# A Multiscale Hypothesis Testing Approach to Anomaly Detection and Localization from Noisy Tomographic Data

Austin B. Frakt, *Student Member, IEEE,* W. Clem Karl, *Member, IEEE,* and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—In this paper, we investigate the problems of anomaly detection and localization from noisy tomographic data. These are characteristic of a class of problems that cannot be optimally solved because they involve hypothesis testing over hypothesis spaces with extremely large cardinality. Our multiscale hypothesis testing approach addresses the key issues associated with this class of problems. A multiscale hypothesis test is a hierarchical sequence of composite hypothesis tests that discards large portions of the hypothesis space with minimal computational burden and zooms in on the likely true hypothesis. For the anomaly detection and localization problems, hypothesis zooming corresponds to spatial zooming—anomalies are successively localized to finer and finer spatial scales. The key challenges we address include how to hierarchically divide a large hypothesis space and how to process the data at each stage of the hierarchy to decide which parts of the hypothesis space deserve more attention. To answer the former we draw on [1] and [7]–[10]. For the latter, we pose and solve a nonlinear optimization problem for a decision statistic that maximally disambiguates composite hypotheses. With no more computational complexity, our optimized statistic shows substantial improvement over conventional approaches. We provide examples that demonstrate this and quantify how much performance is sacrificed by the use of a suboptimal method as compared to that achievable if the optimal approach were computationally feasible.**

***Index Terms*—Anomaly detection, composite hypothesis testing, hypothesis zooming, nonlinear optimization, quadratic programming, tomography.**

## I. Introduction

IN THIS PAPER, we present a new approach to hierarchical, multiresolution anomaly detection and localization from noisy tomographic data. This problem is of interest in several areas of active research. For example, it arises in nondestructive evaluation as well as adaptive tomographic reconstruction from limited data in which one attempts to use degrees of freedom frugally by localizing areas of interest to be imaged at finer resolutions. In addition, the anomaly detection and localization problems raise issues that surface in many other contexts. Thus, a second objective of this paper is to provide some insight into these more general issues.

In particular, a fundamental characteristic of a large class of signal and image analysis problems is that they involve hypothesis testing over hypothesis spaces of extremely large cardinality—so large that enumeration of all hypotheses and exhaustive comparison is computationally infeasible. This characteristic is present in the specific problems considered in this paper, since the enumeration of all possible anomalies leads to a very large hypothesis space. Large hypothesis spaces also arise in other applications including detection of regions of interest for automatic target recognition from wide-area imagery as well as model-based object recognition.

Such problems thus invite efficient hierarchical approaches. Our approach achieves great efficiency by discarding large parts of the hypothesis space with minimal computational burden and zooming in on a smaller part of the hypothesis space to be scrutinized more extensively and, perhaps, exhaustively. We call this hypothesis zooming methodology a *multiscale hypothesis test* (MSHT). As depicted in Fig. 1, an MSHT is a scale-recursive sequence of composite hypothesis tests that increasingly disambiguates hypotheses at finer scales and zooms in on one element of the global set of hypotheses, $\mathcal{H} \triangleq \{H_m\}_{m=0}^{M-1}$. At the top (or coarsest) scale, the MSHT divides $\mathcal{H}$ into a family of possibly overlapping subsets, and then chooses one of these subsets, discarding all the others. Similarly, at subsequent scales, the remaining hypotheses are divided into subsets, all but one of which are discarded. Note that the statistical decision problem at each scale is a composite hypothesis testing problem in which one composite hypothesis[1], $\tilde{H}_n \subset \mathcal{H}$, is selected for finer scale investigation.

There are several questions that must be addressed in formulating and solving problems in the manner we have described. The first is: how should a large hypothesis space be divided hierarchically so that hypothesis zooming can proceed? For anomaly detection in imaging (e.g., tomographic) problems there is a natural choice—grouping hypotheses spatially, i.e., subdividing the region of interest at a succession of resolutions so that anomalies can be localized to finer and finer spatial scales. This is not an original idea, and, in particular, it has

Manuscript received September 21, 1996; revised July 14, 1997. This work was supported by the Air Force Office of Scientific Research under Grants F49620-95-1-0083 and F49620-96-1-0028, by the Defense Advanced Research Projects Agency under Grant F49620-93-1-0604, and by a National Defense Science and Engineering Graduate Fellowship awarded by the Defense Advanced Research Projects Agency. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. C.-C. Jay Kuo.

A. B. Frakt and A. S. Willsky are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: frakt@mit.edu).

W. C. Karl is with the Department of Electrical, Computer, and Systems Engineering, Boston University, Boston, MA 02215 USA.

Publisher Item Identifier S 1057-7149(98)04001-9.

[1] We use a tilde $(\tilde{\cdot})$ to indicate composite hypotheses and quantities and objects associated with them to distinguish these from individual (noncomposite) hypotheses and their associated similar objects.

