May 1996 LIDS-TH-2332

Research Supported By:

Army Research Office
ARO DAAL03-92-G-0115

Office of Naval Research
Grant ONR N00014-91-J-1004

Advanced Research Project Agency
Grant ARPA F49620-93-1-0604

National Defense Science and Engineering
Graduate Fellowship

MULTISCALE HYPOTHESIS TESTING WITH APPLICATION TO ANOMALY CHARACTERIZATION FROM TOMOGRAPHIC PROJECTIONS

Austin B. Frakt

May 1996

LIDS-TH-2332

Sponsor Acknowledgments

Army Research Office
ARO DAAL03-92-G-0115

Office of Naval Research
Grant ONR N00014-91-J-1004

Advanced Research Project Agency
Grant ARPA F49620-93-1-0604

National Defense Science and Engineering
Graduate Fellowship

MULTISCALE HYPOTHESIS TESTING WITH APPLICATION TO ANOMALY CHARACTERIZATION FROM TOMOGRAPHIC PROJECTIONS

Austin B. Frakt

This report is based on the unaltered thesis of Austin B. Frakt submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Master of Science at the Massachusetts Institute of Technology in May 1996.

This research was conducted at the M.I.T. Laboratory for Information and Decision Systems with research support gratefully acknowledged by the above mentioned sponsors.

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Multiscale Hypothesis Testing with Application to Anomaly Characterization from Tomographic Projections

by

Austin B. Frakt

B.S., Applied and Engineering Physics
Cornell University, 1994

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1996

Author ....
Department of Electrical Engineering and Computer Science
May 10, 1996

Certified by ....
Alan S. Willsky
Professor of Electrical Engineering
Thesis Supervisor

Certified by ....
W. Clem Karl
Research Affiliate
Thesis Supervisor

Accepted by ....
Frederick R. Morgenthaler
Chairman
Departmental Committee on Graduate Students

Multiscale Hypothesis Testing with Application to Anomaly Characterization from Tomographic Projections

by

Austin B. Frakt

Submitted to the Department of Electrical Engineering and Computer Science
on May 10, 1996, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering

Abstract

A common objective in many applied problems is to infer properties of the interior of an object based on tomographic (line-integral) projections. In a number of applications the ultimate goal is to characterize (e.g., detect, locate) regions of the interior which are, in some sense, anomalous. A major challenge is to develop methods which can characterize anomalies directly in the data domain (i.e., without image reconstruction). In this thesis we develop data domain techniques for the detection and localization of a single anomaly from tomographic projections. These techniques are based upon a multiscale hypothesis test (MSHT) framework. A MSHT represents an efficient alternative to a very large conventional hypothesis test which may be computationally infeasible due to the overwhelming number of hypotheses which must be considered. Previous application of MSHTs to anomaly localization problems has focussed on the intuitive idea of spatial zooming with natural statistics [19–21]. A major contribution of this thesis is the broader interpretation of multiscale hypothesis testing as statistical zooming on the set of hypotheses rather than spatial zooming in the image domain. This broader interpretation leads naturally to the formulation of an optimization problem, the solution of which provides a MSHT statistic which yields improved performance.

Thesis Supervisor: Alan S. Willsky
Title: Professor of Electrical Engineering

Thesis Supervisor: W. Clem Karl
Title: Research Affiliate

Acknowledgments

The MIT graduate experience is a seemingly infinite sequence of hurdles. These hurdles are made surmountable only with the support and guidance of advisors, friends, and family. I first wish to express my gratitude to my thesis advisors: Professors Alan S. Willsky and W. Clem Karl. I find Alan's intellectual energy and dedication absolutely astounding. In addition to his remarkable skills as an advisor/mentor, he creates a vibrant and supportive research environment in which it is easy to excel. No less astounding is Clem's ability to find time in his busy schedule to maintain a marginally stable computer network, sit in on grouplets, and provide one-on-one assistance to many SSG group members. Thank you Clem for your invaluable insight.

I also appreciate the interaction I've had with members of the SSG and other MIT friends. I especially thank: Dr. Hamid Krim for the countless times he's patiently listened to me struggle to express a new idea or approach; Dr. Paul Fieguth for knowing a bit about everything and for being the group martyr and chef; Dr. Bill Irving, for his ability and willingness to explain complex ideas simply; Mike Daniel for sharing his insight and technical knowledge; Charlie Fosgate for listening to my gripes and taking my displaced abuse; Mike Branicky, Patrick Maurer, Sekhar Tatikonda, Mike Schneider and Stark Draper for helping me prepare for the OQE; the rest of the SSG for their camaraderie and for providing the benefit of their experience.

I wish to recognize MIT's Electrical Engineering and Computer Science Department and the Advanced Research Project Agency (ARPA) for their financial support of my graduate studies. The former provided a full fellowship for my first year funded by generous alumni donations. The latter continues to provide full support through a National Defense Science and Engineering Graduate Fellowship.

Finally, I thank the members of my family for twenty-four years of unwavering trust, emotional support, financial assistance, and unconditional love. Though you may understand little of it, I certainly could not have written this thesis without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

A common objective in many applied problems is to infer properties of the interior of an object based on measurements obtained at the exterior. When these measurements take the form of path-integrals, they are said to be *tomographic* and the problem falls under the purview of *tomography.* Computed tomography (CT) is concerned with the study of the interior of objects based on line-integral projections while in diffraction tomography (DT) the integrals need not be over straight lines. In tomographic problems, a datum is typically obtained by probing the object with a source of energy and measuring the transmitted energy with some detector. Data from different source-detector positions are then combined in the analysis of the interior.

For some problems, tomographic data are used to obtain a detailed image of the cross section of an object. This pixel-by-pixel reconstruction is then used as input to a subsequent processing stage or directly viewed for interpretation. A familiar example is the reconstruction of a cross section of the body from CT scan data for the purposes of medical diagnosis. The problem of image reconstruction from projections arises in a variety of fields other than medicine including geophysical exploration [19–21], astronomy [6], non-destructive testing [1,11], and others [8,14].

The reconstruction problem has received significant attention and, as a result, several

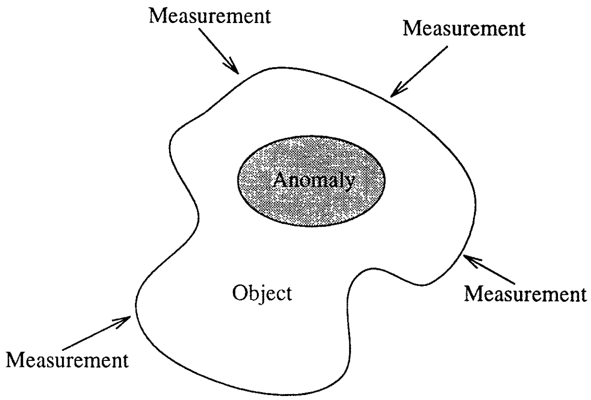


Figure 1-1: This figure illustrates the general applied problem of inferring properties of the interior of some object (which may contain an anomalous region) based on measurements obtained at the exterior.

standard general purpose methods exist for its solution (e.g., convolution back-projection (CBP)) [8,13–17]. A well known drawback to these standard techniques is that they rely crucially on the availability of a full set of low noise data for artifact free reconstructions. In many applications, however, only limited or noisy data are available due to one or several factors: object geometry, physical interference of other objects, time constraints, budget constraints, safety considerations, modeling errors, etc. [30, 31]. In limited data or high noise cases, the reconstruction problem is ill-posed and, unless appropriately regularized, any reconstruction obtained suffers from streaking and other artifacts [7,20,31,33].

In a number of applications the ultimate goal is *not* to reconstruct a cross sectional image. Often the primary objective is to characterize (e.g., detect, locate) regions of the interior which are, in some sense, anomalous. Anomaly characterization is of interest in a range of applications such as medicine (tumor detection) [31], electro-geophysical exploration (conductivity inhomogeneity localization) [19–21], the non-destructive evaluation of industrial parts and machinery like aging aircraft (corrosion detection) [1, 11], oceanography (ocean-acoustic tomography), and spot-light mode synthetic aperture radar (automatic target recognition) [22].

One general approach to anomaly characterization problems begins with reconstructing a cross sectional image [7,20,31,33]. Anomalous regions are then analyzed by post-processing the reconstructed image. This reconstruct and post-process approach is rarely the best way to attack the problem. First, it does not necessarily make optimal (or even good suboptimal) use of the data. Second, as mentioned, in limited data or high noise cases (which arise frequently in practice), reconstruction introduces artifacts which are, by definition, anomalous. In this way, image reconstruction can make the problem of finding the true anomalies more difficult. Finally, reconstruction is a computationally non-trivial task and, in light of the above, is a waste of computational resources when the ultimate goal is the far more modest one of anomaly characterization.

A major challenge, therefore, is to develop methods which can characterize anomalies directly in the data domain (i.e., without image reconstruction). This challenge consists of at least five sub-problems:

1. The anomaly detection problem: Are anomalies present?

2. The anomaly enumeration problem: How many are there?

3. The anomaly localization problem: Where are they in space and scale?

4. The anomaly shape problem: What are their shapes?

5. The anomaly intensity problem: What do they look like?

In this thesis we focus almost exclusively on problems involving detection and localization of a single anomaly. In the single anomaly case, the detection problem is equivalent to the enumeration problem. Given answers to the first four questions, the last one is the problem of local reconstruction which has received a good deal of attention in recent years [23–25,40,43] but will not be addressed in this thesis. Our algorithms may be used, however, as a pre-processing stage to cue a local reconstruction routine as to which areas of the cross section to reconstruct.

In most data domain approaches to anomaly characterization problems it is assumed that the anomaly belongs to a class of objects which are parameterized by one or several parameters [7,31]. These parameters are then estimated based on noisy observations of tomographic projections. Recently, several authors have introduced *non-parametric* approaches to the problem based on hypothesis testing [2, 19–21]. In this thesis we build on this hypothesis testing approach and, in doing so, explore issues which have relevance to a broad range of problems, not just tomographic.

Our focus is on the design of multiscale hypothesis tests (MSHTs) with tomographic anomaly characterization as one particular application. A MSHT attempts to offer an efficient alternative to a very large conventional hypothesis test, which may be computationally infeasible due to the overwhelming number of hypotheses which must be considered. Instead of selecting a single hypothesis on the basis of one large M-ary hypothesis test, the MSHT philosophy is to zoom in on the true hypothesis via a scale-recursive sequence of smaller composite hypothesis tests. Each test in the sequence consists of some decision statistics and several subsets of the global set of hypotheses. Often no optimal (uniformly most powerful (UMP)) tests exist for the composite hypothesis problems comprising a MSHT. In this thesis we introduce criteria and methods for selecting subsets and statistics when no UMP tests exist and apply these methods to the anomaly characterization problem with the aim of obtaining performance near that of the optimal but infeasible M-ary hypothesis test.

Recent work in multiscale hypothesis testing methods for anomaly localization using tomographic data has focussed on the idea of *spatial zooming* using intuitively natural statistics to form decisions as to which regions to zoom in on. In spatial zooming, the anomaly is localized first to a coarse spatial scale (with large area) and then to successively finer scale regions (of smaller area). The areas to which the anomaly is sequentially localized form a nested sequence of regions in the image domain. In this thesis we also consider spatial zooming approaches to the anomaly localization problem but we emphasize that, in general, the scale recursive nature of a multiscale hypothesis test need not have an interpretation as a multiscale search in a spatial sense. It is more appropriately defined on the domain of hypotheses. Rather than zooming in on the anomaly through a nested sequence of spatial regions, we view the MSHT framework as providing a means of zooming in on the true hypothesis through a nested sequence of composite hypotheses. The interpretation of multiscale hypothesis testing as statistical rather than spatial zooming represents a significant generalization of recent work and leads naturally to the consideration of composite hypothesis test statistics with improved performance.

1.2 Previous Work

In the last decade and a half, significant progress has been made in data-domain methods for anomaly characterization. In [30,31], Rossi and Willsky consider the problem of estimating the parameters (e.g., location, shape, and size) of an object superimposed on a *known deterministic* background field using noisy and limited angle CT data. The problem of estimating just the location of an otherwise known object is considered and it is shown that nonlinear maximum likelihood (ML) estimation of the location involves a convolution back-projection operation (CBP) operation where the ML convolution kernel is *not* the standard CBP kernel. Thus it is explicitly seen that performing CBP as a first step in object location would truly be a waste of computational resources.

The estimation-theoretic approach of Rossi and Willsky is extended by Devaney et al. in [10,34,37] to the case of diffraction tomography (DT). And in [7], Bressler et al. consider the estimation of the parameters of an unknown number of objects in a three-dimensional volume from incomplete and noisy CT measurements. The *maximum a posteriori* (MAP) estimate rather than the ML estimate is sought. An algorithm is provided which computes an approximation to the MAP estimate through a sequence of stages. At the first stage an ML estimate of object primitives is obtained. Then feasible objects are constructed by combining primitives using a sequential hypothesis testing scheme. A hypothesis test is performed to choose which combination of the feasible objects is most likely. Finally, object estimates are obtained by smoothing the objects selected in the hypothesis test.

Several other authors have also developed methods for object estimation from noisy and/or incomplete CT data. In [28], Prince computes the MAP estimate of the sinogram (the image of the projection domain information) and then reconstructs an image domain estimate using CBP. In forming the estimate, Prince relies upon a Markov random field prior model for the sinogram and consistency conditions between the object field and its sinogram. In [18], Milanfar exploits the relationship between the moments of an image to those of its sinogram to estimate object parameters from CT data. One advantage of this approach is that it admits estimation of a much less restrictive class of objects than that considered by Rossi and Willsky in [30,31].

In all the aforementioned work on object parameter estimation from CT and DT data it is assumed that the background upon which the object is superimposed is deterministic and known. The problem of data-domain characterization of anomalies which are superimposed on an *unknown* but well modeled random field background has been considered only more recently. In [2], Bhatia introduces a pixel-based method for the detection and localization of a single anomaly superimposed on a fractal field background. The wavelet transform of the CT data associated with each pixel is considered independently and a fixed number of candidate pixels which seem most anomalous are selected on the basis of an approximate chi-square test. From these candidate pixels, a subset are chosen by amplitude thresholding. The chosen pixels represent an estimate of the anomaly's support.

In [19–21], Miller and Willsky consider the problem of detecting and localizing multiple anomalies which are superimposed on a fractal random field background based on DT data. They propose a scale-recursive algorithm for which, at each scale, a composite hypothesis test is conducted using a generalized likelihood ratio test. These composite hypothesis tests are designed to zoom in on the anomalies through a scale recursive search in the image (i.e., spatial) domain. Anomalies are first localized to coarse scale regions and then to successively finer scale regions. This method represents a suboptimal but efficient alternative to implementing the computationally daunting optimal hypothesis test which would include a hypothesis for each combination of anomaly locations and sizes. Additionally, analysis techniques based upon a binary hypothesis testing framework are introduced which provide insight into performance limits of the detection and localization algorithm.

1.3 Contributions

This thesis presents the following main contributions:

1. In Chapter 2 we define the general structure of a multiscale hypothesis test. While use of such a hierarchy of composite hypothesis tests is not novel, abstracting the structure from a particular application is a valuable exercise. The insight which this abstraction provides motivates our development of optimized MSHTs in Chapter 5.

2. In Chapter 3 we investigate the role of the background field statistics in the ultimate performance limits of a MSHT approach to the anomaly detection and localization problems. With a simple one-dimensional example, we show that these performance limits rely on the background covariance in an exceedingly complex and seemingly intractable way.

3. In Chapter 4 we apply methods similar to those in [19–21] to the CT anomaly detection and localization problems. Two algorithms are introduced, one which tests for the presence of an anomaly at a coarse scale and another which does so at a finer scale. We show that delaying this crucial decision to a finer scale results in an improvement in detection performance.

4. As forshadowed in Chapter 3, the performance of the algorithms presented in Chapter 4 depends upon the background covariance. In Chapter 5 we present another way of investigating this dependence which motivates the development of a MSHT statistic optimality condition. We propose several ways of solving (or approximately solving) the resulting optimization problem. Finally, we apply optimized statistics so found to the anomaly detection and localization problems.

1.4 Organization

This thesis has two main themes. They are: (1) the generalization and investigation of multiscale hypothesis testing and (2) the application of multiscale hypothesis testing methods to the tomographic anomaly detection and localization problems. These themes are treated, to some degree, in parallel as they are both developed in each chapter.

We begin our development of the two main themes in Chapter 2 which covers the relevant mathematical background. The mathematics of tomography are developed from the mathematics of the Radon transform and its inverse. Our approach to the anomaly detection and localization problems relies on hypothesis testing; therefore, much of Chapter 2 is devoted to elements of detection theory. Included in Chapter 2 are discussions of composite hypothesis testing and multiscale hypothesis testing. We conclude the chapter with a precise problem statement including a summary of all modeling assumptions.

In Chapter 3 we introduce a binary hypothesis testing framework which admits the computation of a detection performance bound and the investigation of anomaly ambiguity. This framework provides insight into the structures of the detection and localization problems and our results indicate the feasibility of a spatial zooming approach to anomaly localization. A one-dimensional problem is discussed to provide insight into the nature of these results.

Several detection and localization algorithm are discussed in Chapter 4. These algorithms are multiscale hypothesis tests which also have a spatial zooming interpretation. The structure of these tests is motivated by intuition and not by any criterion of optimality. We show that these ad hoc methods are not suitable for a certain class of problems. Chapter 4 concludes with a comparison of the computational complexities of the optimal anomaly detection/localization algorithm and our multiscale methods.

In Chapter 5 we emphasize our broader interpretation of multiscale hypothesis testing as statistical rather than spatial zooming. Criteria for good MSHTs are introduced. These criteria lead to the formulation of a non-linear optimization problem. We show how to solve this non-linear problem exactly and also propose a closely related linear programming problem. The solution of the non-linear optimization problem (or its linear programming approximation) is an optimized MSHT statistic which yields better detection performance than the ad hoc statistics of Chapter 4.

Concluding comments and directions for future research are provided in Chapter 6.

Chapter 2

Preliminaries

The main focus of this chapter is the presentation of the mathematical concepts upon which our anomaly characterization work is based. Clearly the mathematics of tomography play a central role; a discussion of tomography and reconstruction methods is found in Section 2.2. We preface the discussion of tomographic projections and reconstruction methods with a review of their idealizations—the Radon transform and its inverse (Section 2.1). Our work also draws heavily from hypothesis testing theory and the theory of stochastic processes. The relevant elements of these disciplines are introduced in Section 2.3. Also in Section 2.3, we introduce the general notion of a multiscale hypothesis test. A brief discussion of metrics for convex sets is found in Section 2.4. We conclude this chapter with precise statements of the anomaly detection and localization problems and a summary of all modeling assumptions in Section 2.5.

2.1 The Radon Transform and Its Inverse

A discussion of the mathematics of tomography naturally begins with the Radon transform (equation (2.1)). Indeed, we shall see in Section 2.2 that tomographic projections are samples of the smoothed Radon transform of some real function of two dimensions. In this section we define the Radon transform and introduce a few of its basic properties. We also briefly discuss the inverse of the Radon transform.

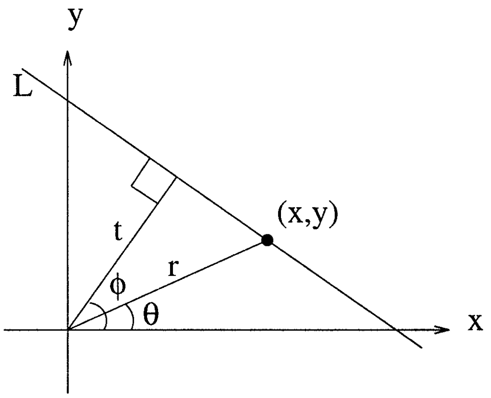


Figure 2-1: This figure illustrates a typical line used in the calculation of the Radon transform.

2.1.1 The Radon Transform

The two-dimensional Radon transform, $f_R(L)$, of a function of two variables, $f(x,y)$, is defined by the line integral of $f(x,y)$ along the line L:

$$f_R(L) \triangleq \mathcal{R}f(x,y) \triangleq \int_L f(x,y)\,ds\,, \tag{2.1}$$

where ds is an increment along the line L and the function f is infinitely differentiable and rapidly decreasing. In many real world problems f has compact support in $\mathbb{R}^2$ and, thus, is rapidly decreasing in a trivial way.

To use equation (2.1), we parameterize the line L by two parameters, t and ϕ. Referring to Figure 2-1, we see that all points (r,θ) on the line L satisfy

$$t = r\cos(\phi - \theta)\,,$$

or, equivalently, in terms of (x,y),

$$t = x\cos\phi + y\sin\phi\,.$$

Therefore, the Radon transform is a function of two variables, t and ϕ, and may be written

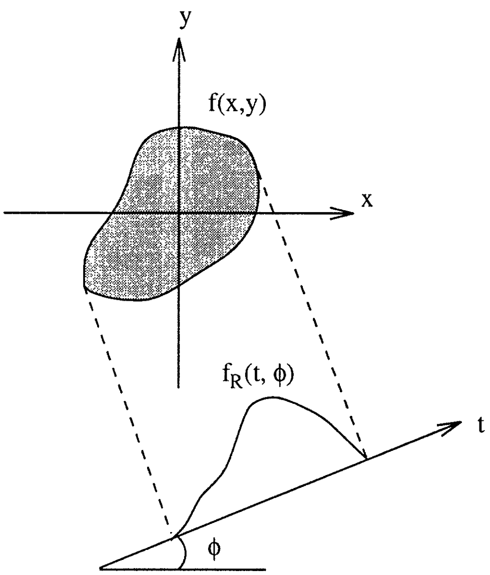


Figure 2-2: This figure represents the projection of a function, $f(x,y)$, with constant intensity over local support at an arbitrary angle ϕ. While the t-axis passes through the origin of the x-y axes, we show it shifted here for clarity.

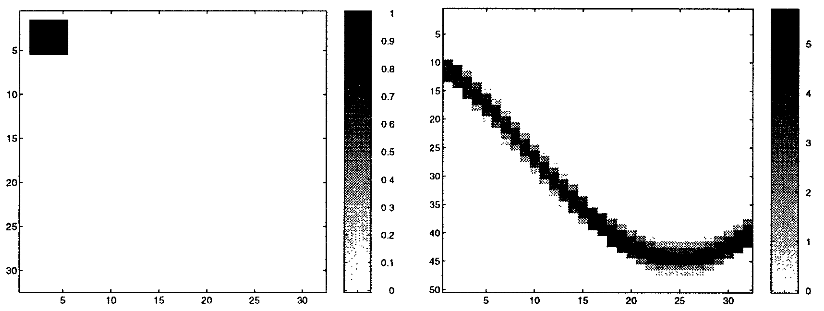
as[1]

$$f_R(t,\phi) = \int_{\mathbb{R}^2} f(x,y)\delta(t - x\cos\phi - y\sin\phi)\,dx\,dy\,. \tag{2.2}$$

From equation (2.2) and Figure 2-1 we see that the Radon transform maps a function which has domain $\mathbb{R}^2$ to one with domain $\mathbb{R} \times [0, 2\pi]$. For a fixed $\phi = \phi_0$, the one-dimensional function $f_R(t,\phi_0)$ is termed a projection of $f(x,y)$. For example, if $\phi_0 = 0$ then $f_R(t,0)$ represents the projection of $f(x,y)$ along the y-direction and, in this case, t coincides with x. Figure 2-2 illustrates a projection at an arbitrary angle ϕ.

There are many equivalent ways of writing the Radon transform and we will find it convenient to adopt a more compact notation. To this end, we define the vectors $\xi \overset{\triangle}{=} [\cos\phi \;\, \sin\phi]^T$ and $\mathbf{x} \overset{\triangle}{=} [x \;\, y]^T$. Now the Radon transform may be written more compactly as

[1] The function $\delta(x)$ is the Dirac delta function which is defined by the properties $\int_{-\epsilon}^{\epsilon} f(x)\delta(x)\,dx = f(0)\,,\ \forall\epsilon > 0$ and $\delta(x) = 0\,,\ \forall x \neq 0$.



(a) (b)

Figure 2-3: Figure (a) is an image domain view of a binary function which has support over a square region. Figure (b) is a sinogram domain view of the Radon transform of the function in (a). The horizontal axis is proportional to the angle ϕ in the range $[0, \pi)$ and the vertical axis is proportional to the offset variable t.

$$f_R(t, \xi) = \int_{I\!R^2} f(\mathbf{x})\delta(t - \xi^T \mathbf{x})\, d\mathbf{x}\,. \tag{2.3}$$

The notation of equation (2.3) is useful because it generalizes for $n \neq 2$ rather nicely. For a general n, ξ is a unit position vector, $\mathbf{x} = [x_1\ x_2 \ldots x_n]^T$, and the integral is over a hyperplane. In this thesis we consider only two-dimensional problems and, when $n = 2$, the Radon transform will be written as $f_R(t, \xi)$ and $f_R(t, \phi)$ interchangeably.

Before we proceed to a discussion of properties of the Radon transform, we present an example function and its Radon transform. Figure 2-3(a) shows a function which is one over a small square region and zero elsewhere. We call the domain of this function the *image domain.* Figure 2-3(b) shows a discrete version of the Radon transform (tomographic projections—to be introduced later) of this function. The domain of the Radon transform is referred to as the *sinogram domain.* Notice that the square function with local support in the image domain maps to a sinusoidal swath with non-local support in the sinogram domain.

2.1.2 Some Properties of the Radon Transform

The Radon transform possesses many useful and interesting properties. In this section we mention only the few which are used in this thesis. A more complete consideration of the

Radon transform and its properties may be found in [13–17].

Property 1 (Linearity) *$\mathcal{R}$ is a linear operator.*

Proof. This follows trivially from the definition of the Radon transform, specifically from the fact that it is an integral transform. □

Property 2 (Symmetry) $f_R(t,\phi) = f_R(-t,\phi+\pi)$.

Proof. This symmetry property follows from the definition of the Radon transform and the fact that $\sin(\phi+\pi) = -\sin\phi$ and $\cos(\phi+\pi) = -\cos\phi$. □

Property 3 (Homogeneity) $f_R(st,s\xi) = \frac{1}{|s|}f_R(t,\xi)$.

Proof. This follows from the property of the Dirac δ-function: $\delta(sx) = \frac{1}{|s|}\delta(x)$. □

There exists a connection between the Radon transform and the Fourier transform. This connection, which we explain next, leads to the well-known projection slice theorem (PST) which is also known as the Fourier slice theorem. We define the n-dimensional continuous Fourier transform $f_F(\mathbf{k})$ of a function $f(\mathbf{x})$ on $I\!R^n$ as

$$f_F(\mathbf{k}) \triangleq \mathcal{F}_n f(\mathbf{x}) \triangleq \int_{I\!R^n} f(\mathbf{x})e^{-2\pi i\mathbf{k}^T\mathbf{x}}\,d\mathbf{x}\,. \tag{2.4}$$

The n-dimensional inverse Fourier transform is

$$f(\mathbf{x}) = \mathcal{F}_n^{-1} f_F(\mathbf{k}) = \int_{I\!R^n} f_F(\mathbf{k})e^{2\pi i\mathbf{k}^T\mathbf{x}}\,d\mathbf{k}\,.$$

To make the connection between the Fourier transform and the Radon transform, rewrite equation (2.4) as

$$f_F(\mathbf{k}) = \int_{-\infty}^{\infty} dp \int_{I\!R^n} f(\mathbf{x})e^{-2\pi i p}\delta(p-\mathbf{k}^T\mathbf{x})\,d\mathbf{x}\,.$$

Notice that the exponential term is only a function of p and, therefore, can be moved outside the integral over $\mathbf{x}$. And if we let $p = st$ and $\mathbf{k} = s\xi$ where $s = ||k||$ then

$$f_F(s\xi) = |s| \int_{-\infty}^{\infty} e^{-2\pi i s t}\, dt \int_{I\!R^n} f(\mathbf{x})\delta(st - s\xi^T\mathbf{x})\, d\mathbf{x}\,.$$

By virtue of the homogeneity property of the Radon transform, it follows that

$$f_F(s\xi) = \mathcal{F}_1\mathcal{R}f(\mathbf{x})\,, \tag{2.5}$$

where the one-dimensional Fourier transform is applied to the variable t (not to ϕ or ξ). In words, for a fixed direction unit vector ξ, the n-dimensional Fourier transform of $f(\mathbf{x})$ evaluated along this direction is the same as one-dimensional Fourier transform of the n-dimensional Radon transform of $f(\mathbf{x})$ taken at this direction. Since we are focusing only on f on $I\!R^2$, $\mathcal{F}_n = \mathcal{F}_2$ and we obtain the famous projection slice theorem (PST). To state and prove the PST, view $f_F(k_x, k_y) = \mathcal{F}_2 f(x, y)$ in polar coordinates: $f_F(k, \phi)$.

Theorem 1 (PST) $f_F(k,\phi) = \mathcal{F}_2 f(x,y) = \mathcal{F}_1\mathcal{R}f(x,y) = \mathcal{F}_1 f_R(t,\phi)$ *where the one-dimensional Fourier transform is over t.*

Proof. This theorem follows directly from equation (2.5). In words, the two-dimensional Fourier transform evaluated on the central slice taken at angle ϕ is equivalent to the one-dimensional Fourier transform of the projection of $f(x,y)$ taken at angle ϕ. □

2.1.3 The Inverse Radon Transform

Formal mathematical development of the inverse of the Radon transform of a function of two variables was first published by Johann Radon in 1917 [29]. Mathematically rigorous treatments of the inverse transform may also be found in [8, 13] and most proofs will be omitted here.

Back-Projection (BP)

Our discussion of the inverse Radon transform begins with the back-projection (BP) operator. The BP operator appears in several factorizations of the Radon transform and, thus, back-projection plays a role in many tomographic reconstruction techniques. The BP operator will be denoted as $\mathcal{B}$ and is defined as

$$h_B(x, y) \triangleq \mathcal{B}h(t, \xi) \triangleq \int_0^\pi h(x \cos\phi + y \sin\phi, \xi)\, d\phi\,,$$

where and $h(t, \xi)$ is an arbitrary function with $t = \xi^T \mathbf{x}$ and $\mathbf{x}$ and ξ are as defined in Section 2.1.1. An alternative definition of the back-projection operator yields h_B in polar coordinates:

$$h_B(r, \theta) = \int_0^\pi h(r \cos(\theta - \phi), \phi)\, d\phi\,.$$

The back-projection operation appears at first glance to be close to the inverse of the Radon transform. By letting $h = f_R$, we see that $\mathcal{B}f_R(t, \phi)$ evaluated at a particular point (x, y) is a summation (integration) of all points of $f_R(t, \phi)$ which correspond to lines L which pass through (x, y). The following analysis, however, shows that back-projection does not exactly invert the Radon transform.

By the PST, $\mathcal{F}_2 f(x, y) = \mathcal{F}_1 \mathcal{R} f(x, y)$. Therefore, the Radon transform can be factored as

$$\mathcal{R} = \mathcal{F}_1^{-1} \mathcal{F}_2\,.$$

Let $h_B(r, \theta)$ be the result of back-projecting $f_R(t, \phi)$. Then

$$h_B(r, \theta) = \mathcal{B}\mathcal{F}_1^{-1}\mathcal{F}_2 f(x, y) = \mathcal{B}\mathcal{F}_1^{-1} f_F(k, \phi)\,, \tag{2.6}$$

where the one-dimensional inverse Fourier transform operates on the variable k. Substituting the definitions of the operators into equation (2.6) we get

$$h_B(r, \theta) = \int_0^\pi d\phi \int_{-\infty}^{\infty} f_F(k, \phi) e^{2\pi i k r \cos(\theta - \phi)}\, dk\,.$$

Finally, by introducing a factor of $\frac{k}{k}$ in the integrand and recognizing that the double integral becomes an integral over the plane in polar coordinates, we have

$$h_B(r,\theta) = \int_0^\pi \int_{-\infty}^\infty k^{-1} f_F(k,\phi) e^{2\pi i k r \cos(\theta-\phi)} k \, dk \, d\phi = \mathcal{F}_2^{-1}\left[k^{-1} f_F(k,\phi)\right] .$$

Since multiplication in the Fourier domain corresponds to convolution in the "time" domain,

$$h_B(r,\theta) = f(r,\theta) ** \frac{1}{r} ,$$

where $\mathcal{F}_2 \frac{1}{r} = \frac{1}{k}$.

So we see clearly that the result of back-projection is related to the original function f through a two-dimensional convolution. The BP operator is the *adjoint* not the inverse of the Radon transform operator. Many techniques have been developed to rid the back-projection of the $1/r$ blurring induced by this convolution. These techniques and their relation to the inverse of the Radon transform are discussed in Section 2.2.

Inverse Radon Transform Factorizations

In this section we introduce the classical inverse Radon transform factorization and show its equivalence to a more useful form. The classical factorization is close to the form of the inverse as expressed by Johann Radon in 1917 but is impractical for numerical implementation. We present this form without proof. From this classical form we derive a factorization which lends itself to numerical implementation and provides the basis for many standard tomographic inversion techniques.

In the classical factorization of the inverse Radon transform, $\mathcal{R}^{-1}$, the BP operator appears with the Hilbert transform operator, $\mathcal{H}_t$:

$$f(x,y) = \mathcal{R}^{-1} f_R(t,\phi) = -\frac{1}{2\pi} \mathcal{B} \mathcal{H}_t \frac{\partial}{\partial t} f_R(t,\phi) , \tag{2.7}$$

where

$$\mathcal{H}_t f(t,\phi) \triangleq \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(t,\phi)}{t-\tau} \, dt = -\frac{1}{\pi} f(t,\phi) * \frac{1}{t} .$$

By inserting the identity $\mathcal{F}_1^{-1}\mathcal{F}_1$ between the $\mathcal{B}$ and the $\mathcal{H}_t$ operators in equation (2.7) and applying the definition of the Hilbert transform we get that

$$f(x,y) = \frac{1}{2\pi^2}\mathcal{B}\mathcal{F}_1^{-1}\mathcal{F}_1\frac{\partial}{\partial t}\left[\frac{1}{t} * f_R(t,\phi)\right] , \tag{2.8}$$

where we have used the fact that the (linear) Hilbert transform operator commutes with differentiation. Before continuing, we will find it useful to recall the following facts:

- $\mathcal{F}_1\frac{\partial}{\partial t}h(t) = 2\pi ik\mathcal{F}_1h(t)$.
- The Fourier transform of $h_1(t) * h_2(t)$ is the product of the Fourier transforms of $h_1(t)$ and $h_2(t)$.
- $\mathcal{F}_1\frac{1}{t} = -i\pi\text{sgn}(k)$.

Applying these facts to equation (2.8) we find that

$$f(x,y) = \mathcal{B}\mathcal{F}_1^{-1}\left[|k|\mathcal{F}_1f_R(t,\phi)\right] . \tag{2.9}$$

Equation (2.9) shows clearly (again) that the BP operator alone does not invert the Radon transform. Instead, the inverse consists of a ramp filtering (with $|k|$) of the projections in the Fourier domain followed by a Fourier inverse and finally back-projection. Notice that $\mathcal{F}_1^{-1}\left[|k|\mathcal{F}_1\right]$ in equation (2.9) corresponds to $-\frac{1}{2\pi}\mathcal{H}_t\frac{\partial}{\partial t}$ in equation (2.7).

2.2 Tomographic Projections and Reconstruction Methods

In the previous section we introduced the Radon transform and its inverse. In this section we relate tomographic projections to the Radon transform and discuss how several factorizations of its inverse are used as bases for different tomographic reconstruction routines. Our treatment of the reconstruction problem is brief and details may be found in [13–17].

2.2.1 Tomographic Projections

Ideal CT projections of an object consist of line integrals through the object of some parameter associated with the object. Prototypical examples of tomographic projections are those obtained from x-rays. When a narrow, mono-energetic x-ray beam passes through an object, the intensity, I, of the beam is attenuated exponentially with distance:

$$\frac{I}{I_0} = e^{-\int_L \mu(x,y)\, ds}, \tag{2.10}$$

where I_0 is the initial intensity of the beam which travels along the straight line L, ds is an incremental distance along line L, and $\mu(x, y)$ is the (possibly) space-varying attenuation coefficient of the medium. The attenuation coefficient is a function of the density, ρ, and the atomic number, Z, of the medium, both of which may vary with space. That is,

$$\mu(x, y) = \mu\left(\rho(x, y), Z(x, y)\right) .$$

Therefore, I contains non-local information about the density and atomic structure of the medium through which the x-ray traveled. By taking the natural logarithm of equation (2.10), we see the connection to the Radon transform of $\mu(x, y)$:

$$\mu_R(L) = -\ln\left(\frac{I}{I_0}\right) = \int_L \mu(x, y)\, ds = \mathcal{R}\mu(x, y) . \tag{2.11}$$

In practice, however, data are not available for all angles ϕ and all offset values t. Further, the x-ray beam is not infinitesimally thin (nor is it exactly mono-energetic). Therefore, while ideal tomographic projections are directly connected with the Radon transform (equation (2.11)), actual tomographic projections are connected to the Radon transform in a more indirect way. Since the x-rays (and any probe) have finite width, the projections obtained in practice correspond to a Radon transform which has been smoothed in the t direction. And, since projections at all angles and all offsets are not obtainable, actual tomographic data correspond to a sampling in ϕ and t of this smoothed Radon transform.

