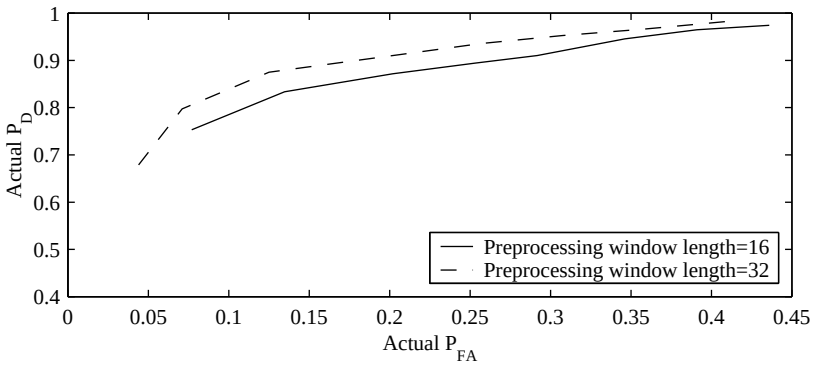

(b) Actual false alarm rate versus desired false alarm rate.



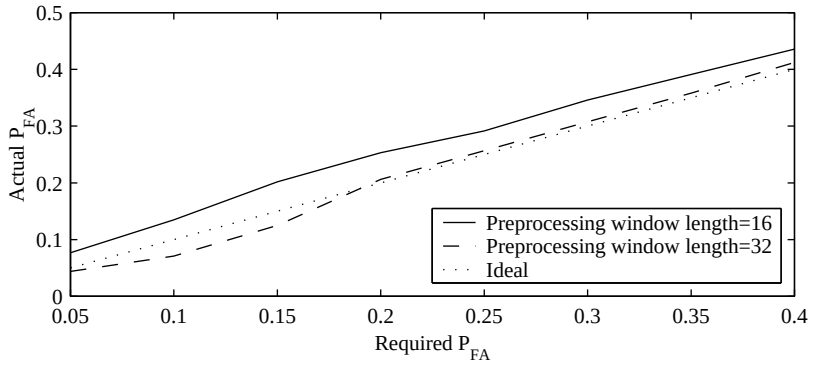

(c) Actual detection probability versus desired false alarm rate.

Figure 6.6: Modified AMF detection results (without an invariance subspace), using an estimation window of length 32.

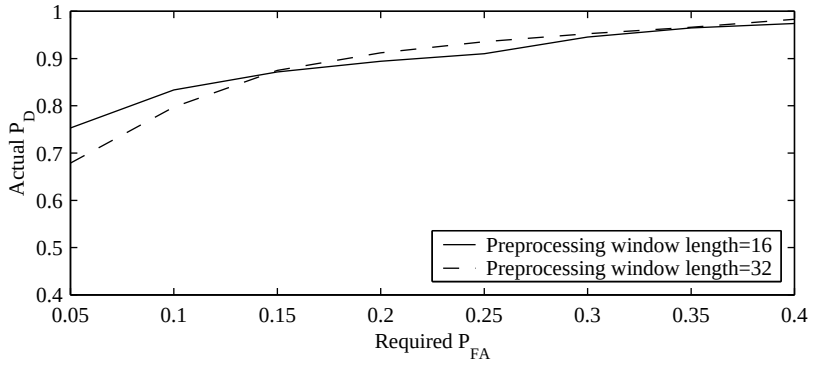
**Section 6.4: Examples of detection performance using single samples**




(a) ROC curves for detection.

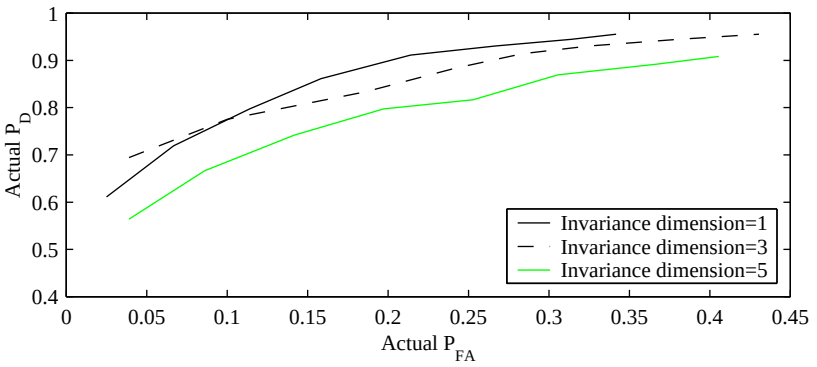



(b) Actual false alarm rate versus desired false alarm rate.

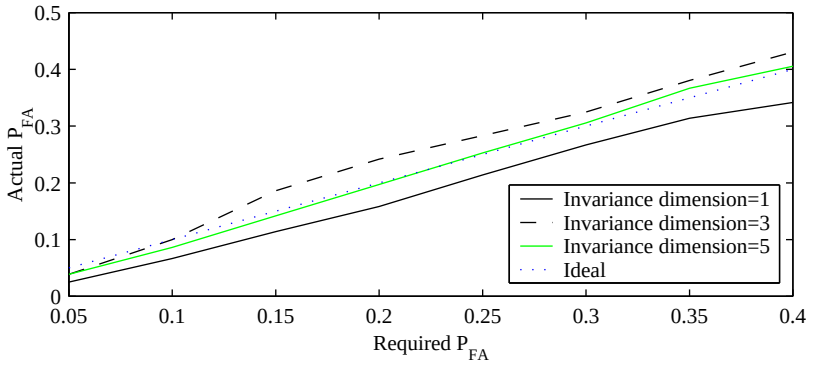



(c) Actual detection probability versus desired false alarm rate.

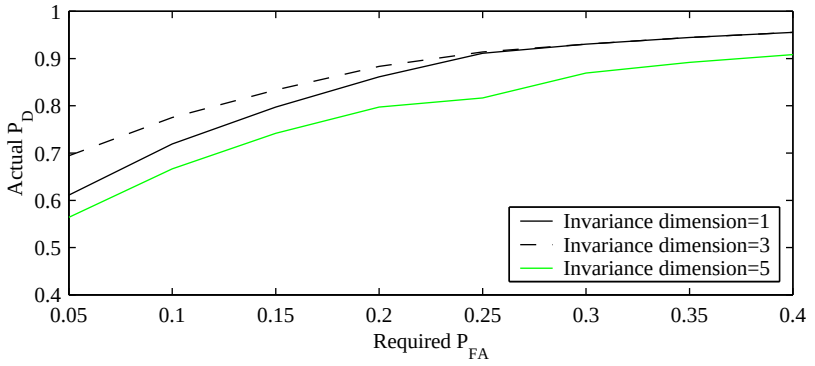
Figure 6.7: Modified AMF detection results (without an invariance subspace), using an estimation window of length 64.