1057–7149/98$10.00 © 1998 IEEE

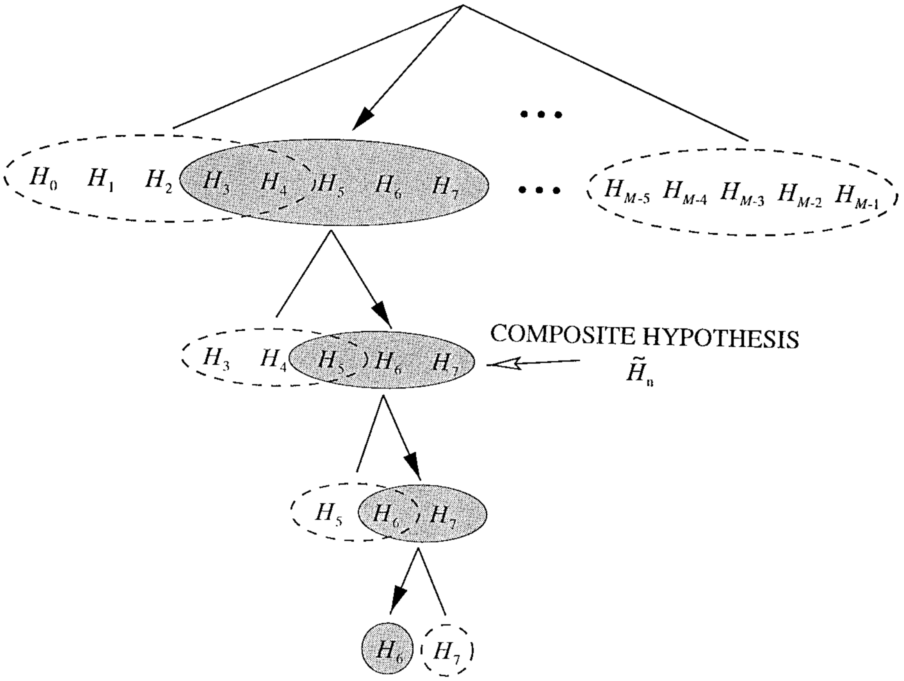

$H_0$ $H_1$ $H_2$ $H_3$ $H_4$ $H_5$ $H_6$ $H_7$
$H_{M-5}$ $H_{M-4}$ $H_{M-3}$ $H_{M-2}$ $H_{M-1}$
COMPOSITE HYPOTHESIS
$\tilde{H}_n$

Fig. 1. MSHT. Composite hypotheses indicated by dashed lines are discarded, shaded ones retained.

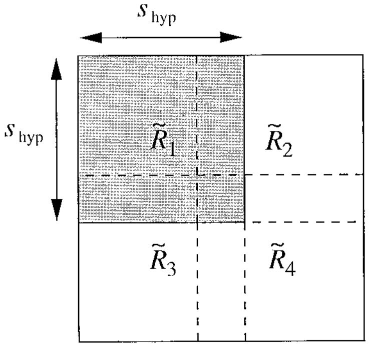


Fig. 2. Coarse-scale subdivisions. Subdivision $\tilde{R}_n$ corresponds to $\tilde{H}_n$. $\tilde{R}_1$ is shaded and with a solid border.

been used in [9] and [10] for the detection and localization of anomalies in inverse scattering data and is standard in the image segmentation community [1], [7], [8]. We also adopt this method for dividing our hypothesis space by defining composite hypothesis $\tilde{H}_n$ as containing all hypotheses, $H_m$, which correspond to anomalies with support within a region of the image domain which we denote $\tilde{R}_n$. Fig. 2 illustrates this at the coarsest scale at which the regions $\tilde{R}_n$ are $s_{\text{hyp}} \times s_{\text{hyp}}$ in size. Composite hypotheses at finer scales have smaller cardinality and, in the anomaly detection and localization problems, naturally correspond to regions of smaller area so that hypothesis zooming and spatial zooming coincide.

The second important question is: Once composite hypotheses are defined, how should the data be processed at each stage in the hierarchy in order to determine which composite hypotheses to discard? We need to address this question in light of the fact that computational complexity is a key concern. In particular, if computational load were not an issue, then we could solve the problem optimally using a generalized likelihood ratio test (GLRT) [14] as follows. At a coarse level in the hierarchy, we would enumerate all hypotheses, $H_m$, process the data to choose among these optimally, and then choose the composite hypothesis, $\tilde{H}_n$, that contains this optimal choice. Since an underlying assumption for problems of interest here is that complete enumeration and comparison is unacceptably complex, we limit our consideration to the class of decision rules that have acceptable computational complexity. By no means do we provide a complete solution to the problem of finding the best decision rule under a complexity constraint, since ranking the relative complexities of different decision rules is difficult and the characterization of all decision rules with complexity less than some specified level is both ill posed and prohibitively complex. What we do provide, however, is a solution within a specific class of decision rules that includes what can be thought of as the "natural" choices. In the process, we highlight issues that are of importance in any composite hypothesis testing problem.

For anomaly detection and localization in imaging applications, there is an obvious first choice for the decision statistic and decision rule to be used at a particular stage in the hypothesis zooming procedure. Specifically, for each composite hypothesis, $\tilde{H}_n$, at that stage, compute a likelihood statistic for the hypothesis that an anomaly has support over the *entire* region $\tilde{R}_n$. The decision rule is simple: keep the composite hypothesis $\tilde{H}_n$ corresponding to the largest likelihood statistic.[2]

[2] As an aside, we note that for problems in which the data are in a different domain from the image (e.g., tomographic problems), the calculation of the likelihood statistic is most easily done directly in the data domain rather then by first forming an image and then calculating the statistic [13].