Suppose, for example, that the beam width is such that the obtained projections correspond to the Radon transform of $f(x,y)$ convolved with a smoothing filter, $S(t)$. The function

$$\tilde{f}_R(t,\phi) = \mathcal{R}f(x,y) * S(t) = \int_{I\!R^2} f(\mathbf{x})\delta(t - \xi^T\mathbf{x})\,d\mathbf{x} * S(t)$$

is the smoothed Radon transform. Since the Radon transform is linear, it commutes with convolution so we can bring the convolution under the integral over $\mathbf{x}$ to get

$$\tilde{f}_R(t,\phi) = \int_{I\!R^2} f(\mathbf{x})\,d\mathbf{x} \int_{-\infty}^{\infty} \delta(t - \xi^T\mathbf{x} - \tau)S(\tau)\,d\tau\,.$$

We can see immediately that the smoothing function, $S(t)$, replaces the delta function as the kernel of the Radon transform so that

$$\tilde{f}_R(t,\phi) = \int_{I\!R^2} f(\mathbf{x})S(t - \xi^T\mathbf{x})\,d\mathbf{x}\,.$$

The smoothing function, $S(t)$, models the finite width of the x-ray beam and, more generally, the finite width of any measurement probe or detector. Therefore, while $S(t)$ is not infinitesimally thin, it is often rapidly decreasing and may have compact support.

As mentioned above, any measurement process must acquire samples of $\tilde{f}_R(t,\phi)$. Here we sample $\tilde{f}_R(t,\phi)$ at constant intervals in t and ϕ. The angle ϕ will be sampled in the interval $[0,\pi)$ with sample spacing $\Delta\phi$. There will be N_ϕ such projections[2]. By virtue of the symmetry of the Radon transform, samples at angles greater than π are redundant. The offset variable t will be sampled at each angle between the values $t = -t_0$ to $t = t_0$ with spacing Δt. There will be N_s such samples per angle. Thus there are a total of $N_s N_\phi$ data samples.

We intend to place the measurement values in a vector and it is irrelevant how this vector is ordered so long as it is consistent with the ordering of other vectors and matrices in the problem. One such ordering (and the one we use, though this is not of any particular

[2] Analogous to the Radon transform of $f(x,y)$, the one-dimensional function $\tilde{f}_R(t,\phi_i)$ is termed the projection of $f(x,y)$ at angle ϕ_i

consequence) is obtained by simply stacking the N_s-length data vectors obtained at each of the N_ϕ projections on top of one another. To do so, we begin sampling at $\phi = 0$ and $t = -t_0$ and denote this sample number 1 and number the samples at $\phi = 0$ in order of increasing offset value t such that the last sample at $\phi = 0$ (the one at $t = t_0$) will be indexed by N_s and sample $N_s + 1$ will be taken from angle $\phi = \Delta\phi$ and $t = -t_0$. Continuing in this way, the i^{th} sample will be at $\phi = \lfloor \frac{i-1}{N_s} \rfloor \Delta\phi$ and $t = -t_0 + (i-1)(\text{mod} N_s)\Delta t$. Denote these as ϕ_i and t_i respectively. The sampled, smoothed Radon transform is

$$\tilde{f}_R(t_i, \phi_i) = \int_{I\!R^2} f(x,y) S(t_i - x\cos\phi_i - y\sin\phi_i)\, dx\, dy\,. \tag{2.12}$$

Let us call the i^{th} sample g_i and the corresponding smoothing function which appears in equation (2.12) $S_i(x,y)$. Also, in general, when each tomographic measurement is made there will be additive measurement noise. Denote the i^{th} sample of this noise by n_i. Putting all this together, we have that

$$g_i = \int_{I\!R^2} f(x,y) S_i(x,y)\, dx\, dy + n_i\,. \tag{2.13}$$

Typically the smoothing function, $S(t)$, turns the line integrals of the Radon transform into strip integrals. That is, $S_i(x,y)$ is an indicator function which is one over the i^{th} strip and zero elsewhere. Data acquisition with these indicator functions is illustrated in Figure 2-4. Finally, for computational purposes, the object, $f(x,y)$, is discretized in a basis. We take as our basis for expansion of the field, $f(x,y)$, the rectangular pixel basis so that

$$f(x,y) = \sum_{j=1}^{N_p} f_j p_j(x,y)\,, \tag{2.14}$$

where $p_j(x,y)$ is one over the j^{th} pixel and zero elsewhere and there are N_p pixels. Combining equations (2.13) and (2.14) we find that

$$\mathbf{g} = \mathbf{T}\mathbf{f} + \mathbf{n}\,, \tag{2.15}$$

where $\mathbf{g}$, $\mathbf{f}$, and $\mathbf{n}$ are vectors containing the measured data values, field pixel values, and

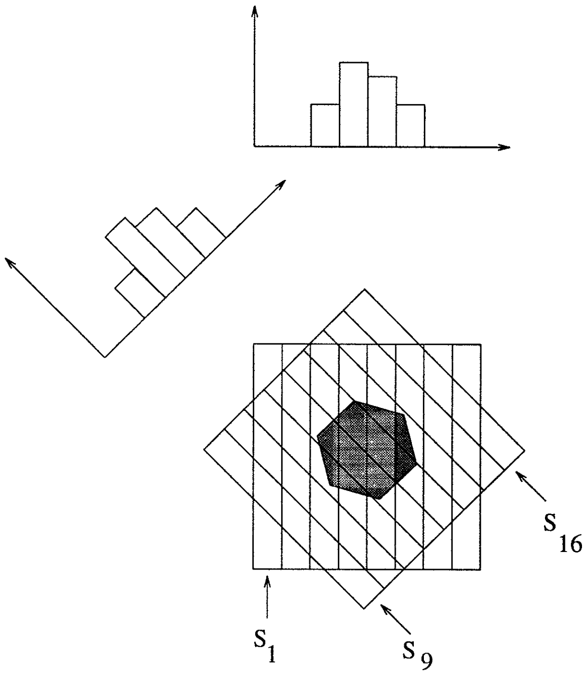


Figure 2-4: This figure is a representation of the projection process with strip indicator functions $S_i(x,y)$. Projections are shown at two different angles ($N_\phi = 2$ with $N_s = 8$). Three of the sixteen strip integral indicator functions (S_1, S_9, and S_{16}) are labeled.

additive noise values, respectively, in some consistent order. That is,

$$\begin{aligned} \mathbf{g} &= [g_1\, g_2 \cdots g_{N_\phi N_s}]^T\,, \\ \mathbf{f} &= [f_1\, f_2 \ldots f_{N_p}]^T\,, \\ \mathbf{n} &= [n_1\, n_2 \ldots n_{N_\phi N_s}]^T\,. \end{aligned}$$

The components of the matrix $\mathbf{T}$ are given by,

$$[\mathbf{T}]_{ij} = \int_{I\!R^2} S_i(x,y) p_j(x,y)\, dx\, dy\,,$$

where

$$
\begin{aligned}
i &= 1,\ldots,N_\phi N_s\,, \\
j &= 1,\ldots,N_p\,.
\end{aligned}
$$

Equation (2.15), coupled with whatever *a priori* knowledge we have about $\mathbf{n}$ and $\mathbf{f}$, represents our observational model. Notice that equation (2.15), like the Radon transform itself, is linear. The tomographic projection matrix, $\mathbf{T}$, captures a discrete representation of the smoothed line integrals. The application of $\mathbf{T}$ to $\mathbf{f}$ is called the projection of $\mathbf{f}$. Recall that the back-projection operator is the adjoint of the Radon transform. Analogously, the discrete back-projection matrix is the transpose (adjoint) of the projection matrix.

Figure 2-5 illustrates the projection matrix used in this thesis (except where explicitly stated otherwise). The number of projections is $N_\phi = 32$ and the number of samples in each projection is $N_s = 50$. Therefore, the number of rows is $N_\phi N_s = 1600$ and $N_p = 1024$ corresponding to a 32×32 pixel field.

2.2.2 Methods for Tomographic Reconstruction

Section 2.1.3 concluded with equation (2.9) which we reproduce here:

$$
f(x,y) = \mathcal{B}\mathcal{F}_1^{-1}\left[|k|\mathcal{F}_1 f_R(t,\phi)\right]\,. \tag{2.16}
$$

Equation (2.16) forms the basis of many reconstruction algorithms and any reconstruction technique which is based upon equation (2.16) is called filtered back-projection (FBP). From the FBP factorization it is simple to derive another widely used factorization which leads to the convolution back-projection (CBP) algorithm. Again making use of the fact that the Fourier transform of a convolution is the product of Fourier transforms and equation (2.16) we get that

$$
f(x,y) = \mathcal{B}\left[v(t) * f_R(t,\phi)\right]\,,
$$

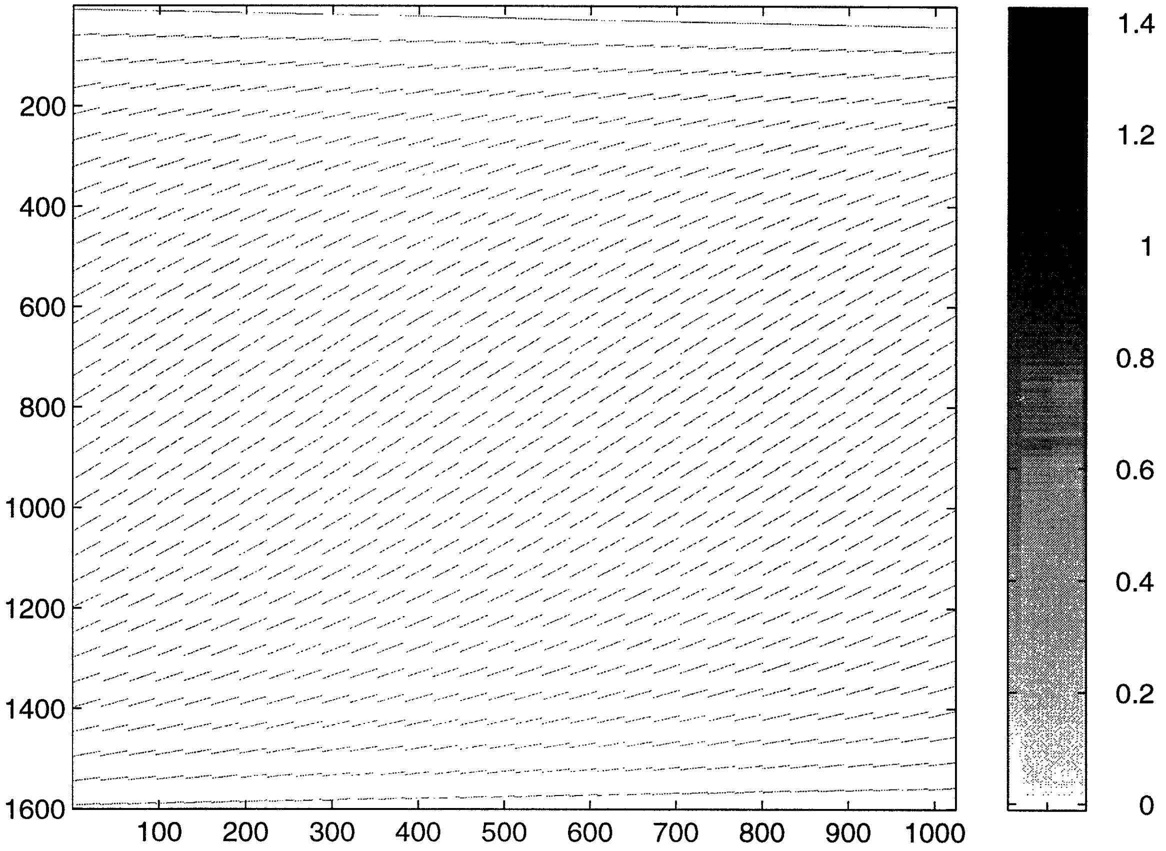


Figure 2-5: This figure illustrates a tomographic projection matrix for which $N_\phi = 32$, $N_s = 50$, and $N_p = 1024$ corresponding to a 32 × 32 pixel field. Notice that this matrix is extremely sparse.

where

$$\mathcal{F}_1 v(t) = |k| w(k) .$$

The window function $w(k)$ has been introduced because $\mathcal{F}_1^{-1}|k|$ is a singular distribution which does not lend itself to numerical computation. Additionally, $w(k)$ is used to attenuate high frequencies which tend to be dominated by noise.

Other tomographic reconstruction techniques follow from other factorizations of the inverse Radon transform. Some of these factorizations involve filtering or convolution *after* back-projection. Other popular techniques, such as the algebraic reconstruction technique (ART) [14,16], are iterative and are based on the discrete representation developed in the previous section. CBP and FBP are the most popular and widely used methods, however.

2.3 Hypothesis Testing

Our anomaly characterization methods are based on hypothesis testing. In this section we review the relevant elements of this theory which is also commonly known as detection theory. Our notation and terminology is consistent with that of [35,38,42]. A hypothesis test is a mapping of the observed data to a decision as to which one of many hypotheses is most likely true. We may write this mapping abstractly as $h: \mathcal{D} \rightarrow \mathcal{H}$, where $\mathcal{D}$ is the data domain (e.g., $\mathbb{R}^n$) and $\mathcal{H}$ is the (finite) set of hypotheses. When there are just two hypotheses, H_0 and H_1, the problem is said to be a *binary hypothesis testing problem.* When there are $M > 2$ hypotheses, the problem is said to be an *M-ary hypothesis testing problem.* Most of the essential aspects of hypothesis testing theory are revealed in the study of the simpler binary hypothesis test (BHT). In Section 2.3.1 we review binary hypothesis testing and indicate how the concepts generalize to the M-ary hypothesis testing (MHT) case. In Section 2.3.2 we discuss the concept of multiscale hypothesis testing which is relevant only to MHT problems.

2.3.1 Binary Hypothesis Testing

In a BHT problem exactly one of the two events (or statements) H_0 and H_1 is true. The two events are interpreted as two hypotheses about reality and the event labeled H_0 is called the *null hypothesis*. In such problems data are available which depend, in some way, on which hypothesis is true. Let us assume that the available data are discrete and comprise a data vector $\mathbf{y}$. In problems where the observed data are continuous (e.g., a continuous waveform) techniques exist to reduce the continuous data to an equivalent set of discrete data (e.g., Karhunen-Loeve expansion). These techniques are beyond the scope of this thesis but we mention them to emphasize that our discrete data assumption is more general than it may appear.

To apply Bayesian detection techniques we must know the probability distributions of $\mathbf{y}$ conditioned on each of the two hypotheses. In addition, we must know the probability that each hypothesis is true. In the absence of this latter information, Bayesian techniques may not be applied and a different type of optimality criterion must be used. Let us assume that this is the case—we do not know the probability that H_0 is true or the probability that H_1 is true. Let us apply the Neyman-Pearson criterion to the BHT problem.

The Neyman-Pearson Criterion

The Neyman-Pearson criterion is stated in terms of the probability of detection, P_d, and probability of false alarm, P_f. These are defined as

$$P_d \triangleq \Pr\left[h(\mathbf{Y}) = H_1 \text{ assuming that } H_1 \text{ is true}\right], \tag{2.17}$$

$$P_f \triangleq \Pr\left[h(\mathbf{Y}) = H_1 \text{ assuming that } H_0 \text{ is true}\right], \tag{2.18}$$

where $\mathbf{Y}$ is the particular observed realization of the random vector $\mathbf{y}$. The Neyman-Pearson criterion is that the decision function, $h(\cdot)$, is such that it maximizes P_d while satisfying the constraint that P_f is below a certain value $\alpha > 0$. The optimal decision rule is the *likelihood ratio test* (LRT) as is stated in the following theorem.

Theorem 2 (Neyman-Pearson Rule) *Let*[3] $L(\mathbf{Y}) \triangleq \frac{p_{\mathbf{y}}(\mathbf{Y};H_1)}{p_{\mathbf{y}}(\mathbf{Y};H_0)}$. *The decision rule* $h(\mathbf{Y})$ *which satisfies the Neyman-Pearson criterion has the form*

$$h(\mathbf{Y}) = \begin{cases} H_0 & \textit{if } L(\mathbf{Y}) \leq \gamma \\ H_1 & \textit{if } L(\mathbf{Y}) > \gamma \end{cases} .$$

We note that the function $L(\mathbf{Y})$ is called the *likelihood ratio function* and the resulting decision rule is called the *likelihood ratio test* (LRT). The LRT happens also to be the optimal Bayesian decision rule under a symmetric cost assignment and when $\Pr[H_0] = \Pr[H_1]$. Also note that any monotonically increasing function, $\ell(\cdot)$, of $L(\mathbf{Y})$ (e.g., the logarithm) may be inserted in place of $L(\mathbf{Y})$ without affecting the outcome of the LRT (as long as γ is also replaced by $\ell(\gamma)$).

Proof. We wish to minimize

$$J = 1 - P_d + \gamma\,(P_f - \alpha) \ ,$$

where γ is the Lagrange multiplier associated with the constraint on P_f. Let $\mathcal{D}_i$ be the domain of $\mathcal{D}$ for which we decide H_i. Then we can rewrite J as

$$\begin{aligned} J &= 1 - \int_{\mathcal{D}_1} p_{\mathbf{y}}(\mathbf{Y};H_1)\,d\mathbf{Y} + \gamma\left[\int_{\mathcal{D}_1} p_{\mathbf{y}}(\mathbf{Y};H_0)\,d\mathbf{Y} - \alpha\right] , \\ &= 1 - \gamma\alpha + \int_{\mathcal{D}_1} \{\gamma p_{\mathbf{y}}(\mathbf{Y};H_0) - p_{\mathbf{y}}(\mathbf{Y};H_1)\}\ d\mathbf{Y} \,. \end{aligned}$$

Therefore, to minimize J we ought to assign $\mathbf{Y}$ to $\mathcal{D}_1$ (i.e., declare H_1 for $\mathbf{Y}$) whenever the integrand is negative. Equivalently, we arrive at the decision rule

[3] The the function $p_{\mathbf{x}}(\mathbf{X};\alpha)$ is the probability density function for the random vector $\mathbf{x}$ and is parameterized by α.

$$L(\mathbf{Y}) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma ,$$

where γ is chosen to satisfy the P_f constraint. This completes the proof. □

The LRT generalizes in a straight forward way in the case of an MHT. A likelihood ratio function is defined for each hypothesis (other than the null hypothesis). These likelihood ratio functions are then compared with each other and with a threshold, γ, in the decision rule

$$h(\mathbf{Y}) = \begin{cases} H_0 & \text{if } \max_j L_j(\mathbf{Y}) \le \gamma \\ H_i & \text{if } \max_j L_j(\mathbf{Y}) > \gamma \text{ where } i = \arg\max_j L_j(\mathbf{Y}) \end{cases} ,$$

where the likelihood ratio functions are defined as

$$L_j(\mathbf{Y}) \triangleq \frac{p_{\mathbf{y}}(\mathbf{Y}; H_j)}{p_{\mathbf{y}}(\mathbf{Y}; H_0)} .$$

Gaussian BHT

When the conditional probabilities $p_{\mathbf{y}}(\cdot;\cdot)$ have certain forms the LRT has a particularly simple structure. One special form is the Gaussian probability density function. A further special case arises when the conditional probabilities share the same covariance but have different means. In this section we derive the form of the LRT and the forms of P_f and P_d as functions of the LRT threshold γ for this special case.

Let the conditional probabilities have the form[4]

$$\begin{aligned} H_0 &: \mathbf{y} \sim \mathcal{N}(\mathbf{m_0}, \mathbf{\Lambda}) , \\ H_1 &: \mathbf{y} \sim \mathcal{N}(\mathbf{m_1}, \mathbf{\Lambda}) . \end{aligned}$$

[4] The notation $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{P})$ means that $\mathbf{x}$ is a Gaussian random vector with mean $\mathbf{m}$ and covariance $\mathbf{P}$, i.e., $p_{\mathbf{X}}(\mathbf{X}) = \exp\{-\frac{1}{2}(\mathbf{X}-\mathbf{m})^T\mathbf{P}^{-1}(\mathbf{X}-\mathbf{m})\}/|2\pi\mathbf{P}|^{1/2}$.

Then the *log-likelihood ratio function* (LRF) is

$$\ell(\mathbf{Y}) \triangleq \ln L(\mathbf{Y}) = (\mathbf{m_1} - \mathbf{m_0})^T \boldsymbol{\Lambda}^{-1}\mathbf{Y} + \frac{1}{2}\mathbf{m_0}^T \boldsymbol{\Lambda}^{-1}\mathbf{m_0} - \frac{1}{2}\mathbf{m_1}^T \boldsymbol{\Lambda}^{-1}\mathbf{m_1}\,. \tag{2.19}$$

The LRF simplifies further if $||\mathbf{m_1}||^2_{\boldsymbol{\Lambda}^{-1}} = ||\mathbf{m_0}||^2_{\boldsymbol{\Lambda}^{-1}}$ where $||\mathbf{x}||^2_{\mathbf{P}} \triangleq \mathbf{x}^T\mathbf{P}\mathbf{x}$. We may assume this to be the case without loss of generality for we may always adjust the LRT threshold γ to cancel the last two terms of equation (2.19). Let us define $\mathbf{A} \triangleq (\mathbf{m_1} - \mathbf{m_0})^T \boldsymbol{\Lambda}^{-1}$. Therefore the LRF is just a linear function of the data and so is itself a Gaussian random variable with the conditional probability density functions

$$\begin{aligned} H_0 &: \ \ell \sim \mathcal{N}(\mathbf{A}\mathbf{m_0}, \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T)\,, \\ H_1 &: \ \ell \sim \mathcal{N}(\mathbf{A}\mathbf{m_1}, \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T)\,. \end{aligned}$$

Since the LRF is a Gaussian random variable it is easy to calculate P_d and P_f as a function of the LRT threshold $\eta \triangleq \ln \gamma$. Expressing equations (2.17) and (2.18) in terms of η we get that

$$\begin{aligned} P_d &= \Pr\left[\ell > \eta \text{ assuming that } H_1 \text{ is true}\right] = Q\left\{\frac{\eta - \mathbf{A}\mathbf{m_1}}{\sqrt{\mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T}}\right\}\,, \\ P_f &= \Pr\left[\ell > \eta \text{ assuming that } H_0 \text{ is true}\right] = Q\left\{\frac{\eta - \mathbf{A}\mathbf{m_0}}{\sqrt{\mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T}}\right\}\,, \end{aligned}$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{y^2}{2}}\,dy = \frac{1}{2}\operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right)\,,$$

and

$$\operatorname{erfc}(u) \triangleq \frac{2}{\sqrt{\pi}}\int_u^\infty e^{-t^2}\,dt\,.$$

In this equal variance, Gaussian BHT problem, performance is completely determined

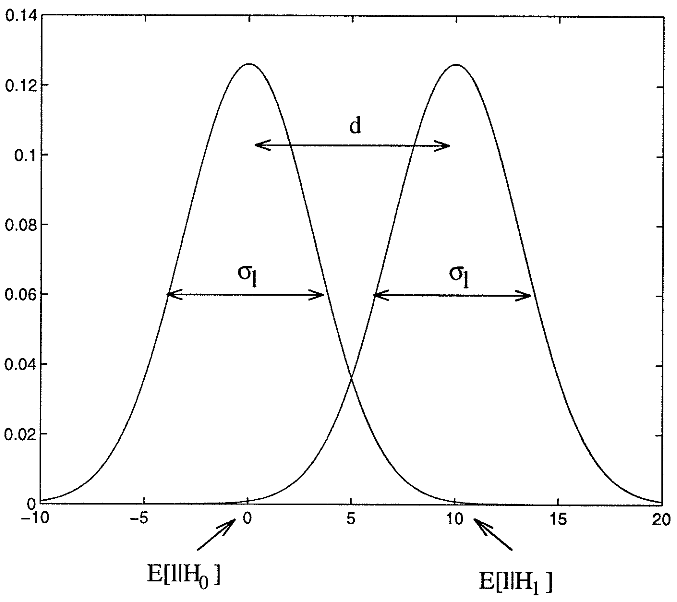


Figure 2-6: This figure depicts two equal variance Gaussian probability density functions with different means. Here $\sigma_\ell = E[\ell|H_1] = 10$ and $E[\ell|H_0] = 0$.

by the d statistic (which, when squared, is known as the d^2 statistic). The d statistic measures the difference in the conditional mean of ℓ relative to its standard deviation (see Figure 2-6):

$$d \triangleq \frac{E[\ell|H_1] - E[\ell|H_0]}{\sigma_\ell}.$$

Plugging in the appropriate expressions, we get that

$$d = \frac{\mathbf{A}(\mathbf{m_1} - \mathbf{m_0})}{\sqrt{\mathbf{A}\mathbf{\Lambda}\mathbf{A}^T}} = \sqrt{(\mathbf{m_1} - \mathbf{m_0})^T \mathbf{\Lambda}^{-1} (\mathbf{m_1} - \mathbf{m_0})}\,. \tag{2.20}$$

Also,

$$d = Q^{-1}(P_f) - Q^{-1}(P_d) \triangleq \Pi(P_f, P_d)\,. \tag{2.21}$$

Equating the square of (2.20) with the square of equation (2.21) yields the equation for an ellipsoid in $\mathbf{m_1} - \mathbf{m_0}$:

$$(\mathbf{m_1} - \mathbf{m_0})^T \boldsymbol{\Lambda}^{-1} (\mathbf{m_1} - \mathbf{m_0}) = \Pi^2(P_f, P_d)\,. \tag{2.22}$$

We shall make use of this ellipsoid in our performance analysis in Chapter 3. Note that the preceding P_d, P_f performance analysis is rather complicated in the MHT case since it involves integration over multidimensional Gaussian functions.

2.3.2 Composite and Multiscale Hypothesis Testing

The number of hypotheses which must be considered in some MHT problems is so large that the optimal hypothesis test cannot be performed in any reasonable amount of time (where reasonableness is determined by the particular application). Examples of such problems arise in computer vision where object recognition is often performed using a suite of matched filters. Each matched filter represents a hypothesis about which object is present and at what orientation and articulation. A similar approach to automatic target recognition problems (e.g., synthetic aperture radar based ATR) results in the formulation of a large number of hypotheses. In many applications in both these disciplines, near-real time performance is crucially important. The anomaly characterization problem considered in this thesis is similar to these two problems and, therefore, suffers the same combinatorial explosion of hypotheses. An attractive alternative to conducting the computationally infeasible optimal MHT is the multiscale hypothesis test (MSHT). In this section we introduce the general form of the MSHT and indicate some of the challenges associated with constructing one. Before discussing the multiscale hypothesis test, however, we first review the notion of a composite hypothesis test. Optimal composite hypothesis tests are called uniformly most powerful (UMP) tests. We will illustrate the concept of a UMP test with an example.

Composite Hypothesis Testing

A MSHT involves a sequence of composite hypothesis tests. In this section we define the notion of a composite hypothesis test. For some composite hypothesis testing problems an optimal test can be formulated. These tests are called *uniformly most powerful* (UMP)

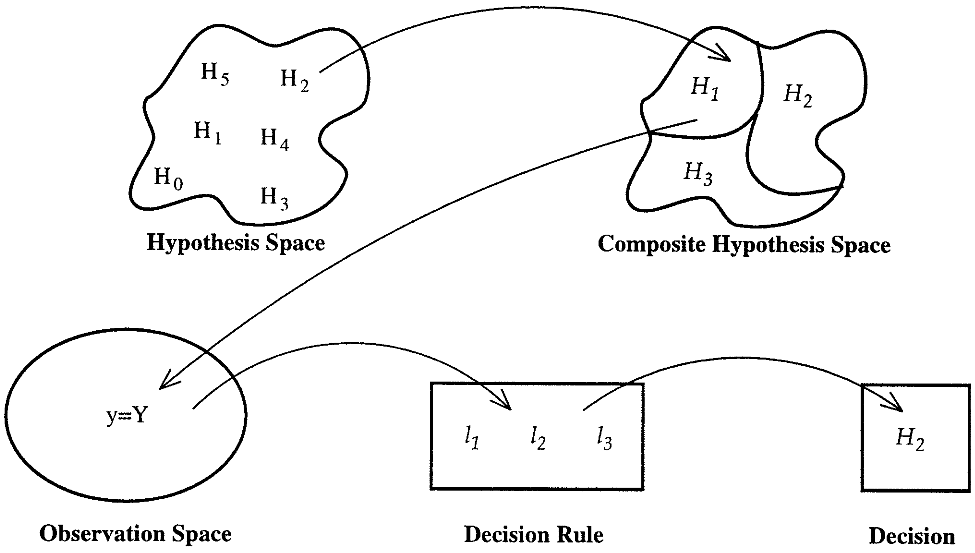


Figure 2-7: This figure illustrates a general composite hypothesis test.

tests. We define and illustrate the UMP test idea with a simple example problem after our introduction of composite hypothesis testing.

Figure 2-7 illustrates a general composite hypothesis test. A composite hypothesis test is *not* defined on the original hypothesis space (the set $\mathcal{H} \triangleq \{H_i\}_{i=0}^{M-1}$) as the optimal MHT would be. Instead, it is defined on elements of a *finite cover* of $\mathcal{H}$.

Definition 1 (Finite Cover) *Let A be a set and N be a finite positive integer. The family of sets $\{A_i\}_{i=1}^{N}$ is said to be a finite cover for A if $A \subseteq \bigcup_{i=1}^{N} A_i$.*

If the A_i of Definition 1 also have the property that $A_i \cap A_j = \emptyset$, $\forall i \neq j$ then the finite cover is said to be a *partition.*

The elements of a finite cover of $\mathcal{H}$ comprise the composite hypothesis space which we denote by $\tilde{\mathcal{H}}$. Denote the j^{th} element of this cover by $\mathcal{H}_j$ (in Figure 2-7 these script letters are italicized). The elements of $\tilde{\mathcal{H}}$ are called composite hypotheses because, in general, they each contain one or more of the original hypotheses H_i. Note that the range of $\mathcal{H}$ and $\tilde{\mathcal{H}}$ are the same, viz., the set $\{H_0, \ldots, H_{M-1}\}$. Also note that the number of composite hypotheses is N where $N \leq M$ and M is the number of original hypotheses.

Exactly one of the $H_i \in \mathcal{H}$ is true. Let's suppose that H_1 is the one which is true. The hypothesis H_1 also belongs to at least one of the composite hypotheses. Let us suppose, for simplicity, that H_1 belongs to exactly one composite hypothesis, $\mathcal{H}_2$, say. Although composite hypothesis $\mathcal{H}_2$ contains many false hypotheses, it also contains the true one (H_1) so we say that $\mathcal{H}_2$ is true. In general, all the composite hypotheses which contain the true hypothesis are said to be true and all the composite hypotheses which do not contain the true hypothesis are said to be false.

One decision statistic is associated with each composite hypothesis. These decision statistics, ℓ_j, are functions of the observed data $\mathbf{Y}$ (e.g., LRFs). A decision rule is defined in terms of these decision statistics (e.g., a LRT) and one composite hypothesis is selected on the basis of this rule.

When considering a composite hypothesis test the following questions arise: what finite cover should be used? What are the decision statistics? What is the decision rule? Once these questions are answered then one may naturally ask: is my composite hypothesis test optimal (i.e., is it a UMP test)? If not, to what degree is it suboptimal? We illustrate the notion of a UMP with an example.

An Example Problem

Consider the following MHT problem with scalar Gaussian data. Under hypothesis H_i the observation y is a scalar Gaussian random variable with mean m_i and variance σ^2. The mean value m_i belongs to the finite set $\{0, m_1, \ldots, m_{M-2}\}$. Suppose we do not wish to formulate the optimal MHT for this problem. In the following composite hypothesis reformulation, two composite hypotheses are defined. Under $\mathcal{H}_0$ the observation, y is a zero-mean Gaussian random variable with variance σ^2. Under $\mathcal{H}_1$, y has mean $m \in \{m_1, \ldots, m_{M-2}\}$ and variance σ^2. Therefore $\mathcal{H}_0$ is a trivial composite hypothesis since it contains only the zero-mean hypothesis and $\mathcal{H}_1$ contains all the $M-1$ other hypotheses.

The composite hypothesis testing problem, therefore, is to choose between $\mathcal{H}_0$ and $\mathcal{H}_1$. The actual value of m is irrelevant and it is called an *unwanted parameter* [38]. There are three cases to consider depending on the possible values of the unwanted parameter (i.e., the set $\{0, m_1, \ldots, m_{M-2}\}$). We consider each of these three cases in turn below.

Case I Suppose that $\min\{m_1, \ldots, m_{M-2}\} > 0$ and consider the decision rule satisfying the Neyman-Pearson criterion. The optimal decision rule is [38]

$$Y \mathop{\gtrless}_{\mathcal{H}_0}^{\mathcal{H}_1} \gamma^+ ,$$

where $P_f = Q\{\frac{\gamma^+}{\sigma}\}$. Thus, the optimal test can be specified without knowledge of the true value of m.

Case II Suppose that $\max\{m_1, \ldots, m_{M-2}\} < 0$ and consider the decision rule satisfying the Neyman-Pearson criterion. The optimal decision rule is [38]

$$Y \mathop{\gtrless}_{\mathcal{H}_1}^{\mathcal{H}_0} \gamma^- ,$$

where $P_f = 1 - Q\{\frac{\gamma^-}{\sigma}\}$. Thus, the optimal test can be specified without knowledge of the true value of m.

Case III Suppose that $\max\{m_1, \ldots, m_{M-2}\} > 0$ and $\min\{m_1, \ldots, m_{M-2}\} < 0$. If the true value of m is positive then the optimal decision rule is given in Case I above. If the true value of m is negative, the optimal decision rule is given in Case II above. Therefore, without knowledge of the sign of m the optimal decision rule cannot be specified in this case.

In Case I and Case II in the above example, knowing the value of the unwanted parameter m does not increase the performance of the composite hypothesis test. For this reason, the designed test in these cases is said to be *uniformly most powerful* (UMP). In Case III, however, any test designed without knowledge of at least the sign of m is necessarily suboptimal. Such a test is, therefore, not a UMP test.

Definition 2 (UMP Test) *A UMP test is one which is as good as the optimal test one could design if the values of all unwanted parameters were known.*

Property 4 (UMP Test Existence) *A UMP test exists if and only if the LRT for every value of the unwanted parameters can be completely specified without knowledge of the unwanted parameters' values.*

Proof. This property follows from the definition of a UMP test and the optimality of the LRT. □

In many practical problems, finding a UMP test or proving its existence is extremely difficult. If one cannot be found then a test must be designed based on some other convenient optimality criterion or using heuristics. One common technique is the generalized likelihood ratio test (GLRT). In a GLRT the unwanted parameters are first estimated under the assumption that each composite hypothesis is true. These estimates are then used in a likelihood ratio test. The form of the GLRT for a binary composite hypothesis test with a scalar unwanted parameter and scalar data is

$$\frac{\max_{m\in\mathcal{H}_1} p_y(Y;m)}{\max_{m\in\mathcal{H}_0} p_y(Y;m)} \begin{array}{c}\mathcal{H}_1\\ >\\ <\\ \mathcal{H}_0\end{array} \gamma,$$

where $m \in \mathcal{H}_i$ means that m spans the values allowed under $\mathcal{H}_i$. The GLRT generalizes in a straightforward way for a vector of unwanted parameters and vector data.

In the example above the unwanted parameter is discrete so the estimation part of the GLRT is really a hypothesis test to decide the value of m. But this hypothesis test is precisely the thing we wished to avoid! So other means are sought.

Multiscale Hypothesis Testing

A multiscale hypothesis consists of a sequence of composite hypothesis tests. Each test in the sequence is associated with a finite cover of some subset of $\mathcal{H}$ and decision statistics. And each test in the sequence is given an index which we call *scale*. The range of the covers (i.e., the subset of $\mathcal{H}$ which they cover) becomes smaller in cardinality as the scale index increases. At the coarsest scale a finite cover is defined for all of $\mathcal{H}$. The finest scale consists of a finite cover for just a subset of $\mathcal{H}$. If the MSHT continues to the finest possible

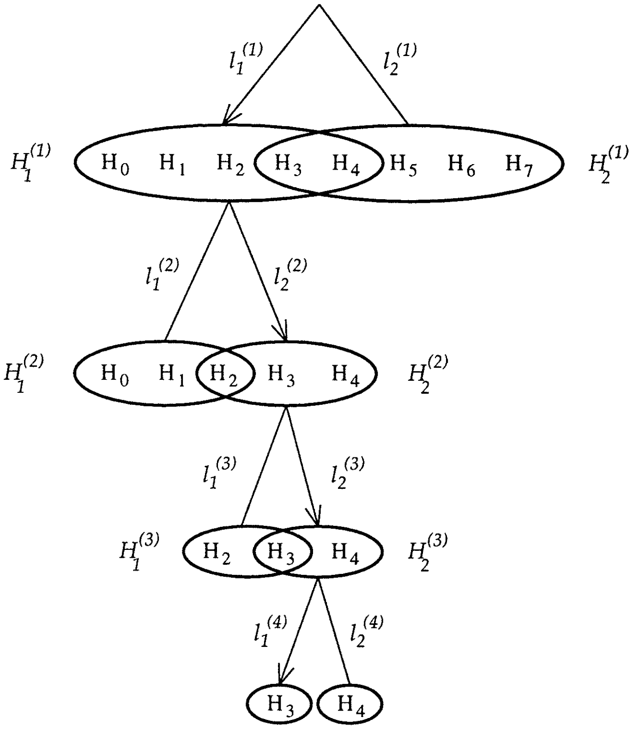


Figure 2-8: This figure illustrates a multiscale hypothesis test.

scale, the hypothesis test at the finest scale is just a BHT for two elements of $\mathcal{H}$. Some of the $H_i \in \mathcal{H}$ are not included in the composite hypotheses at an intermediate scale. Those which are not included are said to have been *discarded.* Any hypothesis, H_i, which has been discarded cannot ultimately be selected as the one which we think is true. The efficiency of a MSHT is achieved by discarding many H_i at each scale.