(a) ROC curves for detection.



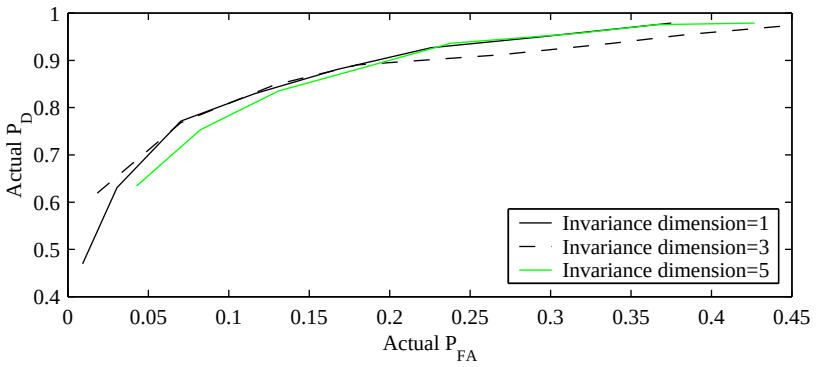

(b) Actual false alarm rate versus desired false alarm rate.



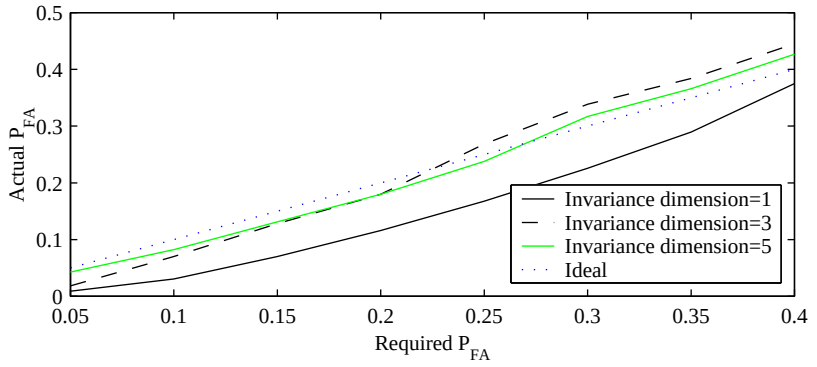

(c) Actual detection probability versus desired false alarm rate.

Figure 6.8: Invariant modified AMF detection results, using an estimation window of length 32.

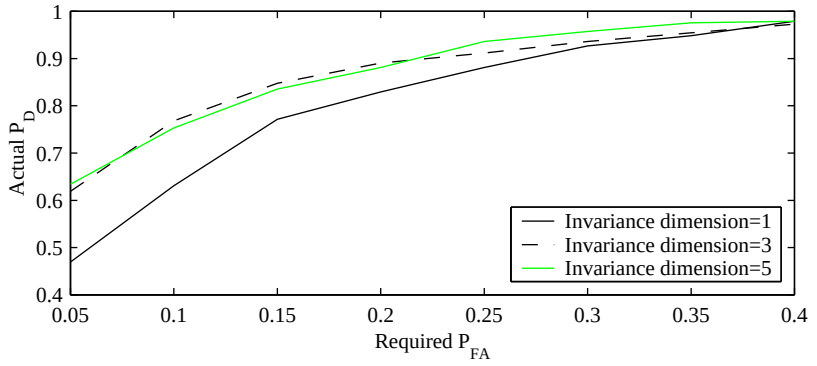
**Section 6.4: Examples of detection performance using single samples**




(a) ROC curves for detection.




(b) Actual false alarm rate versus desired false alarm rate.




(c) Actual detection probability versus desired false alarm rate.

Figure 6.9: Invariant modified AMF detection results, using an estimation window of length 64.

works well in the case of the data under analysis here, which are quite homogeneous. For more variable data the lack of adaptivity will incur a performance penalty. Nonetheless, the results indicate that adaptivity in general causes a decrease in performance, as a result of having to make estimates of the required parameters. Thus detectors should only be made as adaptive as necessary.

The invariant detector works quite well in this case, especially with regard to predictability. The detection performance does suffer, however, again as a result of part of the target being ignored in the detection. The observation can once again be made that the longer estimation window lengths improve the overall performance, a factor which can be directly attributed to the invariance subspace making the assumed model appropriate over longer data lengths.

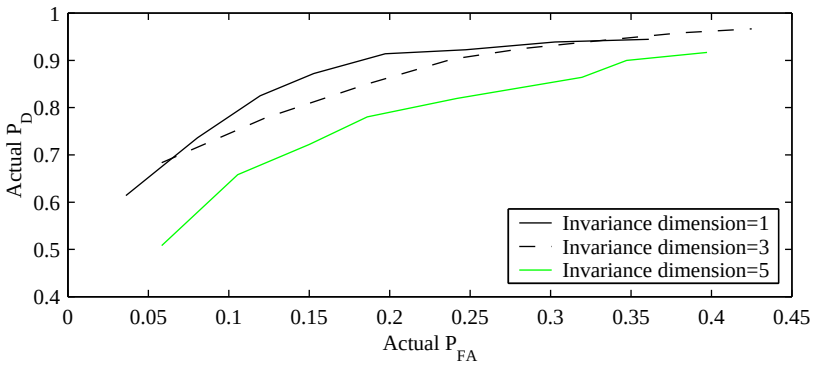
Finally, the last plot in Figure 6.9 demonstrates how improved predictability can lead directly to improved detectability, purely as a result of being able to set the test threshold more accurately. As the invariance subspace increases the detectability decreases, but through better threshold specification the more invariant detector can ultimately produce higher detectability as a function of the required false alarm rate.

### 6.4.3 Target-invariant AMF results using single samples

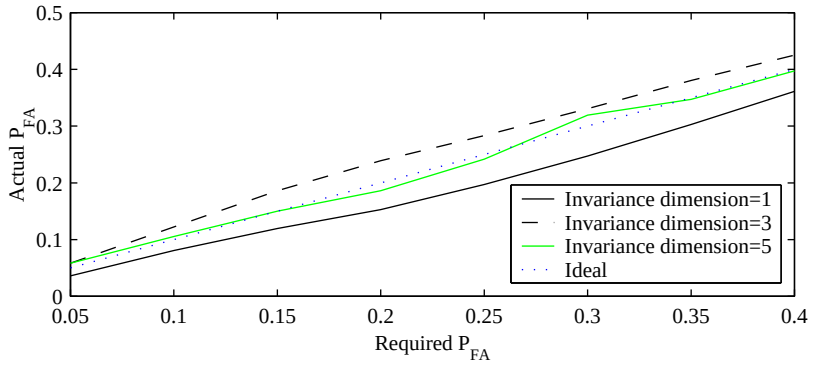
The final results demonstrate the use of the target-invariant method for estimating the covariance matrix used in the invariant AMF. These are provided in Figures 6.10 and 6.11. Here the covariance estimate is completely independent of the possible presence of a target. The results are marginally better than the test which assumes only noise present in the estimation data, but the difference is slight.