Approaches of this type have been proposed and used in the literature [9]–[12]. A principle objective of this paper is to demonstrate that we can achieve much better performance with the same level of complexity by choosing the decision statistics in a significantly different manner. To understand this, note that the statistic mentioned in the preceding paragraph corresponds to a hypothesis that, in some sense, "covers" all of the hypotheses that comprise a composite hypothesis $\tilde{H}_n$. Thus, this statistic sacrifices sensitivity to each individual hypothesis in $\tilde{H}_n$ in order to achieve *some* sensitivity to *all* of the hypothesis in $\tilde{H}_n$. In contrast, the statistics used in the computationally intractable but optimal GLRT have, as we shall see, *significantly* greater sensitivity to each of hypotheses in a composite hypothesis. Our aim is to find statistics that approach the sensitivity of those used in the GLRT but with the same computational simplicity as that required for the "natural" decision rule described previously.

In this paper, we limit ourselves to a class of decision rules based on the computation of a single affine function of the measured data for each composite hypothesis. However, we design these affine statistics to maximize their utility for the job for which they will be used. In particular, since each such statistic will be used to decide whether or not to discard its associated composite hypothesis, we aim to design a statistic that maintains strong sensitivity to the hypotheses that comprise the corresponding composite hypothesis and has minimal sensitivity to all other hypotheses. Using this philosophy, we are led to an interesting optimization problem, the solution of which is a decision statistic that is quite different from and superior to the natural choices. In particular, these optimized statistics do a far better job of increasing separation between composite hypotheses.

In this paper, we develop the MSHT methodology and optimized statistic design for the tomographic anomaly detection and localization problems and also present results to quantify how well our methods work. Of particular interest to us are quantifying i) how much performance is sacrificed by using a suboptimal method as compared to that achievable if exhaustive hypothesis enumeration and comparison were computationally feasible; and ii) how much better our approach is compared to one based on the natural choice of decision statistics. In addition, as we have stated, one of our objectives in this paper is to lay out what we believe to be the key issues in large hypothesis testing problems more generally, and in the conclusion to this paper we provide a discussion of some of the issues and questions which we believe our work brings into sharper focus.

This paper is organized as follows. In Section II we outline all modeling assumptions and set up the problems considered in this paper. We discuss two types of decision statistics in Section III—a conventional likelihood statistic and our optimized statistic. Section IV includes pseudo-code for our MSHT algorithm and considers the computational complexity of this and the optimal approach. Examples are provided in Section V. Closing remarks and a discussion of some complements and extensions are found in Section VI.

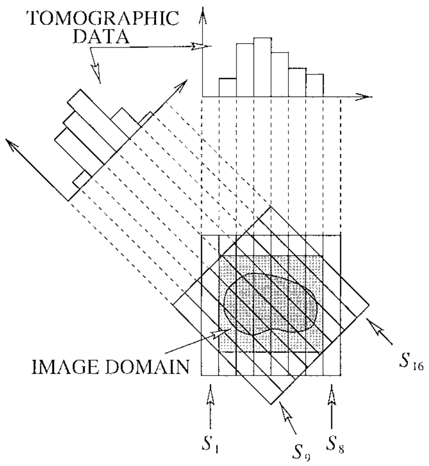


Fig. 3. Tomographic projections with strip indicator functions $S_i(x, y)$.

## II. Problem Formulation

### A. Observation Equation

We model tomographic data collection with the equation

$$g_i = \int_{\mathbb{R}^2} f(x,y)S_i(x,y)\,dx\,dy + \nu_i \tag{1}$$

where $g_i$ is the $i$th datum, $f(x,y)$ is a real function of two spatial variables representing the object, $\nu_i$ is the $i$th sample measurement noise. The function $S_i(x,y)$ is one over the $i$th strip and zero elsewhere. Data acquisition with these functions is illustrated in Fig. 3. We let $N_\phi$ be the number of projections (angular positions) and we assume these positions are equally spaced in the interval $[0, \pi)$. The number of equally spaced samples per projection angle is $N_{\mathrm{s}}$. Finally, for computational purposes, the object, $f(x,y)$, is discretized in a rectangular pixel basis so that

$$f(x,y) = \sum_{j=1}^{N_{\mathrm{p}}} f_j p_j(x,y) \tag{2}$$

where $p_j(x,y)$ is one over the $j$th pixel and zero elsewhere and there are $N_{\mathrm{p}}$ pixels corresponding to a $N \times N$ field.