A particular example is illustrated in Figure 2-8. At each scale in the tree illustrated in Figure 2-8 a choice is made between two composite hypotheses (written in italics) based on two statistics. The chosen composite hypothesis is indicated with an arc with an arrow. The superscripts on the composite hypotheses and statistics indicate the scale. Notice that the subset of $\mathcal{H}$ for which a finite cover is defined at scale k is precisely the subset which is contained in the composite hypothesis which has been selected at scale $k-1$. The elements of a MSHT discussed above and illustrated in Figure 2-8 apply equally to the case where an arbitrary finite number of composite hypotheses $N^{(k)}$ are defined at scale k.

The challenge in designing a MSHT is to select good covers and statistics for every scale and for every possible sequence of decisions. These covers and statistics should be selected so that the resulting sequence of composite hypothesis tests effectively zooms in on the true anomaly via the multiscale search on the domain $\mathcal{H}$. We discuss criteria for good covers and statistics in Chapter 5.

Notice that in our discussion of MSHT we have not associated the scale of a composite hypothesis with a spatial or temporal scale. In fact, it need not be the case that the multiscale zooming effected by a MSHT corresponds to spatial or temporal zooming. In this sense the zooming is truly a statistical one, defined wholly in terms of the hypothesis space $\mathcal{H}$.

2.4 Measures for Convex Sets

The anomalies for which we search in our examples have support on a convex subset of $\mathbb{R}^2$. The regions to which we localize these anomalies are also convex. One means of evaluating our anomaly localization algorithms is to measure, in some sense, the difference between the true anomaly's support and the chosen region. In this section we describe two ways of measuring this difference. One way is to use the *Hausdorff distance* which is a true distance metric between two convex sets. Another way is to use the *one-sided Hausdorff measure* which is not a distance at all but proves useful nonetheless.

A common measure between convex sets is the Hausdorff distance [12]. Denote the set of all convex sets which are subsets of $\mathbb{R}^2$ by $\mathcal{K}$. To define the Hausdorff distance we first define the distance between a point $x \in \mathbb{R}^2$ and a convex set $A \in \mathcal{K}$ as

$$d(x, A) = \min_{a \in A} ||x - a|| .$$

Then the Hausdorff distance $h(A, B)$ between two sets $A, B \in \mathcal{K}$ is defined as

$$h(A, B) = \max \left(\max_{b \in B} d(b, A), \max_{a \in A} d(a, B) \right) .$$

The Hausdorff distance has the following intuitive interpretation: $\max_{b \in B} d(b, A)$ is the

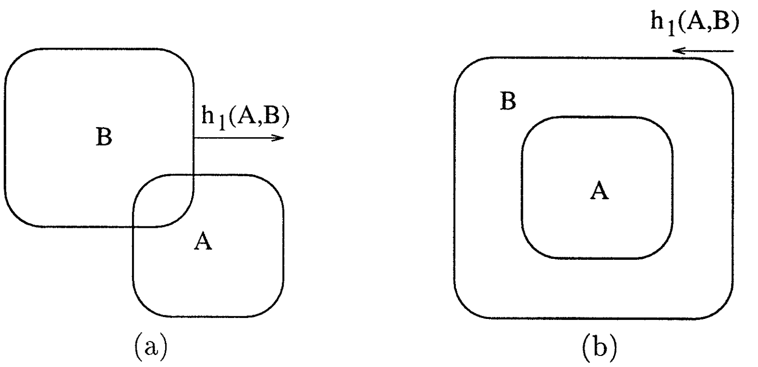


Figure 2-9: This figure illustrates the one-sided Hausdorff metric. In (a) $h_1(A,B)$ is positive since $A \not\subseteq B$ whereas in (b) $h_1(A,B)$ is negative.

minimum amount by which set A must grow uniformly in all directions to include set B. The expression $\max_{a \in A} d(a,B)$ has a similar interpretation as the minimum amount by which set B must grow uniformly in all directions to include set A. The Hausdorff distance is the maximum of these two numbers.

While the Hausdorff distance is a standard measure of the distance between two convex sets, we shall see that it is not well suited to our problem. The following measure is, however, well suited to our problem. We will define the one-sided Hausdorff metric, $h_1(A,B)$, as

$$h_1(A,B) = \begin{cases} \max_{a \in A} d(a,B) & \text{if } A \not\subseteq B \\ -\min_{a \in A} d(a,(A \cup B)^c) & \text{otherwise} \end{cases},$$

where superscript c denotes complement. The one-sided Hausdorff metric is illustrated in Figure 2-9 and has the interpretation as the amount by which set B must grow or shrink uniformly in all directions to just enclose set A. If B encloses A (i.e., A is a subset of B) then $h_1(A,B)$ tells us how much we must shrink B (this number is negative). If B does not enclose A then $h_1(A,B)$ tells us by how much we must grow B (this number is positive).

2.5 Problem Statement

The main theoretical problems considered in this thesis are concerned with the development of the multiscale hypothesis testing framework. We introduced the notion of multiscale hypothesis testing in Section 2.3.2 and consider criteria for good MSHTs in Chapter 5. The main applied problems addressed in this thesis are single anomaly detection and localization from tomographic measurements. In this section we outline all modeling assumptions and state the anomaly detection and localization problems precisely.

The form of the measurement model is given in Section 2.2.1:

$$\mathbf{g} = \mathbf{T}\mathbf{f} + \mathbf{n} \,. \tag{2.23}$$

The field $\mathbf{f}$ is modeled as a superposition of a field with at most a single anomaly, $\mathbf{f}_a$, and an anomaly-free background field, $\mathbf{f}_b$:

$$\mathbf{f} = \mathbf{f}_a + \mathbf{f}_b \,. \tag{2.24}$$

The anomaly field is zero everywhere except over a square patch. We write

$$\mathbf{f}_a = c\mathbf{b}_{s,N}(i,j) \,,$$

where c is the unknown non-negative anomaly intensity and $\mathbf{b}_{s,N}(i,j)$ is the lexicographically ordered vector associated with the $N \times N$ field which is zero everywhere except over the $s \times s$ area with upper left corner at pixel (i,j) where $\mathbf{b}_{s,N}(i,j)$ takes the value one. The exact size s and location (i,j) of this square anomaly are unknown. We assume knowledge, however, of the maximum possible size, s_{max}, the anomaly can be where $s_{max} \ll N$.

The background field is modeled as a zero-mean Gaussian random field with known covariance $\mathbf{\Lambda}$. In this thesis, the background covariance is either the diagonal matrix $\mathbf{\Lambda} = \mathbf{\Lambda}_w \overset{\triangle}{=} \sigma^2\mathbf{I}$ (so the background field is white) or it is a fractal field covariance matrix $\mathbf{\Lambda}_f$ (so the background is correlated). A fractal background has been chosen as a comparison to the white background because fractal fields have been found to accurately model a wide range of natural textures [41]. The $32^2 \times 32^2$ fractal field covariance matrix used in this thesis (except where stated) is illustrated in Figure 2-10. It has a spectral exponent of two. Details regarding the the structure of the fractal field covariance matrix are found in Appendix A. The additive measurement noise $\mathbf{n}$ is assumed to be a zero-mean white Gaussian random

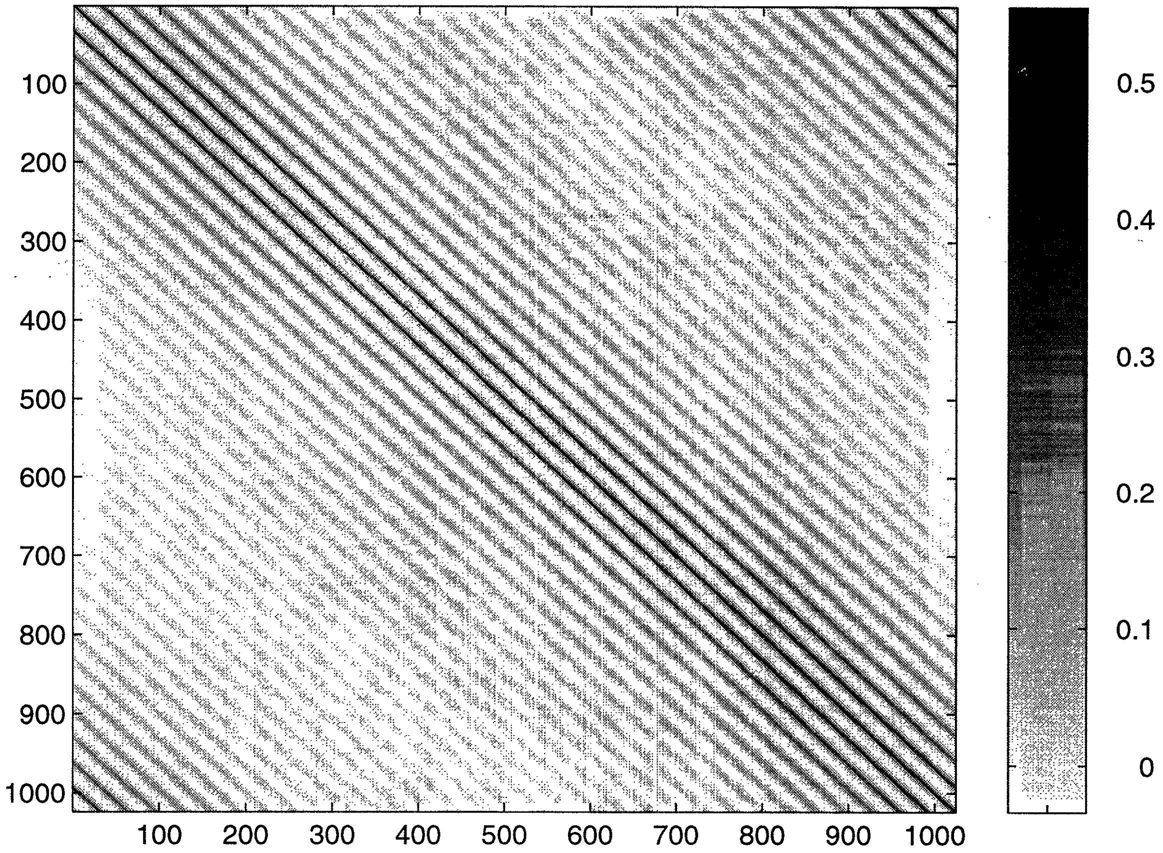


Figure 2-10: This figure illustrates the fractal field covariance matrix associated with a 32×32 pixel field. The spectral exponent is two.

vector with intensity λ and is independent of the background and anomaly fields. Therefore, the data are conditionally Gaussian:

$$\mathbf{g} \sim \mathcal{N}(c\mathbf{T}\mathbf{b}_{s,N}(i,j), \mathbf{\Lambda_g}) ,$$

where $\mathbf{\Lambda_g} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^T + \lambda\mathbf{I}$.

Before proceeding to an example anomaly and background field, we present the definitions of *signal-to-noise ratio* (SNR) and *anomaly-to-background ratio* (ABR):

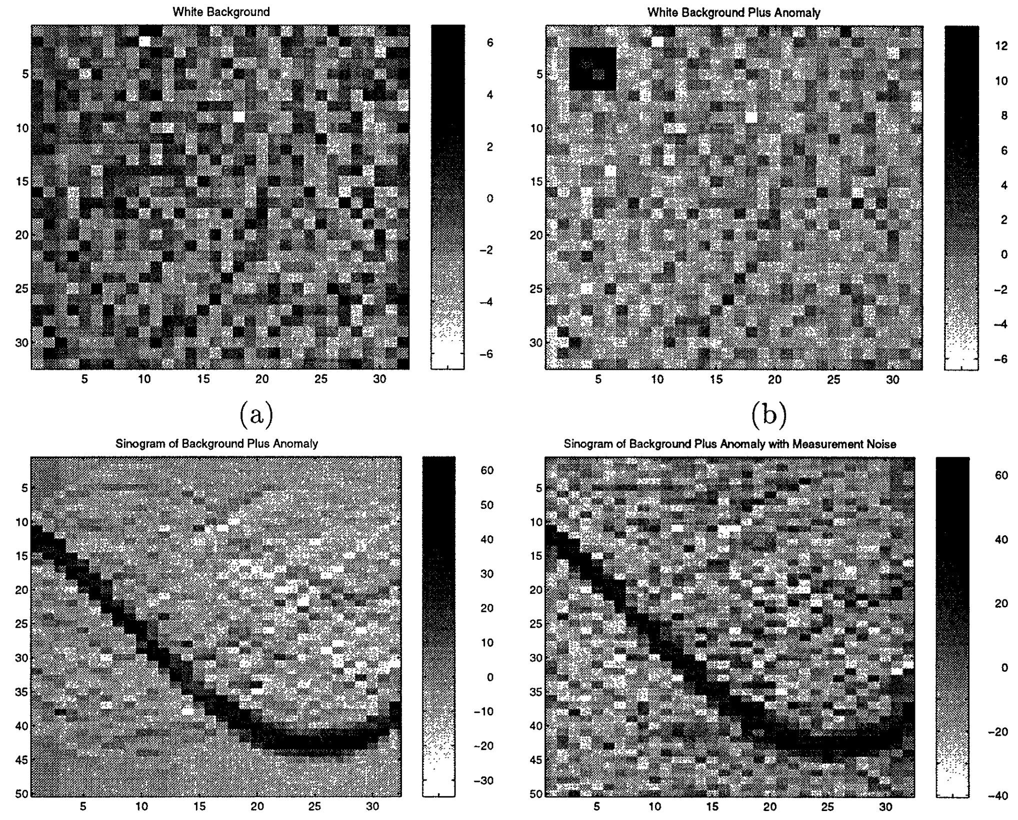
$$\begin{aligned}
\mathrm{SNR(dB)} &\triangleq 10\log\left[\frac{\mathrm{trace}\left(\mathbf{T}\mathbf{\Lambda}\mathbf{T}^T\right)}{\mathrm{trace}\left(\lambda\mathbf{I}\right)}\right], \\
\mathrm{ABR(dB)} &\triangleq 10\log\left[\frac{\mathbf{f}_a^T\mathbf{f}_a}{\mathrm{trace}\left(\mathbf{\Lambda}\right)}\right].
\end{aligned}$$

These two quantities measure the relative power between the projected background and the additive noise and the relative power between the anomaly field and the background, respectively.

Figure 2-11 illustrates an example of the kind of anomaly and background field which are considered in this thesis. The projection of $\mathbf{f}_a + \mathbf{f}_b$ is also shown with and without the addition of noise. Throughout this thesis the domain in which the background and anomalies are defined is referred to as the *image domain* or *object domain.* Figures (a) and (b) of Figure 2-11 are views of the image domain. The range of the projection matrix $\mathbf{T}$ is called the *data domain* or the *sinogram domain.* Figures (c) and (d) of Figure 2-11 are views of the data domain and are referred to as *sinograms.*

Figure 2-11 illustrates some important relationships between the image domain and the sinogram domain. Comparing (a) with (c) we see that one effect of the projection matrix has been to color the white noise present in (a). Comparing (b) with (c) (or (d)), we see that the square anomaly (with local support) has a sinusoidal signature in the data domain (with nonlocal support). For the purposes of illustration the anomaly in this example has a particularly large intensity. The signature of lower intensity anomalies is much more difficult to discern (especially with a fractal instead of white background).

The anomaly detection problem is to determine whether or not $\mathbf{f}_a$ is identically zero. The anomaly localization problem is to determine the values of the size $s \leq s_{max}$ and location (i, j) of the anomaly if indeed one is present. The goal is to solve these problems directly in the data domain. Referring to Figure 2-11, solving these problems in the data domain means that we use the information present in figure (d) directly without first attempting to reconstruct the information information provided by figure (b). Therefore the detection and localization problems are essentially ones of detecting the presence of and determining



(a) (b)

(c) (d)

Figure 2-11: Figure (a) illustrates a zero-mean white Gaussian background field. The pixels are independent and identically distributed with variance about 3. Figure (b) is a superposition of the background shown in (a) and a constant intensity square anomaly near the upper left corner. The anomaly intensity is 10. Figure (c) illustrates the projection of the anomaly plus background field. The horizontal axis is the projection number (there are 32 projections equally spaced between zero and π). The vertical axis is the sample offset (there are 50 samples per projection). In figure (d), zero-mean white Gaussian measurement noise has been added to the projections. The variance of the measurement noise is about 63.

the properties (e.g., the width, amplitude, and phase) of a sinusoidal signature in the data domain. In the next chapter we begin discussing our methods for attacking these problems.

Chapter 3

Performance Bound and Ambiguity Analysis

In this chapter we employ a relatively simple binary hypothesis testing framework to investigate the structures of the anomaly detection and localization problems. Specifically, we derive a performance bound for any anomaly detection algorithm and characterize anomaly ambiguity (i.e., we measure, in some sense, the degree to which an anomaly localization algorithm may confuse the location and scale of the anomaly). The techniques presented here are similar to those applied in [19–21]. We carry the analysis a bit further, however, by studying the dependence of the performance bound and anomaly ambiguity on the background field statistics. We shall show that a change in the background covariance has a drastic effect on the structure of the problem. The dependence of the performance bound and anomaly ambiguity on the background covariance is an exceedingly complicated one, which we attempt to elucidate with the analysis of a simple one-dimensional signal.

This chapter is organized as follows. In Section 3.1 we introduce the principal tool with which we measure detection performance and anomaly ambiguity. In Sections 3.2 and 3.3 we apply this tool to the investigation of a detection performance bound and anomaly ambiguity, respectively. Finally, in Section 3.4 we analyze a simple one-dimensional signal to illustrate the complex way the background field covariance structure enters the problem. Throughout this chapter we rely on the definitions and notation established in Chapter 2.

3.1 The Ambiguity Ellipse

In this section we introduce the principal tool with which we measure detection performance and anomaly ambiguity. This tool is called the *ambiguity ellipse.* The ambiguity ellipse is a special case of the ellipsoid encountered at the end of Section 2.3.1. To derive the form of the ambiguity ellipse, we consider the BHT where each hypothesis, H_0 and H_1, corresponds to the presence of a different anomaly field. That is,

$$\begin{aligned} H_0 &: \mathbf{f}_a = c_0 \mathbf{b}_{s_0,N}(i_0, j_0) = c_0 \mathbf{b}_0 \,, \\ H_1 &: \mathbf{f}_a = c_1 \mathbf{b}_{s_1,N}(i_1, j_1) = c_1 \mathbf{b}_1 \,, \end{aligned}$$

where we have introduced $\mathbf{b}_k$ as shorthand for the indicator function $\mathbf{b}_{s_k,N}(i_k, j_k)$. This formulation includes the special case that one of the hypotheses corresponds to the absence of an anomaly by setting either intensity c_0 or c_1 to zero. Applying equations (2.23) and (2.24), the conditional observations are

$$\begin{aligned} H_0 &: \mathbf{g} = c_0 \mathbf{T} \mathbf{b}_0 + \mathbf{T} \mathbf{f}_b + \mathbf{n} \,, \\ H_1 &: \mathbf{g} = c_1 \mathbf{T} \mathbf{b}_1 + \mathbf{T} \mathbf{f}_b + \mathbf{n} \,. \end{aligned}$$

The data are conditionally Gaussian since they are linear combinations of independent Gaussian random vectors. The background field and noise vector have covariance $\mathbf{\Lambda}$ and $\lambda \mathbf{I}$, respectively; therefore, under either hypothesis the data, $\mathbf{g}$, have covariance

$$\mathbf{\Lambda}_{\mathbf{g}} = \mathbf{T} \mathbf{\Lambda} \mathbf{T}^T + \lambda \mathbf{I} \,.$$

The conditional means of the data differ so that

$$H_0 : \mathbf{g} \sim \mathcal{N}(c_0 \mathbf{T} \mathbf{b}_0, \mathbf{\Lambda}_{\mathbf{g}}) \,,$$

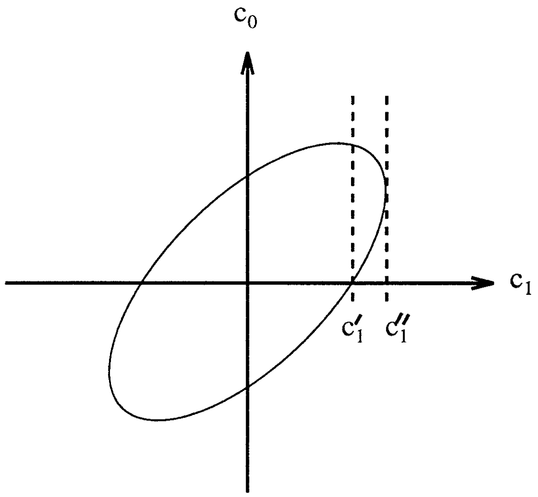


Figure 3-1: This figure shows the ambiguity ellipse. Only the portion of the ellipse in the first quadrant is relevant.

$$H_1 \quad : \quad \mathbf{g} \sim \mathcal{N}(c_1\mathbf{T}\mathbf{b}_1, \mathbf{\Lambda_g}) \, .$$

We have, therefore, satisfied all the conditions assumed in deriving equation (2.22). Specializing equation (2.22) we arrive at an equation for an ellipse in the (c_0, c_1) plane:

$$\delta_1 c_1^2 - 2\delta_{10} c_1 c_0 + \delta_0 c_0^2 - \Pi^2(P_f, P_d) = 0 \, , \tag{3.1}$$

where $\delta_j = \mathbf{b}_j^T \mathbf{T}^T \mathbf{\Lambda_g}^{-1} \mathbf{T} \mathbf{b}_j$ for $j = 1, 2$ and $\delta_{10} = \mathbf{b}_1^T \mathbf{T}^T \mathbf{\Lambda_g}^{-1} \mathbf{T} \mathbf{b}_0$. Since c_1 and c_0 are restricted to be non-negative, only values in the first quadrant are valid. We call this ellipse the *ambiguity ellipse*, an example of which is depicted in Figure 3-1. For all points on or outside this ellipse, the specified (P_f, P_d) performance is achieved or exceeded. For all points within the ellipse, this performance level is not achieved. Two points on the c_1 axis are labeled in Figure 3-1. The significance of the points c_1' and c_1'' will be discussed in Sections 3.2 and 3.3, respectively.

3.2 A Detection Performance Bound

In this section we derive an anomaly detection performance bound and investigate how this bound behaves as a function of background field statistics and anomaly location. To this end we consider the special case of the BHT discussed above. Specifically, we assume that $c_0 = 0$ so that the BHT is a test for the presence of the anomaly $c_1\mathbf{b}_1$. If H_1 is true then this particular anomaly is present. If, on the other hand, H_0 is true then no anomaly is present. In essence, this BHT represents a vastly simplified instance of the full anomaly detection problem for which the anomaly's structure and intensity are *not* known. Since we assume this additional knowledge, the detection performance of this optimal BHT represents a bound on the performance of any more general anomaly detection method.

There are several equivalent ways of measuring the performance of this BHT. One way is to compute P_d as a function of P_f, c_1, and $\mathbf{b}_1$. Equivalently, one could compute the d or d^2 statistic which simply combines P_d and P_f through equation (2.21). The structure of the ambiguity ellipse, however, admits a more elegant measure of performance.

Consider a particular choice of values in the range $[0, 1]$ for P_d and P_f. We call the chosen (P_f, P_d) pair the performance benchmark. Having chosen this benchmark, $\Pi(P_f, P_d)$ is set. Letting $c_0 = 0$, equation (3.1) provides a formula for determining the minimal value of c_1 required to achieve the performance benchmark. This minimal value is denoted c_1' where

$$c_1' = \frac{\Pi(P_f, P_d)}{\sqrt{\mathbf{b}_1^T\mathbf{T}^T\mathbf{\Lambda_g}^{-1}\mathbf{T}\mathbf{b}_1}}. \tag{3.2}$$

The value of c_1' is the value of anomaly intensity for which d is equal to $\Pi(P_f, P_d)$. Higher c_1' values indicate poorer detection performance since a larger anomaly intensity is required to meet the given (P_f, P_d) benchmark. From equation (3.2) it is clear that c_1' depends on, among other things, the data covariance matrix and the anomaly location and size given by $\mathbf{b}_1$. We shall explore these dependencies by considering two data covariance matrices and several families of anomaly indicator functions. It is assumed that the performance benchmark and the projection matrix are fixed. For a given data covariance matrix, $\mathbf{\Lambda_g}$, and (P_f, P_d) performance benchmark, c_1' is a function only of $\mathbf{b}_1$ and it varies with $\mathbf{b}_1$

precisely so that the d statistic is a constant viz., $\Pi(P_f, P_d)$, $\forall \mathbf{b}_1$.

The difference between the two data covariance matrices considered here is that one models a white background (with covariance $\mathbf{\Lambda}_w = \sigma^2\mathbf{I}$) and the other models a fractal background (with covariance $\mathbf{\Lambda}_f$). So the data covariance is either

$$\mathbf{\Lambda}_\mathbf{g} = \mathbf{T}\mathbf{\Lambda}_w\mathbf{T}^T + \lambda\mathbf{I}\,,$$

in the white background case or

$$\mathbf{\Lambda}_\mathbf{g} = \mathbf{T}\mathbf{\Lambda}_f\mathbf{T}^T + \lambda\mathbf{I}\,,$$

in the fractal background case. Throughout this thesis we consider only fractal covariance matrices with a spectral exponent of two (see Appendix A for further details).

For each covariance structure, we consider several families of anomaly structures. Each family corresponds to a dyadic tessellation of the $N \times N$ image domain field (N is assumed to be an integral power of 2 and in the particular case considered here $N = 32$). We let $\mathcal{T}_1$ be the set of all single pixel indicator functions which take on the value one over a pixel of the image field. Similarly, $\mathcal{T}_2$ is the set of indicator functions which are one over 2×2 regions within the image field which, together, tile the field. More generally, the set $\mathcal{T}_k$ consists of $(N/k)^2$ indicator functions, each of which is one over the unique $k \times k$ region in the image field with upper left corner at pixel $(ki+1, kj+1)$ where $i, j \in \{0, 1, 2, \ldots, \frac{N}{k} - 1\}$ and k is a non-negative integral power of 2. The elements of $\mathcal{T}_k$ are, therefore, $\mathbf{b}_{k,N}(ki+1, kj+1)$.

Figure 3-2 illustrates the values of c_1' for four families of anomaly structures with the white background. Figure 3-3 illustrates the values of c_1' for the same four families of anomaly structures but with the fractal background. To understand exactly what is plotted in these figures, consider just one of them. Figure 3-2(c) is a view of the image domain. Each 4×4 region in the image with upper left corner at pixel $(4i + 1, 4j + 1)$ (where $i, j \in \{0, 1, \ldots, 7\}$) corresponds to the element $\mathbf{b}_1 = \mathbf{b}_{4,32}(4i + 1, 4j + 1) \in \mathcal{T}_4$ which is one over that region. The intensity of the region is the value of c_1' associated with that element of $\mathcal{T}_4$ (i.e., the value obtained when $\mathbf{b}_1 = \mathbf{b}_{4,32}(4i + 1, 4j + 1)$ in equation (3.2)). All the

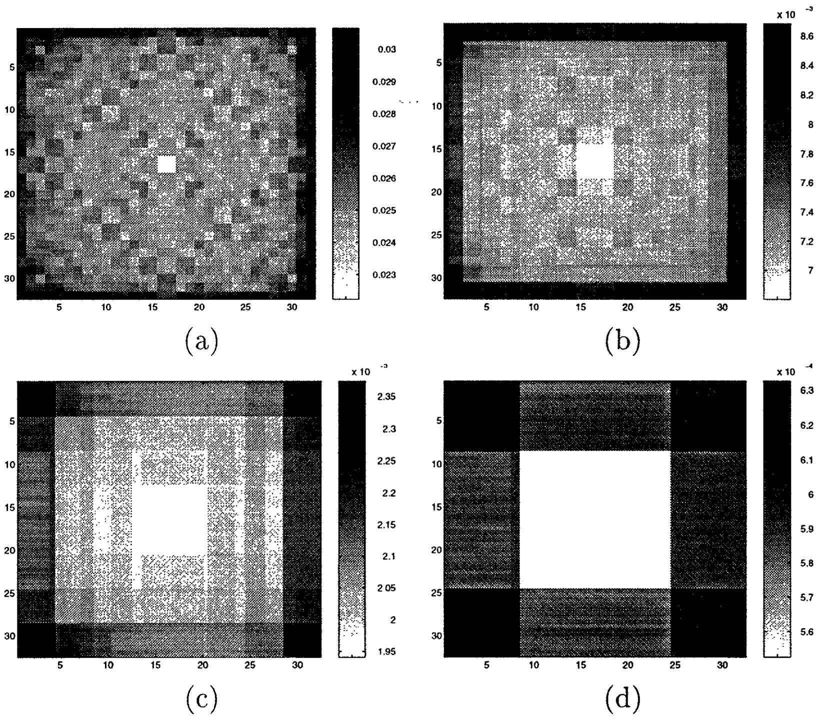
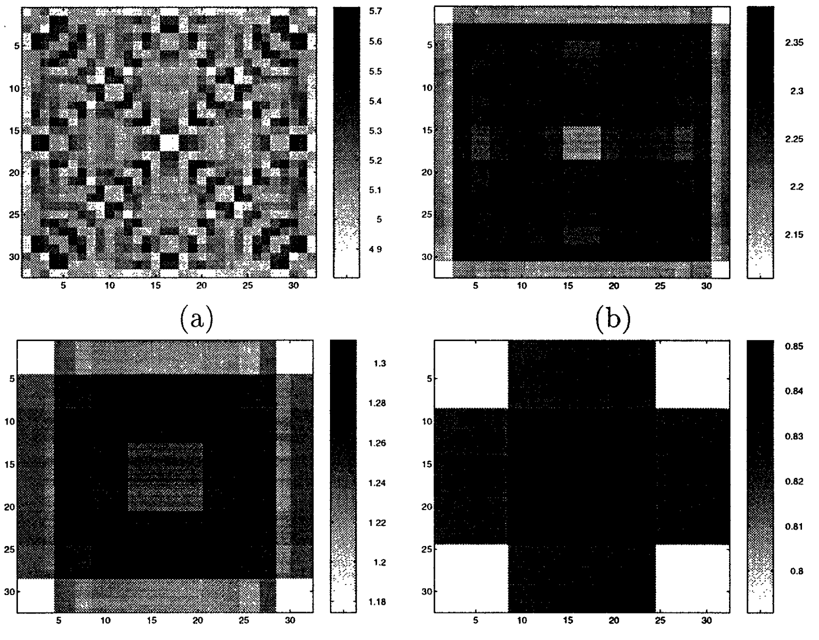


Figure 3-2: Figures (a)–(d) illustrate the values of c_1' for anomalies associated with indicator functions in $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_4$, and $\mathcal{T}_8$, respectively. The background field covariance is white. The SNR is about 3dB and $P_d = 0.95$, $P_f = 0.1$.

other plots in Figures 3-2 and 3-3 are similar.

There are several similarities but also some striking differences between the plots in these two figures. Observe that all plots exhibit quadrantal symmetry. This is due to the symmetry of the data collection, viz., equispaced projections between zero and π. In general, the value of c_1' does not seem to vary in a smooth way from tessellation element to tessellation element. One might expect the value of c_1' at one tessellation element to be between the values of c_1' for its neighboring elements. But this seems not to be the case. For example, consider the plot in Figure 3-3(c). Look along one of the main diagonals of the plot (it makes no difference which one). Notice that the value of c_1' is relatively low at a corner, then increases toward the center, but then decreases again at the center of the plot. This type of variation is most likely due to the detailed structure of the projection matrix (i.e., exactly which pixels are intersected and by how much).



(a) (b) (c) (d)

Figure 3-3: Figures (a)–(d) illustrate the values of c_1' for anomalies associated with indicator functions in $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_4$, and $\mathcal{T}_8$, respectively. The background field covariance is fractal. The SNR is about 3dB and $P_d = 0.95$, $P_f = 0.1$.

The differences between the plots in the Figures 3-2 and 3-3 is startling. The values of c_1' in Figure 3-2 are much smaller than those in Figure 3-3. This indicates that it is easier to detect an anomaly superimposed on a white background than on a colored one which is consistent with intuition. In general, it appears that anomalies at the center of the image are easier to detect than ones at the corners when the background is white. The opposite is true for the fractal background case. And the values of c_1' at the edges are between those at the corner and center. It almost seems that the color map has been reversed from Figure 3-2 to Figure 3-3 (but it has not been).

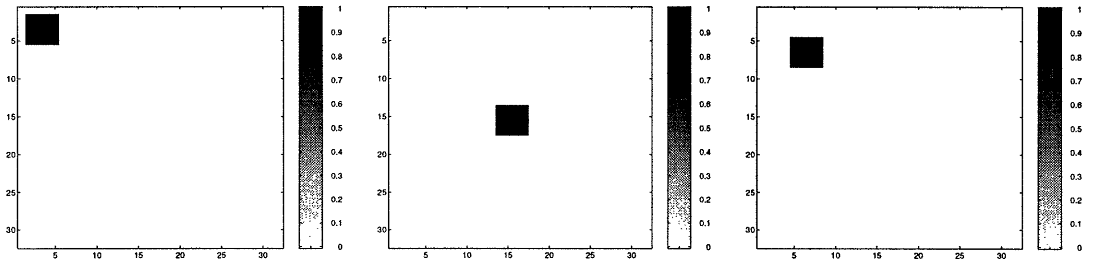
Clearly the structure of the anomaly detection problem is intimately linked to the structure of the background covariance matrix. The nature of this link seems to be quite a complicated one as we illustrate in Section 3.4 with a one-dimensional example. Regardless of the spatial pattern, however, these anomaly intensity values (c_1') are lower bounds for any anomaly detection method with performance benchmark $P_d = 0.95$, $P_f = 0.1$.

3.3 Ambiguity Analysis

In this section we use the ambiguity ellipse to investigate anomaly ambiguity. We wish to measure, in some sense, the degree to which an anomaly associated with indicator function $\mathbf{b}_1$ may be confused with some other structure, $\mathbf{b}_0$, by an anomaly localization algorithm. The results of this ambiguity analysis will indicate whether a scale-recursive spatial zooming technique such as that employed by Miller and Willsky in [19–21] is feasible. (In spatial zooming the anomaly is localized with a multiscale hypothesis test for which the sequence of selected composite hypotheses corresponds to a nested sequence of regions in the image domain.) If anomalies are most confused with structures which spatially overlap or are adjacent to them then a spatial zooming technique seems appropriate. On the other hand, if anomalies are most confused with structures which are spatially disjoint and far away from them then a spatial zooming approach seems dubious. We shall see that, as was the case with the detection bound, the nature of anomaly ambiguity depends critically on the nature of the background covariance structure.

Consider a particular choice of indicator functions $\mathbf{b}_0$ and $\mathbf{b}_1$ and intensities c_0 and c_1. The BHT of Section 3.1 is a test to decide which is present, $c_0\mathbf{b}_0$ or $c_1\mathbf{b}_1$. If the probability of detection is relatively high while the probability of false alarm is relatively low (i.e., the d^2 statistic is high) then the two structures $c_0\mathbf{b}_0$, $c_1\mathbf{b}_1$ are disambiguated to a high degree. On the other hand, if P_d is relatively low and P_f is relatively high (i.e., the d^2 statistic is low) then the two structures are highly confused with one another. Clearly P_f, P_d, $\mathbf{b}_1$, $\mathbf{b}_0$, c_0, and c_1 are not all independent. In our investigation of anomaly ambiguity, we choose a particular (P_f, P_d) performance benchmark and consider several different choices of $\mathbf{b}_1$ and several families of $\mathbf{b}_0$.