This improvement is small primarily because of the tight constraint on the noise model. That is, the presence of a short duration transient target in the signal is almost completely ignored in the estimation, since it conflicts very strongly with the assumed set of possible distributions. The noninvariant estimation therefore does not suffer greatly from the presence of this target. For different targets, which more closely resemble realisations of the low-order random process, this may no longer be the case. Also, less restrictive noise models may be more affected by the presence of the target.

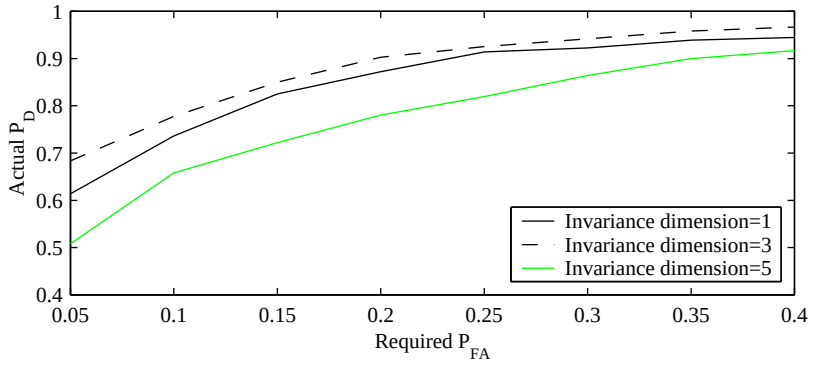
## Section 6.4: Examples of detection performance using single samples




(a) ROC curves for detection.

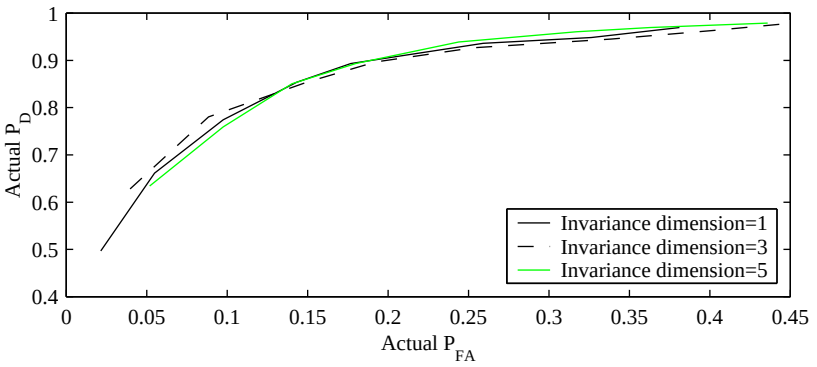



(b) Actual false alarm rate versus desired false alarm rate.

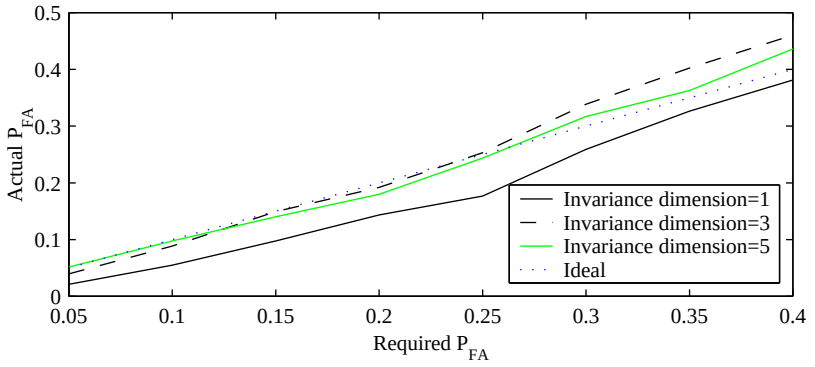



(c) Actual detection probability versus desired false alarm rate.

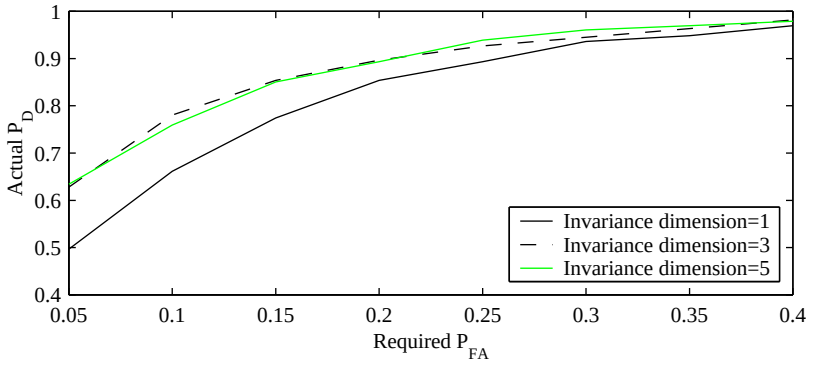
Figure 6.10: Target-invariant AMF detection results, using an estimation window of length 32.




(a) ROC curves for detection.




(b) Actual false alarm rate versus desired false alarm rate.




(c) Actual detection probability versus desired false alarm rate.

Figure 6.11: Target-invariant AMF detection results, using an estimation window of length 64.

**Section 6.4: Examples of detection performance using single samples**

# Chapter 7

# Conclusion

Section 7.1 summarises the main topics and results in this thesis. Section 7.2 then makes some general conclusions based on the work.

## 7.1 Summary of main topics

The primary goal of this thesis is to demonstrate some of the roles that invariance and constraints can play in detection problems. Two types of invariance are considered, namely invariance to cyclic permutations, and invariance to components of the data contained in linear subspaces of the observation space.

The GLRT plays an important part in applied target detection, and this work is no exception. Chapter 3 addresses the issue of suboptimality of the GLRT for simple detection of targets with unknown location in certain types of noise. Specifically, a UMPI test is developed for the case where the target is known and the covariance circulant. This test is most powerful in the class of all cyclic permutation invariant tests. Since the GLRT falls into this class, a direct comparison of the two tests is possible. It is shown that for circulant covariances the differences between the GLRT and the optimal UMPI test are small. This justifies the use of the more conventional GLRT for the testing problem.

Another issue that is addressed is the use of invariance for eliminating nuisance parameters. For certain unknown target location problems, it is demonstrated that cyclic permutation invariance is a natural restriction to impose, and even when additional unknown parameters are present, the resulting test can be more powerful than the corresponding noninvariant test. The specific case of detection in noise with unknown power is discussed in detail. A cyclic permutation invariance condition is imposed, and an invariant estimate of the noise power used to specify the test. The test which includes invariance is again shown to be more powerful than the conventional GLRT for the problem.