Combining (1) and (2) we find that

$$\mathbf{g} = \mathbf{T}\mathbf{f} + \boldsymbol{\nu} \tag{3}$$

where $\mathbf{g}$, $\mathbf{f}$, and $\boldsymbol{\nu}$ are vectors containing the data values, field pixel values, and noise values, respectively. The field vector, $\mathbf{f}$, is lexicographically ordered and $\mathbf{g}$ and $\boldsymbol{\nu}$ are formed by stacking the values obtained at each projection. For example, $\mathbf{g} = [\mathbf{g}_1^T \;\; \mathbf{g}_2^T \;\; \cdots \;\; \mathbf{g}_{N_\phi}^T]^T$ where $\mathbf{g}_k$ contains the data values at the $k$th projection angle. The components of the matrix $\mathbf{T}$ are given by

$$[\mathbf{T}]_{ij} = \int_{\mathbb{R}^2} S_i(x,y)p_j(x,y)\,dx\,dy \tag{4}$$

for $i = 1, \cdots, N_\phi N_{\mathrm{s}}$ and $j = 1, \cdots, N_{\mathrm{p}}$. Equation (3), coupled with whatever *a priori* knowledge we have about $\boldsymbol{\nu}$ and $\mathbf{f}$, represents our observational model. The tomographic projection matrix, $\mathbf{T}$, captures a discrete representation of the strip integrals. The application of $\mathbf{T}$ to $\mathbf{f}$ is called the projection of $\mathbf{f}$.

### B. Models

Our notion of an anomaly is a localized region of the image domain which differs statistically from our prior set of expectations of the image. Therefore, we statistically characterize an anomaly-free background and define a parameterized class of anomalies. We model the field, $\mathbf{f}$, as a superposition of a field, $\mathbf{f}_{\rm a}$, which contains, at most, a single anomaly from this class and an anomaly-free background field, $\mathbf{f}_{\rm b}$. That is,

$$\mathbf{f} = \mathbf{f}_{\rm a} + \mathbf{f}_{\rm b}. \qquad (5)$$

The anomaly field and background field are statistically independent.

The class of anomalies we consider is parameterized by an intensity, $c$, a size, $s$, and a position, $(i, j)$ as follows. The anomaly field, $\mathbf{f}_{\rm a}$, is zero everywhere except over a square patch where it is constant. Our notation is

$$\mathbf{f}_{\rm a} = c\mathbf{b}(s, i, j) \qquad (6)$$

where $c \geq 0$ and $\mathbf{b}(s, i, j)$ is the lexicographically ordered indicator vector associated with an $N \times N$ field which is zero everywhere except over the $s \times s$ support area with upper left corner at pixel $(i, j)$ where $\mathbf{b}(s, i, j)$ takes the value one. The size, $s$, and location, $(i, j)$, are unknown. We assume knowledge, however, of the minimum possible size, $s_{\rm min}$, and the maximum possible size, $s_{\rm max}$, the anomaly can be where $1 \leq s_{\rm min} \leq s \leq s_{\rm max} \ll N$.

While the class of anomalies we consider is restrictive, the methodology we present in this paper can be applied to other enumerable anomaly classes. We focus on anomalies that are constant intensity squares in some known size range for clarity of presentation, since our main focus is on presenting a new methodology for large hypothesis testing problems and not to accurately model any particular type of anomaly.

The background field, $\mathbf{f}_{\rm b}$, is a zero-mean, wide sense stationary (WSS), Gaussian random field with known covariance $\mathbf{\Lambda}$. We consider two types of background covariance statistics in our examples in Section V—white and fractal. The fractal background has a power spectral density of the form $1/\omega^2$ where $\omega$ is frequency. Additional structural details are found in [3] and [5]. We consider a fractal background because fractal fields accurately model a wide range of natural textures [15].

The additive measurement noise, $\boldsymbol{\nu}$, is assumed to be a zero-mean, white, WSS, Gaussian random vector with intensity $\lambda$ and is statistically independent of the background and anomaly fields. Therefore, the data are jointly Gaussian, as follows:[3]

$$\mathbf{g} \sim \mathcal{N}(c\mathbf{T}\mathbf{b}(s, i, j), \mathbf{\Lambda}_{\mathbf{g}}) \qquad (7)$$

where

$$\mathbf{\Lambda}_{\mathbf{g}} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}^T + \lambda\mathbf{I}. \qquad (8)$$

Notice that, even when the background is white ($\mathbf{\Lambda}$ is diagonal), the data are correlated due to the structure of the tomographic projection matrix, $\mathbf{T}$.

Fig. 4 illustrates an example of the kind of anomaly and background field that are considered in this paper. Projections of $\mathbf{f}$ are also shown with and without the addition of noise. Fig. 4(a) and (b) are views of the image domain while Fig. 4(c) and (d) are views of the data domain.

### C. Problem Statement

The anomaly detection problem is to determine whether or not $\mathbf{f}_{\rm a}$ is identically zero. The anomaly localization problem is to determine the values of the size $s_{\rm min} \leq s \leq s_{\rm max}$ and location $(i, j)$ of the anomaly if indeed one is present. The optimal solution to these problems includes one hypothesis, $H_m$, for each possible anomaly size, $s_m$, and each possible location of the anomaly's upper left hand corner, $(i_m, j_m)$, where

$$s_m \in \{s_{\rm min}, s_{\rm min}+1, \cdots, s_{\rm max}-1, s_{\rm max}\}, \qquad (9)$$

$$(i_m, j_m) \in \{1, 2, \cdots, N - s_m + 1\}^2. \qquad (10)$$

The optimal decision statistics are likelihood ratios and the optimal decision rule is a likelihood ratio test. While the optimal test is straightforward, it is computationally infeasible for all but trivial-sized problems. Therefore, we propose the MSHT as an efficient and effective alternative.