The structure specified by $\mathbf{b}_1$ is the anomaly's support. The indicator function $\mathbf{b}_0$ specifies the structure with which we wish to compare the anomaly. Having specified the performance benchmark and indicator functions, the ambiguity ellipse is determined. Referring to Figure 3-1, for all values of the anomaly intensity, c_1, greater than or equal to c_1'' the performance benchmark is achieved or exceeded independent of intensity, c_0, of the comparison structure. Higher values of c_1'' indicates higher degree of ambiguity (lower d^2)



(a) (b) (c)

Figure 3-4: Figure (a) shows $\mathbf{b}_1 = \mathbf{b}_{4,32}(2,2)$ which is a 32×32 field which is zero everywhere except over the 4×4 area with upper left pixel at $(2,2)$ where it is one. Figure (b) shows $\mathbf{b}_1 = \mathbf{b}_{4,32}(14,14)$. Figure (c) shows $\mathbf{b}_1 = \mathbf{b}_{4,32}(5,5)$.

between the structures $\mathbf{b}_0$ and $\mathbf{b}_1$.

The three anomaly structures ($\mathbf{b}_1$) for which we present ambiguity analysis are shown in Figure 3-4. We compare the structures depicted in Figure 3-4(a) and (b) to the families of structures contained in several tessellation sets in Figures 3-5 and 3-6, respectively, with a white background. In Figures 3-7 and 3-8 we compare the structures depicted in Figure 3-4(a) and (b), respectively, to several tessellation sets with a fractal background.

To understand what is plotted in these figures, consider just one of them, Figure 3-5(a), say. Figure 3-5(a) illustrates the ambiguity between the anomaly with support $\mathbf{b}_1 = \mathbf{b}_{4,32}(2,2)$ with all elements of $\mathcal{T}_4$. Each element of $\mathcal{T}_4$ is associated with the value of c_1'' which achieves the performance benchmark $(P_f, P_d) = (0.1, 0.95)$. Let the k^{th} element of $\mathcal{T}_4$ be associated with the value $c_1'' = \alpha_k$. Then Figure 3-5(a) is an image domain view of the function $\sum_{k,\beta_k\in\mathcal{T}_4} \alpha_k\beta_k$. That is, the region of the image domain associated with $\beta_k \in \mathcal{T}_4$ has the value α_k. The other plots are similar but for different tessellation sets and/or anomaly structures.

Figures 3-5, 3-6, 3-7, and 3-8 indicate that spatial zooming may indeed be a reasonable way to localize the anomaly since the small scale anomalies are most confused with larger structures which overlap them for both the fractal and white background cases. Comparing Figures 3-5 and 3-6 with Figures 3-7 and 3-8 we see that, in the presence of a white background, a smaller value of c_1'' is required than in the presence of a fractal background. This indicates that it is easier to distinguish the anomaly structure from other structures

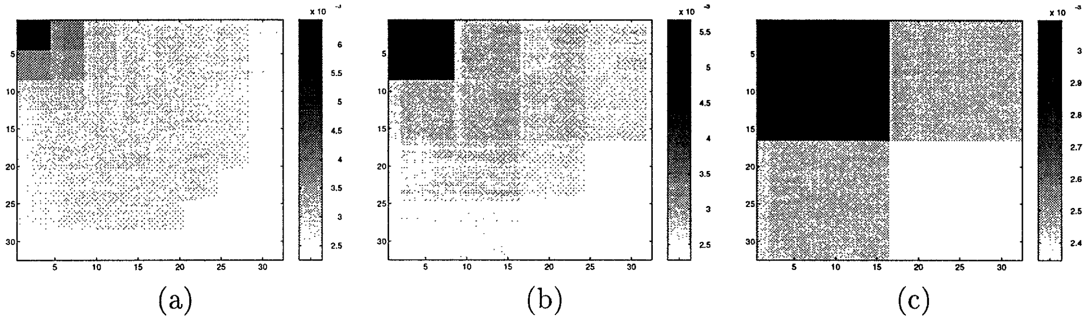


Figure 3-5: Figures (a)–(c) show the value of c_1'' for each structure, $\mathbf{b}_0$, in $\mathcal{T}_4$, $\mathcal{T}_8$, and $\mathcal{T}_{16}$, respectively. The anomaly structure is $\mathbf{b}_1 = \mathbf{b}_{4,32}(2,2)$ and the background is white. The performance benchmark is $(P_f, P_d) = (0.1, 0.95)$ and the SNR is about 3dB.

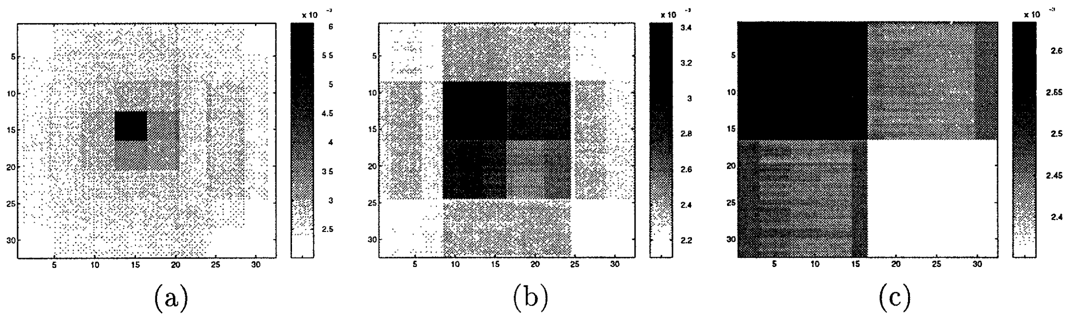


Figure 3-6: Figures (a)–(c) show the value of c_1'' for each structure, $\mathbf{b}_0$, in $\mathcal{T}_4$, $\mathcal{T}_8$, and $\mathcal{T}_{16}$, respectively. The anomaly structure is $\mathbf{b}_1 = \mathbf{b}_{4,32}(14,14)$ and the background is white. The performance benchmark is $(P_f, P_d) = (0.1, 0.95)$ and the SNR is about 3dB.

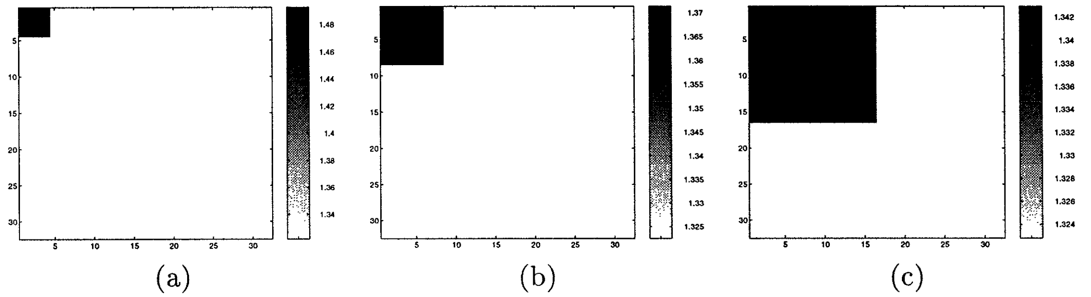
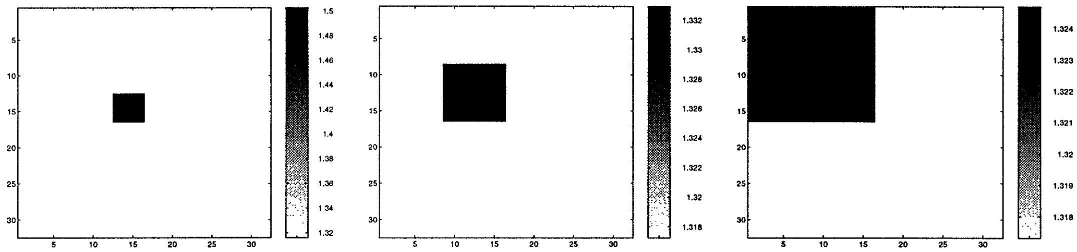


Figure 3-7: Figures (a)–(c) show the value of c_1'' for each structure, $\mathbf{b}_0$, in $\mathcal{T}_4$, $\mathcal{T}_8$, and $\mathcal{T}_{16}$, respectively. The anomaly structure is $\mathbf{b}_1 = \mathbf{b}_{4,32}(2,2)$ and the background is fractal. The performance benchmark is $(P_f, P_d) = (0.1, 0.95)$ and the SNR is about 3dB.



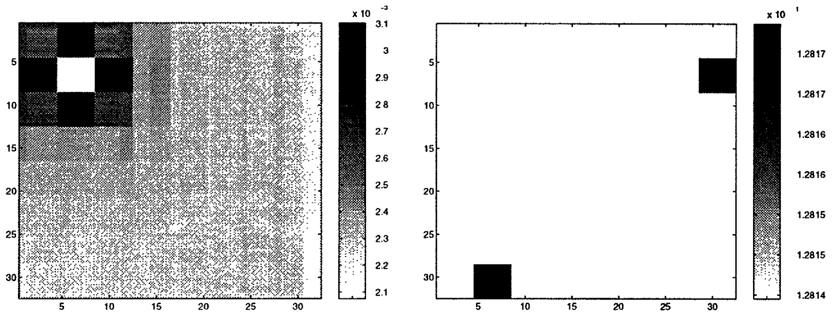
(a) (b) (c)

Figure 3-8: Figures (a)–(c) show the value of c_1'' for each structure, $\mathbf{b}_0$, in $\mathcal{T}_4$, $\mathcal{T}_8$, and $\mathcal{T}_{16}$, respectively. The anomaly structure is $\mathbf{b}_1 = \mathbf{b}_{4,32}(14,14)$ and the background is fractal. The performance benchmark is $(P_f, P_d) = (0.1, 0.95)$ and the SNR is about 3dB.

when the background is white. This is consistent with intuition. For the fractal background cases (Figures 3-7 and 3-8), the dynamic range of the c_1'' values is quite small (see the values on the color bar to the right of the plots). From these plots it seems that there is hardly any variation in ambiguity throughout the image domain.

We have seen that anomalies tend to be most confused with coarser scale structures which overlap them. Another question to consider is whether or not anomalies are most confused with structures which are of the same scale and nearby. Figure 3-9 addresses this question with the anomaly structure of Figure 3-4(c). This anomaly overlaps with exactly one element of $\mathcal{T}_4$ since $\mathbf{b}_{4,32}(5,5)$ is an element of $\mathcal{T}_4$. Figure 3-9 illustrates the ambiguity of this anomaly with all other elements of $\mathcal{T}_4$ both for the white and fractal background cases.

Notice that in the case of the white background the maximal ambiguity (corresponding to the maximum value of c_1'') is local while for the fractal background the maximal ambiguity is with structures which are on opposite sides of the image field. The results for the white background provide further support for a spatial zooming approach since it seems to indicate that anomalies in the same area of the image tend to look more like one another (in the "eyes" of a detection algorithm) than anomalies which are spatially separated. The results for the fractal background, on the other hand, raise some doubt as to whether spatial zooming is appropriate when the background is not white. The implication of Figure 3-9(b) is that distant, not adjacent, anomalies look most similar. If this is indeed true then a



(a) (b)

Figure 3-9: Figure (a) shows the value of c_1'' for each structure, $\mathbf{b}_0$, in $\mathcal{T}_4$ for the anomaly with indicator function $\mathbf{b}_1 = \mathbf{b}_{4,32}(5,5)$ and white background. Figure (b) illustrates the same thing but with a fractal background. The SNR is about 3dB.

spatial zooming algorithm could easily zoom in on an area which is quite far from and does not overlap with the anomaly. On the other hand, the dynamic range of Figure 3-9(b) is very small, indicating that there is no significant variation in ambiguity across the image domain. However, in Chapter 5 we shall see from another point of view that the fractal background does indeed give rise to non-local ambiguity as suggested by Figure 3-9(b).

In the following section we address the curious phenomenon illustrated in Figure 3-9. And in Chapter 5 we shall again consider anomaly ambiguity but from a different perspective.

3.4 Analysis of a One-Dimensional Signal

There are two aspects of the foregoing performance bound investigation and ambiguity analysis which seem puzzling. One is the difference in nature of the plots in Figure 3-2 and Figure 3-3. Recall that these figures illustrate the detection performance bound as measured by the value c_1', the minimum value of the anomaly intensity for which the performance benchmark is achieved. Figure 3-2 illustrates values of c_1' for many anomalies with a white background and Figure 3-2 does so for a fractal background. For the former, anomalies at the center are easier to detect than ones at the corners. The exact opposite is true for the latter. The other is the difference in nature of Figure 3-9(a) and Figure 3-9(b). Recall that these figures illustrate ambiguity (as measured by c_1'', the minimum value of anomaly intensity for which the performance benchmark is achieved independent of the comparison structure's intensity) between the anomaly with indicator function $\mathbf{b}_{4,32}(5,5)$ as compared with structures in the set $\mathcal{T}_4$ for the white and fractal background cases, respectively. For the white case, the maximal ambiguity is local while for the fractal case the maximal ambiguity is non-local.

In this section we aim to answer the question: what is it about the background covariance structure which causes c_1' (performance bound measure) or c_1'' (anomaly ambiguity measure) to be high or low for a particular tessellation element? We seek intuition for these curiosities through the analysis of a one-dimensional signal. The relative simplicity of this signal admits closed form analytical analysis of the dependence of performance and ambiguity on background covariance structure. Even for this simple signal, however, the relationships are complex and, seemingly, without pattern. This suggests that an analytical analysis of these relationships for the full two-dimensional tomography problem would be extremely difficult or impossible. We do not attempt such an analysis here.

In the next section we introduce the one-dimensional problem and an associated binary hypothesis test. The structure of this one-dimensional problem is similar to that of the two-dimensional tomography problem. Therefore, many intermediate steps and explanatory comments are omitted. In Sections 3.4.2 and 3.4.3 we investigate the relationship between the background covariance and detection performance and ambiguity, respectively.

3.4.1 A One-Dimensional Problem

Let $\mathbf{x}$ be a length N vector which is composed of the sum of a statistically known background $\mathbf{x}_b$ and an unknown deterministic signal (anomaly) $\mathbf{x}_a$. That is,

$$\mathbf{x} = \mathbf{x}_a + \mathbf{x}_b \,.$$

The background is a zero-mean Gaussian random vector with known covariance:

$$\mathbf{x}_b \sim \mathcal{N}(0, \mathbf{P})\,.$$

It is assumed that $\mathbf{P}$ is Toeplitz (not circulant) so that $\mathbf{x}_b$ is wide sense stationary (WSS) but not necessarily periodic. The anomaly has constant intensity over part of the vector and is zero elsewhere. We write $\mathbf{x}_a = c\mathbf{b}_{s,N}(j)$, where c is the intensity and $\mathbf{b}_{s,N}(j)$ is the anomaly indicator function. It takes on the value one over the length s portion of the length N vector beginning at position j and is zero elsewhere. For example,

$$\mathbf{b}_{2,5}(2) = [0\ 1\ 1\ 0\ 0]^T\ .$$

Consider the noisy observation of the vector $\mathbf{x}$:

$$\begin{aligned}\mathbf{y} &= \mathbf{x} + \mathbf{n} \\ &= \mathbf{x}_a + \mathbf{x}_b + \mathbf{n} \\ &= c\mathbf{b}_{s,N}(j) + \mathbf{x}_b + \mathbf{n}\,.\end{aligned}$$

The vector $\mathbf{n}$ represents zero-mean additive white Gaussian noise which is independent of the background and anomaly vectors and has covariance $\lambda\mathbf{I}$. Therefore,

$$\mathbf{y} \sim \mathcal{N}(c\mathbf{b}_{s,N}(j), \mathbf{P} + \lambda\mathbf{I})\,.$$

We define $\mathbf{P}_\mathbf{y} \triangleq \mathbf{P} + \lambda\mathbf{I}$.

Consider a binary hypothesis test to decide which is true

$$\begin{aligned}H_0 &: \ \mathbf{x}_a = c_0\mathbf{b}_{s_0,N}(j_0)\,, \\ H_1 &: \ \mathbf{x}_a = c_1\mathbf{b}_{s_1,N}(j_1)\,.\end{aligned}$$

The Neyman-Pearson optimal decision rule for such a binary hypothesis testing problem is the log-likelihood ratio test (LRT) which has the form,

$$\ell(\mathbf{Y}) \triangleq (c_1\mathbf{b}_{s_1,N}(j_1) - c_0\mathbf{b}_{s_0,N}(j_0))^T\mathbf{P_y}^{-1}\mathbf{Y} \underset{H_0}{\overset{H_1}{\gtrless}} \gamma\,,$$

The log-likelihood ratio function (LRF), $\ell(\mathbf{Y})$, is conditionally Gaussian and has the same variance (but different mean) under each hypothesis. Specializing equation (2.20), the d^2 statistic is

$$\begin{aligned} d^2 &= \frac{(E(\ell|H_1) - E(\ell|H_0))^2}{\sigma_\ell^2} \\ &= c_1^2\mathbf{b}_1^T\mathbf{P_y}^{-1}\mathbf{b}_1 + c_0^2\mathbf{b}_0^T\mathbf{P_y}^{-1}\mathbf{b}_0 - 2c_0c_1\mathbf{b}_0^T\mathbf{P_y}^{-1}\mathbf{b}_1\,, \end{aligned} \tag{3.3}$$

where we have made the notational simplification $\mathbf{b}_k = \mathbf{b}_{s_k,N}(j_k)$.

Since the d^2 statistic completely characterizes performance, we may use it to analyze performance and anomaly ambiguity. In the following sections we relate the d^2 statistic to the elements of the background covariance matrix $\mathbf{P}$ for a length three ($N = 3$) signal. For such a signal, the background covariance matrix, $\mathbf{P}$, and inverse data covariance matrix, $\mathbf{P_y}^{-1}$, are related as follows. Define

$$\mathbf{P} \triangleq \begin{pmatrix} p_1 & p_2 & p_3 \\ p_2 & p_1 & p_2 \\ p_3 & p_2 & p_1 \end{pmatrix}.$$

Therefore,

$$\mathbf{P_y}^{-1} = \begin{pmatrix} p_1+\lambda & p_2 & p_3 \\ p_2 & p_1+\lambda & p_2 \\ p_3 & p_2 & p_1+\lambda \end{pmatrix}^{-1} \triangleq \frac{1}{|\mathbf{P_y}|}\begin{pmatrix} a_1 & a_3 & a_4 \\ a_3 & a_2 & a_3 \\ a_4 & a_3 & a_1 \end{pmatrix}.$$

The elements of $\mathbf{P_y} = \mathbf{P} + \lambda \mathbf{I}$ may be related to those of $\mathbf{P_y}^{-1}$ using Cramer's rule. Doing so yields

$$\begin{aligned} a_1 &= p_1^2 - p_2^2 + 2p_1\lambda + \lambda^2 \,, & (3.4)\\ a_2 &= p_1^2 - p_3^2 + 2p_1\lambda + \lambda^2 \,, & (3.5)\\ a_3 &= p_2p_3 - p_1p_2 - p_2\lambda \,, & (3.6)\\ a_4 &= p_2^2 - p_1p_3 - p_3\lambda \,. & (3.7) \end{aligned}$$

3.4.2 Detection Performance

As in the two-dimensional tomography problem, we investigate detection performance by setting $c_0 = 0$. Rather than selecting a performance benchmark and studying the behavior of c_1 for various anomaly structures, $\mathbf{b}_1$, we instead consider the d^2 statistic which is an equivalent performance measure. In this section, we fix $c_1 = 1$ and interpret high values of d^2 as indications of higher detection performance and low values of d^2 as indications of lower detection performance.

We are interested in understanding what causes d^2 to be relatively high or relatively low. Specifically, we wish to study how the structure of d^2 (as a function of $\mathbf{b}_1$) depends on the background covariance matrix, $\mathbf{P}$. To do so we consider a special class of anomaly indicator functions, viz., the set $\mathcal{S} \triangleq \{\mathbf{b}_{1,3}(j)\}$, for $j \in \{1,2,3\}$. The set $\mathcal{S}$ is the set of all length three vectors which take the value one over one element and are zero elsewhere. Since, throughout this section and the next, our indicator functions will always be from the set $\mathcal{S}$, we may drop the subscript $1,3$ and write the anomaly indicator function as $\mathbf{b}_1(j)$. (In the next section, $c_0 \neq 0$ and we shall use the analogous notation $\mathbf{b}_0(j)$ for the comparison structure since it also will always be an element of $\mathcal{S}$.)

The d^2 statistic, therefore, is only a function of j, the position of the anomaly. Thus, using equations (3.3) and (3.4)–(3.7), we write

$$\begin{aligned} d^2(j) &= \mathbf{b}_1^T(j)\mathbf{P_y}^{-1}\mathbf{b}_1(j) \\ &= \begin{cases} \frac{a_1}{|\mathbf{Py}|} & \text{if } j = 1,3 \\ \frac{a_2}{|\mathbf{Py}|} & \text{if } j = 2 \end{cases} \\ &= \begin{cases} \frac{p_1^2 - p_2^2 + p_1\lambda + \lambda^2}{|\mathbf{Py}|} & \text{if } j = 1,3 \\ \frac{p_1^2 - p_3^2 + 2p_1\lambda + \lambda^2}{|\mathbf{Py}|} & \text{if } j = 2 \end{cases} . \end{aligned}$$

Having expressed the d^2 statistic in terms of the elements of the background covariance matrix, we can readily understand what, precisely, determines its structure. We can state under what conditions a particular anomaly ($\mathbf{b}_1(j)$) is easier to detect than another. Specifically, $\mathbf{b}_1(2)$ is easier to detect than either $\mathbf{b}_1(1)$ or $\mathbf{b}_1(3)$ if and only if $d^2(2)$ is larger than either $d^2(1)$ or $d^2(3)$ which is true exactly when p_2 is larger than p_3. Notice that neither the background variance, p_1, nor the noise intensity, λ, play a role in the relative structure of d^2.

3.4.3 Ambiguity

To investigate anomaly ambiguity, we imagine that the anomaly structure $\mathbf{b}_1 \in \mathcal{S}$ is set to $\mathbf{b}_1(1)$ and we wish to determine to what degree (as measured by the d^2 statistic) the two other other structures in $\mathcal{S}$ are confused with it. That is, we are interested in investigating the structure of d^2 as a function of $\mathbf{b}_0 \in \mathcal{S} \setminus \mathbf{b}_1(1)$ and we also wish to relate this structure to the background covariance matrix $\mathbf{P}$. A smaller d^2 value indicates a higher ambiguity. The converse holds for higher d^2 values. Throughout this section we assume that $c_1 = c_0 = 1$.

There are only two possible d^2 values in this problem since $N = 3$ and we have fixed $\mathbf{b}_1 = \mathbf{b}_1(1)$. One corresponds to the d^2 statistic using $\mathbf{b}_1(1)$ and $\mathbf{b}_0(2)$. Using equations (3.3) this d^2 statistic has the form

$$d^2 = d_{1,2}^2 \triangleq a_2 + a_1 - 2a_3 ,$$

where we have included the subscript 1, 2 to remind us that this value corresponds to the binary hypothesis test between an anomaly at the first element and the second element. The remaining d^2 statistic measures the performance of the binary hypothesis test between $\mathbf{b}_1(1)$ and $\mathbf{b}_0(3)$:

$$d^2 = d_{1,3}^2 \triangleq a_1 + a_1 - 2a_4 \,.$$

If $d_{1,3}^2 > d_{1,2}^2$ then the anomaly (at the first entry) is harder to distinguish from the structure which is one at the second entry than the structure which is one at the third entry. In other words, the maximal ambiguity is local (i.e., it is adjacent to the anomaly). On the other hand, if $d_{1,3}^2 < d_{1,2}^2$ then the opposite holds and the maximal ambiguity is nonlocal. For flat ambiguity (neither local nor nonlocal) we require the two d^2 values to be equal. We define

$$f(p_1, p_2, p_3, \lambda) \triangleq d_{1,2}^2 - d_{1,3}^2 = 3p_2^2 - p_3^2 - 2p_2p_3 + 2p_1p_2 + 2p_2\lambda - p_1p_3 - 2p_3\lambda \,.$$

We have the following conditions on the elements of the data covariance matrix, $\mathbf{P_y}$: for nonlocal maximal ambiguity we require that $f(p_1, p_2, p_3, \lambda) > 0$. This is satisfied, for example, with $p_1 = 3$, $p_2 = 3$, $p_3 = 1$, and $\lambda = 1$. For local maximal ambiguity we require that $f(p_1, p_2, p_3, \lambda) < 0$. This is satisfied, for example, with $p_1 = 4$, $p_2 = 2$, $p_3 = 3$, and $\lambda = 1$. For flat ambiguity we require that $f(p_1, p_2, p_3, \lambda) = 0$. This is achieved, for example, with $p_1 = p_2 = p_3 = \lambda = 1$. (Note: it is not enough just to satisfy the sign constraint on $f(p_1, p_2, p_3, \lambda)$, but the p_i and λ values must also form a valid invertible covariance matrix, $\mathbf{P_y}$—a matrix which is symmetric positive definite. The values provided above also satisfy this additional constraint.)

We see that, for this small, structured, one-dimensional problem, any kind of ambiguity may arise by proper choice of the background covariance matrix. While it is difficult to analyze in as precise a manner as we have done here for this small one-dimensional problem, we have seen that the same is true in the two-dimensional tomography problem. What this one-dimensional case suggests, however, is that a great variety of background covariance matrices (not just, say, fractal) satisfying certain conditions gives rise to nonlocal ambiguity.

This, no doubt, holds in higher dimensions as well.

In the foregoing analysis we have assumed direct measurements of the underlying vector $\mathbf{x}$. The measurement model we used does not contain a term analogous to the projection matrix in the two-dimensional anomaly problem. This suggests that if we were to replace the projection matrix with the identity matrix and repeat the performance bound and anomaly ambiguity analysis we would continue to witness the variety of "strange" behavior which we saw in Sections 3.2 and 3.3. Indeed, we have performed a few preliminary experiments which indicate that non-local maximal ambiguity may arise without the projection matrix. We have not conducted enough analysis on this direct measurement case to conclude under what precise conditions non-local maximal ambiguity is achieved. This remains an open problem.

Chapter 4

Simple Approach to Detection and Localization

In Chapter 2 we introduced the concept of a multiscale hypothesis test (MSHT) as an efficient but suboptimal alternative to the optimal M-ary hypothesis test (MHT). In this chapter we apply MSHTs to the anomaly detection and localization problems. In Section 4.1 we formulate these problems as MHT problems. These MHTs require the consideration of a prohibitively large number of hypotheses even for the simple class of anomalies considered in this thesis. Therefore, in Section 4.2 we develop two types of MSHTs. These MSHTs are applied to several instances of the anomaly detection and localization problems in Section 4.3. In Section 4.4 we compare the computational complexity of the optimal MHT with several MSHT formulations.

4.1 M-ary Hypothesis Testing Problem Formulation

In this section we formulate the optimal MHT for the anomaly detection and localization problems. This optimal test includes one hypothesis for every possible anomaly size, position, and intensity as these are the only unknown parameters associated with anomalies considered in this thesis. As mentioned in Section 2.5, we assume that the anomaly has the form

$$\mathbf{f}_a = c\mathbf{b}_{s,N}(i,j)\,,$$

where c is the unknown non-negative anomaly intensity and $\mathbf{b}_{s,N}(i,j)$ represents the field which is zero everywhere except over the $s \times s$ area with upper left corner at pixel (i,j) where it is one. The exact size, s, and location, (i,j), of this square anomaly are unknown.

While $c \in I\!\!R_+ \cup \{0\}$, s, i, and j are elements of finite sets:

$$\begin{aligned} s &\in \{1, 2, \ldots, s_{max}\} \overset{\triangle}{=} \mathcal{S}\,, \\ i,j &\in \{1, 2, \ldots, N-s+1\} \overset{\triangle}{=} \mathcal{J}_s\,. \end{aligned}$$

If we restrict the range of c to the finite set

$$c \in \{a_1, a_2, \ldots, a_{N_c}\} \overset{\triangle}{=} \mathcal{C}$$

then an optimal MHT may be formulated as follows. Let H_k represent the hypothesis that $\mathbf{f}_a = c_k\mathbf{b}_{s_k,N}(i_k,j_k)$ where $c_k \in \mathcal{C}$, $s_k \in \mathcal{S}$, and $i_k, j_k \in \mathcal{J}_s$. For each allowable anomaly size and intensity there is one hypothesis for every possible anomaly location. We shall find it convenient to use the simplified notation

$$H_k : \mathbf{f}_a = c_k\mathbf{b}_k\,,$$

where $\mathbf{b}_k \overset{\triangle}{=} \mathbf{b}_{s_k,N}(i_k,j_k)$. Therefore, the conditional probability function for the data is Gaussian (cf., equation (2.5)):

$$H_k : \mathbf{g} \sim \mathcal{N}(c_k\mathbf{T}\mathbf{b}_k, \mathbf{\Lambda}_\mathbf{g})\,,$$

where $\mathbf{\Lambda}_\mathbf{g} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^T + \lambda\mathbf{I}$, and $\mathbf{\Lambda}$ is the covariance matrix for the background field $\mathbf{f}_b$.

Letting H_0 represent the hypothesis that there is no anomaly (i.e., that $\mathbf{f}_a \equiv 0$), the log-likelihood ratio functions (LRFs) have the form

$$\begin{aligned} \ell_k(\mathbf{G}) &\triangleq \ln\left[L_k(\mathbf{G})\right] \\ &= \ln\left[\frac{p_{\mathbf{g}}(\mathbf{G};H_k)}{p_{\mathbf{g}}(\mathbf{G};H_0)}\right] \\ &= c_k(\mathbf{T}\mathbf{b}_k)^T\mathbf{\Lambda}_{\mathbf{g}}{}^{-1}\mathbf{G} - \frac{1}{2}c_k^2(\mathbf{T}\mathbf{b}_k)^T\mathbf{\Lambda}_{\mathbf{g}}{}^{-1}(\mathbf{T}\mathbf{b}_k)\,, \end{aligned}$$

where $\mathbf{G}$ is the particular observed realization of the random vector $\mathbf{g}$. The optimal decision rule, $h(\mathbf{G})$, is the log-likelihood ratio test

$$h(\mathbf{G}) = \begin{cases} H_0 & \text{if } \max_j \ell_j(\mathbf{G}) \le \eta \\ H_i & \text{if } \max_j \ell_j(\mathbf{G}) > \eta \text{ where } i = \arg\max_j \ell_j(\mathbf{G}) \end{cases} .$$

If c is not restricted to a finite set as above, then suboptimal methods must be used. The generalized likelihood ratio test (GLRT) described in Section 2.3.2 is one such method. Another possible approach is to estimate the anomaly field $\mathbf{f}_a$ from the data $\mathbf{g}$. The maximum likelihood (ML) estimator is considered in Appendix B and it is shown that such an estimator is not well suited to the problems posed in this thesis.

4.2 Multiscale Hypothesis Testing Formulations

A multiscale hypothesis test possesses three main high level characteristics: the form of the covers (the composite hypotheses), the form of the statistics, and the form of the decision structure. In this section we formulate MSHTs for the anomaly detection and localization problems. The form of the covers and statistics we use are motivated by the work of Miller and Willsky in [19–21] and are chosen to effect spatial zooming. In particular, the composite hypotheses of the MSHTs introduced in this chapter are associated with spatially contiguous regions of the image domain. The particular statistics we use have an interpretation as log-likelihood ratios and are, in some sense, intuitively natural. They are not, in any sense, optimal. We discuss optimal statistics in Chapter 5. The decision

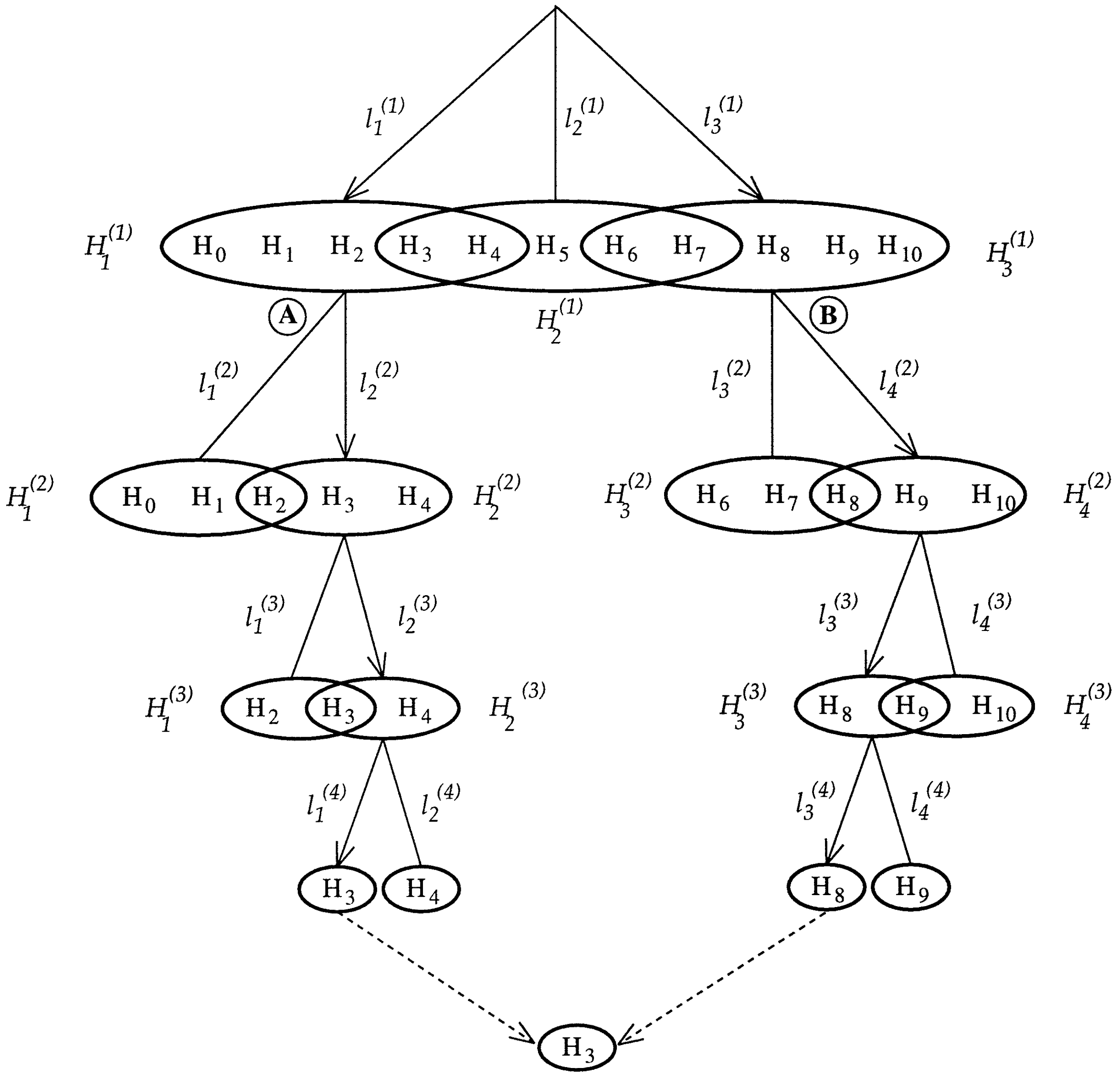


Figure 4-1: This figure illustrates a multiscale hypothesis test with a decision structure which is different from that of Figure 2-8. The selected composite hypotheses are indicated with directed arcs. The MSHT tree terminates after four scales and the dashed lines indicate the post-processing stage which selects one of the remaining N_{leaf} candidate hypotheses.

structure reflects how the covers and associated statistics are used to test for anomaly presence and to deduce localization. Figure 2-8 illustrates one particular decision structure: one composite hypothesis, $\mathcal{H}_i^{(k)}$, is selected at each scale k.

Figure 4-1 illustrates another decision structure in which two of three composite hypotheses are selected at the coarsest scale: $\mathcal{H}_1^{(1)}$ and $\mathcal{H}_3^{(1)}$ are selected. These two composite hypotheses are then considered independently in subtrees A and B, respectively. At each scale in each of these subtrees only one of two composite hypotheses is selected. The scale recursive selection continues until only one hypothesis remains at the bottom of each subtree. The processing in subtrees A and B ends with the choice of hypothesis H_3 and H_8, respectively. All other hypotheses have been discarded. The final processing stage is indicated with dashed lines and compares these two hypotheses. In this example, H_3 is selected and H_8 is discarded. Alternatively, the scale recursion may be terminated at a coarser scale in which case one of the *composite* hypotheses is ultimately selected. For example, if in Figure 4-1 the scale recursion were terminated at scale $k = 3$ then either $\mathcal{H}_2^{(3)}$ or $\mathcal{H}_3^{(3)}$ will be selected in a post-processing stage.

A general decision structure selects $r^{(k)}$ composite hypotheses at each scale k where $r^{(k)} \in \{1, 2, \ldots, N^{(k)}\}$ and $N^{(k)}$ is the number of composite hypotheses at scale k. Each selected composite hypothesis spawns a subtree and the finer scale processing along a subtree is independent of processing along other subtrees (e.g., the processing along subtrees A and B in Figure 4-1 are independent of one another and may be done in parallel). At the finest scale of the *entire* MSHT tree there are N_{leaf} hypotheses which have not been discarded (e.g., in Figure 4-1 $N_{leaf} = 2$ since H_3 and H_8 remain at the bottom of the tree). The N_{leaf} hypotheses are called *candidates*, one of which is selected in a post-processing stage.