**Section 7.1: Summary of main topics**

The second application of invariance concerns invariance to data components contained in linear subspaces of the observation space. In Chapter 4 subspace invariant tests are formulated, and methods provided for calculating the resulting test statistics. This form of invariant testing is applied to two problems, namely that of detection in subspace interference, and detection in noise which is difficult to characterise.

For both these situations, methods are presented for estimating candidate invariance subspaces under the assumption that the covariance matrix is known. In the first instance a maximum likelihood formulation is proposed, where the invariance subspace serves to eliminate those components of the data which can be modelled least accurately. In the second instance, a method based on the Frobenius distance between covariances contained in certain subspaces is used, where the invariance subspace serves to ignore the dimensions where the conflict between the assumed model and the actual data is greatest. It is demonstrated that both of these methods lead to useful estimates.

In particular, simulated results are provided for the case of interference subspace identification. The estimates are shown to be accurate. However, on account of approximations that had to be made in the derivation of the estimates, they are not ideal because they are not consistent. Nevertheless, they work well in instances where the interference components are large. A demonstration of the use of an invariance subspace for reducing model mismatch is presented for a real-data application, and it is shown that the resulting estimate improves both the detectability and predictability of the hypothesis test.

Chapter 5 discusses the problem of estimating covariance matrices for use in invariant detectors. It is argued that when the covariance matrices are constrained, the estimates are required to be invariant to the same subspace as the detector. For the subspace invariance problem, an invariant maximum likelihood paradigm is proposed, where only the data contained in subspaces complementary to the invariance subspace are used in the estimation. Detailed methods are presented for calculating the covariance estimates under circulant, Toeplitz, doubly block circulant, doubly block Toeplitz, and ARMA constraints.

It is shown that when the covariance matrix is unconstrained, explicit invariance in the estimate is not required. When constraints are present, however, this is no longer true, since the data components in the invariance subspace corrupt the estimate in a way which is not ignored in the subsequent detection.

The problem of simultaneous covariance and invariance subspace estimation is then addressed, and methods provided for obtaining such estimates under the assumption that the noise largely obeys an ARMA model. Practical examples are provided demonstrating the use of these estimates for actual problems. Simulations are again provided showing the performance for the subspace interference interpretation, and the results shown to be good. The methods are also applied to actual data, both in the context of subspace interference and subspace invariance for model mismatch reduction. The use of an invariance subspace is shown to improve the predictability of the resulting detectors, and in some

cases to improve the detectability. Perhaps more importantly, explicit use of an invariance subspace can improve the ability of the detector to generalise beyond the set of training data.

The improved predictability is of particular importance in adaptive detection, where the noise estimates are used to yield approximately CFAR properties. The use of subspace invariance for this application is discussed in Chapter 6. The chapter opens by providing a decomposition of the subspace invariant testing problem into relevant subspaces, which has application in interpreting the roles of the various components of the observations. The AMF and the GLRT, which assume the presence of secondary noise-only observations, are then modified for the case of a known target with unknown amplitude, where the amplitude is restricted to be positive. This is a more natural formulation for many problems involving detection in real data, and the resulting tests are shown to be more powerful than those which consider the amplitude parameter completely unknown. The tests have a CFAR property, even when a known invariance subspace is included.

It is argued that for some applications it is reasonable to make a single invariance subspace estimate, and hold this fixed for the remainder of the adaptive testing. The invariance subspace then just serves to improve the modelling on the average. The modified AMF and GLRT formulations are discussed for the case where constraints in the covariance are enforced, and the special case is discussed where the same data are used for parameter estimation and detection. In these cases the tests are not strictly CFAR, but tend to become so as the parameter estimation improves.

The use of invariance to possible target presence is then proposed as a method of improving performance, or reducing computational requirements, particularly in the case of constrained covariances. It is shown that if target-invariant estimates are used in the tests, then the problem of invalid assumptions in the estimation is overcome. This should lead to improved detectability, expecially if the target amplitude is large.

Partially adaptive detectors are also proposed, where the test threshold is adapted but the basic statistic is held constant. These tests have an advantage in that a costly matrix inversion is eliminated. They also lead to CFAR tests in the unconstrained covariance case, and asymptotically CFAR tests otherwise.

The specific case of simultaneous estimation and detection from a single sample of data is discussed in detail, particularly with regard to the AMF, GLRT, and simplified AMF. A covariance constraint is essential in these cases, and the tests become CFAR as the parameter estimates improve. The AMF and modified AMF are also applied to a sample of real data, both in noninvariant and invariant forms. The invariant test formulations are shown to have improved predictability, which comes at the cost of slightly reduced detectability. In some cases this improved predictability can, however, lead to tests which perform better on the whole.

## 7.2 General conclusions

The results of Chapter 3 partially validate the use of the GLRT for detection problems where the location of the target is unknown. For some very simple situations, the performance of the GLRT is almost as good as the optimal test for the problem. It is only possible to compare the GLRT to the optimal test under a restricted set of circumstances, but these are sufficiently similar to more common situations to make the comparison useful as a baseline.

In practice, unknown target location problems arise frequently, and are usually dealt with in an ad hoc manner. By explicitly introducing an unknown location parameter, however, the assumed model represents the actual situation being modelled far more accurately. If the problem formulation then has a reasonable solution, it is of considerably more value than a solution which assumes the location known. Admittedly, complications arising from the introduction of such a parameter may greatly hinder the search for a solution. On the other hand, it forces the system designer to at least consider the effect that the unknown location actually has on the problem. In the literature, detectors designed for known target location problems are often applied in a sliding window framework, with little or no consideration given to associated problems which might arise.

More importantly, in this thesis, is the role that invariance can play in the problem. It is demonstrated that when a nuisance parameter can be eliminated though a legitimate invariance criterion, the resulting detection performance almost always improves. This stems directly from the fact that the parameter otherwise has to be estimated, which increases the uncertainty in the test statistic. However, invariant detectors introduce other difficulties, namely the need for invariant estimators. Under a maximum likelihood framework these detectors may be quite easy to specify, but their calculation may not be trivial.

The results of Chapters 4 and 5 address a related issue. In particular, the use of invariance to a certain subspace of the data is proposed. The first application is to the problem of detection in subspace interference. Under the condition that the interference lies in a known subspace but is otherwise unknown, an invariant detector can be derived which is optimal in a UMPI sense. In practice, however, the interference subspace is seldom known, and even the assumption that the components contained in this subspace are completely unknown is questionable. Nevertheless, the subspace interference assumption may still be adopted, albeit in a non-optimal context.

To this end, methods of estimating candidate interference subspaces are presented. These methods are more general than required, in that they do not take into account the fact that the interference components must always enter into the sample covariance through a positive definite matrix. Rather than considering this a shortcoming, however, it can be argued that it improves the applicability of the subspace invariance formalism to applications beyond those of strict subspace interference. The methods also do not require an assumption of multivariate normality of the covariance in the noise-

plus-interference process, which extends the applicability.