### D. Composite Hypothesis Structure

An MSHT has two main high-level characteristics: the form of the composite hypotheses and the form of the statistics used to decide which composite hypotheses to discard. In this section, we define the form of the composite hypotheses. We defer discussion of decision statistics to Section III. For clarity of presentation and notational simplicity, we specify in detail only the coarsest scale composite hypothesis test of the MSHT. The processing at other scales follows by analogy.

Fig. 2 provides an interpretation of the composite hypotheses. Composite hypothesis $\tilde{H}_n$ contains hypotheses that correspond to anomalies with support entirely within the $s_{\rm hyp} \times s_{\rm hyp}$ region denoted as $\tilde{R}_n$. For example, all hypotheses associated with anomalies with support *entirely* within the shaded region labeled $\tilde{R}_1$ belong to $\tilde{H}_1$. We associate each composite hypothesis with an indicator function $\tilde{\mathbf{b}}_n \triangleq \mathbf{b}(s_{\rm hyp}, i_n, j_n)$, which is one over the region $\tilde{R}_n$, and zero elsewhere. The composite hypothesis regions overlap by at least $s_{\rm max} - 1$ pixels so that each possible anomaly lies entirely within at least one region. This ensures that $\mathcal{H} = \tilde{H}_1 \cup \tilde{H}_2 \cup \tilde{H}_3 \cup \tilde{H}_4$.

## III. Decision Statistics

In this section, we specify the form of the decision statistics for the coarsest scale composite hypothesis test of a MSHT. The statistics at subsequent scales are easily understood by analogy. We will discuss two types of decision statistic. The first type, discussed in Section III-A, is a coarse-scale likelihood statistic of the form used in [9], [10] for problems similar to the ones addressed here. As discussed in Section I, while

[3] 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 $\mathbf{P}$.

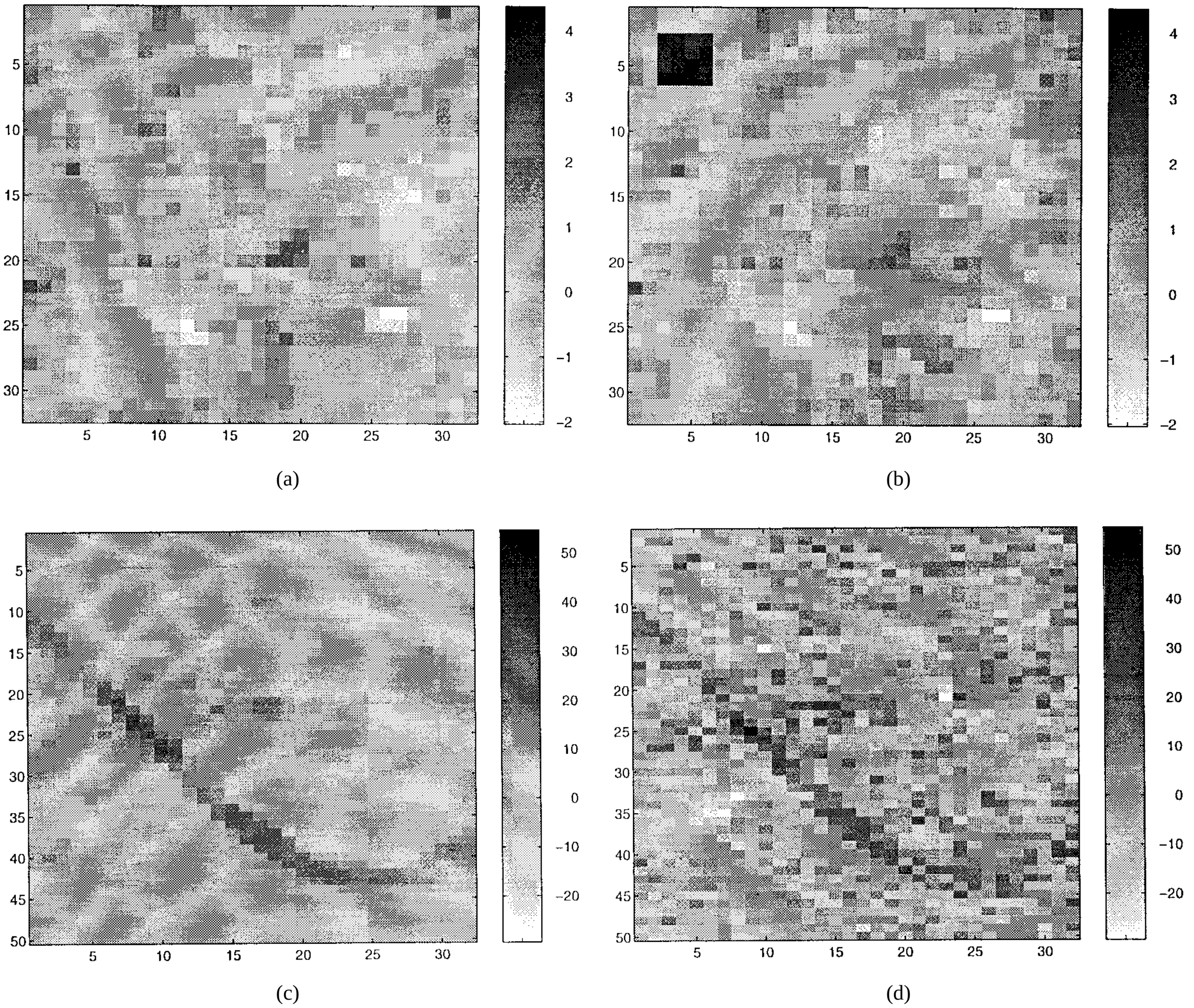


Fig. 4. (a) Fractal background, $\mathbf{f}_b$. (b) Superposition of the background shown in (a) and an anomaly near the upper left corner. (c) Projection of the anomaly plus background field. The horizontal axis is the projection number ($N_\phi = 32$). The vertical axis is the sample number ($N_s = 50$). (d) Measurement noise has been added to the projections.