In this section we develop two types of MSHTs for the anomaly detection and localization problems. These two types differ in their decision structure. One aspect of this difference is how the two algorithms conduct the test for the presence of an anomaly (the detection test). Another difference is the number of composite hypotheses selected at each scale (this relates to localization determination). We shall first discuss the *single candidate (SC) algorithm* which tests for the presence of an anomaly at the coarsest scale before attempting localization. Then, localization is only attempted if it is determined that an anomaly exists. The SC algorithm, like the example depicted in Figure 2-8, selects one composite hypothesis per scale and, thus, terminates with a single candidate hypothesis. In contrast, the *multiple candidate (MC) algorithm*, like the example depicted in Figure 4-1, terminates with several candidates, at most one of which is selected in a post-processing stage. In the MC algorithm, a detection is *not* conducted before localization as is done in the SC algorithm. Instead, detection is conducted at a fine scale by comparing the log-likelihood ratio function values associated with the N_{leaf} selected regions with a threshold. In other words the MC algorithm assumes that an anomaly exists and then computes many (N_{leaf}) estimates of the anomaly's location (candidates) using a MSHT. The most likely candidate is then selected only if the LRF values are sufficiently high. If none is sufficiently high then it is determined that no anomaly exists.

4.2.1 Single Candidate Algorithm

As emphasized in Chapter 2, the scale recursion of a MSHT need not have any interpretation in a spatial domain. As a starting point for our application of MSHTs to the anomaly detection and localization problems, however, we consider MSHTs which *do* have a spatial zooming interpretation. We begin with a discussion of the composite hypothesis test conducted at the coarsest scale of the SC MSHT. The processing at other scales is similar.

Recall that the hypothesis $H_k \in \mathcal{H} \stackrel{\triangle}{=} \{H_0, \ldots, H_{M-1}\}$ is associated with the indicator function $\mathbf{b}_k \stackrel{\triangle}{=} \mathbf{b}_{s_k,N}(i_k, j_k)$ and the anomaly intensity c_k. We make two additional assumptions. The first is that $s_k \in \{1, 2, \ldots, s_{max}\}$ where s_{max}, the maximum possible size of the anomaly, is much less than N, the linear dimension of the image domain field. The value of s_{max} is known. The second is that c_k is known and is independent of k. In Appendix C we show that this latter assumption results in no loss of generality.

Figure 4-2 provides an image domain interpretation of the composite hypotheses at the coarsest scale. There are four composite hypotheses: $\mathcal{H}_i^{(1)}$, for $i \in \{1, 2, 3, 4\}$. Each composite hypothesis corresponds to a square $s_{hyp} \times s_{hyp}$ region of the image domain as shown. We associate each composite hypothesis with an indicator function $\mathbf{b}_i^{(1)}$ which is one over the $s_{hyp} \times s_{hyp}$ region corresponding to $\mathcal{H}_i^{(1)}$ and zero elsewhere. The hypothesis H_k belongs to composite hypothesis $\mathcal{H}_i^{(1)}$ if and only if $\mathbf{b}_k^T \mathbf{b}_i^{(1)} = s_k^2$. For example, composite hypothesis $\mathcal{H}_1^{(1)}$ corresponds to the shaded region in Figure 4-2. All and only hypotheses associated with anomalies with support *entirely* within this shaded region belong to $\mathcal{H}_1^{(1)}$.

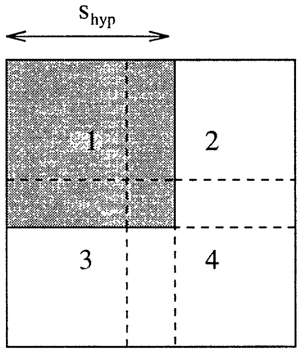


Figure 4-2: The field is divided into four subdivisions at the coarsest scale. Subdivision i corresponds to $\mathcal{H}_i^{(1)}$. Subdivision 1 is shown shaded and with a solid border; the other subdivisions are unshaded with dashed borders. The subdivisions overlap so that the chosen anomalous subdivision contains the entire anomaly.

The composite hypothesis regions overlap by at least $s_{max} - 1$ pixels so that each possible anomaly lies entirely within at least one region.

The composite hypothesis test conducted at the coarsest scale selects one of $\mathcal{H}_0^{(1)}$, $\mathcal{H}_1^{(1)}$, $\mathcal{H}_2^{(1)}$, $\mathcal{H}_3^{(1)}$, and $\mathcal{H}_4^{(1)}$ where

$$\begin{aligned}\mathcal{H}_0^{(1)} &: \text{ no anomaly}\,, \\ \mathcal{H}_i^{(1)} &: \text{ anomaly has support in region } i\,.\end{aligned}$$

One of these composite hypotheses is selected on the basis of a comparison of four LRF statistics, $\ell_1^{(1)}$, $\ell_2^{(1)}$, $\ell_3^{(1)}$, $\ell_4^{(1)}$. These statistics are derived by imagining that the composite hypothesis test is really the test

$$\mathcal{H}_0^{(1)} \;:\; \mathbf{f}_a \equiv 0\,, \tag{4.1}$$

$$\mathcal{H}_i^{(1)} \;:\; \mathbf{f}_a = c\mathbf{b}_i^{(1)}\,, \tag{4.2}$$

where $i \in \{1, 2, 3, 4\}$.

In this test, the null hypothesis, $\mathcal{H}_0^{(1)}$, is that no anomaly exists and composite hypothesis $\mathcal{H}_i^{(1)}$ for $i \in \{1,2,3,4\}$ is that the anomaly has intensity c and support over the *entire* region i indicated in Figure 4-2. Note that, in general, there does not exist a hypothesis $H_k \in \mathcal{H}$ which corresponds to an anomaly with support over the entire region i (the overlapping ensures this). While $\mathcal{H}_i^{(1)}$ does not exactly match any $H_k \in \mathcal{H}_1^{(1)}$, the hope is that the statistic derived from this formulation may be interpreted as the likelihood that the associated region contains the anomaly.

A log-likelihood ratio test is used to choose which one of the five hypotheses is most likely true. The form of this test is

$$i = \begin{cases} 0 & \text{for } \max_j \left[\ell_j^{(1)}(\mathbf{G})\right] < \eta^{(1)} \\ \arg\max_j \left[\ell_j^{(1)}(\mathbf{G})\right] & \text{otherwise} \end{cases} ,$$

where

$$\ell_j^{(1)}(\mathbf{G}) = c(\mathbf{T}\mathbf{b}_j^{(1)})^T \mathbf{\Lambda_g}^{-1}\mathbf{G}\,, \tag{4.3}$$

and $j \in \{1,2,3,4\}$, $\eta^{(1)}$ is a constant threshold. So, after this coarse-scale test, either no subdivision is selected (i.e., $\mathcal{H}_0^{(1)}$ is chosen) or one of the four overlapping regions of Figure 4-2 is selected as the anomalous subdivision. The statistics $\ell_j^{(1)}$ are the log-likelihood ratio functions associated with the hypothesis test specified by equations (4.1) and (4.2). In deriving equation (4.3), we have used the fact that

$$(\mathbf{b}_i^{(1)})^T \mathbf{T}^T \mathbf{\Lambda_g}^{-1} \mathbf{T}\mathbf{b}_i^{(1)} = (\mathbf{b}_j^{(1)})^T \mathbf{T}^T \mathbf{\Lambda_g}^{-1} \mathbf{T}\mathbf{b}_j^{(1)}$$

for all $i, j \in \{1,2,3,4\}$. This follows from the symmetry of the composite hypothesis regions, the fact that we have a complete set of data, and the wide sense stationarity of $\mathbf{g}$. Note, however, that this type of relation need not hold at subsequent scales since the composite hypothesis regions do not have the requisite symmetry at other than the coarsest scale. Despite this, we have found that, to a good degree of approximation, such a relation holds at all scales. Notice that the scalar c appears in all the statistics $\ell_j^{(1)}(\mathbf{G})$ which are compared. Since the result of a comparison is not affected by the value of c, we may remove it from all of the $\ell_j^{(1)}(\mathbf{G})$ statistics.

The composite hypothesis test at subsequent scales is similar to the one just specified for the coarsest scale. At the second scale, for example, the region associated with the composite hypothesis selected at the first scale is subdivided in the same way that the entire image domain was subdivided at the coarsest scale—with four overlapping squares. Each square is associated with a composite hypothesis, one of which is selected on the basis of a log-likelihood ratio test. This scale-recursive, decision-directed process continues until the anomaly (if any) is localized.

One difference between the first and subsequent levels of the SC algorithm is that the null hypothesis takes on a slightly different meaning after the first step. In the first step, if the null hypothesis is chosen then it is assumed no anomaly exists in the field and no further processing is conducted—the search ends. If the null hypothesis is not declared in the first step then an anomaly is assumed to exist and further processing is done. Therefore, in subsequent steps, when the null hypothesis is declared it does not mean that no anomaly exists. It means that none of the corresponding log-likelihood ratio function values are larger than the threshold at that step. Possible interpretations of such a result include: (1) the anomaly is larger than any of the subregions being tested; (2) the data do not warrant finer scale localization. (Note that a slightly different algorithm could include a "no anomaly" hypothesis at every scale so that the detection determination can be deferred to finer scales.)

The following is high level pseudo-code for the SC algorithm. The inputs to the algorithm are the region to investigate (initialized to the entire image domain) and the scale number (initialized to one). The output is the region corresponding to the estimate of the anomaly's support.

```
Algorithm 1 (Single Candidate)
  $H = SingleCandidate(R, scale)$
  Step 1. If scale is the maximum possible scale, $H = R$, stop.
  Step 2. Otherwise, subdivide the region $R$ into four overlapping squares
          where the amount of overlap is at least $s_{max} - 1$. Denote these
          squares $R_i$ for $i \in \{1, 2, 3, 4\}$.
  Step 3. For each subdivision $R_i$ compute the LRF value $\ell_i$.
  Step 4. Let $k \triangleq \arg\max_i \ell_i$. If $\ell_k < \eta$, the threshold at the current scale,
          $H = R$, stop.
  Step 5. Otherwise call the SC algorithm again with
          $H = SingleCandidate(R_k, scale + 1)$.
```

4.2.2 Multiple Candidate Algorithm

The MC algorithm is a simple extension of the SC algorithm. There are some important differences however. The main difference is that more than one composite hypothesis may be selected at each scale. Another difference is that the test for the existence of an anomaly (the detection test) is conducted at a fine rather than coarse scale. We discuss these differences in detail in this section by specifying the coarsest scale processing. Finer scale processing is similar.

At the coarsest scale the image domain is again subdivided as shown in Figure 4-2. It is assumed that an anomaly exists so the hypothesis test conducted at this coarsest scale consists of only four composite hypotheses. The test has a similar form to the one used in the SC algorithm: $\mathcal{H}_i^{(1)}$ is that $\mathbf{f}_a = c\mathbf{b}_i^{(1)}$, where $\mathbf{b}_i^{(1)}$ is as defined in the previous section.

Any number $r^{(1)} \in \{1, 2, 3, 4\}$ of these composite hypotheses may be selected for finer scale investigation (in the examples provided in Section 4.3, $r^{(k)}$ at each scale k is a program parameter chosen by the user). The ones selected are those with the $r^{(1)}$ highest LRF values. As show in Figure 4-1, each selected composite hypothesis spawns a tree along which finer processing is conducted. For example, if it is decided that $\mathcal{H}_1^{(1)}$ and $\mathcal{H}_4^{(1)}$ are most likely to contain the anomaly then each is subdivided just as the image domain was subdivided at the coarsest scale—with four overlapping squares. Any number of these squares may be selected and the ones which are selected are themselves subdivided.

This process continues until the finest scale has been reached at which point there are N_{leaf} hypotheses to be compared in a post-processing step. It is at this stage that the anomaly detection test is conducted. If the LRF values of the N_{leaf} hypotheses are too low, it is determined that no anomaly exists. Otherwise, the hypothesis corresponding to the highest LRF value is chosen.

The following is high level pseudo-code for the MC algorithm (the post-processing step is not included). The inputs to the algorithm are the region to investigate (initialized to the entire image domain) and the scale number (initialized to one). The output is a list of regions corresponding to the N_{leaf} candidates.

Algorithm 2 (Multiple Candidate)

```
$H = MultipleCandidate(R, scale)$
Step 1. If scale is the maximum possible scale, $H = [H\ R]$, stop.
Step 2. Otherwise, subdivide the region $R$ into four overlapping squares
        where the amount of overlap is at least $s_{max} - 1$. Denote these
        squares $R_i$ for $i \in \{1, 2, 3, 4\}$.
Step 3. For each subdivision $R_i$ compute the LRF value $\ell_i$.
Step 4. Let $\mathcal{L}$ be the set of $R_i$ associated with the $\ell_i$ with the $r^{(scale)}$ highest
        values. Call the MC algorithm with
        $H = MultipleCandidate(R_i, scale + 1)$ for each $R_i \in \mathcal{L}$.
```

4.3 Examples

In this section we provide examples and evaluate the performance and complexity of the SC and MC algorithms.

4.3.1 Single Candidate Algorithm Examples

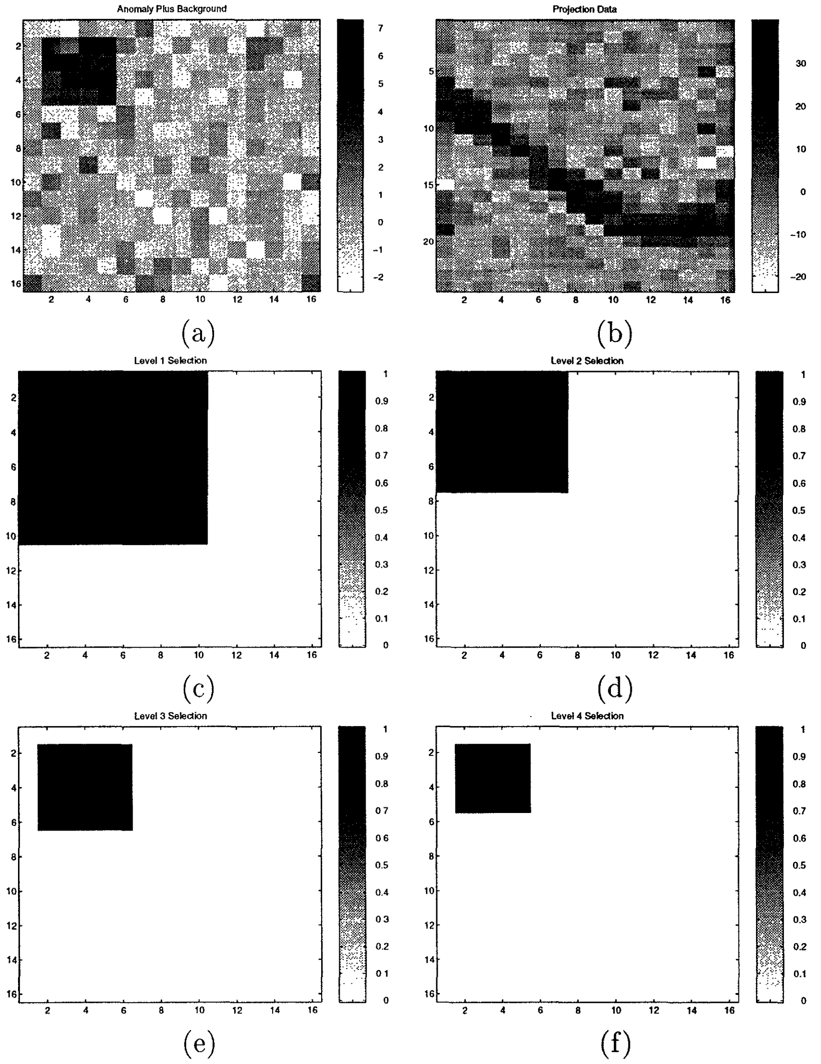
Example Output

Our first example illustrates the output of the SC algorithm at each scale. The anomaly considered in this example is $\mathbf{f}_a = 5\mathbf{b}_{4,32}(2,2)$. The background is white with a variance of about 1.5 and the white additive measurement noise has a variance of about 14 (SNR = 0dB, ABR = -8.9dB). The value of s_{max} has been set to 4 and the minimum size region considered by the algorithm is 4×4. In other words, the algorithm continues to localize the anomaly to finer scales until the scale corresponding to regions of size 4×4 has been reached. Figure 4-3 illustrates the output at each scale.

Coarse Scale Performance

Intuition suggests (and experimentation supports) that the larger the difference between s_{hyp}, the size of the hypothesized regions, and s, the size of the anomaly, the worse the performance. In the following example we explore this relationship by considering different types of coarse scale tests. Four tests are considered and they differ in the number of composite hypothesis regions formulated. One of the four tests is the coarsest scale of the SC algorithm. The other three tests are slight variations of the coarsest scale of the SC algorithm. The composite hypothesis regions in each of these three variations are squares but there are more than four of them and they differ in size. We adopt the following name convention for these tests: SC refers to the SC algorithm and SCm refers to the variation which has m rather than four coarse scale regions. Table 4.1 summarizes the dimension and number of regions defined at the coarse scale (the only scale considered) for each of these algorithms.

In this example, a 2×2 anomaly in a 16×16 field is considered. The anomaly has its upper



(a) (b) (c) (d) (e) (f)

Figure 4-3: These figures illustrate the output of the SC algorithm. Figure (a) is a superposition of a zero-mean white background (variance about 1.5) and a constant intensity anomaly (size 4 × 4, intensity 5) near the upper left corner. The ABR is about -8.9dB. Figure (b) is the sinogram of the anomaly plus background field shown in (a) with zero-mean Gaussian additive measurement noise (variance about 14, SNR about 0dB). Figures (c)–(f) illustrate the regions selected by the SC algorithm at each scale. Figure (f) represents the final selection and it corresponds precisely to the region of support of the anomaly.

Test Name	*# Coarsest Scale Regions*	*Linear Dimension (s_{hyp})*
SC	4	10
SC9	9	8
SC16	16	6
SC49	49	4

Table 4.1: This table lists the number of composite hypotheses defined at the coarsest scale and the respective linear dimension for several variations of the SC algorithm.

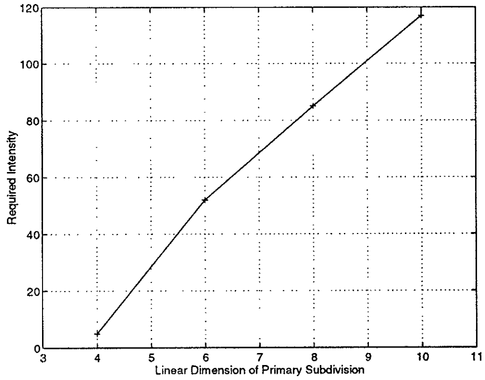


Figure 4-4: This figure indicates the required intensity for a 2×2 anomaly with upper left corner at $(3,3)$ to achieve (P_f, P_d) of roughly $(0.1, 0.95)$ using the coarsest scale test of four variations of the SC algorithm. The variations differ in the number and size of the composite hypothesis test regions formulated. The horizontal axis is the size of the composite hypothesis regions and the vertical axis is the required intensity to achieve the performance benchmark. The noise intensity is 14.4 corresponding to an SNR of 0dB.

left corner at pixel $(3,3)$ and the performance benchmark is set to $(P_f, P_d) = (0.1, 0.95)$. The background field covariance is fractal, $\mathbf{\Lambda} = \mathbf{\Lambda}_f$. Figure 4-4 illustrates the anomaly intensity required to achieve this performance benchmark for each of the four coarse scale tests. A detection is declared if the anomaly exists and if a composite hypothesis region is chosen which entirely overlaps the anomaly. A false alarm is declared if the anomaly is not present but the null hypothesis is not selected.

Note that as the linear dimension of the composite hypothesis regions increase, so does the required intensity. For all of the tests, the required intensity is larger than that required by the BHT of Section 3.2 (around 1.8). For example, for the SC49 test the required intensity is 5.

As suggested by the analysis in Chapter 3, the structure of the background covariance

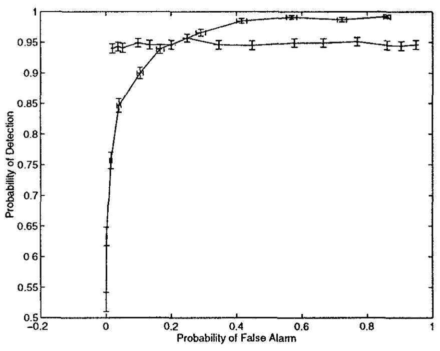


Figure 4-5: These ROCs illustrate the performance of the SC algorithm and the optimal M-ary hypothesis test when the background is white. The field size is 16 × 16 and the anomaly is 4×4 with upper left hand pixel at $(2, 2)$ and intensity 2.5 (SNR about 0dB and ABR about -5.7dB). The ROC for the SC algorithm seems asymptotic to $P_d = 1$ while the ROC for the M-ary hypothesis test seems roughly constant at $P_d = 0.95$. Each data point corresponds to one-thousand Monte Carlo runs. Error bars are drawn plus and minus one standard deviation.

matrix affects performance. In the next example, we compare the performance of the SC algorithm to the optimal M-ary hypothesis test (in which all hypotheses are exhaustively searched) both with a fractal and white background. Again, only the coarsest scale test is considered. The field size is 16×16 and the anomaly is 4×4 with upper left hand corner at pixel $(2, 2)$ and intensity 2.5. The noise intensity is 14.4 (SNR = 0dB). Figure 4-5 illustrates performance comparison for the white background (ABR = -5.7dB). Note that for $P_f < 0.3$ the optimal test outperforms the SC algorithm. However, for $P_f > 0.3$ it seems that the SC algorithm outperforms the optimal test. This is slightly misleading since the SC algorithm has, in some sense, an easier task: it is only localizing the anomaly to a coarse scale region (10 × 10) while the optimal test attempts to select exactly the right 4 × 4 region. The ROC for the optimal test indicates a probability of a miss (i.e., selecting the wrong region when an anomaly exists) is about 0.05 while the ROC for the SC algorithm seems to have a probability of a miss asymptotic to zero.

Figure 4-6 illustrates performance comparison for the fractal background. The top curve is the ROC for the optimal test and the bottom curve is the ROC for the coarse scale SC algorithm. We see that, in the case of a fractal background and for this particular anomaly (4 × 4 with intensity 2.5 (ABR = -0.51dB, SNR = 0dB) and upper left pixel at $(2, 2)$), the

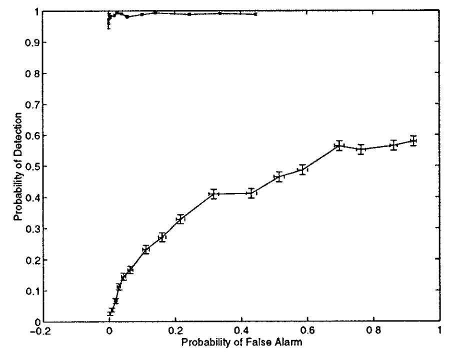


Figure 4-6: These ROCs illustrate the performance of the SC algorithm (bottom curve) and the optimal M-ary hypothesis test (top curve) when the background is fractal. The field size is 16×16 and the anomaly is 4×4 with upper left hand pixel at $(2, 2)$ and intensity 2.5 (SNR about 0dB and ABR about -0.51dB). Each data point corresponds to one-thousand Monte Carlo runs. Error bars are drawn plus and minus one standard deviation.

optimal test outperforms the SC algorithm by a wide margin.

Full Algorithm Performance

In this section we illustrate the performance of the full SC algorithm (not just the coarsest scale). The two anomalies depicted in Figure 3-4(a) and (b) are each considered superimposed on a white noise background with variance 3 and field size 32×32. Both anomalies have intensity 7 and the additive measurement noise has variance 63. The SNR is 3dB and the ABR is -5.7dB. The maximum linear size, s_{max}, is set to 4 and the smallest region considered in the SC algorithm is 4×4. That is, the algorithm continues to localize the anomaly to finer scales until the scale corresponding to regions of size 4×4 has been reached. A detection is declared if the anomaly is present and the chosen region overlaps at least one-quarter of the anomaly's area. A false alarm is declared if an anomaly is not present and the null hypothesis is not selected at the coarsest scale. In the examples provided in this section, there is no null hypothesis at any scale other than the coarsest one. Therefore, once it is determined that an anomaly is present, the algorithm localizes the anomaly to a 4×4 region.

Figure 4-7 is a ROC curve corresponding to the anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(2, 2)$. Figure 4-8 contains two histograms which indicate how well the chosen region matches the anomalous

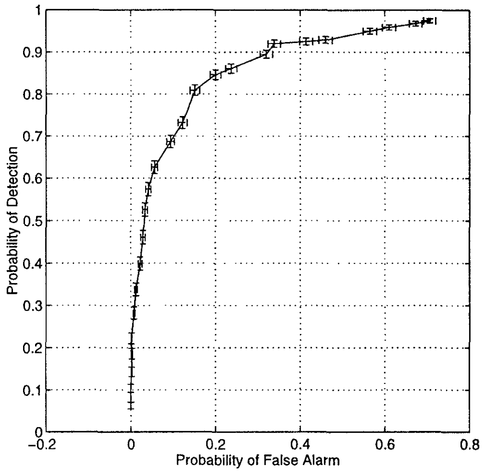
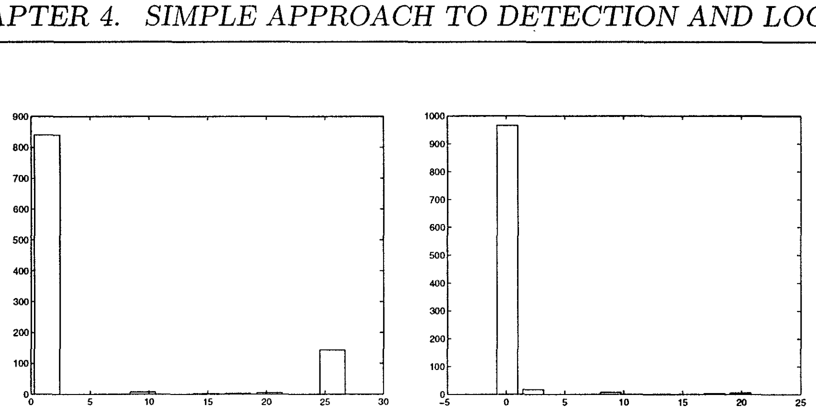


Figure 4-7: This figure contains an ROC which illustrates the performance of the SC algorithm with anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(2,2)$. The background is white, the SNR is about 3dB and the ABR is about -5.7dB. One-thousand Monte-Carlo runs were conducted for each data point and the error bars are drawn plus and minus one standard deviation.

region. The Hausdorff distance and the one-sided Hausdorff measure discussed in Chapter 2 are used for this purpose. The histograms of the values of these two metrics appear in (a) and (b), respectively where the parameters of the SC algorithm are set so that P_d and P_f are about 0.85 and 0.2, respectively. The argument of the Hausdorff distance, $h(A, H)$, are the convex set representing the support of the anomaly, A, and the convex set representing the support of the estimate, H. These are also supplied to the one-sided Hausdorff measure, $h_1(A, H)$.

What is of interest in this problem is knowledge of how much larger (or smaller) one could make the estimate's support in order to completely enclose that of the anomaly's. Such information is useful, for example, if one used the anomaly localization algorithm to cue a local reconstruction method as to which area to reconstruct. The reconstruction routine might then reconstruct not just the area specified but also a buffer region if it is known that the localization algorithm tends to under-estimate the anomaly's size by a certain amount, for example. This information is provided by the one-sided Hausdorff measure and not by the Hausdorff distance.

We can conclude from Figure 4-7 that the particular anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(2,2)$ can be



(a) (b)

Figure 4-8: This figure contains two histograms which illustrate the performance of the SC algorithm with anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(2,2)$. The background is white, the SNR is about 3dB and the ABR is about -5.7dB. Figure (a) is a histogram of the Hausdorff distance and figure (b) is a histogram of the one-sided Hausdorff measure. Both were calculated during 1000 runs of the algorithm. Thresholds were chosen so that P_d and P_f are about 0.85 and 0.2 respectively.

detected with relatively high probability of detection and low probability of false alarm (e.g., the ROC curve passes through $(P_f, P_d) = (0.2, 0.85)$). What is more telling, however, is Figure 4-8 which shows that the SC algorithm does a good job of localizing this anomaly at the operating point $(P_f, P_d) = (0.2, 0.85)$. From Figure 4-8(a) we can see that only roughly twenty percent of the time the anomaly's support and estimate's support are quite far off (by about 25 as measured by the Hausdorff distance). However, by looking at Figure 4-8(b) we see that, even when the Hausdorff distance is large, the amount by which we would need to grow or shrink the estimate's support to contain that of the anomaly's is rarely non-zero.

Figures 4-9 and 4-10 are similar to Figures 4-7 and 4-8 except that the anomaly is now $7\mathbf{b}_{4,32}(14,14)$, which is illustrated in Figure 3-4(b). We see from Figure 4-10 that this anomaly, like the last, is localized with little error nearly all of the time. Notice that performance is a bit better for this anomaly as compared to the one considered previously. This is consistent with the performance bound results of Chapter 3 which indicated that anomalies closer to the center of the field are easier to detect in the presence of a white noise background.

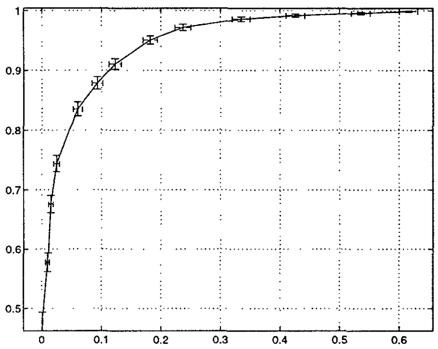


Figure 4-9: This figure contains an ROC which illustrates the performance of the SC algorithm with anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(14, 14)$. The background is white, the SNR is about 3dB and the ABR is about -5.7dB. One-thousand Monte-Carlo runs were conducted for each data point and the error bars are drawn plus and minus one standard deviation.

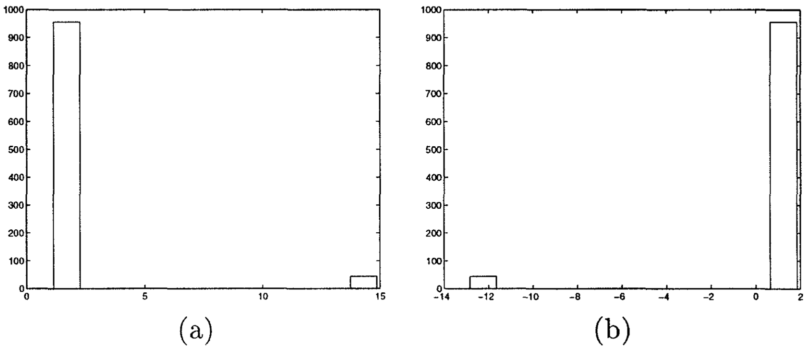
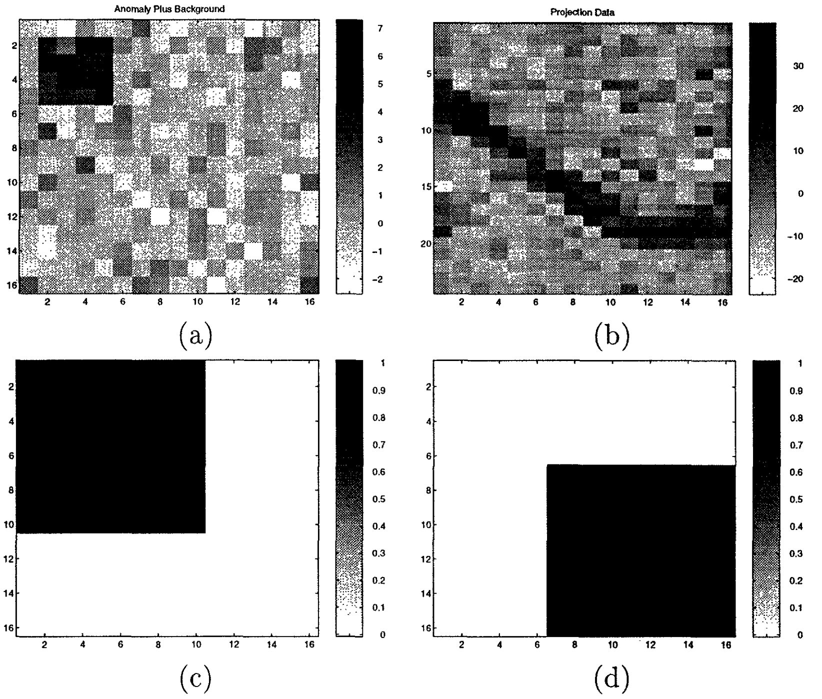


Figure 4-10: This figure contains two histograms which illustrate the performance of the SC algorithm with anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(14, 14)$. The background is white, the SNR is about 3dB and the ABR is about -5.7dB. Figure (a) is a histogram of the Hausdorff distance and figure (b) is a histogram of the one-sided Hausdorff measure. Both were calculated during 1000 runs of the algorithm. Thresholds were chosen so that P_d and P_f are about 0.85 and 0.2 respectively.



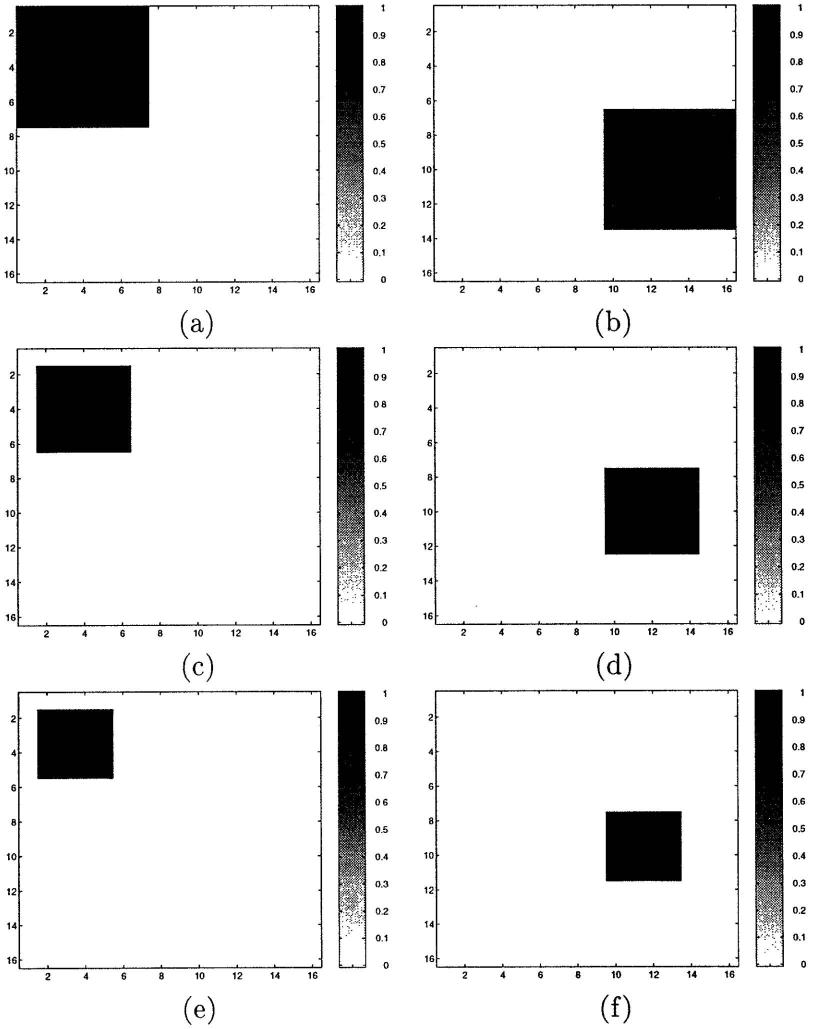
(a) (b) (c) (d)

Figure 4-11: This figure illustrates output from the MC algorithm. Figure (a) is a superposition of a zero-mean white background (variance about 1.5) and a constant intensity anomaly (size 4 × 4, intensity 5) near the upper left corner. The ABR is about -8.9dB. Figure (b) is the sinogram of the anomaly plus background field shown in (a) with zero-mean Gaussian additive measurement noise (variance about 14, SNR about 0dB). Figures (c) and (d) show the regions selected at the coarsest scale.

4.3.2 Multiple Candidate Algorithm Examples

Example Output

Our first example is a sample output of the MC algorithm. The set up is the same as that for the sample output for the SC algorithm provided in Section 4.3.1. Namely, the anomaly considered in this first example is $\mathbf{f}_a = 5\mathbf{b}_{4,32}(2,2)$. The background is white with a variance of about 1.5 and the white additive measurement noise has a variance of about 14 (SNR = 0dB, ABR = −8.9dB). The value of s_{max} has been set to 4 and the minimum size region considered by the algorithm is 4×4. Two composite hypotheses are selected at the coarsest scale and then one is selected at each scale along each subtree defined by the two selected coarse scale regions. Figures 4-11 and 4-12 illustrate the output at each scale.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4-12: This figure illustrates the MC algorithm and is a continuation of the previous one. Figures (a) and (b) show the chosen regions for the second level. Figures (c) and (d) show the chosen regions for the third level. Finally, figures (e) and (f) show the chosen regions for the fourth level. Figure (e) is the correct region.