The fact that an exact maximum likelihood method has not been found for the interference subspace estimation problem is problematic, particularly when joint estimation of the subspace and the covariance matrix is required. Nevertheless, the approximate methods appear to work when applied to real-data problems. Indeed, the fact that they work so well is surprising, particularly when a subspace interference interpretation on the data is questionable. In short, a noise plus subspace interference model seems quite appropriate for modelling a wide variety of data.

The use of invariance to explicitly reduce model mismatch is unique, and could have application to problems far beyond those discussed in this thesis. In particular, apart from computational convenience there is no reason to restrict the notion to subspace invariance. By imposing equivalence classes on those portions of the data which are in conflict with some convenient modelling assumption, better performance may be achieved than if the mismatch is simply ignored. This improvement does however come at the risk of reduced detectability, if the application of invariance substantially eliminates part of the target.

The use of subspace invariance for model mismatch is quite tractable, on account of the simple relationships between MVN distributions and linear transformations. In the example given in Section 4.5.2 it is quite clear that the presence of such an invariance subspace leads to a considerably more predictable detector.

Once again the presence of invariance raises the need for invariant estimators. Conceptually there is nothing complicated about such detectors — they can simply be obtained by performing the estimation based on the maximal invariant statistic rather than on the original data. They may however be difficult to calculate. The complexity of the methods presented in Chapter 5 attribute to this fact.

The invariant detectors proposed are particularly useful in situations where computational efficiency is paramount. Using the methods discussed, a single simple detection statistic is obtained which is robust to certain expected deviations of the data from the assumed model. The resulting detection statistic can be calculated using a simple convolution product, if used in a sliding window context. Thus the need for adaptivity or preprocessing can sometimes be avoided. Avoiding adaptivity is desirable from a computational perspective. Avoiding preprocessing is desirable since it introduces a second stage of modelling and computation, which increases the uncertainty in the results.

The methods presented for simultaneous estimation of invariance subspaces and covariance matrices are quite limiting, in that explicit maximisation of the likelihood is indicated. Nevertheless, for many applications ARMA models are sufficiently rich to successfully model data, particularly when an invariance subspace is included to accommodate any mismatch. Being stationary, they can also successfully identify primary causes of nonstationarity in data, which to a large extent is what the invariance subspace is required for. Additionally, since the invariance subspace identification procedure

can be performed off-line, methods with high computational complexity are not inappropriate.

The Toeplitz and circulant constraints on the covariance are primarily of interest in applications where there is a large amount of data from which to make estimates. In many practical problems, this condition is not met. When detectors are made adaptive, the computational requirements can also be prohibitive, at least with computing technology where it now stands. Nonetheless, as far as multiple-observation detection is concerned, the use of constraints is important and can improve detectability substantially.

Chapter 6 outlines many of the uses of subspace invariance and invariant covariance estimation in a detection context. The opening analysis regarding the roles of the target, invariance, and noise-only subspaces is conceptually invaluable, particularly when detection and estimation are to be performed using the same data sample.

Subspace invariance can extend the application of multiple-observation detector formulations to cases where the observations do not come from identical processes, as long as the processes differ only in their distribution in a restricted linear subspace. Under both the AMF and GLRT formulations the invariant tests continue to be CFAR.

Also, subspace invariance can extend the lengths of data over which a stationarity assumption can be assumed valid, by ignoring the primary causes of nonstationarity. This is particularly important when tests are only asymptotically CFAR, with an intractable or unknown distribution for the test statistic. Parameter estimates should then be as accurate as possible for asymptotic results to be used effectively.

The improved modelling accuracy brought about by the inclusion of an invariance subspace translates directly to more predictable tests. That is, the actual attained false alarm rates correspond far more closely to the false alarm rates specified in the design. This is important in any application where a false alarm can be considered costly, perhaps in the form of a subsequent action which then has to be taken. In these contexts the reduced detectability may be fully justified.

The presence of an invariance subspace also partially eliminates the need for preprocessing of the data, particularly when highly-constrained models are considered. As demonstrated in the results, a method of preprocessing which is appropriate for one detector may not be appropriate for a different detector, even if the two are quite similar. Also, preprocessing is usually fairly ad hoc, with no systematic method available for designing a good solution. The invariant detectors, however, compensate for modelling inaccuracies in the calculation of the test statistic itself, where they can be analysed as part of the detector specification. In general, when it can be achieved, this provides a far better solution.

The analysis of the modified AMF, which does not involve adaptivity in the numerator term, provides insight into the price that has to be paid for adaptivity in a detector. In effect, every term in the test expression that has to be estimated contributes uncertainty into the resulting value. Thus tests should only be made as adaptive is necessary.

The subspace methods presented in this work do not provide a final solution to any given testing problem. Rather, they provide additional tools which may sometimes be used to design detectors for certain problems, which otherwise would be difficult to solve. In this sense, they just increase the set of possible detectors which may be considered for a problem. The methods cannot be applicable in all problems, however. It is up to the system designer to ascertain whether the methods have merit for any particular problem which is under investigation.

Finally, the methods presented in this work have been based on maximum likelihood principles primarily in the interests of rigour. The maximum likelihood framework is a good paradigm for investigating ideas, but the computational complexity can limit the application somewhat. When applied in real situations, it may often be expected that less complicated estimates will suffice for certain components of the detection problems.

# Bibliography


[1] A. A. ABU-DAYYA AND N. C. BEAULIEU. Outage probabilities in the presence of correlated lognormal interferers. *IEEE Transactions on Vehicular Technology*, 43(1):164–173, February 1994.

[2] H. AKAIKE. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60(2):255–265, 1973.

[3] C. F. ANSLEY. An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika*, 66(1):59–65, 1979.

[4] H. ANTON. *Elementary Linear Algebra*. John Wiley & Sons, sixth edition, 1991.

[5] P. M. BAGGENSTOSS AND S. M. KAY. An adaptive detector for deterministic signals in noise of unknown spectra using the rao test. *IEEE Transactions on Signal Processing*, 40(6):1460–1468, June 1992.

[6] A. BASILEVSKY. *Applied Matrix Algebra in the Statistical Sciences*. Elsevier Science Publishing Company, 1983.

[7] N. C. BEAULIEU, A. A. ABU-DAYYA, AND P. J. MCLANE. Estimating the distribution of a sum of independent lognormal random variables. *IEEE Transaction on Communications*, 43(12):2869–2873, December 1995.

[8] A. A. L. BEEX AND M. P. FARGUES. Highly parallel recursive/iterative Toeplitz eigenspace decomposition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1765–1768, November 1989.

[9] R. T. BEHRENS AND L. L. SCHARF. Signal processing applications of oblique projection operators. *IEEE Transactions on Signal Processing*, 42(6):1412–1424, June 1994.