this statistic is natural and intuitive, it sacrifices considerable sensitivity to achieve computational simplicity. We use this statistic as a benchmark against which to compare what we call an *optimized statistic* which we discuss in Section III-B.

### A. Coarse-Scale Likelihood Statistics

As discussed in the introduction, the coarse-scale likelihood statistic associated with composite hypothesis $\tilde{H}_n$ is the log-likelihood ratio to discriminate between a single hypothesis for a coarse-scale anomaly with support over the entire region $\tilde{R}_n$ and the hypothesis that no anomaly exists. To derive the coarse-scale likelihood statistic, we associate with each $\tilde{H}_n$ a coarse-scale anomaly: $\mathbf{f}_a = \tilde{\mathbf{b}}_n$ where $\tilde{\mathbf{b}}_n$, as defined in Section II-D, is the indicator function for $\tilde{R}_n$.

From (7) and standard results in hypothesis testing [14], we find that the log-likelihood for each of these four hypotheses is given by an affine operation on the observed data, $\mathbf{g}$, namely

$$\tilde{\ell}_n(\mathbf{g}) = (\mathbf{T}\tilde{\mathbf{b}}_n)^T \mathbf{\Lambda}_g^{-1}\mathbf{g} - \frac{1}{2}(\mathbf{T}\tilde{\mathbf{b}}_n)^T \mathbf{\Lambda}_g^{-1}\mathbf{T}\tilde{\mathbf{b}}_n, \qquad n = 1,2,3,4 \quad (11)$$

and the resulting decision rule consists of choosing the largest of these four values.

Note that at this coarsest scale in the MSHT, we have that

$$(\mathbf{T}\tilde{\mathbf{b}}_i)^T \mathbf{\Lambda}_g^{-1}\mathbf{T}\tilde{\mathbf{b}}_i = (\mathbf{T}\tilde{\mathbf{b}}_j)^T \mathbf{\Lambda}_g^{-1}\mathbf{T}\tilde{\mathbf{b}}_j, \quad i,j = 1,2,3,4. \quad (12)$$

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 $\mathbf{g}$. Consequently, at this level we can drop the second term in (11), as it has no influence on the decision rule. Note, however, that relations such as (12) need not hold at subsequent scales, since the composite hypothesis regions do not have the requisite symmetry. Despite this, we have found by simulation that the two sides of (12) are approximately equal. They differ by only 5–10% depending on the exact parameters of the problem (i.e., whether the background is fractal or white, the size of the regions, etc.). While nothing in the sequel relies upon the approximate equality of the two sides of (12) at finer scales, we mention it to illustrate that one can, to some degree of approximation, use a linear rather than an affine statistic at all scales.

### B. Optimized Statistics

In this section, we design a statistic that is no more computationally complex than the coarse-scale likelihood ratio but is much more sensitive and discriminating. The statistic we design is affine in the observed data and has the form

$$\tilde{\ell}_i(\mathbf{g}) = \mathbf{a}_i^T\mathbf{g} + d_i, \quad i = 1, 2, 3, 4. \tag{13}$$

Roughly, our objective in designing such a statistic is to choose the vector $\mathbf{a}_i$ and the constant $d_i$ to force $\tilde{\ell}_i$ to be significantly larger, on average, when $\tilde{H}_i$ is true than when it is false. That is, we would like $\tilde{\ell}_i$ to have a large separation between its mean when any $H_j \in \tilde{H}_i$ is true and its mean when any $H_k \notin \tilde{H}_i$ is true. Since doubling the magnitude of $\mathbf{a}_i$ will double this difference in mean values, we normalize this difference by the standard deviation of $\tilde{\ell}_i$.

More precisely, we define

$$m_{ij} \triangleq E[\tilde{\ell}_i \mid H_j] = c\mathbf{a}_i^T\mathbf{T}\mathbf{b}_j + d_i \tag{14}$$

$$\sigma_i^2 \triangleq \mathrm{var}[\tilde{\ell}_i \mid H_j] = \mathbf{a}_i^T\boldsymbol{\Lambda}_{\mathbf{g}}\mathbf{a}_i \tag{15}$$

where we have introduced the shorthand notation $\mathbf{b}_j \triangleq \mathbf{b}(s_j, i_j, k_j)$ for the anomaly indicator function associated with hypothesis $H_j$. Notice that the conditional mean is affine in $\mathbf{a}_i$ while the conditional variance is quadratic in $\mathbf{a}_i$. Also note that the conditional variance is independent of $j$.