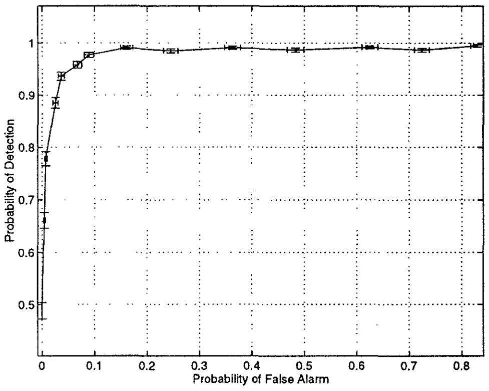
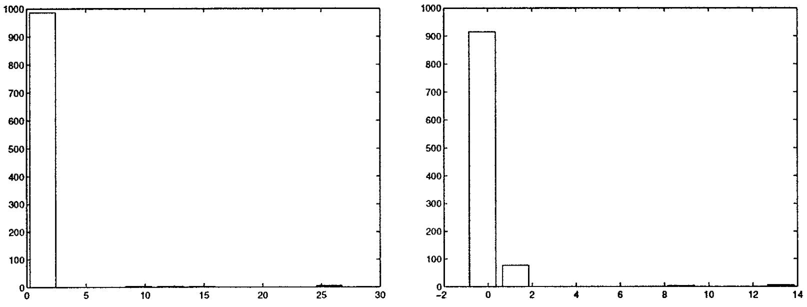


Figure 4-13: This figure contains an ROC which illustrates the performance of the MC algorithm with anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(2,2)$. The background is white. The SNR is about 3dB and the ABR is about -5.7dB. One-thousand Monte-Carlo runs were conducted for each data point and the error bars are drawn plus and minus one standard deviation.

Full Algorithm Performance

The exact same analysis as is presented above for the SC algorithm is presented here for the MC algorithm. In this case, the MC algorithm selects two composite hypotheses at the coarsest scale and then just one at at each scale along each subtree defined by the two selected coarse scale regions. Unless explicitly stated, all parameters remain the same as specified above for the SC algorithm analysis. Figures 4-13 and 4-14 correspond to $\mathbf{f}_a = 7\mathbf{b}_{4,32}(2,2)$ and Figures 4-15 and 4-16 correspond to $\mathbf{f}_a = 7\mathbf{b}_{4,32}(14,14)$.

Essentially the same statements made about the SC algorithm can be made about the MC algorithm when considering the anomalies $\mathbf{f}_a = 7\mathbf{b}_{4,32}(2,2)$ and $\mathbf{f}_a = 7\mathbf{b}_{4,32}(14,14)$. The performance for the latter is a bit better than that for the former which is consistent with the performance bound results of Chapter 3. Comparing the results of the SC algorithm with those of the MC algorithm we see that the latter performs better (i.e., with higher probability of detection). This is due to the fact that the detection decision is delayed until a finer scale in the MC algorithm. For example, comparing the results depicted in Figure 4-13 with those depicted in Figure 4-7 we see that for a probability of false alarm of 0.2 the SC algorithm yields a probability of detection of 0.85 while the MC algorithm yields one of 0.99.



(a)

(b)

Figure 4-14: This figure contains two histograms which illustrate the performance of the MC algorithm with anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(2,2)$. The background is white. The SNR is about 3dB and the ABR is about -5.7dB. Figure (a) is a histogram of the Hausdorff distance and figure (b) is a histogram of the one-sided Hausdorff measure. Both were calculated during 1000 runs of the algorithm. Thresholds were chosen so that P_d and P_f are about 0.99 and 0.2 respectively.

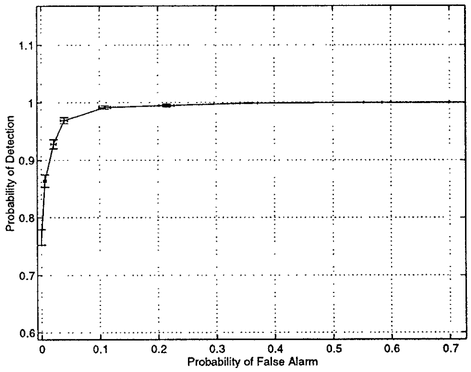
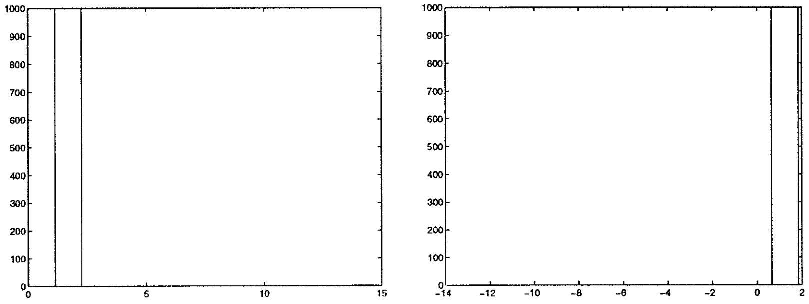


Figure 4-15: This figure contains an ROC which illustrates the performance of the MC algorithm with anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(14,14)$. The background is white. The SNR is about 3dB and the ABR is about -5.7dB. One-thousand Monte-Carlo runs were conducted for each data point and the error bars are drawn plus and minus one standard deviation.



(a) (b)

Figure 4-16: This figure contains two histograms which illustrate the performance of the MC algorithm with anomaly $\mathbf{f}_a = 7\mathbf{b}_{4,32}(14, 14)$. The background is white. The SNR is about 3dB and the ABR is about -5.7dB. Figure (a) is a histogram of the Hausdorff distance and figure (b) is a histogram of the one-sided Hausdorff measure. Both were calculated during 1000 runs of the algorithm. Thresholds were chosen so that P_d and P_f are about 0.99 and 0.2 respectively.

4.4 Complexity Analysis

Our primary motivation for considering anomaly detection and localization algorithms based on multiscale hypothesis testing is that the optimal hypothesis test (for which each possible combination of anomaly intensity, location, and size is represented by a hypothesis) is too computationally costly. Our multiscale algorithms formulate fewer hypotheses than the optimal test and are therefore more efficient. Here we measure and compare the computational complexity of the optimal algorithm with our sub-optimal ones. To do so, we will compute the number of hypotheses formulated in each algorithm and the number of operations per hypothesis.

First consider the optimal test. Suppose the linear size of the square field is N—there are $N_p = N^2$ pixels in the field. Suppose that we know that the maximum size of the anomaly is s_{max}. We require

$$M_{opt} = \sum_{r=1}^{s_{max}} (N - r + 1)^2 = \sum_{k=N-s_{max}+1}^{N} k^2$$

hypotheses for each possible intensity value (hereafter, we assume that there is just one intensity value)—one for every possible shift and every possible anomaly size. Consulting

[36], we find that the sum is

$$M_{opt} = \frac{1}{3}\left\{\left[\frac{3}{2} + (a+b+c)\right]N^2 + \left[\frac{1}{2} - (ab+ac+bc)\right]N + abc\right\}, \tag{4.4}$$

where

$$\begin{aligned} a &= s_{max}\,, \\ b &= s_{max} - 1\,, \\ c &= s_{max} - \frac{1}{2}\,. \end{aligned}$$

It is easily verified that equation (4.4) gives $M_{opt} = N^2$ when $s_{max} = 1$ as it should. Also notice that, for s_{max} constant, M_{opt} grows as N^2.

Now consider the SC algorithm in which four overlapping subdivisions are defined at each level. The amount of overlap is at least $s_{max} - 1$, one pixel less than the maximum possible size of the anomaly. Since we know there are four hypotheses at each level, all we have to figure out is how many levels are required. For simplicity, we neglect the overlapping. The accuracy of our answer will not suffer much from this simplifying assumption so long as the overlap is small relative to the size of the field.

To compute the number of hypotheses for the SC algorithm, we need to make one additional assumption. Let us assume that our algorithm localizes the anomaly to an area which is $s_{max} \times s_{max}$, though this is not necessarily always the case in practice. Such an assumption will give us a reasonable estimate which will be of the correct order of magnitude.

At each level, we divide the area of our search by one quarter. First our subdivisions are $\frac{N}{2} \times \frac{N}{2}$, then they are $\frac{N}{4} \times \frac{N}{4}$, etc. We stop searching when

$$\frac{N}{2^k} = s_{max}\,,$$

where k is the number of levels we need to search. Therefore the number of hypotheses is

$$M_{SC} = 4 \log_2 \frac{N}{s_{max}} . \tag{4.5}$$

Notice that when $s_{max} = \frac{N}{2}$ equation (4.5) gives $M_{SC} = 4$ which makes sense. In contrast to the N^2 growth in the number of hypotheses required by the optimal algorithm, the number required by the SC algorithm grows with the logarithm of N.

Finally, the MC algorithm will require the same number of scales as the SC algorithm but will formulate at least as many (and usually more) hypotheses at each scale. Recall that in the MC algorithm $r^{(k)} \leq N^{(k)}$ composite hypotheses are not discarded at scale k where $N^{(k)}$ is the number of composite hypotheses formulated and is at most 4^k. Consider first the worst case in which we do not discard any hypotheses. In such a case, the number of hypotheses formulated in the MC algorithm is

$$\begin{aligned} M_{MC,worst} &= \sum_{j=1}^{\log_2 \frac{N}{s_{max}}} 4^j \\ &= \frac{4^{\log_2 \left(\frac{N}{s_{max}}\right)+1} - 4}{3} \\ &= \frac{4\left(\frac{N^2}{s_{max}^2}\right) - 4}{3} . \end{aligned}$$

Thus $M_{MC,worst}$ is quadratic in N.

Now consider a more realistic case where we do not retain all 4^k composite hypotheses at scale k of the MC algorithm. Suppose we instead retain only half of the available composite hypotheses at each scale. At the first scale four composite hypotheses are formulated but only two are retained. Having retained only two at the first scale, eight are formulated at the second scale of which four are retained. Continuing, at scale k we formulate 2^{k+1} hypotheses. Therefore the total number of hypotheses which are formulated in this case is

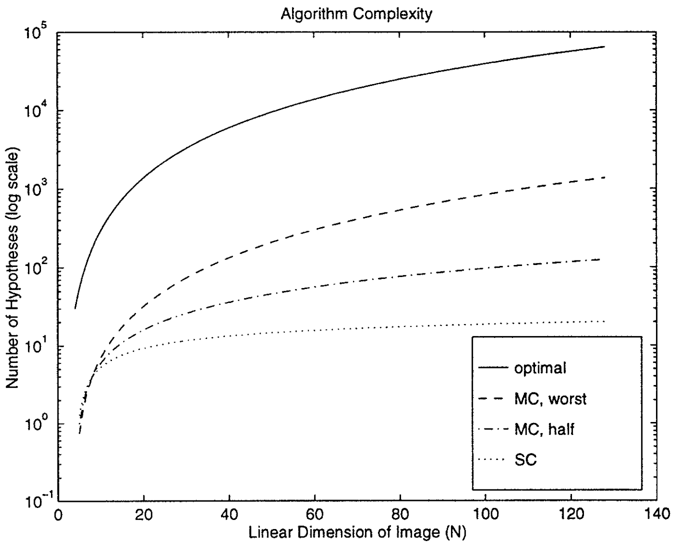


Figure 4-17: This plot shows the complexity of the optimal algorithm (top curve), the SC algorithm (bottom curve), the the MC worst case algorithm (curve second from top), and the MC algorithm for which half the available hypotheses are retained at each scale (curve second from bottom). The horizontal-axis is the number of hypotheses formulated (on a log scale) and the vertical-axis is the linear dimension of the field size. Here $s_{max} = 4$.

$$\begin{aligned} M_{MC,half} = &= 2 \sum_{j=1}^{\log_2 \frac{N}{s_{max}}} 2^j \\ &= 2\left\{2^{\log_2\left(\frac{N}{s_{max}}\right)+1} - 2\right\} \\ &= 4\frac{N}{s_{max}} - 4\,. \end{aligned}$$

Therefore, $M_{MC,half}$ is linear in N. In Figure 4-17 plots of M_{opt}, M_{SC}, $M_{MC,worst}$, and $M_{MC,half}$ are shown as a function of N for $s_{max} = 4$. We see clearly that the SC algorithm is far more efficient than the optimal test. The worst case MC algorithm, while still quadratic in N, also requires fewer hypotheses than the optimal test. Finally, the MC algorithm for which half of the available hypotheses are retained, while linear in N, is more complex than the SC algorithm.

We now consider the number of operations required per hypothesis. We have considered only the log-likelihood ratio statistic which is linear in the data and, thus, requires $2N_\phi N_s - 1$ operations (adds and multiplies). This result follows from the fact that the linear statistic is an inner product between two length $N_\phi N_s$ vectors. Since each hypothesis requires a constant amount of work ($2N_\phi N_s - 1$ operations), the overall complexity of an algorithm (SC, MC, or optimal) is proportional to the number of hypotheses formulated. Hence we take the number of hypotheses formulated as a measure of algorithmic complexity.

Chapter 5

Optimized Multiscale Hypothesis Tests

In Chapter 4 we introduced two algorithms for anomaly detection and localization. These algorithms are based on multiscale hypothesis testing and, therefore, consist of sequences of composite hypothesis tests. Each of these tests is associated with a set of composite hypotheses and statistics. The composite hypotheses are chosen to effect spatial zooming and the statistics are coarse scale log-likelihood ratio functions. While these choices of composite hypotheses and statistics prove useful (both for our work and that of [19–21]), they are by no means the "best" choices. In fact, we saw in Section 4.3.1 that the SC algorithm performs poorly relative to the optimal MHT when the background field has a fractal covariance.

In this chapter we take a broader view of multiscale hypothesis testing by recognizing that the statistics employed need not be of the form introduced in Chapter 4. (Additionally, a MSHT need not correspond to spatial zooming however we do not explore this degree of freedom in significant depth in this thesis.) Our more general view of multiscale hypothesis testing naturally leads to the consideration of MSHTs with improved performance. We begin our broader consideration of MSHTs in Section 5.1 with another look at anomaly ambiguity. We shall see more precisely why the SC algorithm is not well suited to the case for which the background field has a fractal covariance structure. This new notion of anomaly ambiguity, which we call *statistical sensitivity*, will lead us to a set of quality criteria for MSHTs which we discuss in Section 5.2. This set of criteria motivates an optimization problem (Section 5.3), the solution of which is an optimized MSHT statistic. Sub-optimal solutions to the optimization problem are presented in Section 5.4 and optimal solutions are discussed in Section 5.5. A revised algorithm and examples are provided in Sections 5.6 and 5.7, respectively.

5.1 Ambiguity Revisited: Statistical Sensitivity

The ambiguity analysis of Chapter 3 suggests spatial zooming is easier in the case of a white noise background relative to the case of a fractal background. Indeed, in Chapter 4 we show that the SC algorithm performs reasonably well in the former case but relatively poorly in the latter relative to the optimal MHT. In this section we endeavor to discover more precisely why spatial zooming is more difficult in the fractal background case. To do so, we develop another measure of anomaly ambiguity which we call *statistical sensitivity.* This statistical sensitivity analysis is based upon the composite hypothesis testing scheme which underlies multiscale hypothesis testing. The results of our analysis will, therefore, directly provide insight into the performance of the algorithms developed in Chapter 4.

Recall that at each scale k of a MSHT there are a set of statistics, $\{\ell_i^{(k)}\}_{i=1}^{N^{(k)}}$, which we wish to interpret as likelihoods. In other words, we wish $\ell_i^{(k)}$ to have a relatively large value whenever $\mathcal{H}_i^{(k)}$ is true (i.e., it contains the true hypothesis H_j for some j). And we wish $\ell_i^{(k)}$ to have a relatively small value whenever $\mathcal{H}_i^{(k)}$ is false (i.e., it does not contain the true hypothesis H_j for some j).

The statistics, $\{\ell_i^{(k)}\}_{i=1}^{N^{(k)}}$, are random variables. We can get a sense of their magnitude by taking their expected values. We define the conditional expected value of $\ell_i^{(k)}$ given that the true hypothesis is H_j as

$$m_{ij}^{(k)} \triangleq E\left[\ell_i^{(k)} | H_j\right] .$$

The conditional expected value of $\ell_i^{(k)}$ alone is not so meaningful if the corresponding

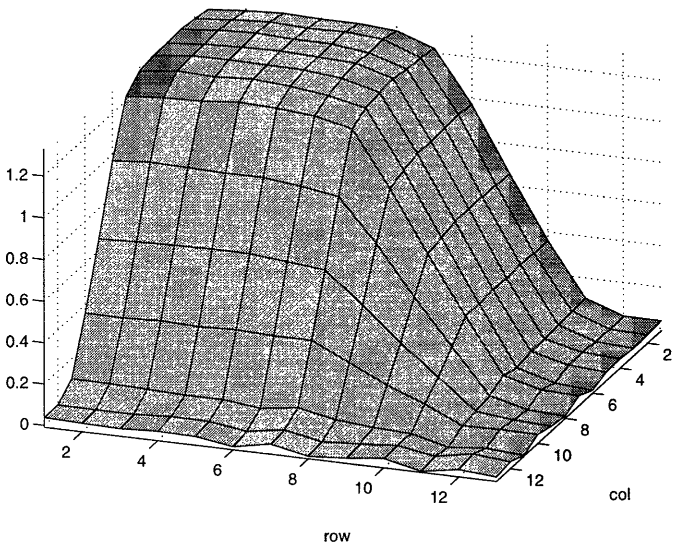


Figure 5-1: This figures illustrates $\tilde{m}_{1j}^{(1)}$ with a white background (SNR about 0dB). Pixel (m, n) corresponds to $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$.

variance is not also considered. Therefore, we define the standard-deviation-normalized mean as

$$\tilde{m}_{ij}^{(k)} \triangleq \frac{m_{ij}^{(k)}}{\sigma_{\ell_i^{(k)}|H_j}} .$$

Figures 5-1 and 5-2 illustrate values of $\tilde{m}_{1j}^{(1)}$ (the normalized conditional mean of statistic one at the coarsest scale) for the case of a white and fractal background respectively (SNR = 0dB). Recall that $\ell_1^{(1)}$ is associated with the upper left-hand region of the image domain (see Figure 4-2) and is given by

$$\ell_j^{(1)}(\mathbf{G}) = (\mathbf{T}\mathbf{b}_j^{(1)})^T \mathbf{\Lambda_g}^{-1}\mathbf{G} .$$

The hypotheses, H_j, considered in these figures is the set of all 4×4 unit intensity anomalies in a 16×16 field. The point (m, n) in the plots of Figures 5-1 and 5-2 corresponds to the hypothesis that $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$ where $m, n \in \{1, 2, \ldots, 13\}$.

The values of $\tilde{m}_{1j}^{(1)}$ in Figure 5-1 exhibit precisely the type of behavior we desire. The

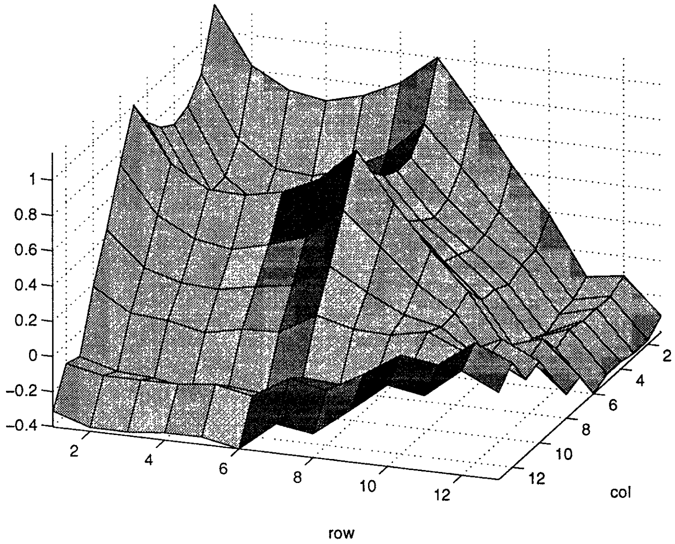


Figure 5-2: This figures illustrates $\tilde{m}_{1j}^{(1)}$ with a fractal background (SNR about 0dB). Pixel (m, n) corresponds to $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$.

mean value of $\ell_1^{(1)}|H_j$ is relatively high for $H_j \in \mathcal{H}_1^{(1)}$ and relatively low for $H_j \notin \mathcal{H}_1^{(1)}$. The values of $\tilde{m}_{1j}^{(1)}$ in Figure 5-2, on the other hand, do not exhibit good behavior. In fact there exist $H_k \notin \mathcal{H}_1^{(1)}$ and $H_j \in \mathcal{H}_1^{(1)}$ for which $\tilde{m}_{1k}^{(1)} > \tilde{m}_{1j}^{(1)}$. This means that, in the case of a fractal background, the statistic $\ell_1^{(1)}$ cannot be reasonably interpreted as the likelihood that $\mathcal{H}_1^{(1)}$ contains the anomaly. It is sensitive to the wrong hypotheses.

The above sensitivity analysis suggests that, in the case of the fractal background, either the type of statistic or the type of composite hypotheses ought to be changed. For example, the same statistics may be used if we redefine the composite hypothesis sets $\mathcal{H}_i^{(1)}$ to consist of the hypotheses to which the log-likelihood ratio function statistics are sensitive. Or, perhaps the spatial zooming approach may be feasible if the statistics are chosen differently. In the next section we consider these issues in more detail.

5.2 Composite Hypothesis Test Quality Criteria

Having seen that the composite hypotheses and/or the statistics introduced in Chapter 4 are inadequate, we wish to change them to achieve better performance. But what criteria ought to guide our choices? What properties do good composite hypotheses and statistics have? We shall address these questions in this section for the composite hypothesis test at the coarsest scale of a MSHT. Analogous conclusions may be made for the composite hypothesis tests at other scales. Therefore, we shall drop the superscript (k), the scale index, in our notation.

5.2.1 Criteria for Composite Hypotheses

First consider the composite hypotheses. While associating composite hypotheses with contiguous regions (as was done in Chapter 4) is intuitively appealing, they need not be defined in this way. Indeed, in many large M-ary hypothesis testing problems for which a multiscale hypothesis testing scheme is of interest, there may be no natural way to group hypotheses (e.g., in computer vision problems in which object recognition is of interest it may not be clear which objects and orientations ought to be grouped). Even when a natural grouping exists, as does for the anomaly detection and localization problems, such a grouping may not yield good results for a particular choice of statistics.

No matter what the structure of composite hypotheses, there are several criteria which we would like them to satisfy: (1) the number of composite hypotheses, N, at the coarse scale must be much less than the number of original hypotheses, M; (2) the composite hypothesis sets at every scale must form a cover for the subset of $\mathcal{H}$ which has not been discarded; and (3) the amount by which the composite hypothesis sets overlap should be as small as possible. The reasons for criteria (1) and (2) are clear. Criterion (3) strives for maximum efficiency because if the overlap is small then discarding a composite hypothesis (i.e., not selecting it for finer scale investigation) discards many hypotheses $H_j \in \mathcal{H}$. Below we shall make these criteria precise for the coarsest scale. Generalization to other scales is conceptually straight forward (although notationally cumbersome).

The cardinality constraint is simply $N \ll M$. The covering constraint is

$$\mathcal{H} = \bigcup_{i=1}^{N} \mathcal{H}_i \,. \tag{5.1}$$

To make the overlap criterion precise we define

$$\begin{aligned} n_{ij} &\triangleq |\mathcal{H}_i \cap \mathcal{H}_j|\,, \\ \Delta_i &\triangleq \frac{\sum_{\forall j \neq i} n_{ij}}{n_{ii}}\,. \end{aligned}$$

The quantity Δ_i is the number of hypotheses H_j which are in $\mathcal{H}_i$ and also in some other composite hypothesis set divided by the cardinality of $\mathcal{H}_i$. In other words, it is the relative amount of $\mathcal{H}_i$ which is shared with other composite hypotheses. It is the quantity Δ_i which we wish to make small for all i subject to the constraints $N \ll M$ and equation (5.1).

Of course having composite hypotheses with small Δ_i is not enough to ensure a good multiscale hypothesis test. This is clear from our sensitivity analysis above. It is also important that the statistics associated with the composite hypotheses have good properties. We elaborate on these properties next.

5.2.2 Criteria for Statistics

As a starting point for our discussion of criteria for statistics we restrict attention to statistics which are linear in the data: $\ell_i(\mathbf{G}) = \mathbf{a}_i^T\mathbf{G}$. The coarse scale log-likelihood ratio functions considered previously are of this form but may not be the "best" linear statistics. In this section we define criteria that good linear statistics ought to satisfy.

There are two criteria which apply to the statistics. First, we wish the conditional means to have the property that

$$m_{ij} > m_{ik}\,, \tag{5.2}$$

$\forall j, k$ such that $H_j \in \mathcal{H}_i$ and $H_k \notin \mathcal{H}_i$, $\forall i \in \{1, \ldots, N\}$. Ensuring a spread in conditional means as described by equation (5.2) is not enough, however. For if the standard deviation of the statistics is sufficiently large, any spread in conditional means will be overshadowed. Therefore, the second criterion is that the standard deviation remain small. The conditional standard deviation is

$$\sigma_i \triangleq \sqrt{\operatorname{var}(\ell_i|H_j)} = \sqrt{\mathbf{a}_i^T \mathbf{\Lambda_g} \mathbf{a}_i}\,.$$

The standard deviation is independent of the true hypothesis H_j and we wish it to be small relative to the spread in means. At the same time, we wish the spread in means to be large. These criteria form the basis for an optimization problem introduced next.

5.3 An Optimization Problem

Having specified the criteria by which we measure the quality of the composite hypotheses and statistics in a multiscale hypothesis test, we are in a position to formulate an optimization problem. The solution to the optimization problem we formulate in this section is an optimized[1] MSHT statistic. We shall only consider the optimization problem at the coarsest scale as it is similar at other scales. Therefore, we shall have no need for the superscripts (k) which refer to scale; these will be dropped.

While one could in principle combine the quality criteria for the composite hypotheses and statistics into one optimization problem, such a problem would be extremely complex and likely intractable. Therefore, we begin with a simpler optimization problem which we arrive at by fixing the composite hypotheses to be the ones with which we have been working all along (the contiguous square regions depicted in Figure 4-2). The quality criteria associated with composite hypotheses discussed in Section 5.2.1 are, therefore, automatically satisfied.

The optimization problem is, therefore, concerned with the statistics. We formulate the problem as follows. The generalized notion of the d statistic in this problem is

$$d_{ijk} \triangleq \frac{m_{ij} - m_{ik}}{\sigma_i} = \frac{\mathbf{a}_i^T \mathbf{T} \mathbf{b}_j - \mathbf{a}_i^T \mathbf{T} \mathbf{b}_k}{\sqrt{\mathbf{a}_i^T \mathbf{\Lambda_g} \mathbf{a}_i}}\,,$$

where $H_j \in \mathcal{H}_i$ and $H_k \notin \mathcal{H}_i$, $\forall i \in \{1, \ldots, N\}$. (We have assumed unit intensity anomalies

[1] By "optimized" we mean with respect to the optimization problem formulated.

for convenience.) This generalized d statistic measures the normalized spread in conditional means of the statistic ℓ_i conditioned on a hypotheses H_j associated with the composite hypotheses $\mathcal{H}_i$ and conditioned on a hypotheses H_k not associated with $\mathcal{H}_i$. It is this generalized d statistic which we want to be large for all i, j, k.

Continuing, we define the modified d statistic for which we use the notation $\tilde{d}$ as

$$\tilde{d}_i \overset{\triangle}{=} \min_{(j,k)\in\mathcal{A}_i} d_{ijk} \,,$$

where $\mathcal{A}_i \overset{\triangle}{=} \{(j,k)|H_j \in \mathcal{H}_i \text{ and } H_k \notin \mathcal{H}_i\}$. The $\tilde{d}_i$ statistic measures the smallest difference between two sets of numbers. One set consists of all standard-deviation-normalized mean values of ℓ_i each of which is conditioned on a hypothesis H_j which is in the composite hypothesis set associated with ℓ_i, namely $\mathcal{H}_i$. The other set consists of all standard-deviation-normalized mean values of ℓ_i each of which is conditioned on a hypothesis H_j which is *not* in the composite hypothesis set associated with ℓ_i, i.e., it is *not* in $\mathcal{H}_i$. Finally, it is the quantity $\tilde{d}_i$ which we want to maximize as a function of $\mathbf{a}_i$, the linear statistic weights. Therefore, the optimization problem is

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}_i} \min_{(j,k)\in\mathcal{A}_i} \frac{\mathbf{a}_i^T\mathbf{T}\mathbf{b}_j - \mathbf{a}_i^T\mathbf{T}\mathbf{b}_k}{\sqrt{\mathbf{a}_i^T\mathbf{\Lambda}_\mathbf{g}\mathbf{a}_i}} \,.$$

We shall find it useful to adopt a more compact, but more abstract, notation for this optimization problem. To this end, we transpose the numerator and write the problem as

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}_i} \frac{\min_{m\in\mathcal{A}_i} \mathbf{q}_m^T\mathbf{a}_i}{\sqrt{\mathbf{a}_i^T\mathbf{\Lambda}_\mathbf{g}\mathbf{a}_i}} \,,$$

where $\mathbf{q}_m = \mathbf{q}_{(i,j)} \overset{\triangle}{=} \mathbf{T}\mathbf{b}_j - \mathbf{T}\mathbf{b}_k$. It is clear from this latter formulation that the numerator is a piecewise linear concave function of $\mathbf{a}_i$. The denominator is the square root of a quadratic form and, since the covariance matrix is positive definite, is strictly positive.

An additional simplification may be made by recognizing the fact that it is sufficient to consider linear weights $\mathbf{a}_i$ for which $\mathbf{a}_i^T\mathbf{\Lambda}_\mathbf{g}\mathbf{a}_i = 1$. Therefore we may write the optimization problem in the form

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}_i} z$$
$$\text{subject to} \begin{cases} z \leq \mathbf{q}_m^T \mathbf{a}_i, \ \forall m \in \mathcal{A}_i \\ ||\mathbf{a}_i||^2_{\mathbf{\Lambda_g}} \leq 1 \end{cases}, \tag{5.3}$$

where, since we are maximizing z, we may substitute an inequality for an equality in the quadratic constraint of equation (5.3). From this formulation of the optimization problem (which we label P_i) we see that the problem is one of maximizing a linear cost subject to many linear constraints and one quadratic constraint. Further, the set $\mathcal{C}_i \triangleq \{(\mathbf{a}, z)|z \leq \mathbf{q}_m^T\mathbf{a}, \ \forall m \in \mathcal{A}_i$ and $||\mathbf{a}||^2_{\mathbf{\Lambda_g}} \leq 1\}$ is the intersection of two convex sets and is itself convex.

There is one such convex optimization problem, P_i, for each $i \in \{1, 2, \ldots, N\}$ where N is the number of coarse scale composite hypotheses. To find the optimized statistics at the second scale, another set of optimization problems must be solved, etc. Therefore, there are many such optimization problems to be solved to determine the statistics for an entire multiscale hypothesis test. In the next section we discuss a sub-optimal solution to one such problem.

5.4 Sub-Optimal Solutions

The optimization problem, P_i, posed in the previous section would be a simple linear programming problem if it were not for the single quadratic constraint. The presence of this constraint turns an otherwise relatively simple optimization problem into one which is apparently quite difficult. In this section we derive an approximation to this optimization problem which *is* a linear programming problem. The optimal solutions to this linear program, however, are not necessarily optimal solutions to P_i, the original non-linear problem.

The linear programming (LP) formulation we propose here simply replaces the bothersome non-linear constraint of the problem P_i posed in Section 5.3 with many linear constraints. This is done as follows. Consider a particular instance of P_i. Let the length of the vector[2] $\mathbf{y} \triangleq \mathbf{\Lambda_g}^{1/2}\mathbf{a}_i$ be N_y. The problem $\tilde{P}_i$ is defined as

$$\tilde{\mathbf{a}}_i = \arg\max_{\mathbf{a}_i} z$$

$$\text{subject to} \begin{cases} z \leq \mathbf{q}_m^T\mathbf{a}_i, \ \forall m \in \mathcal{A}_i \\ \mathbf{\Lambda_g}^{1/2}\mathbf{a}_i \leq \frac{1}{\sqrt{N_y}}\mathbf{e} \\ \mathbf{\Lambda_g}^{1/2}\mathbf{a}_i \geq -\frac{1}{\sqrt{N_y}}\mathbf{e} \end{cases}, \tag{5.4}$$

where $\mathbf{e}$ is the length N_y vector with a one in every entry. The inequalities of equation (5.4) are understood to be component-wise. Denote the feasible set as $\tilde{\mathcal{C}}_i \triangleq \{(\mathbf{a}, z) | z \leq \mathbf{q}_m^T\mathbf{a}, \ \forall m \in \mathcal{A}_i \text{ and } \sqrt{N_y}\mathbf{\Lambda_g}^{1/2}\mathbf{a} \leq \mathbf{e} \text{ and } \sqrt{N_y}\mathbf{\Lambda_g}^{1/2}\mathbf{a} \geq -\mathbf{e}\}$. The following theorem demonstrates that any feasible solution to problem $\tilde{P}_i$ is also a feasible solution to P_i.

Theorem 3 *$\tilde{\mathcal{C}}_i \subset \mathcal{C}_i$ but the converse does not hold in general.*

Proof. Let $\mathbf{x} = [\mathbf{a} \ z]^T \in \tilde{\mathcal{C}}_i$. Therefore $z \leq \mathbf{q}_m^T\mathbf{a}$, $\forall m \in \mathcal{A}_i$, and $N_y\mathbf{\Lambda_g}^{1/2}\mathbf{a} \leq \mathbf{e}$, and $\sqrt{N_y}\mathbf{\Lambda_g}^{1/2}\mathbf{a} \geq -\mathbf{e}$. The latter two constraints imply that $\mathbf{a}^T\mathbf{\Lambda_g}\mathbf{a} \leq 1$. It follows that $\mathbf{x} \in \mathcal{C}_i$. To show that the converse does not hold in general, consider the following trivial counter example. Let $\mathbf{a} = [1 \ 0]^T$ and $\mathbf{\Lambda_g} = \mathbf{I}$. Then, while $\mathbf{a}^T\mathbf{\Lambda_g}\mathbf{a} \leq 1$, $\mathbf{\Lambda_g}^{1/2}\mathbf{a} \not\leq [1/\sqrt{2} \ \ 1/\sqrt{2}]^T$.□

The relationship of the feasible set $\mathcal{C}_i$ for the original non-linear optimization problem P_i to the feasible set $\tilde{\mathcal{C}}_i$ for the LP problem $\tilde{P}_i$ is illustrated abstractly in Figure 5-3. Qualitatively, the difference between the two sets is that $\mathcal{C}_i$ restricts the linear weights $\mathbf{a}_i$ to be inside a N_y-dimensional ellipsoid while $\tilde{\mathcal{C}}_i$ restricts them to be within a N_y-dimensional box which is inscribed in the ellipsoid.

Since the feasible set for problem $\tilde{P}_i$ is a subset of that for problem P_i, the optimal cost of the former is no higher (and most likely lower) than that of the latter. The optimal solution for $\tilde{P}_i$ is a suboptimal one for P_i. Nevertheless, the linear weight vectors which are solutions to $\tilde{P}_i$ are better than the log-likelihood ratio function statistics previously

[2] $\mathbf{A}^{1/2}$ is the unique positive (semi) definite symmetric square root of matrix $\mathbf{A}$ when $\mathbf{A}$ is itself symmetric positive (semi) definite. That is, $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$.

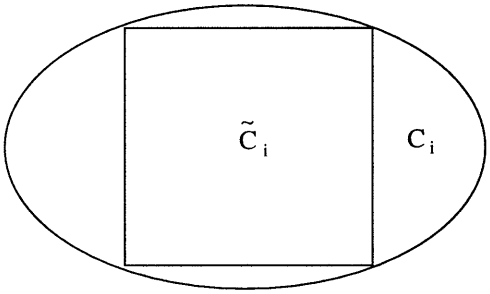


Figure 5-3: This figure abstractly illustrates the relationship between $\mathcal{C}_i$ and $\tilde{\mathcal{C}}_i$.

introduced. We shall demonstrate this in Section 5.7.

We conclude this section with a few technical details regarding the LP problem $\tilde{P}_i$. First, it is *not* an infeasible problem because the zero vector is an element of $\tilde{\mathcal{C}}_i$. Second, the optimal cost is bounded since the elements of $\mathbf{a}_i$ are bounded. Therefore, an optimal solution exists with bounded optimal cost. Such a solution may be found exactly using, for example, the simplex method (e.g., using MATLAB's `lp()` function), or approximately using, for example, an interior point method. It is important to keep in mind, however, that exact solutions to the LP problem $\tilde{P}_i$ are *not* exact solutions to the original non-linear problem P_i.