[10] J. O. BERGER. *Statistical Decision Theory: Foundations, concepts, and methods*. Springer Series in Statistics. Springer-Verlag, 1980.

[11] C. M. BISHOP. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[12] S. BOSE AND A. O. STEINHARDT. A maximal invariant framework for adaptive detection with structured and unstructured covariance matrices. *IEEE Transactions on Signal Processing*, 43(9):2164–2175, September 1995.

[13] S. BOSE AND A. O. STEINHARDT. Adaptive array detection of uncertain rank one waveforms. *IEEE Transactions on Signal Processing*, 44(11):2801–2809, November 1996.

[14] D. E. BOWYER, P. K. RAJASEKARAN, AND W. W. GEBHART. Adaptive clutter filtering using autoregressive spectral estimation. *IEEE Transactions on Aerospace and Electronic Systems*, AES-15(4):538–546, July 1979.

[15] G. E. P. BOX, G. M. JENKINS, AND G. C. REINSEL. *Time Series Analysis: Forecasting and Control*. Prentice-Hall, third edition, 1994.

[16] L. E. BRENNAN AND I. S. REED. Theory of adaptive radar. *IEEE Transactions on Aerospace and Electronic Systems*, AES-9(2):237–252, March 1973.

[17] P. J. BROCKWELL AND R. A. DAVIS. *Time series: theory and methods*. Springer Series in Statistics. Springer-Verlag, second edition, 1991.

[18] P. J. BROCKWELL AND R. A. DAVIS. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics. Springer-Verlag, 1996.

[19] J. P. BURG, D. G. LUENBERGER, AND D. L. WENGER. Estimation of structured covariance matrices. *Proceedings of the IEEE*, 70(9):963–974, September 1982.

[20] K. A. BURGESS AND B. D. VAN VEEN. Subspace-based adaptive generalized likelihood ratio detection. *IEEE Transactions on Signal Processing*, 44(4):912–927, April 1996.

[21] D. BURSHTEIN. An efficient algorithm for calculating the likelihood and likelihood gradient of ARMA models. *IEEE Transactions on Automatic Control*, 38(2):336–340, February 1993.

[22] J. Y. CHEN AND I. S. REED. A detection algorithm for optical targets in clutter. *IEEE Transactions on Aerospace and Electronic Systems*, AES-23(1):46–59, January 1987.

[23] G. CORTELAZZO, G. A. MIAN, AND R. RINALDO. Toeplitz properties of the block matrices encountered in the processing of spatiotemporal signals. *IEEE Transactions on Signal Processing*, 39(7):1672–1674, July 1991.

[24] P. J. DAVIS. *Circulant Matrices*. Pure & Applied Mathematics: A Wiley-Interscience Series of Texts, Monographs & Tracts. John Wiley & Sons, 1979.

[25] M. H. DEGROOT. *Probability and Statistics*. Addison-Wesley Series in Statistics. Addison-Wesley, second edition, 1986.

[26] A. S. DEIF. *Advanced Matrix Theory for Scientists and Engineers*. John Wiley & Sons, 1982.

[27] A. DEMBO, C. L. MALLOWS, AND L. A. SHEPP. Embedding nonnegative definite Toeplitz matrices in nonnegative definite circulant matrices, with application to covariance estimation. *IEEE Transactions on Information Theory*, 35(6):1206–1212, November 1989.

[28] A. P. DEMPSTER, N. M. LAIRD, AND D. B. RUBIN. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: B*, 39:1–38, 1977.

[29] G. M. DILLARD. Mean-level detection of nonfluctuating signals. *IEEE Transactions on Aerospace and Electronic Systems*, AES-10(6):795–799, November 1984.

[30] M. L. EATON. *Multivariate Statistics: A Vector Space Approach*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1983.

[31] M. H. EL AYADI. Generalised likelihood adaptive detection of signals deformed by unknown linear filtering in noise with slowly fluctuating power. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(2):401–405, April 1985.

[32] M. H. EL AYADI. NAR estimators of spatial covariance matrices for adaptive array detection. *IEEE Transactions on Signal Processing*, 39(7):1682–1686, July 1991.

[33] M. H. EL AYADI AND B. PICINBONO. NAR AGC adaptive detection of nonoverlapping signals in noise with fluctuating power. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(5):952–963, October 1981.

[34] T. S. FERGUSON. *Mathematical Statistics: A decision theoretic approach*. Probability in Mathematical statistics. Academic Press, 1967.

[35] B. FRIEDLANDER AND B. PORAT. Performance analysis of transient detectors based on a class of linear data transforms. *IEEE Transactions on Information Theory*, 38(2):665–673, March 1992.

[36] M. FRISCH AND H. MESSER. The use of the wavelet transform in the detection of an unknown transient signal. *IEEE Transactions on Information Theory*, 38(2):892–897, March 1992.

[37] M. FRISCH AND H. MESSER. Transient signal detection using prior information in the likelihood ratio test. *IEEE Transactions on Signal Processing*, 41(6):2177–2192, June 1993.

[38] M. FRISCH AND H. MESSER. Detection of a known transient signal of unknown scaling and arrival time. *IEEE Transactions on Signal Processing*, 42(7):1859–1863, July 1994.

[39] D. R. FUHRMANN. Application of Toeplitz covariance estimation to adaptive beamforming and detection. *IEEE Transactions on Signal Processing*, 39(10):2194–2198, October 1991.

[40] K. FUKUNAGA. *Introduction to Statistical Pattern Recognition*. Computer Science and Scientific Computing. Academic Press, second edition, 1990.

[41] W. A. FULLER. *Introduction to Statistical Time Series*. Wiley Series in Probability and Statistics. John Wiley & Sons, second edition, 1996.

[42] W. A. GARDNER AND C. M. SPOONER. The cumulant theory of cyclostationary time-series, part I: foundations. *IEEE Transactions on Signal Processing*, 42(12):3387–3408, December 1994.

[43] N. GASTINEL. *Linear Numerical Analysis*. Hermann, 1970. Translated from the original French text *Analyse numérique linéaire*, first published by Hermann in 1966.

[44] Y.-L. GAU AND I. S. REED. An improved reduced-rank CFAR space-time adaptive radar detection algorithm. *IEEE Transactions on Signal Processing*, 46(8):2139–2146, August 1998.

[45] N. C. GEÇKINLI AND D. YAVUZ. *Discrete Fourier Transformation and Its Applications to Power Spectra Estimation*, volume 8 of *Studies in Electrical and Electronic Engineering*. Elsevier, 1983.

[46] D. F. GINGRAS AND E. MASRY. Autoregressive spectral estimation in additive noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(4):490–501, April 1988.

[47] G. H. GOLUB AND C. F. VAN LOAN. *Matrix Computations*. Johns Hopkins University Press, 1983.