The criterion we adopt is to maximize the worst-case (i.e., smallest) normalized difference between $m_{ij}$ for $H_j \in \tilde{H}_i$ and $m_{ik}$ for $H_k \notin \tilde{H}_i$ where $\sigma_i$ is used for normalization. That is, we choose $\hat{\mathbf{a}}_i$ and $\hat{d}_i$ as the solution to the optimization problem

$$[\hat{\mathbf{a}}_i^T \quad \hat{d}_i]^T = \arg\max_{\mathbf{a},d} \min_{(j,k)\in\mathcal{A}_i} \frac{m_{ij} - m_{ik}}{\sigma_i} \tag{16}$$

where $\mathcal{A}_i \triangleq \{(j,k) \mid H_j \in \tilde{H}_i \text{ and } H_k \notin \tilde{H}_i\}$. Substituting in the definitions of (14) and (15), we find that the constant, $d_i$, cancels and the optimization problem reduces to

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}} \min_{(j,k)\in\mathcal{A}_i} \frac{\mathbf{a}^T\mathbf{T}\mathbf{b}_j - \mathbf{a}^T\mathbf{T}\mathbf{b}_k}{\sqrt{\mathbf{a}^T\boldsymbol{\Lambda}_{\mathbf{g}}\mathbf{a}}}. \tag{17}$$

Since we are free to choose $\hat{d}_i$ to be any value whatsoever we shall set $\hat{d}_i = 0$ in the sequel. Also notice that the anomaly intensity $c$ is independent of $i$, $j$, $k$, $d$, and $\mathbf{a}$ and so can be dropped from the optimization problem. As is shown in Appendix A, using Lagrange duality theory, this problem can be reformulated as the quadratic program

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}} \mathbf{y}^T\mathbf{Q}\boldsymbol{\Lambda}_{\mathbf{g}}^{-1}\mathbf{Q}^T\mathbf{y} \tag{18}$$

$$\text{subject to} \begin{cases} \mathbf{e}^T\mathbf{y} = 1 \\ \mathbf{y} \geq 0 \end{cases} \tag{19}$$

where

$$\mathbf{Q} \triangleq \begin{bmatrix} (\mathbf{T}\mathbf{b}_{j_1} - \mathbf{T}\mathbf{b}_{k_1})^T \\ (\mathbf{T}\mathbf{b}_{j_2} - \mathbf{T}\mathbf{b}_{k_2})^T \\ \vdots \end{bmatrix} \quad \text{and} \quad \mathbf{e} \triangleq [1 \quad 1 \quad 1 \quad \cdots \quad 1]^T \tag{20}$$

and all pairs $(j_n, k_n) \in \mathcal{A}_i$. This quadratic program may be solved using, for example, methods described in [6]. Since the solution depends only on the data covariance matrix and the structure of the hypothesis space, but not on the data itself, solving the quadratic program is an off-line procedure. Once $\hat{\mathbf{y}}$ is found, $\hat{\mathbf{a}}_i$ is given by

$$\hat{\mathbf{a}}_i = \frac{\boldsymbol{\Lambda}_{\mathbf{g}}^{-1}\mathbf{Q}^T\hat{\mathbf{y}}}{\sqrt{\hat{\mathbf{y}}^T\mathbf{Q}\boldsymbol{\Lambda}_{\mathbf{g}}^{-1}\mathbf{Q}^T\hat{\mathbf{y}}}}. \tag{21}$$

Note that in optimizing over all affine functions of the tomographic data, we implicitly consider the coarse-scale likelihood statistic as well as all affine functions of any affine reconstruction of the image. Therefore, preprocessing the data with any affine reconstruction routine (like convolution backprojection) cannot result in a better statistic.

## IV. MSHT Algorithm and Computational Complexity

### A. The MSHT Algorithm

Given the hierarchical hypothesis space decomposition described in Section II-D and either choice of decision statistic discussed in Section III, the MSHT algorithm detects and localizes an anomaly in the coarse-to-fine manner described in Section I. At each scale, other than the finest, one of four subdivisions of the remaining part of the image domain is selected for finer-scale investigation. The scale-recursion terminates at a scale at which the optimal test *is* feasible and at which the regions $\tilde{R}_n$ are no smaller than $s_{\max} \times s_{\max}$. Then, having localized the anomaly to an area significantly smaller than the entire image domain, the optimal test is performed, which includes only hypotheses associated with anomalies in that area. Finally, the statistic value associated with the selected hypothesis is compared with a threshold. If it is larger than the threshold, the associated hypothesis is selected, otherwise it is declared that no anomaly exists.

The following pseudocode summarizes the 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 a hypothesis in $\mathcal{H}$ or the decision that no anomaly exists.