Finally, we arrived at a LP formulation by approximating an ellipsoid by an inscribed high-dimensional box. The solution to the LP is one of the corners of this box. In the $\mathbf{y} = \mathbf{\Lambda_g}^{1/2}\mathbf{a}_i$ coordinate system, the edges of this box are of equal length since each element of the right-hand sides of equation (5.4) has the same value $(1/\sqrt{N_y})$. Changing these right-hand side values (so that they are not all equal) changes the box to a N_y-cell (for $N_y = 2$ this is like changing a square to a rectangle). Such a change would move the corners of the cell around and, hence, change the optimal cost of the LP. In principle, for any given point on the original ellipsoid, there exists a set of right-hand side values of equation (5.4) so that a corner of the cell coincides with this ellipsoid point. Therefore, if right-hand side values could be found such that a corner of the inscribed N_y-cell coincided with the point on the ellipsoid which corresponds to the optimal solution to P_i then the optimal solution to $\tilde{P}_i$ would also be optimal for P_i.

5.5 Optimal Solutions

We have shown how to find sub-optimal solutions to the problem P_i using linear programming. In this section we show how to solve P_i exactly using quadratic programming. We leave the details of this quadratic program (QP) formulation to Appendix D and merely state the results here. Unfortunately, the work presented in this section was completed too late to include many examples in Section 5.7 with these optimal solutions to P_i.

As is shown in Appendix D, using Lagrange duality theory it can be shown that P_i is equivalent to the QP

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\min_{\mathbf{y}} \mathbf{y}^T \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^T \mathbf{y} \\ & \quad \text{subject to} \begin{cases} \mathbf{e}^T \mathbf{y} = 1 \\ \mathbf{y} \geq 0 \end{cases}, \end{aligned}$$

where

$$\mathbf{e} \triangleq [1\ 1\ 1\ \ldots\ 1]^T ,$$

and

$$\mathbf{Q} \triangleq \begin{bmatrix} (\mathbf{Tb}_{j_1} - \mathbf{Tb}_{k_1})^T \\ (\mathbf{Tb}_{j_2} - \mathbf{Tb}_{k_2})^T \\ \vdots \end{bmatrix} ,$$

and all pairs $(j_m, k_m) \in \mathcal{A}_i$. The optimized weight vector is given by

$$\hat{\mathbf{a}}_i = \frac{\mathbf{\Lambda}^{-1} \mathbf{Q}^T \hat{\mathbf{y}}}{\sqrt{\hat{\mathbf{y}}^T \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^T \hat{\mathbf{y}}}} .$$

5.6 A Revised Algorithm

Earlier in this chapter we discussed desirable criteria for good covers and statistics. In the previous three sections we focussed only on finding optimized statistics, not optimized covers. Our algorithms of Chapter 4 are easily modified to accommodate these optimized statistics—we replace the simple statistics of Chapter 4 with the ones found by solving either the LP problem $\tilde{P}_i$ (sub-optimal) or the non-linear problem P_i. Since the form of the covers stay the same, the interpretation of the composite hypotheses in terms of contiguous regions in the image domain is still valid for our revised algorithm using the optimized statistics. In particular, the scale recursion of the revised SC algorithm (called SC-revised) begins as shown in Figure 4-2. At most one of the four overlapping regions illustrated is selected on the basis of a comparison among the optimized statistics and the chosen region (if any) is further subdivided in the scale-recursive, decision-directed way described in Chapter 4.

5.7 Examples

5.7.1 Solutions to $\tilde{P}_i$

Figures 5-1 and 5-2 illustrate the value of $\tilde{m}_{1j}^{(1)}$ corresponding to the statistic introduced in Chapter 4 for the set of all 4×4 unit intensity anomalies in a 16×16 field. For comparison we illustrate $\tilde{m}_{1j}^{(1)}$ for the same set of anomalies but corresponding to the statistic associated with the solution of $\tilde{P}_1$. Figure 5-4 is for the white background and Figure 5-5 is for the fractal background.

Comparing Figure 5-4 with Figure 5-1 we see that there is a very modest amount of improvement in the behavior of the statistic. The best that can be said is that the transition between regions for which the statistic is designed to be high and where it is designed to be low is sharper. Generally speaking, however, it appears that when the background is white, the statistic of Chapter 4 seem to be comparable in quality to the one found by solving $\tilde{P}_i$.

Comparing Figure 5-5 with Figure 5-2 we see a dramatic improvement in statistic performance. The value of $\tilde{m}_{1j}^{(1)}$ is consistently high where we wish it to be and low elsewhere. Additionally, the transition between high and low values is sharper in Figure 5-5. This

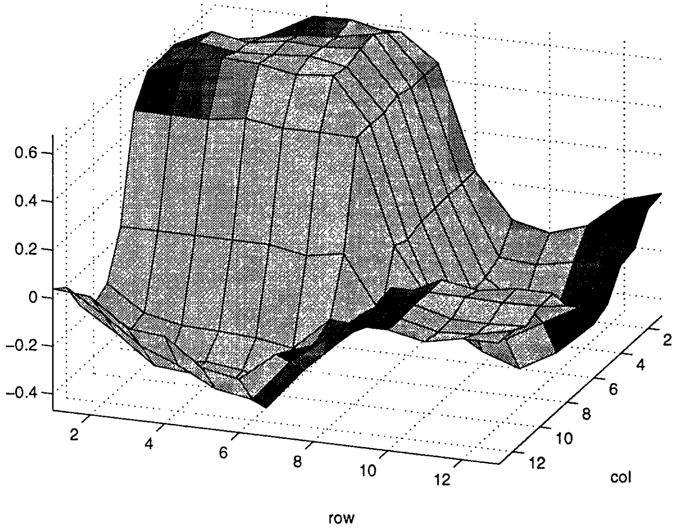


Figure 5-4: This figure illustrates $\tilde{m}_{1j}^{(1)}$ corresponding to the statistic found by solving $\tilde{P}_1$ with a white background (SNR about 0dB). Pixel (m, n) corresponds to $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$.

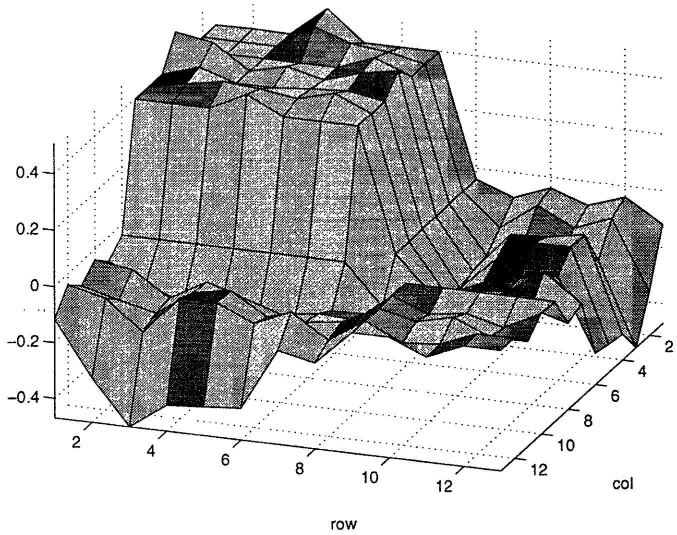


Figure 5-5: This figure illustrates $\tilde{m}_{1j}^{(1)}$ corresponding to the statistic found by solving $\tilde{P}_1$ with a fractal background (SNR about 0dB). Pixel (m, n) corresponds to $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$.

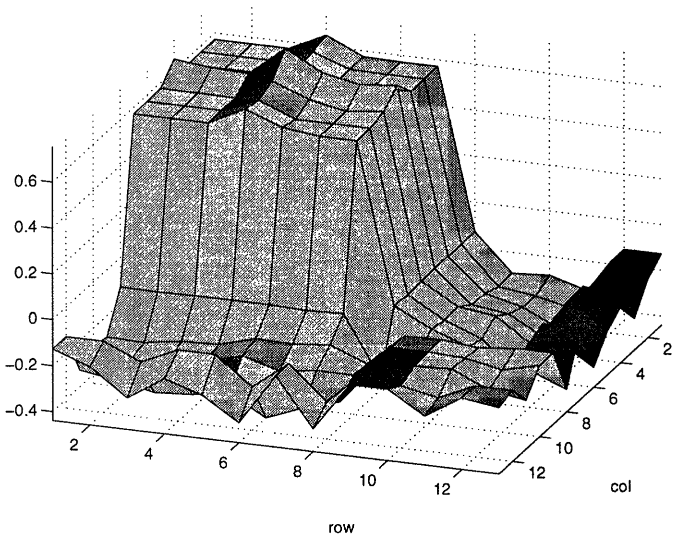


Figure 5-6: This figure illustrates $\tilde{m}_{1j}^{(1)}$ corresponding to the statistic found by solving P_1 with a fractal background (SNR about 0dB). Pixel (m, n) corresponds to $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$.

result is encouraging and, as we shall see, translates into a real gain in performance.

5.7.2 Solutions to P_i

In this example we illustrate results of sensitivity analysis similar to that shown in Figures 5-5 and 5-2 but now with the statistic corresponding to the exact solution of P_1. Figure 5-6 illustrates the value of $\tilde{m}_{1j}^{(1)}$.

It is a bit difficult to see the difference between Figures 5-6 and 5-5 so in Figure 5-7 we have plotted the difference between these plots. That is, we plot

$$\tilde{m}_{1j}^{(1)}\Big|_{P_1} - \tilde{m}_{1j}^{(1)}\Big|_{\tilde{P}_1} .$$

We see from Figure 5-7 that the exact solution to P_1 provides a higher standard-deviation-normalized conditional mean when the composite hypothesis associated with the statistic is true and a lower one when other composite hypotheses are true (the seemingly high "wings" at the right and bottom of the figure stay below zero).

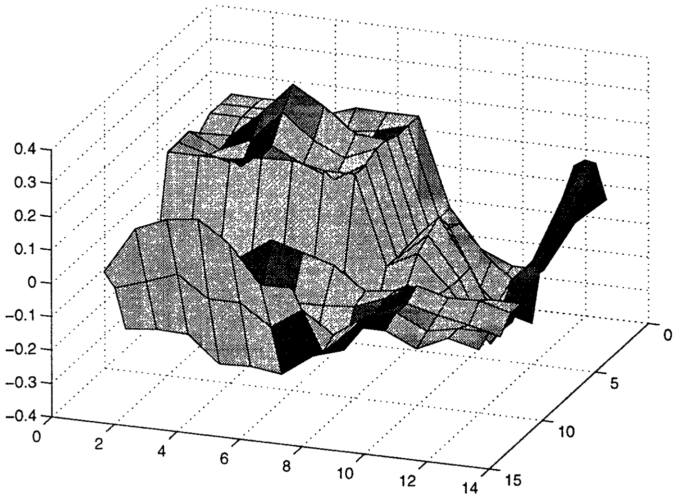


Figure 5-7: This figure illustrates the difference $\tilde{m}_{1j}^{(1)}\Big|_{P_1} - \tilde{m}_{1j}^{(1)}\Big|_{\tilde{P}_1}$ with a fractal background (SNR about 0dB). Pixel (m, n) corresponds to $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$.

5.7.3 Coarse Scale Comparison

In this section we compare the performance of the SC-revised algorithm with the SC algorithm of Chapter 4. We shall only do this comparison for the coarsest scale composite hypothesis test. That is, both algorithms are terminated after the first scale. For the following comparison we define a detection to mean that the region selected by the algorithm is the one which contains the anomaly. In Figures 5-8 and 5-9 we illustrate the probability of detection for each possible 4 × 4 anomaly in a 16 × 16 field. Pixel (i, j) in each of these figures corresponds to the 4×4 anomaly with upper left corner at (i, j) in the image domain. The value associated with that pixel is the probability of detection. Figure 5-8 illustrates the probability of detection values for the SC algorithm (using the simple statistic of Chapter 4). Figure 5-9 illustrates the probability of detection values for the SC-revised algorithm (using the optimized statistic found by solving $\tilde{P}_i$). The background for both is fractal, the SNR is 0dB and the ABR is -0.5dB. Carefully examining the color bar, we can see that the optimized statistics have higher worst case performance. The optimized statistic seems to trade off the excellent performance at the corners for better worst case performance. This is consistent with the form of the optimization problem P_i and its approximation $\tilde{P}_i$, namely,

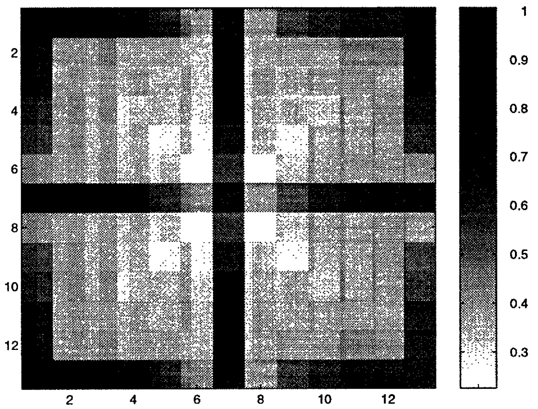


Figure 5-8: This figure illustrates the probability of detection for 4 × 4 anomalies using the simple statistic with a fractal background (SNR about 0dB, ABR about -0.5dB). The standard deviation for each P_d value is less than 0.03 and 250 Monte Carlo runs were conducted for each data point. Pixel (m, n) corresponds to $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$.

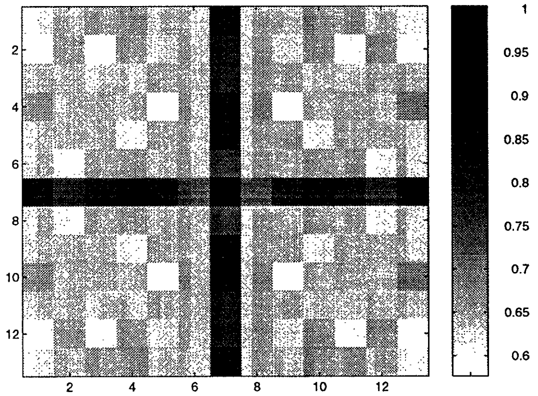


Figure 5-9: This figure illustrates the probability of detection for 4 × 4 anomalies using the optimized statistic (solutions to $\tilde{P}_i$) with a fractal background (SNR about 0dB, ABR about -0.5dB). The standard deviation for each P_d value is less than 0.03 and 250 Monte Carlo runs were conducted for each data point. Pixel (m, n) corresponds to $\mathbf{f}_a = \mathbf{b}_{4,16}(m, n)$.

that they are max-min problems.

We have repeated the analysis illustrated in Figures 5-8 and 5-9 for both a white and fractal background and at a number of different ABR values (to vary the ABR we simple changed the anomaly intensity). This analysis is summarized in Figure 5-10 which compares the performance of the coarse scale test for four different cases. What is plotted in this figure is the minimum probability of detection as a function of ABR. The minimum is taken over

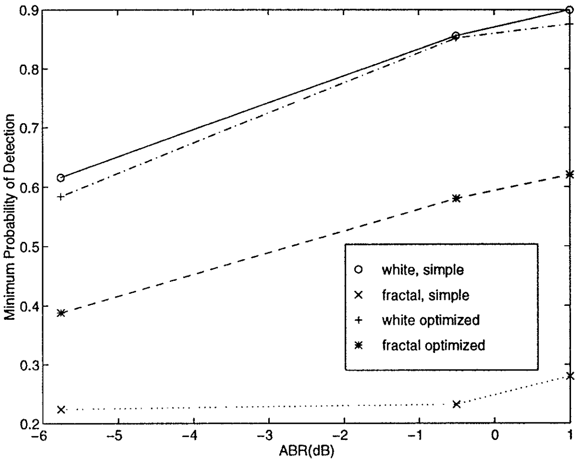


Figure 5-10: This figure compares the minimum probability of detection as a function of ABR for each of four cases: white background and simple statistic (top curve), white background and optimized statistic (solutions to $\tilde{P}_i$, curve second from top), fractal background and optimized statistic (solutions to $\tilde{P}_i$, curve second from bottom), fractal background and simple statistic (bottom curve). The standard deviation for each P_d value is less than 0.06 and 250 Monte Carlo runs were conducted for each data point. In all cases SNR=0dB.

all possible 4 × 4 anomalies (e.g., taking the minimum value of Figure 5-8). The top curve corresponds to a white background and the simple statistic. The curve just below the top corresponds to a white background and the optimized statistic (solutions to $\tilde{P}_i$). The curve second from the bottom corresponds to the fractal background and the optimized statistic (solutions to $\tilde{P}_i$). The bottom curve corresponds to the fractal background and the simple statistic.

Taking into consideration the standard deviation error bars (all less than 0.06 but not shown in Figure 5-10 for visual clarity), the top two curves are essentially identical. This implies that the simple statistic is about as good as the optimized ($\tilde{P}_i$) one for the white background case. Turning to the bottom two curves we see that the optimized statistic is significantly better than the simple one for the fractal background case. For all curves, $P_{d,min}$ increases with increasing ABR which is as expected.

5.8 Another Optimization Problem

The solution to the optimization problem P_i introduced in Section 5.3 is a set of linear weights for one statistic, ℓ_i, corresponding to one composite hypothesis, $\mathcal{H}_i$. There is no guarantee, however, that the set of statistics so found will work well together. For example, consider the following situation. Suppose we have a set of optimized statistics found by solving P_i for $i \in \{1,2,3,4\}$. Suppose ℓ_1 tends to be small when a particular hypothesis, say $H_3 \notin \mathcal{H}_1$, is true. This is good since H_3 is not in the set of hypotheses for which we want ℓ_1 to be large. But, what if none of the other statistic is particularly large when H_3 is true either? The general difficulty is that each statistic may be optimized with respect to P_i but there may still exist some hypothesis to which none seem particularly sensitive. Put another way, there may be a hypothesis which does not seem to belong in *any* composite hypothesis set. In this section we formulate a different optimization problem to address this potential difficulty. We do not attempt to solve this new problem but merely formulate it to indicate another type of optimization procedure one might consider when designing a MSHT.

Recall the definition of the standard-deviation-normalized conditional mean of ℓ_i:

$$\tilde{m}_{ij} \triangleq \frac{m_{ij}}{\sigma_{\ell_i|H_j}} = \frac{\mathbf{a}_i^T(\mathbf{T}\mathbf{b}_j)}{\sqrt{\mathbf{a}_i^T \mathbf{\Lambda_g} \mathbf{a}_i}}. \tag{5.5}$$

Again, it will be sufficient to consider only $\mathbf{a}_i$ for which $||\mathbf{a}_i||_{\mathbf{\Lambda_g}} = 1$. We wish $\tilde{m}_i j$ of equation (5.5) to be large for the values of j such that $H_j \in \mathcal{H}_i$. Ensuring that $\tilde{m}_{ij}$ to be large for the appropriate set of hypotheses H_j is obviously desirable but we need something more to avoid the difficulty alluded to in the introduction of this section. We don't just need ℓ_i to be large, we need it to be larger than *all* the other statistics when H_j is true and is an element of $\mathcal{H}_i$. This desire may be captured by maximizing the cost function

$$\min_k \left\{ \mathbf{a}_{i_k}^T \mathbf{T}\mathbf{b}_k - \mathbf{a}_i^T \mathbf{T}\mathbf{b}_k \right\}, \tag{5.6}$$

where $i_k = r$ if k is such that $H_k \in \mathcal{H}_r$ for $r \neq i$. In words, the argument of the min is the difference in the conditional mean value of two statistics conditioned on the same hypothesis, H_k. The right term of this difference is the conditional mean of ℓ_i. The left term of this difference is the conditional mean of ℓ_{i_k}. The index i_k takes on the value r such that $H_k \in \mathcal{H}_r$. The hypotheses H_k range over all hypotheses which are *not* in $\mathcal{H}_i$. This cost function attempts to measure how much more sensitive ℓ_{i_k} is to hypothesis H_k than ℓ_i.

We wish to maximize equation (5.6) as a function of $\mathbf{a}_i$. Rather than do so for each i separately, we may take advantage of the symmetry of the composite hypothesis test structure and measurements at the coarsest scale and find the optimal weights for all the coarse scale statistics at once. Since our data are wide sense stationary, our measurements are equispaced in angle in $[0, \pi)$ and taken at regular offset intervals, and since our composite hypotheses have a quadrantal symmetry, there is no need to do this maximization for each i. The $\mathbf{a}_i$ will simply be permuted versions of one another: $\mathbf{a}_i = \mathbf{R}_i \mathbf{a}_1$ for some permutation matrix $\mathbf{R}_i$. Putting all this together, we arrive at the optimization problem

$$\begin{aligned} &\hat{\mathbf{a}}_1 = \arg\max_{\mathbf{a}_1}\min_{k} \left\{ (\mathbf{T}\mathbf{b}_k)^T (\mathbf{R}_{i_k} - \mathbf{I})\mathbf{a}_1 \right\} \ , \ \forall H_k \notin \mathcal{H}_1 \\ &\text{where } i_k = r \Leftrightarrow H_k \in \mathcal{H}_r \ , \ r \in \{2,3,4\} \\ &\text{subject to } ||\mathbf{a}_i||^2_{\mathbf{\Lambda_g}} \leq 1 \, . \end{aligned}$$

It is a straight forward exercise to recast this problem as a linear programming one as we have done for P_i above.

Chapter 6

Conclusion

In Section 6.1 of this chapter we highlight the most significant contributions of this thesis. The work discussed in the main body of this thesis (Chapters 3 through 5) raises a variety of questions and issues for future research. Some of these are outlined in Section 6.2. We conclude this thesis with some closing remarks in Section 6.3.

6.1 Thesis Contributions

The primary applied goal in this thesis was to develop computationally efficient data domain methods for the single anomaly detection and localization problems from tomographic data. The essence of these problems is to characterize a region of an image which differs statistically from a well modeled background field. A secondary, but no less important goal, was to understand the structure of these problems—specifically, how the nature of the problems change with background covariance. A binary hypothesis testing framework was employed for this purpose.

In our approach to the anomaly detection and localization problems we developed the abstract framework of the multiscale hypothesis test. The notion of a MSHT is conceptually simple. It is a sequence of composite hypothesis tests where the range of the composite hypotheses considered at a stage in the sequence is a strict subset of the range of the composite hypotheses considered at the previous stage. The main difficulty associated with a MSHT is in determining the detailed structure of the test. Challenges include how to choose the form of the composite hypotheses and statistics so that the resulting test effectively zooms in on the correct hypothesis. In the following subsections we emphasize the main results of our efforts.

6.1.1 Impact of Background Covariance

At each stage of our analysis of the anomaly detection and localization problems we compared results for two different types of background field statistics: white and fractal. Our main analytical results in this comparison are found in Chapter 3 where we showed, using a binary hypothesis framework, that the structure of the problem depends crucially on the background covariance matrix. Our performance bound results in that chapter indicate that detection is easier in the presence of a white, rather than fractal, background. Our ambiguity analysis suggest that a spatial zooming approach to anomaly localization may be feasible in the white background case but may prove difficult in the fractal background case. (Analysis later in the thesis, which we review below, showed that spatial zooming is indeed feasible in both cases provided the statistic is chosen appropriately.) The analysis of a simple one-dimensional problem provided some additional insight as to how the background covariance relates to performance and ambiguity.

6.1.2 Computationally Efficient Detection and Localization

In this thesis we have viewed the anomaly detection and localization problems as M-ary hypothesis testing problems. The optimal M-ary hypothesis test is easy to formulate but computationally infeasible even for a relatively restricted class of anomalies due to the overwhelming number of hypotheses which must be considered. One main contribution of this thesis has been the development of computationally efficient alternatives to the full M-ary hypothesis test. Our idea, introduced in Chapter 4, is to localize the anomaly in a spatial scale-recursive manner. The efficiency of such a method is achieved by discarding many hypotheses with a few small composite hypothesis tests. One such method (the SC algorithm) first localizes the anomaly to a coarse scale (large area) region and then to successively finer scale (smaller area) regions. We also introduced another algorithm (the MC algorithm) which retains more hypotheses at each scale, effectively delaying the difficult decision as to which ones to discard until a finer scale. We showed that by retaining more hypotheses at each scale, the MC algorithm achieves better performance at the expense of higher computational complexity.

Both the SC and the MC algorithms of Chapter 4 are based on a sequence of composite hypothesis tests. At each stage a set of statistic values (each value associated with a different region) are compared and one (SC) or several (MC) regions are selected based on these values. Both algorithms rely upon an intuitively natural, but not necessarily good, choice of statistic. We showed that this statistic yields reasonably good performance compared to the optimal M-ary hypothesis test for the case of a white background but poor performance in the fractal background case. Finding a good statistic is one of the main challenges of multiscale hypothesis testing. We elaborate on this, and other, challenges in the next section.

6.1.3 Multiscale Hypothesis Testing

The main conceptual contribution of this thesis is the general multiscale hypothesis testing framework, first introduced in Chapter 2. We have discussed the three main aspects of a MSHT: the form of the composite hypothesis sets, the form of the statistics, the decision structure. We emphasized that the multiscale nature of a MSHT need not have an interpretation in a spatial or temporal domain. The view that MSHTs effect statistical rather than spatial (or temporal) zooming led to the issue of how to define the composite hypothesis sets which comprise a MSHT. We did not exploit this degree of freedom in this thesis, however, and restricted attention to composite hypotheses which are associated with contiguous regions of the image domain.

The issue of how to choose appropriate statistics for a MSHT was addressed in Chapter 5. We introduced an optimization problem based on natural criteria. We solved this problem both approximately (with a linear programming approximation) and exactly (using Lagrange duality and quadratic programming). The solution to this optimization problem is an improved MSHT statistic. We explored the structure of optimized statistics and compared them to the ad hoc statistics introduced in Chapter 4. A significant conclusion is that the intuitively natural statistic introduced in Chapter 4 is close to the optimized statistic for a white background case but quite different for the fractal background case.

Finally, we used these optimized statistics in place of natural ones of Chapter 4 in the SC algorithm. Our results showed a vast improvement in the fractal background case and no improvement in the white background case. This indicates that the statistics of Chapter 4 are not necessarily good ones for all background types.

6.2 Directions for Future Work

In this thesis we have made some significant contributions to the anomaly characterization problem and established the general multiscale hypothesis testing framework. Our work in these areas only scratch the surface of possible research directions, however. And where we have made our mark, we have opened up a host of issues and questions. In this section we outline some of these issues and questions as we indicate directions for possible future work.

6.2.1 Anomaly Characterization

There are a variety of ways to build on the anomaly detection and localization methods developed in this thesis. In this section we mention a few.

Performance and Ambiguity Structure

In Chapter 3 we investigated a performance bound and anomaly ambiguity using a BHT framework. Analysis of a one-dimensional signal provided some insight into our results. Our one-dimensional analysis was quite simple in several respects: the signal length was small (length three) and we assumed direct signal measurements (not something analogous to tomographic measurements). Extensions of this work in one dimension might focus on lifting these simplifications. And extending the analysis to the full two-dimensional tomography problem would be difficult but likely insightful.

Multiple Anomaly Detection and Localization

In this thesis we have considered only the single anomaly case. There are at least two main ways to apply multiscale hypothesis tests to the multiple anomaly problem: sequential and parallel. A sequential method might first localize one anomaly and then subtract its estimated contribution from the data before localizing a second anomaly. This is repeated until all anomalies are localized.

A parallel method attempts to localize all the anomalies simultaneously. One such method is discussed in [19–21] for geophysical inverse problems. This method is similar to the MC algorithm presented in Chapter 4 which represents a reasonable way to approach the multiple anomaly problem. Key difficulties in augmenting the MC algorithm include finding data-driven ways to decide how many composite hypotheses ought to be further subdivided and how many ought to be selected in the post-processing stage.

Non-Constant Intensity Anomalies

We have assumed that the anomaly has constant non-negative intensity. This assumption may be relaxed in two stages. Non-constant but still non-negative anomalies ought to be considered first. It is likely that such anomalies may be detected and localized quite well with the methods presented in this theses. The next step is to consider arbitrary intensity anomalies. Such anomalies, however, most likely will not be easily detected and localized by our methods and new techniques will be needed.

A model based approach for the arbitrary intensity anomaly seems promising. With a probabilistic model for the intensity, the generalized likelihood ratio test (GLRT) may be a useful tool: the anomaly intensity field is estimated assuming a certain support (indicator function) yielding a generalized likelihood. Ratios of these likelihoods are then compared as discussed in Section 2.3.2.

Arbitrary Shaped Anomalies

Arbitrary shaped anomalies pose a bigger challenge than arbitrary intensity square ones. One possibility is to view arbitrary shaped anomalies as many square (or rectangular) primitive anomalies and apply a multiple anomaly localization method (see above). Then, in a post-processing stage, the primitives may be merged when appropriate.

6.2.2 Multiscale Hypothesis Testing Theory

In this thesis we have defined the general notion of a MSHT and indicated how optimized ones might be discovered. While we have treated MSHTs independently from the anomaly detection and localization problems, we have developed multiscale hypothesis testing theory with these problems in mind. Therefore, the range of applicability of our methods (e.g., those of Chapter 5) is limited to problems with similar structure (e.g., linear, Gaussian). Since multiscale hypothesis testing is of interest in many disciplines, a more general theory would be of great value.

6.2.3 Other Extensions and Issues

Optimal Experiment Design

In our work we have assumed that the projection angles and spacing are fixed. Clearly better performance may be achieved if the projection positions are not fixed. There are two general ways of incorporating such flexibility. With a prior model of the anomaly location and size, one could determine the optimal positions for all the projections. Without a prior model, one could do so sequentially. For example, a few judiciously place projections could be used to make an initial crude estimate of the anomaly's location and size. Then, based on this estimate, the optimal positions of the next few projections may be set. Using this new data, the anomaly localization may be re-estimated, etc.

Multiscale Reconstruction and Imaging

A long-term goal in tomography research is the development of reconstruction methods with are sensitive to the quantity and quality of the available data in a space-adaptive way. Standard reconstruction methods (e.g., convolution back-projection) reconstruct the entire image at the finest scale regardless of the available data. Hence when data are noisy, sparse, irregularly sampled, or angle-limited, the reconstruction suffers from severe streaking artifacts. Spatially-varying, multiscale reconstruction, however, would control this reconstruction greed by estimating regions of the image at varying scales as warranted by the data. Several authors have already begun work towards this goal. See, for example, [2–5, 9, 26, 27, 32].

The methods presented in this thesis represent means of multiscale *detection*. Extending these methods to multiscale estimation might be done in stages, first considering multiple anomalies and then considering multiple anomalies with arbitrary intensity. Each link in this research chain is a step toward the goal of multiscale imaging.

6.3 Closing Remarks

In this thesis we have investigated the structure of the anomaly detection and localization problems from tomographic data, developed efficient methods for solving these problems, introduced the general multiscale hypothesis testing framework, and provided ways of finding good statistics for a MSHT. Our investigation of the structure of the anomaly detection and localization problems has shown that performance of detection and localization algorithms relies crucially on the nature of the background field covariance.

In a first step toward the anomaly detection and localization problems, we developed methods which do not take into consideration the unique difficulties associated with a particular background structure; our first detection and localization methods used intuitively natural statistics which were not well suited for all classes of background fields. A deeper consideration of the flexibility of the MSHT framework led naturally to the consideration of an optimization problem whose solution provided a better statistic given the particular background structure of the problem.

One additional and significant aspect of our application of multiscale hypothesis testing to the anomaly detection and localization problems is that all processing is conducted entirely in the data domain. Image reconstruction plays no part in our techniques. Therefore, we have addressed a major challenge associated with the tomographic anomaly characterization problem. We have presented efficient and effective data domain anomaly characterization methods.

The most general of our contributions is the elucidation of the flexibility of the MSHT framework. Multiscale hypothesis testing may be applied to a host of problems in a wide range of fields including computer vision and remote sensing. Problems in these fields often take the form of M-ary hypothesis testing problems where M is prohibitively large. As we have shown in the tomographic anomaly characterization problem, a MSHT is an efficient and effective alternative to the daunting optimal M-ary hypothesis test.

Appendix A

Fractal Field Covariance Matrix

In this appendix we describe the structure of the fractal field covariance matrix and show how it is used to generate a fractal field background. We assume that this background is wide sense stationary (WSS) and periodic. That is, we imagine that the field is on a periodic toroidal lattice. The WSS and periodicity assumptions imply that the fractal field covariance matrix is doubly circulant or, equivalently, that it is diagonalized by $\mathbf{F} \triangleq \mathbf{D} \otimes \mathbf{D}$ where $\mathbf{D}$ is the discrete Fourier transform (DFT) matrix and $\otimes$ represents the Kronecker product. We show here that this DFT diagonalization may be exploited for fast computation of the fractal field covariance matrix, its positive definite symmetric square root, and sample paths.

The background, f_b, is discretized on a $N \times N$ grid ($N_p = N^2$) so that

$$f_b(x,y) = \sum_{j=1}^{N^2} f_{b_j} p_j(x,y) \, ,$$

where $p_j(x,y)$ takes on the value one over the j^{th} pixel and zero elsewhere. The f_{b_j} are ordered in a vector denoted by $\mathbf{f}_b$ and the pixel located at (a_1, b_1) is mapped to the j^{th} location in $\mathbf{f}_b$ by

$$j = (b_1 - 1)N + a_1 \, ,$$

where $b_1, a_1 \in \{1, 2, \ldots, N\}$.

If $\mathbf{\Lambda}_f$ is the covariance matrix for the fractal-field background with spectral parameter γ then the correlation between the pixel (a_1, b_1) and (a_2, b_2) is given by the mn^{th} element of $\mathbf{\Lambda}_f$ where

$$\begin{aligned} m &= (b_1 - 1)N + a_1 \,, \\ n &= (b_2 - 1)N + a_2 \,. \end{aligned}$$

The entries of the covariance matrix $\mathbf{\Lambda}_f$ are given by

$$[\mathbf{\Lambda}_f]_{mn} = \left(\frac{1}{N^2}\right) \left[\sum_{(c_u, c_v) \in \chi} \frac{e^{j(\omega_u a + \omega_v b)}}{(\omega_u^2 + \omega_v^2)^{\gamma/2}} \right] + \left(\frac{1}{N^2}\right) \left(\frac{N}{2\pi}\right)^{\gamma} , \tag{A.1}$$

where

$$\begin{aligned} a &= a_1 - a_2 \,, \\ b &= b_1 - b_2 \,, \\ \omega_u &= \left(\frac{2\pi}{N}\right) c_u \,, \\ \omega_v &= \left(\frac{2\pi}{N}\right) c_v \,, \end{aligned}$$

and

$$\chi = \left\{ (c_u, c_v) \mid c_u = -\frac{N}{2} + 1, \ldots, \frac{N}{2}; c_v = -\frac{N}{2} + 1, \ldots, \frac{N}{2} \right\} \setminus \{(0,0)\} \,.$$

Once $\mathbf{\Lambda}_f$ is obtained, the fractal field is generated as follows. Let ν be a zero-mean, white Gaussian random vector with unit intensity. That is,

$$\nu \sim \mathcal{N}(0, \mathbf{I}) \,.$$

Then a fractal field background sample path is obtained by computing

$$\mathbf{f}_b = \mathbf{\Lambda}_f^{1/2} \nu \,,$$

where $\mathbf{\Lambda}_f^{1/2}$ is the positive definite symmetric square root of $\mathbf{\Lambda}_f$.

While equation (A.1) provides a straight forward way to compute the mn^{th} of $\mathbf{\Lambda}_f$, direct application of it does not take advantage of the structure of $\mathbf{\Lambda}_f$. Specifically, since $\mathbf{\Lambda}_f$ is diagonalized by $\mathbf{F} \triangleq \mathbf{D} \otimes \mathbf{D}$ and since we know the eigenvalues of $\mathbf{\Lambda}_f$, we can specify $\mathbf{\Lambda}_f$ in the Fourier and then transform to the spatial domain.

Let the diagonal matrix $\mathbf{A}$ contain the eigenvalues of $\mathbf{\Lambda}_f$ on its diagonal. These eigenvalues are the coefficients in the sum of equation (A.1). That is, the eigenvalues, $\lambda_{u,v}$, are

$$\lambda_{u,v} = \frac{1}{(\omega_u^2 + \omega_v^2)^{\gamma/2}} \,,$$

where ω_u and ω_v take on the same values as above. We must also include the eigenvalue for the DC value,

$$\lambda_{0,0} = \left(\frac{N}{2\pi}\right)^{\gamma} .$$

Therefore, $\mathbf{\Lambda}_f$ is obtained by

$$\mathbf{\Lambda}_f = \frac{1}{N^2} \mathbf{F}^* \mathbf{A} \mathbf{F} \,,$$

where the superscript $*$ indicates Hermitian (conjugate) transpose.

With no more work in the Fourier domain, $\mathbf{\Lambda}_f$ can be obtained using the inverse two-dimensional FFT. To do so, one must reshape the eigenvalue matrix, $\mathbf{A}$, so that it is $N \times N$. This reshaped matrix, $\tilde{\mathbf{A}}$, has an eigenvalue of $\mathbf{\Lambda}_f$ in each entry. Applying the two-dimensional FFT to $\tilde{\mathbf{A}}$ yields a matrix $\mathbf{H}$ which contains all the information necessary to build $\mathbf{\Lambda}_f$. To build $\mathbf{\Lambda}_f$ out of $\mathbf{H}$, one must take each column of $\mathbf{H}$ and build a circulant matrix. This yields N circulant matrices. The fractal field covariance matrix is obtained by arranging these circulant matrices in a block circulant matrix, $\mathbf{\Lambda}_f$. Notice that direct application of equation (A.1) requires $\mathcal{O}(N^4)$ operations while the FFT implementation is $\mathcal{O}(N^2 \log N^2)$.