[48] G. H. GOLUB AND C. F. VAN LOAN. *Matrix Computations*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, third edition, 1996.

[49] A. GRAHAM. *Kronecker Products and Matrix Calculus*. Ellis Horwood Series in Mathematics and its Applications. Halsted Press, 1981.

[50] R. M. GRAY. On the asymptotic eigenvalue distribution of Toeplitz matrices. *IEEE Transactions on Information Theory*, IT-18(6):725–730, November 1972.

[51] U. GRENANDER AND G. SZEGÖ. *Toeplitz Forms and their Applications*. University of California Press, 1958.

[52] M. GUIDA, M. LONGO, AND M. LOPS. Biparametric CFAR procedures for lognormal clutter. *IEEE Transactions on Aerospace and Electronic Systems*, 29(3):798–809, July 1993.

[53] D. V. GUPTA, J. F. VETELINO, T. J. CURRY, AND J. T. FRANCIS. An adaptive threshold system for nonstationary noise backgrounds. *IEEE Transactions on Aerospace and Electronic Systems*, AES-13(1):11–16, January 1977.

[54] M. A. HAMDAN. The logarithm of the sum of two correlated log-normal variates. *Journal of the American Statistical Association*, 66(333):105–106, March 1971.

[55] M. HAMERMESH. *Group Theory and its Application to Physical Problems*. Addison-Wesley, 1962.

[56] M. H. HAYES. The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-30(2):140–154, April 1982.

[57] S. S. HAYKIN, editor. *Array Signal Processing*. Prentice-Hall, 1985.

[58] S. S. HAYKIN. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, second edition, 1999.

[59] C. W. HELSTROM. *Statistical Theory of Signal Detection*. Pergamon Press, 1960.

[60] A. HERO. Hilights of statistical signal and array processing. *IEEE Signal Processing Magazine*, pages 21–64, September 1998.

[61] S. D. HIMONAS AND M. BARKAT. Adaptive CFAR detection in partially correlated clutter. *IEE Proceedings, Pt F*, 137(5):387–394, October 1990.

[62] R. V. HOGG AND A. T. CRAIG. *Introduction to Mathematical Statistics*. Macmillan, third edition, 1970.

[63] R. A. HORN AND C. R. JOHNSON. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

[64] D. HUANG. Symmetric solutions and eigenvalue problems of Toeplitz systems. *IEEE Transactions on Signal Processing*, 40(12):3069–3074, December 1992.

[65] A. K. JAIN. *Fundamentals of Digital Image Processing*. Prentice-Hall, 1989.

[66] W. A. JANOS. Tail of the distribution of sums of log-normal variates. *IEEE Transactions on Information Theory*, IT-16(3):299–302, May 1970.

[67] W. B. JONES AND O. NJØASTAD. Applications of Szegö polymonials to digital signal processing. *Rocky Mountain Journal of Mathematics*, 21(1):387–436, 1991.

[68] W. B. JONES, O. NJØASTAD, AND E. B. SAFF. Szegö polynomials associated with Wiener-Levinson filters. *Journal of Computational and Applied Mathematics*, 32:387–406, 1990.

[69] G. G. JUDGE, W. E. GRIFFITHS, R. CARTER HILL, H. LÜTKEPOHL, AND T.-C. LEE. *The Theory and Practice of Econometrics*. Wiley Series in Probability and Statistics. John Wiley & Sons, second edition, 1985.

[70] J. H. JUSTICE. An algorithm for inverting positive definite Toeplitz matrices. *SIAM Journal on Applied Mathematics*, 23(3):289–291, November 1972.

[71] J. H. JUSTICE. A Levinson-type algorithm for two-dimensional Weiner filtering using bivariate Szegö polynomials. *Proceedings of the IEEE*, 65(6):882–886, June 1977.

[72] S. A. KASSAM AND H. V. POOR. Robust techniques for signal processing: a survey. *Proceedings of the IEEE*, 73(3):433–481, March 1985.

[73] S. M. KAY. Noise compensation for autoregressive spectral estimates. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28(3):292–303, June 1980.

[74] S. M. KAY. Asymptotically optimal detection in unknown colored noise via autoregressive modeling. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-31(4):927–940, August 1983.

[75] S. M. KAY. *Modern Spectral Estimation: Theory and Applications*. Prentice-Hall, 1988.

[76] S. M. KAY. Asymptotically optimal detection in incompletely characterised non-Gaussian noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(5):627–633, May 1989.

[77] S. M. KAY. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall Signal Processing Series. Prentice-Hall, 1993.

[78] S. M. KAY AND S. L. MARPLE, JR. Spectrum analysis — a modern perspective. *Proceedings of the IEEE*, 69(11):1380–1419, November 1981.

[79] D. KAZAKOS AND P. PAPANTONI-KAZAKOS. *Detection and Estimation*. Computer Science Press, 1990.

[80] E. J. KELLY. An adaptive detection algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, AES-22(1):115–127, March 1986.

[81] M. KENDALL AND A. STUART. *The Advanced Theory of Statistics*, volume 2. Charles Griffin & Company, fourth edition, 1979.

[82] C. G. KHATRI AND C. RADHAKRISHNA RAO. Effects of estimated noise covariance matrix in optimal signal detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-35(5):671–679, May 1987.

[83] I. P. KIRSTEINS AND D. W. TUFTS. Adaptive detection using low rank approximation to a data matrix. *IEEE Transactions on Aerospace and Electronic Systems*, 30(1):55–67, January 1994.

[84] H. KRIM AND M. VIBERG. Two decades of array signal processing research. *IEEE Signal Processing Magazine*, pages 67–94, July 1996.

[85] P. LANCASTER. *Theory of Matrices*. Academic Press, 1969.

[86] H. J. LARSON. *Introduction to Probability Theory and Statistical Inference*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, third edition, 1982.

[87] E. L. LEHMANN. *Testing Statistical Hypotheses*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, second edition, 1986.

[88] R. LENZ. *Group Theoretical Methods in Image Processing*, volume 413 of *Lecture Notes in Computer Science*. Springer-Verlag, 1990.

[89] R. J. A. LITTLE AND D. B. RUBIN. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1987.

[90] J. MAKHOUL. Linear prediction: a tutorial review. *Proceedings of the IEEE*, 63(4):561–580, April 1975.

[91] J. MAKHOUL. On the eigenvectors of symmetric Toeplitz matrices. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-29(4):868–872, August 1981.

[92] A. MARGALIT, I. S. REED, AND R. M. GAGLIARDI. Adaptive optical target detection using correlated images. *IEEE Transactions on Aerospace and Electronic Systems*, AES-21(3):394–405, May 1985.

[93] G. J. MCLACHLAN. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1992.

[94] G. MÉLARD. A fast algorithm for the exact likelihood of autoregressive-moving average models. *Applied Statistics*, 33(1):104–114, 1984.

[95] M. I. MILLER, D. R. FUHRMANN, J. A. O'SULLIVAN, AND D. L. SNYDER. Maximum-likelihood methods for Toeplitz covariance estimation and radar imaging. In S. HAYKIN, editor, *Advances in Spectrum Analysis and Array Processing*, volume II, pages 145–172. Prentice-Hall, 1991.

[96] M. I. MILLER AND D. L. SNYDER. The role of likelihood and entropy in incomplete-data problems: applications to estimating point-process intensities and Toeplitz constrained covariances. *Proceedings of the IEEE*, 75(7):892–907, July 1987.

[97] T. K. MOON. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, pages 47–60, November 1996.

[98] D. F. MORRISON. *Multivariate Statistical Methods*. McGraw-Hill Series in Probability and Statistics. McGraw-Hill, 1967.

[99] R. J. MUIRHEAD. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1982.

[100] R. NITZBERG. Constant-false-alarm-rate processors for locally nonstationary clutter. *IEEE Transactions on Aerospace and Electronic Systems*, AES-9(3):399–405, May 1973.

[101] F. NOOR AND S. D. MORGERA. Recursive and iterative algorithms for computing eigenvalues of hermitian Toeplitz matrices. *IEEE Transactions on Signal Processing*, 41(3):1272–1280, March 1993.

[102] A. V. OPPENHEIM AND R. W. SCHAFER. *Digital Signal Processing*. Prentice-Hall, 1975.

[103] A. PAPOULIS. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill series in System Science. McGraw-Hill, 1965.

[104] B. PORAT. *Digital Processing of Random Signals: theory and methods*. Prentice Hall Information and System Sciences Series. Prentice-Hall, 1994.

[105] B. PORAT AND B. FRIEDLANDER. Adaptive detection of transient signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(6):1410–1418, December 1986.

[106] B. PORAT AND B. FRIEDLANDER. Performance analysis of a class of transient detection algorithms — a unified framework. *IEEE Transactions on Signal Processing*, 40(10):2536–2546, October 1992.

[107] B. PORAT AND B. FRIEDLANDER. On the generalised likelihood ratio test for a class of nonlinear detection problems. *IEEE Transactions on Signal Processing*, 41(11):3186–3190, November 1993.

[108] W. H. PRESS, B. P. FLANNERY, S. A. TEUKOLSKY, AND W. T. VETTERLING. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 1988.

[109] R. S. RAGHAVAN, N. PULSONE, AND D. J. MCLAUGHLIN. Performance of the GLRT for adaptive vector subspace detection. *IEEE Transactions on Aerospace and Electronic Systems*, 32(4):1473–1487, October 1996.

[110] R. S. RAGHAVAN, H. F. QIU, AND D. J. MCLAUGHLIN. CFAR detection in clutter with unknown correlation properties. *IEEE Transactions on Aerospace and Electronic Systems*, 31(2):647–657, April 1995.

[111] I. S. REED, J. D. MALLETT, AND L. E. BRENNAN. Rapid convergence rate in adaptive arrays. *IEEE Transactions on Aerospace and Electronic Systems*, AES-10(6):853–863, November 1974.

[112] I. S. REED AND X. YU. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(10):1760–1770, October 1990.

[113] I. B. RHODES. A tutorial introduction to estimation and filtering. *IEEE Transactions on Automatic Control*, AC-16(6):688–706, December 1971.

[114] J. T. RICKARD AND G. M. DILLARD. Adaptive detection algorithms for multiple-target situations. *IEEE Transactions on Aerospace and Electronic Systems*, AES-13(4):338–343, July 1977.

[115] B. D. RIPLEY. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[116] F. C. ROBEY, D. R. FUHRMANN, E. J. KELLY, AND R. NITZBERG. A CFAR adaptive matched filter detector. *IEEE Transactions on Aerospace and Electronic Systems*, 28(1):208–216, January 1992.

[117] E. A. ROBINSON. A historical perspective of spectrum estimation. *Proceedings of the IEEE*, 70(9):885–907, September 1982.

[118] L. L. SCHARF AND B. FRIEDLANDER. Corrections to "Matched subspace detectors". *IEEE Transactions on Signal Processing*, 45(6):1669, June 1997.

[119] L. L. SCHARF. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Addison-Weseley Series in Electrical and Computer Engineering: Digital Signal Processing. Addison-Wesley, 1991.

[120] L. L. SCHARF AND B. FRIEDLANDER. Matched subspace detectors. *IEEE Transactions on Signal Processing*, 42(8):2146–2157, 1994.

[121] L. L. SCHARF AND D. W. LYTLE. Signal detection in gaussian noise of unknown level: an invariance application. *IEEE Transactions on Information Theory*, IT-17(4):404–411, July 1971.

[122] S. C. SCHWARTZ AND Y. S. YEH. On the distribution function and moments of power sums with log-normal components. *Bell Systems Technical Journal*, 61(7):1441–1462, September 1982.

[123] S. R. SEARLE. *Matrix Algebra useful for Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, 1982.

[124] D. SENGUPTA AND S. M. KAY. Parameter estimation and GLRT detection in colored non-gaussian autoregressive processes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(10):1661–1676, October 1990.

[125] D. J. THOMSON. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70(9):1055–1096, September 1982.

[126] H. L. VAN TREES. *Detection, Estimation, and Modulation Theory (Part I)*. John Wiley & Sons, 1968.

[127] P. A. VOOIS. A theorem on the asymptotic eigenvalue distribution of Toeplitz-block-Toeplitz matrices. *IEEE Transactions on Signal Processing*, 44(7):1837–1841, July 1996.

[128] G. R. WALSH. *Methods of Optimisation*. John Wiley & Sons, 1975.

[129] D. S. WATKINS. *Fundamentals of Matrix Computations*. John Wiley & Sons, 1991.

[130] M. WAX AND T. KAILATH. Efficient inversion of Toeplitz-block Toeplitz matrices. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-31(5):1218–1221, October 1983.

[131] H. WIDOM. *Studies in Real and Complex Analysis*, volume 3, chapter Toeplitz Matrices, pages 179–209. The Mathematical Association of America, 1965.

[132] C. F. J. WU. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.

[133] G. K. YEUNG AND W. A. GARDNER. Search-efficient methods of detection of cyclostationary signals. *IEEE Transactions on Signal Processing*, 44(5):1214–1223, May 1996.

[134] O. ZEITOUNI, J. ZIV, AND N. MERHAV. When is the generalized likelihood ratio test optimal? *IEEE Transactions on Information Theory*, 38(5):1597–1602, September 1992.

[135] S. ZOHAR. Toeplitz matrix inversion: The algorithm of W. F. Trench. *Journal of the Association for Computing Machinery*, 16(4):592–601, October 1969.