Finally, to generate sample paths we require the positive definite symmetric square root of $\mathbf{\Lambda}_f$. While this can be obtained by blind application of any number of matrix factorization algorithms (e.g., MATLAB's `sqrtm()` function), it can be computed very quickly using the DFT factorization. Assume we have generated the eigenvalue matrix $\mathbf{A}$. Then let $\mathbf{B}$ be the diagonal matrix with each diagonal entry the square root of the corresponding diagonal entry of $\mathbf{A}$. Then the positive definite symmetric square root of $\mathbf{\Lambda}_f$ is given by

$$\mathbf{\Lambda}_f^{1/2} = \frac{1}{N^2}\mathbf{F}^*\mathbf{B}\mathbf{F}$$

Again, this may be computed using the two-dimensional FFT.

Appendix B

Maximum Likelihood Anomaly Estimation

In this appendix we consider the maximum likelihood (ML) estimation of the anomaly field $\mathbf{f}_a$. Recall that the observational model is

$$\mathbf{g} = \mathbf{T}\mathbf{f}_a + \mathbf{T}\mathbf{f}_b + \mathbf{n}\,,$$

where

$$\begin{aligned} \mathbf{f}_b &\sim \mathcal{N}(0, \mathbf{\Lambda})\,, \\ \mathbf{n} &\sim \mathcal{N}(0, \lambda\mathbf{I})\,. \end{aligned}$$

Viewing the anomaly field as a vector of non-random parameters, the data are jointly Gaussian:

$$p_{\mathbf{g}}(\mathbf{G}; \mathbf{f}_b) = \frac{1}{|2\pi\mathbf{\Lambda}_{\mathbf{g}}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{G} - \mathbf{T}\mathbf{f}_a)^T {\mathbf{\Lambda}_{\mathbf{g}}}^{-1}(\mathbf{G} - \mathbf{T}\mathbf{f}_a)\right\}\,.$$

Therefore, the ML estimate of $\mathbf{f}_a$ is

$$\begin{aligned}\hat{\mathbf{f}}_a &= \arg\max_{\mathbf{y}} p_{\mathbf{g}}(\mathbf{G};\mathbf{y}) \\ &= \arg\min_{\mathbf{y}} (\mathbf{G}-\mathbf{T}\mathbf{y})^T \mathbf{\Lambda_g}^{-1}(\mathbf{G}-\mathbf{T}\mathbf{y})\end{aligned}$$

The minimum is found by setting the gradient of $(\mathbf{G}-\mathbf{T}\mathbf{y})^T \mathbf{\Lambda_g}^{-1}(\mathbf{G}-\mathbf{T}\mathbf{y})$ with respect to $\mathbf{y}$ to zero. Doing so yields

$$\hat{\mathbf{f}}_a = (\mathbf{T}^T \mathbf{\Lambda_g}^{-1}\mathbf{T})^{-1}\mathbf{T}^T \mathbf{\Lambda_g}^{-1}\mathbf{G}\,.$$

There are several problems associated with this ML estimate. First, recall that the dimension of the tomographic projection matrix $\mathbf{T}$ is $N_\phi N_s \times N_p$, i.e., the number of projection angles times the number of samples per angle by the number of image domain pixels. In low data cases (e.g., limited angle or low offset sampling rate) it may be the case that $N_\phi N_s < N_p$ in which case the rank of $\mathbf{T}$ is less than N_p. In such a case, when the data have fewer degrees of freedom than the dimensionality of the estimate, $\mathbf{T}^T \mathbf{\Lambda_g}^{-1}\mathbf{T}$ is not invertible and the ML estimate must be approximated in some way. Even if the ML estimate may be computed, it will not exhibit the assumed structure of the anomaly field, namely that $\mathbf{f}_a = c\mathbf{b}_{s,N}(i,j)$ because no such structure has been imposed on the ML estimate. Further, the ML estimate does not directly provide an indication of whether or not an anomaly exists and, so, it cannot be used to directly solve the anomaly detection problem.

Appendix C

Anomaly Intensity Assumption

In this appendix we justify the statement made in Section 4.2.1 that assuming knowledge of the anomaly intensity, c, results in no loss of generality. The hypothesis test considered in Section 4.2.1 is of the form

$$\begin{aligned} H_0 &: \mathbf{f}_a \equiv 0\,, \\ H_i &: \mathbf{f}_a = c\mathbf{b}_i\,, \end{aligned}$$

where $i \in \{1, 2, 3, 4\}$. And the form of the statistic is

$$\ell_i(\mathbf{G}) = \mathbf{T}\mathbf{b}_i^T \boldsymbol{\Lambda}_{\mathbf{g}}{}^{-1}\mathbf{G}\,. \tag{C.1}$$

If the anomaly's intensity were not known, we would have to estimate it in some way and, in that case we rewrite the above hypothesis test as

$$\begin{aligned} H_0 &: \mathbf{f}_a \equiv 0\,, \\ H_i &: \mathbf{f}_a = \hat{c}_i\mathbf{b}_i\,, \end{aligned}$$

where $\hat{c}_i$ is an estimate of the anomaly's intensity assuming H_i is true. The form of the statistic would also include a $\hat{c}_i$ term:

$$\ell_i(\mathbf{G}) = \hat{c}_i \mathbf{T}\mathbf{b}_i^T \mathbf{\Lambda_g}^{-1}\mathbf{G}\,. \tag{C.2}$$

Here we consider the maximum likelihood (ML) estimate of $\hat{c}_i$ (see Appendix B for a discussion of ML estimation). Conditioned on H_i, the data are jointly Gaussian:

$$H_i : \mathbf{g} \sim \mathcal{N}(c_i \mathbf{T}\mathbf{b}_i, \mathbf{\Lambda_g})$$

Therefore, the ML estimate is

$$\hat{c}_i = \frac{(\mathbf{T}\mathbf{b}_i)^T \mathbf{\Lambda_g}^{-1}\mathbf{G}}{(\mathbf{T}\mathbf{b}_i)^T \mathbf{\Lambda_g}^{-1}\mathbf{T}\mathbf{b}_i}\,.$$

Plugging this into equation (C.2) yields

$$\ell_i(\mathbf{G}) = \frac{\left((\mathbf{T}\mathbf{b}_i)^T \mathbf{\Lambda_g}^{-1}\mathbf{G}\right)^2}{(\mathbf{T}\mathbf{b}_i)^T \mathbf{\Lambda_g}^{-1}\mathbf{T}\mathbf{b}_i}\,. \tag{C.3}$$

Except at the coarsest scale, where the $\mathbf{b}_i$ have quadrantal symmetry, the denominator of equation (C.3) is not independent of i. However, we have seen in our experiments that the denominator is, to a good approximation, nearly independent of i at each scale. So, for the purposes of comparing statistics at the same scale, the denominator may be ignored. Therefore, the only difference between the statistic employed in Section 4.2.1 and that of equation (C.3) is that the later is the square of the former. However, since the anomaly intensity is known to be non-negative, $\hat{c}$ will be non-negative so that comparing statistics of the form of equation (C.3) is equivalent to comparing their square root (i.e., equation (C.1) and statistics of the form used in Chapter 4). Hence assuming knowledge of the anomaly intensity results in no loss of generality.

Appendix D

Formulations for Exact Solutions of P_i

In this appendix we describe in detail how to exactly solve the non-linear optimization problem, P_i posed in Section 5.3. The results of the analysis in this appendix are presented in Section 5.5 without proof.

In composite hypothesis testing problems for which there is no uniformly most powerful test it is not clear what choice of statistics will yield good performance. Yet performance relies critically on the ability of the statistics to distinguish between composite hypotheses. Therefore, finding good statistics is of crucial importance. In this appendix we are concerned with finding a statistic which is well matched to a composite hypothesis in the following sense. We want the statistic to be as large as possible when its associated composite hypothesis contains the true hypothesis and as small as possible otherwise, in some sense to be made precise. Toward this goal, we formulate a convex non-linear optimization problem whose solution provides a statistic with maximal composite hypothesis distinguishability. We will provide several reformulations of this optimization problem, each providing its own insight. We shall show that the dual of this problem is a quadratic programming problem which can be solved with off-the-shelf software.

D.1 Assumptions and Definitions

Consider a problem in which, under hypothesis H_k, the observed data vector, $\mathbf{g}$ has the form

$$H_k : \mathbf{g} = \mathbf{T}\mathbf{b}_k + \vartheta\,,$$

where $\mathbf{T} \in I\!R^{m\times n}$ and ϑ is zero mean Gaussian additive noise which is independent of k and has positive definite covariance $\mathbf{\Lambda}$. (Note, the anomaly detection and localization problems have this form.) This measurement equation and associated modeling assumptions are all that are required to apply our method for finding optimized statistics for composite hypothesis testing problems.

Denote the global set of hypotheses by $\mathcal{H}$:

$$\mathcal{H} = \{H_0, H_1, \ldots, H_{M-1}\}\,.$$

In a composite hypothesis formulation, the elements of $\mathcal{H}$ are grouped into subsets $\mathcal{H}_i$ with the property that

$$\mathcal{H} = \bigcup_{i=0}^{N-1} \mathcal{H}_i\,,$$

where $N \leq M$. Note that these subsets need not be mutually exclusive but they must be collectively exhaustive as is indicated by the above property.

We are concerned with finding a *linear* statistic $\ell_i = \mathbf{a}_i^T\mathbf{g}$ to associate with composite hypothesis $\mathcal{H}_i$ for each i. With the intention of using these statistics in a comparison test (i.e., we shall declare $\mathcal{H}_i$ true if $\ell_i \geq \ell_j$ for all j) we will formulate an optimization problem to choose the linear weight vector $\mathbf{a}_i$ so that ℓ_i is maximally sensitive to $\mathcal{H}_i$. Put simply, we will impose a condition that forces ℓ_i to be, on average, large when $\mathcal{H}_i$ is true (i.e., it contains the true hypothesis) and small otherwise.

Before introducing the problem, we make the following definitions and observations. Define the conditional mean and variance of the statistic ℓ_i as

$$\begin{aligned} m_{ij} &\triangleq E[\ell_i | H_j] = \mathbf{a}_i^T \mathbf{T} \mathbf{b}_j \,, \\ \sigma_i^2 &\triangleq \mathrm{var}[\ell_i | H_j] = \mathbf{a}_i^T \mathbf{\Lambda} \mathbf{a}_i \,. \end{aligned}$$

Notice that the conditional mean is linear in $\mathbf{a}_i$ while the conditional variance is quadratic in $\mathbf{a}_i$. Also note that the conditional variance is independent of j.

D.2 The Primal Problem

The optimization problem we consider is

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}} \min_{(j,k) \in \mathcal{A}_i} \frac{m_{ij} - m_{ik}}{\sigma_i} \,,$$

where $\mathcal{A}_i \triangleq \{(j,k) | H_j \in \mathcal{H}_i \text{ and } H_k \notin \mathcal{H}_i\}$. Reading from right to left, we see that in this optimization problem we are considering the difference between two standard-deviation-normalized conditional means. One element in the difference is m_{ij} and we want this element to be large since we constrain $H_j \in \mathcal{H}_i$. The other element in the difference is m_{ik} and we want this element to be small since we constrain $H_k \notin \mathcal{H}_i$. Therefore, we want the difference to be large. Taking the worst case difference (with the min), we maximize this with respect to $\mathbf{a}$.

Plugging in definitions, the optimization problem is

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}} \min_{(j,k) \in \mathcal{A}_i} \frac{\mathbf{a}^T \mathbf{T} \mathbf{b}_j - \mathbf{a}^T \mathbf{T} \mathbf{b}_k}{\sqrt{\mathbf{a}^T \mathbf{\Lambda} \mathbf{a}}} \,.$$

We shall find it useful to adopt a more compact, but more abstract, notation for this optimization problem. To this end, we transpose the numerator and write the problem as

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}} \frac{\min_{m \in \mathcal{A}_i} \mathbf{q}_m^T \mathbf{a}}{\sqrt{\mathbf{a}^T \mathbf{\Lambda} \mathbf{a}}} \,,$$

where $\mathbf{q}_m = \mathbf{q}_{(i,j)} \triangleq \mathbf{Tb}_j - \mathbf{Tb}_k$. It is clear from this latter formulation that the numerator is a piecewise linear concave function of $\mathbf{a}$. The denominator is the square root of a quadratic form and, since the covariance matrix is positive definite, is strictly positive for non-zero $\mathbf{a}$.

An additional simplification may be made by recognizing the fact that it is sufficient to consider linear weights $\mathbf{a}$ for which $\mathbf{a}^T\mathbf{\Lambda a} = 1$. Therefore we may write the optimization problem in the form

$$\begin{aligned} \hat{\mathbf{a}}_i &= \arg\max_{\mathbf{a}} z \\ \text{subject to} & \begin{cases} z \leq \mathbf{q}_m^T\mathbf{a}, \ \forall m \in \mathcal{A}_i \\ \mathbf{a}^T\mathbf{\Lambda a} \leq 1 \end{cases} . \end{aligned}$$

From this formulation of the optimization problem we see that the problem is one of maximizing a linear cost subject to many linear constraints and one quadratic constraint. Further, the set $\mathcal{C}_i \triangleq \{(\mathbf{a}, z) | z \leq \mathbf{q}_m^T\mathbf{a}, \ \forall m \in \mathcal{A}_i \text{ and } \mathbf{a}^T\mathbf{\Lambda a} \leq 1\}$ is the intersection of convex sets and so is itself convex.

We now make a few minor notational modifications to the optimization problem posed. First note that we may write the problem as

$$\begin{aligned} \hat{\mathbf{a}}_i &= \arg\max_{\mathbf{a}} z \\ \text{subject to} & \begin{cases} z\mathbf{e} \leq \mathbf{Qa} \\ \mathbf{a}^T\mathbf{\Lambda a} \leq 1 \end{cases} , \end{aligned}$$

where we define

$$\mathbf{Q} \triangleq \begin{bmatrix} \mathbf{q}_1^T \\ \mathbf{q}_2^T \\ \vdots \end{bmatrix},$$

and

$$\mathbf{e} \triangleq [1\ 1\ 1\ \ldots\ 1]^T .$$

Finally, we make a change of variables by defining

$$\begin{aligned} \mathbf{x} &\triangleq \mathbf{\Lambda}^{1/2}\mathbf{a} , \\ \mathbf{P} &\triangleq \mathbf{Q}\mathbf{\Lambda}^{-1/2} . \end{aligned}$$

Our final formulation of the problem, which we shall call the primal problem, is, therefore

$$\begin{aligned} &\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} z \\ &\text{subject to} \begin{cases} z\mathbf{e} \leq \mathbf{P}\mathbf{x} \\ \mathbf{x}^T\mathbf{x} \leq 1 \end{cases} . \end{aligned}$$

D.3 The Dual Problem

The primal problem posed at the end of the previous section is a convex optimization problem. In particular it belongs to the class of problems known as quadratically constrained quadratic programs (QCQPs). While methods exist for directly solving QCQPs (e.g., semidefinite programming, see [39]), we shall find it convenient to solve the primal problem by first solving a different (and simpler) problem. The simpler problem we consider is the *dual* of the primal problem. In this section we pose the dual problem and show how to obtain the optimal primal solution from the optimal dual solution. We then prove that the problem we pose in this section is indeed the dual to the one posed in the previous section.

We will later show that the dual is

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}} \mathbf{y}^T \mathbf{P}\mathbf{P}^T \mathbf{y}$$

$$\text{subject to} \begin{cases} \mathbf{e}^T \mathbf{y} = 1 \\ \mathbf{y} \geq 0 \end{cases} .$$

From the optimal dual solution, the optimal primal solution is obtained by

$$\hat{\mathbf{x}} = \frac{\mathbf{P}^T \hat{\mathbf{y}}}{\sqrt{\hat{\mathbf{y}}^T \mathbf{P}\mathbf{P}^T \hat{\mathbf{y}}}} ,$$

assuming that $\hat{\mathbf{y}}$ is not identically zero. In the case that it is then $\hat{\mathbf{x}}$ is also identically zero. Finally, $\hat{\mathbf{a}}$ is trivially recovered from $\hat{\mathbf{x}}$ by multiplying by $\mathbf{\Lambda}^{-1/2}$. Notice that this dual problem is a quadratic program (QP) problem. It has quadratic cost with linear constraints. Duality has moved the quadratic constraint of the primal problem to a quadratic cost in the dual. We next show how this is done.

The rough idea behind duality is simple. The primal problem is made difficult due to the presence of constraints. These constraints can be removed if they are incorporated into the cost function with Lagrange multipliers (turning the cost function into what we shall call a Lagrangian cost function). These Lagrange multipliers penalize deviation from the constraints. The key point is that if appropriate values of the Lagrange multipliers (i.e., appropriate penalties) could be found then the optimizing values for the unconstrained Lagrangian cost function are the same as the optimizing values of the original cost function subject to the original constraints. The dual problem aims to find the appropriate values for the Lagrange multipliers. In other words, solving the dual amounts to finding the right multipliers (penalties) so that solving the unconstrained problem (with the Lagrangian cost function) is the same as solving the original constrained problem (with the original cost function). We now show how to form the dual to the primal problem posed in the previous section.

Making minor modifications, the primal problem is

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} z$$
$$\text{subject to} \begin{cases} \mathbf{Px} - z\mathbf{e} \geq 0 \\ 1 - \mathbf{x}^T\mathbf{x} \geq 0 \end{cases} .$$

Let us call the optimal cost to the primal problem $\hat{z}$. In other words, $z = \hat{z}$ when $\mathbf{x} = \hat{\mathbf{x}}$, the optimal solution to the primal. Let us also assume that $\hat{\mathbf{x}}$ exists and that $\hat{z} < \infty$, i.e., the primal problem is feasible and has bounded optimal cost. These assumptions are not crucial but are reasonable and will allow us to sidestep some distracting technicalities. Introducing Lagrange multipliers $\mathbf{y}$ and μ, we define the Lagrangian cost function as

$$L(z, \mathbf{x}, \mu, \mathbf{y}) \stackrel{\triangle}{=} z + \mathbf{y}^T(\mathbf{Px} - z\mathbf{e}) + \mu(1 - \mathbf{x}^T\mathbf{x}) . \tag{D.1}$$

Our aim is to find values for the Lagrange multipliers such that maximizing L is the same as solving the primal problem. Toward this end we define

$$J(\mu, \mathbf{y}) \stackrel{\triangle}{=} \max_{z,\mathbf{x}} L(z, \mathbf{x}, \mu, \mathbf{y}) .$$

The function J is the maximum of the Lagrangian cost as a function of the Lagrange multipliers. For the right values of Lagrange multipliers (which we do not yet know but shall call $\hat{\mathbf{y}}$ and $\hat{\mu}$), $J(\hat{\mu}, \hat{\mathbf{y}}) = \hat{z}$, the optimal cost of the primal problem.

The dual problem, in essence, is a search over $\mathbf{y}$ and μ space to find the right values $\hat{\mathbf{y}}$ and $\hat{\mu}$. We now come to the first subtlety. What is this space? Can we allow any component of $\mathbf{y}$ and μ to be any real value? The answer is no, we must constrain the Lagrange multipliers to be non-negative. To see this, suppose we allow one of the Lagrange multipliers to be negative. For simplicity, let's consider letting μ be negative (though this argument applies equally to all the multipliers). Focus on the last term of equation (D.1). If μ is negative then we can make the Lagrangian cost arbitrarily large by making $1 - \mathbf{x}^T\mathbf{x}$ arbitrarily small. However, doing so would violate the original quadratic primal constraint.

If we allowed μ to be negative then maximizing L would never be the same as solving the original primal problem because we would always end up violating a constraint (and, furthermore, we would end up with unbounded optimal cost which violates our original assumption that $\hat{z}$ is bounded). What we have just argued is that, in searching for the $\hat{\mathbf{y}}$ and $\hat{\mu}$, we must consider only non-negative values.

Having defined the space of our search for Lagrangian multipliers, we will now introduce a function (the dual cost function) which is optimized when $\mathbf{y} = \hat{\mathbf{y}}$ and $\mu = \hat{\mu}$. This brings us the the second subtle point. For an arbitrary choice of $\mathbf{y}$ and μ is there a relationship between $J(\mu, \mathbf{y})$ and $\hat{z}$? Yes, that relationship is

$$J(\mu, \mathbf{y}) \geq \hat{z} ,$$

which holds for all values of $\mathbf{y}$ and μ. This is so because the problem of maximizing L is an unconstrained one, while the primal is a constrained problem. We may trivially achieve $J(\mu, \mathbf{y}) > \hat{z}$ by letting the Lagrange multipliers all be zero (see equation (D.1)). What we have just argued is that $J(\mu, \mathbf{y})$ is an upper bound on $\hat{z}$. This fact is know as *weak duality.*

The dual problem attempts to find the smallest such upper bound. In our case this smallest upper bound is, in fact, tight. In other words

$$\min_{\mu, \mathbf{y}} J(\mu, \mathbf{y}) = \hat{z} .$$

This fact is known as *strong duality.* We shall prove strong duality in the last section of this appendix. Therefore, the dual cost function we seek is $J(\mu, \mathbf{y})$. The Lagrange multipliers which minimize this cost function will be the ones which also cause the problem of maximizing the Lagrangian cost function, L, to be the same as solving the primal problem.

We have now found both the search space for Lagrange multipliers and the cost function which is minimized when the right multipliers are found. Putting these together we get the dual problem:

$$[\hat{\mu}\ \hat{\mathbf{y}}^T] = \arg\min_{\mu,\mathbf{y}} J(\mu,\mathbf{y})$$
$$\text{subject to} \begin{cases} \mu \geq 0 \\ \mathbf{y} \geq 0 \end{cases} .$$

All that remains is to put the dual problem into a more useful form. To begin doing so, recall that J is the maximum of L over all z and $\mathbf{x}$. A necessary condition at the maximum of L is that the gradient of L is zero. Setting the partial derivative of L with respect to z and the gradient of L with respect to $\mathbf{x}$ to zero yields the conditions

$$\left.\frac{\partial L}{\partial z}\right|_{z=\hat{z}} = 1 - \mathbf{y}^T\mathbf{e} = 0 \Longrightarrow \mathbf{y}^T\mathbf{e} = 1 \,,$$

and

$$\left.\frac{\partial L}{\partial \mathbf{x}}\right|_{\mathbf{x}=\hat{\mathbf{x}}} = \mathbf{P}^T\mathbf{y} - 2\mu\hat{\mathbf{x}} = 0 \Longrightarrow \hat{\mathbf{x}} = \frac{1}{2\mu}\mathbf{P}^T\mathbf{y} \,. \tag{D.2}$$

These are conditions, as functions of the Lagrange multipliers, which must be satisfied at the optimum of L. Plugging these conditions back into L yields

$$\begin{aligned} J(\mu,\mathbf{y}) &= \max_{z,\mathbf{x}} L(z,\mathbf{x},\mu,\mathbf{y}) \triangleq L(\hat{z},\hat{\mathbf{x}},\mu,\mathbf{y}) \\ &= \hat{z} + \mathbf{y}^T(\mathbf{P}\hat{\mathbf{x}} - \hat{z}\mathbf{e}) + \mu(1 - \hat{\mathbf{x}}^T\hat{\mathbf{x}}) \\ &= \hat{z} + \mathbf{y}^T\left(\frac{\mathbf{P}\mathbf{P}^T\mathbf{y}}{2\mu} - \hat{z}\mathbf{e}\right) + \mu\left(1 - \frac{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}{4\mu^2}\right) \\ &= \mu + \frac{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}{4\mu} \,, \end{aligned}$$

where in the last equality we have used the fact that $\mathbf{y}^T\mathbf{e} = 1$.

Having found a workable expression for J, the dual problem is to minimize it. We shall first find a necessary condition for μ at the minimum which will still leave the problem of finding $\mathbf{y}$. We will return to this latter problem in a moment. First note from equation

(D.2) that if $\mu = 0$ then $\hat{\mathbf{x}}$ is unbounded. From the form of the primal problem, such an $\hat{\mathbf{x}}$ is not a feasible solution. Therefore, the optimal value of μ cannot be zero. A necessary condition for μ at the optimum is

$$\left.\frac{\partial J}{\partial \mu}\right|_{\mu=\hat{\mu}} = 1 - \frac{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}{4\hat{\mu}^2} = 0 \Longrightarrow \hat{\mu} = \frac{1}{2}\sqrt{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}\,. \tag{D.3}$$

It remains, therefore, to formulate the problem whose solution is the optimal value for $\mathbf{y}$. But now this is easy. Having found the optimal μ we plug this into J to get

$$J(\hat{\mu}, \mathbf{y}) = \sqrt{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}\,. \tag{D.4}$$

Putting all this together, the dual problem is

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\min_{\mathbf{y}} \mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y} \\ \text{subject to} &\left\{ \begin{array}{l} \mathbf{e}^T\mathbf{y} = 1 \\ \mathbf{y} \geq 0 \end{array} \right. , \end{aligned}$$

which is the problem proposed at the start of this section. (Minimizing $\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}$ is equivalent to minimizing $\sqrt{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}$.) Recalling the fact that $\mathbf{P} \triangleq \mathbf{Q}\mathbf{\Lambda}^{-1/2}$ and $\mathbf{x} \triangleq \mathbf{\Lambda}^{1/2}\mathbf{a}$, we may rewrite the dual problem and the primal optimal solution as

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\min_{\mathbf{y}} \mathbf{y}^T\mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^T\mathbf{y} \\ \text{subject to} &\left\{ \begin{array}{l} \mathbf{e}^T\mathbf{y} = 1 \\ \mathbf{y} \geq 0 \end{array} \right. , \end{aligned}$$

and

$$\hat{\mathbf{a}} = \frac{\mathbf{\Lambda}^{-1}\mathbf{Q}^T\hat{\mathbf{y}}}{\sqrt{\hat{\mathbf{y}}^T\mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^T\hat{\mathbf{y}}}}\,.$$

Therefore, solution of the problem does not require the computation of a matrix square root and may be found by solving the dual quadratic program (e.g., with MATLAB's `qp()` function).

D.4 Strong Duality

Recall earlier we showed weak duality, i.e., that $J(\mu, \mathbf{y})$, the cost obtained when optimizing the Lagrangian cost function with Lagrange multipliers μ and $\mathbf{y}$, is an upper bound for $\hat{z}$, the optimal cost of the primal problem. In this section we prove strong duality; we show that the least such upper bound is tight, i.e., that $J(\hat{\mu}, \hat{\mathbf{y}}) = \hat{z}$.

We begin by considering again equation (D.1) but now evaluated at the optimal values of its argument:

$$L(\hat{z}, \hat{\mathbf{x}}, \hat{\mu}, \hat{\mathbf{y}}) = \hat{z} + \hat{\mathbf{y}}^T(\mathbf{P}\hat{\mathbf{x}} - \hat{z}\mathbf{e}) + \hat{\mu}(1 - \hat{\mathbf{x}}^T\hat{\mathbf{x}}) . \tag{D.5}$$

By the definition of the function J we also have that

$$J(\hat{\mu}, \hat{\mathbf{y}}) = L(\hat{z}, \hat{\mathbf{x}}, \hat{\mu}, \hat{\mathbf{y}}) .$$

To show strong duality, therefore, we need only show that $J(\hat{\mu}, \hat{\mathbf{y}}) = L(\hat{z}, \hat{\mathbf{x}}, \hat{\mu}, \hat{\mathbf{y}}) = \hat{z}$. Therefore, we will argue that the last two terms of equation (D.5) are zero.

It is clear from the form of $\hat{\mathbf{x}}$ (see equations (D.2) and (D.3)) that $1 - \hat{\mathbf{x}}^T\hat{\mathbf{x}} = 0$. So we turn our attention to the term $\hat{\mathbf{y}}^T(\mathbf{P}\hat{\mathbf{x}} - \hat{z}\mathbf{e})$. Consider one term in the sum:

$$\hat{y}_i(\mathbf{p}_i^T\hat{\mathbf{x}} - \hat{z}) ,$$

where $\mathbf{p}_i^T$ is the i^{th} row of $\mathbf{P}$. We shall show that each such term must be zero and so the entire sum of such terms ($\hat{\mathbf{y}}^T(\mathbf{P}\hat{\mathbf{x}} - \hat{z}\mathbf{e})$) must be zero.

There are two cases, either $\mathbf{p}_i^T\hat{\mathbf{x}} - \hat{z} = 0$ or $\mathbf{p}_i^T\hat{\mathbf{x}} - \hat{z} > 0$. In the former case we are done. If the latter is true then $\hat{y}_i = 0$. This is so because to obtain $J(\hat{\mu}, \hat{\mathbf{y}})$ we have minimized over all $\hat{\mathbf{y}}$ and $\hat{\mu}$. Therefore if it were the case that $\mathbf{p}_i^T\hat{\mathbf{x}} - \hat{z} > 0$ and $\hat{y}_i > 0$ then we would not be at a minimum (recall that we argued previously that $\hat{y}_i \geq 0$). We have just argued that $\hat{\mathbf{y}}^T(\mathbf{P}\hat{\mathbf{x}} - \hat{z}\mathbf{e}) = 0$ and, hence, $J(\hat{\mu}, \hat{\mathbf{y}}) = L(\hat{z}, \hat{\mathbf{x}}, \hat{\mu}, \hat{\mathbf{y}}) = \hat{z}$. The optimal dual cost is the same as the optimal primal cost: the least upper bound is tight. (Incidentally, the condition that $\hat{y}_i(\mathbf{p}_i^T\hat{\mathbf{x}} - \hat{z}) = 0\,,\ \forall i$ is known as *complementary slackness.*)

Bibliography

[1] S. Azevedo, H. E. Martz, and D. J. Schneberk. Potential of computed tomography for inspection of aircraft components. In *Nondestructive Inspection of Aging Aircraft*, volume 2001, pages 47–57. SPIE, 1993.

[2] M. Bhatia. *Wavelet Transform-Based Multi-Resolution Techniques For Tomographic Reconstruction and Detection.* PhD thesis, Massachusetts Institute of Technology, August 1994.

[3] M. Bhatia, W. C. Karl, and A. S. Willsky. Tomographic reconstruction and estimation based on multiscale natural-pixel bases. In press.

[4] M. Bhatia, W. C. Karl, and A. S. Willsky. A wavelet-based method for multiscale tomographic reconstruction. *IEEE Transactions on Medical Imaging*, 15(1):92–101, 1996.

[5] M. Bhatia, W.C. Karl, and A.S. Willsky. A multiscale method for tomographic reconstruction. In *Proceedings of the SPIE—The International Society for Optical Engineering*, volume 2034, pages 58–69, 1993.

[6] R.N. Bracewell. Strip integration in radio astronomy. *Aust. J. Phys.*, 9:198–217, 1956.

[7] Y. Bresler, J.A. Fessler, and A. Macovski. A Bayesian approach to reconstruction from incomplete projections of a multiple object 3d domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(8):840–858, August 1989.

[8] S. R. Deans. *The Radon Transform and Some of Its Applications.* John Wiley and Sons, Inc., 1983.

[9] A.H. Delaney and Y. Bresler. Multiresolution tomographic reconstruction using wavelets. *IEEE Transactions on Image Processing*, 4, June 1995.

[10] A.J. Devaney and G.A. Tsihrintzis. Maximum likelihood estimation of object location in diffraction tomography. *IEEE Transactions on Signal Processing*, 39(3):672–682, March 1991.

[11] D. A. Froom, J. P. Barton, and J. W. Bader. Neutron and radiography at sacramento ALC. In *Nondestructive Inspection of Aging Aircraft*, volume 2001, pages 38–46. SPIE, 1993.

[12] C.R. Giardina and R.D. Edward. *Morphological Methods in Image and Signal Processing.* Prentice Hall, 1988.

[13] S. Helgason. *The Radon Transform.* Birkhauser, 1980.

[14] G. T. Herman. *Image Reconstruction From Projections.* Academic Press, 1980.

[15] A. K. Jain. *Fundamentals of Digital Image Processing.* Prentice Hall, 1989.

[16] A. C. Kak and M. Slaney. *Principles of Computerized Tomographic Imaging.* IEEE Press, 1988.

[17] A. Macovski. *Medical Imaging Systems.* Prentice-Hall, 1983.

[18] P. Milanfar. *Geometric Estimation and Reconstruction From Tomographic Data.* PhD thesis, Massachusetts Institute of Technology, June 1993.

[19] E. L. Miller. *The Application of Multiscale and Stochastic Techniques to the Solution of Inverse Problems.* PhD thesis, Massachusetts Institute of Technology, August 1994.

[20] E. L. Miller and A. S. Willsky. Multiscale, statistical anomaly detection analysis and algorithms for linearized inverse scattering problems, July 1995. Submitted to *Multidimensional Signals and Systems.*

[21] E.L. Miller and A.S. Willsky. A multiscale, decision-theoretic algorithm for anomaly detection in images based upon scattered radiation. In *First International Converence on Image Processing*, Austin, Texas, November 1994.

[22] D.C. Munson, J.D. O'Brien, and W.K. Jenkins. A tomographic formulation of spotlight-mode synthetic aperture radar. *Proceedings of the IEEE*, 71:917–925, August 1983.

[23] T. Olson and J. DeStefano. Wavelet localization of the Radon transform. *IEEE Transactions on Signal Processing*, 42:2055–2067, August 1994.

[24] T. Olson, D. Healy, J. Weaver, and J. DeStefano. Fast updating in MRI via multiscale localization. In *Proceedings of the SPIE—The International Society for Optical Engineering*, volume 2034, pages 70–83, 1993.

[25] P. Oskoui and H. Stark. A comparative study of three reconstruction methods for a limited-view computer tomography problem. *IEEE Transactions on Medical Imaging*, 8:43–49, March 1989.

[26] F. Peyrin, M. Zaim, and R. Goutte. Multiscale reconstruction of tomographic images. In *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, pages 219–222, 1992.

[27] F. Peyrin, M. Zaim, and R. Goutte. Construction of wavelet decompositions for tomographic images. *Journal of Mathematical Imaging and Vision*, 3:105–122, 1993.

[28] J. L. Prince. *Geometric Model-Based Estimation from Projections.* PhD thesis, Massachusetts Institute of Technology, January 1988.

[29] J. Radon. On the determination of functions from their integral values along certain manifolds. *Berichte der Sächsischen Akadamie der Wissenschaft*, 69:262–277, April 1917. English translations can be found in *IEEE Transactions on Medical Imaging*, vol. MI-5, December, 1986 and in the appendix of *The Radon Transform and Some of Its Applications*, by S.R. Deans, John Wiley and Sons, Inc., 1983.

[30] D. J. Rossi. *Reconstruction From Projections Based on Detection and Estimation of Objects.* PhD thesis, Massachusetts Institute of Technology, October 1982.

[31] D.J. Rossi and A.S. Willsky. Reconstruction from projections based on detection and estimation of objects—parts I and II: Performance analysis and robustness analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(4):886–906, 1984.

[32] B. Sahiner and A.E. Yagle. Image reconstruction from projections under wavelet constraints. *IEEE Transactions on Signal Processing*, 41:3579–3584, December 1993.

[33] K. Sauer, J. Sachs, and C. Klifa. Bayesian stimation of 3-d objects from few radiographs. *IEEE Transactions on Nuclear Science*, 41(5):1780–1790, October 1994.

[34] A. Schatzberg, A.J. Devaney, and A.J. Witten. Estimating target location from scattered field data. *Signal Processing*, 40:227–237, 1994.

[35] K. S. Shanmugan and A. M. Breipohl. *Random Signals: Detection, Estimation and Data Analysis.* John Wiley and Sons, Inc., 1988.

[36] G. Thomas and I. Finney. *Calculus and Analytic Geometry.* Addison-Wesley Publishing Company, 1988.

[37] G.A. Tsihrintzis and A.J. Devaney. Maximum likelihood estimation of object location in diffraction tomography, part II: Strongly scattering objects. *IEEE Transactions on Signal Processing*, 39(6):1466–1470, June 1991.

[38] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I.* John Wiley and Sons, Inc., 1968.

[39] L. Vandenberghe and S. Boyd. Semidefinite programming. *Siam Reiview*, 38(1):49–95, March 1996.

[40] D. Walnut. Local inversion of the Radon transform in the plane using wavelets. In *SPIE Mathematical Imaging*, volume 2034, pages 84–90, 1993.

[41] B. J. West and A. L. Goldberger. Physiology in fractal dimensions. *American Scientist*, 75:354–365, 1987.

[42] A. S. Willsky and G. Wornell. Stochastic processes, detection, and estimation. 6.432 Course Notes. MIT, 1995.

[43] A.E. Yagle. Region-of-interest tomography using the wavelet transform and angular harmonics. *IEEE Signal Processing Letters*, 1:134–135, September 1994.