*Pseudocode.* $\hat{H} = \mathit{MSHT}(\tilde{\ell}, \tilde{R}, \mathit{scale})$:

```
1) If scale is finest possible, perform optimal test on anomalies in $\tilde{R}$. Call the selected hypothesis $H$ and the associated statistic value $\ell$. Then
       $\hat{H} = \begin{cases} H, & \text{if } \ell > \eta \\ \text{“no anomaly,”} & \text{otherwise.} \end{cases}$   (22)
2) Otherwise, subdivide the region $\tilde{R}$ into four overlapping squares where the amount of overlap is at least $s_{\max} - 1$. Denote these squares $\tilde{R}_n$ for $n = 1, 2, 3, 4$.
3) For each subdivision $\tilde{R}_n$ compute the statistic $\tilde{\ell}_n$.
4) Let $k$ be such that $\tilde{\ell}_k > \tilde{\ell}_n$, $\forall k \neq n$. Call $\hat{H} = \mathit{MSHT}(\tilde{\ell}_k, \tilde{R}_k, \mathit{scale} + 1)$.
```

### B. Computational Complexity

Our primary motivation for applying an MSHT to the anomaly detection and localization problems is that the optimal hypothesis test is too computationally costly. The MSHT formulates fewer hypotheses than the optimal test and is therefore more efficient. In this section we quantify this claim by calculating the computational complexity of the optimal algorithm and the MSHT algorithm. To do so, we will compute the number of operations per hypothesis and the number of hypotheses formulated in each algorithm.

Both the optimal test and the MSHT formulate affine statistics that require $O(N_\phi N_\mathrm{s})$ operations (adds and multiplies). This result follows from the fact that an affine statistic is an inner product between two length $N_\phi N_\mathrm{s}$ vectors plus the addition of an offset term. Since each hypothesis requires the same constant amount of work for *any* affine statistic, the overall complexity of either algorithm scales with the number of hypotheses formulated regardless of what affine statistic is used. Hence, we take the number of hypotheses formulated as a measure of algorithmic complexity.

First consider the optimal test. Suppose the linear size of the square field is $N$ and that we know the minimum and maximum possible size of the anomaly: $s_{\min}$ and $s_{\max}$, respectively. The number of hypotheses formulated is $M_{\mathrm{opt}} = \sum_{r=s_{\min}}^{s_{\max}} (N - r + 1)^2 = O(N^2)$.

Now consider the MSHT algorithm. Its computational complexity is a function of, among other parameters, the scale at which the hypothesis zoom terminates and exhaustive enumeration of all remaining hypotheses is conducted (Step 1 of the above pseudocode) and the amount by which the regions $\tilde{R}_n$ overlap. We may neglect the effect of overlapping in our order-of-magnitude calculation, since the overlapping is on the order of $s_{\max}$ and $s_{\max} \ll N$. Therefore, the number of hypotheses formulated is $M_{\mathrm{MSHT}} = O(\log N) + C$ where the constant, $C$, accounts for the number of hypotheses formulated in the finest-scale exhaustive enumeration step. If the finest-scale exhaustive enumeration step is conducted at a scale at which the $\tilde{R}_n$ are on the order of $s_{\max} \times s_{\max}$ then $C$ is negligible since $s_{\max} \ll N$.

Fig. 5 displays the number of hypotheses for the optimal and MSHT algorithms as a function of $N$ for $s_{\min} = 1$ and $s_{\max} = 4$. For the case illustrated, it is assumed that the exhaustive enumeration step of the MSHT is conducted at a scale at which the $\tilde{R}_n$ are $s_{\max} \times s_{\max}$. As can be seen in Fig. 5, the difference in the number of hypotheses considered is quite large. For example, for a 512 × 512 image (i.e., $N = 512$), the MSHT considers about 30 hypotheses[4] while the optimal algorithm considers just over $10^6$ hypotheses.

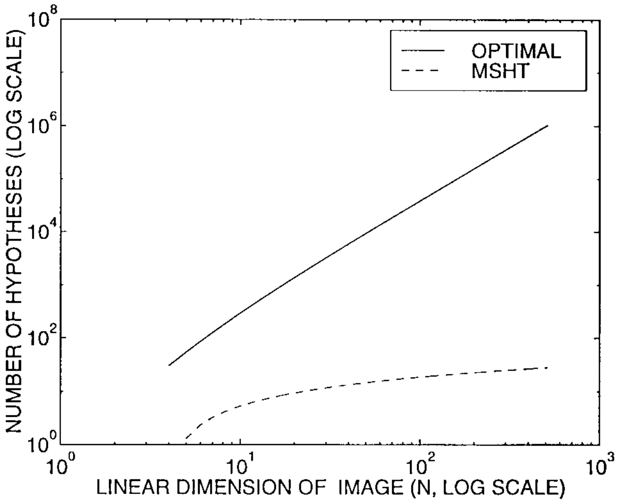


Fig. 5. Complexity of the optimal algorithm (top curve), and an MSHT (bottom curve). The vertical-axis is the number of hypotheses (log scale), the horizontal-axis is the linear dimension of the field (log scale). Here $s_{\max} = 4$ and $s_{\min} = 1$.

TABLE I
DETAILS RELATING TO THE EXAMPLES PRESENTED IN SECTION V

| *Section* | *Field Size* | *Background Field* | $N_\phi$ | $N_\mathrm{s}$ | $s_{\min}$ | $s_{\max}$ | $s_{\mathrm{hyp}}$ *scale 1* | $s_{\mathrm{hyp}}$ *scale 2* |
|---|---|---|---|---|---|---|---|---|
| V-A | 16 × 16 | fractal & white | 16 | 24 | 4 | 4 | 10 | N/A |
| V-B | 16 × 16 | fractal | 16 | 24 | 3 | 4 | 10 | N/A |
| V-C | 32 × 32 | fractal | 32 | 50 | 4 | 4 | 18 | 11 |

## V. EXAMPLES

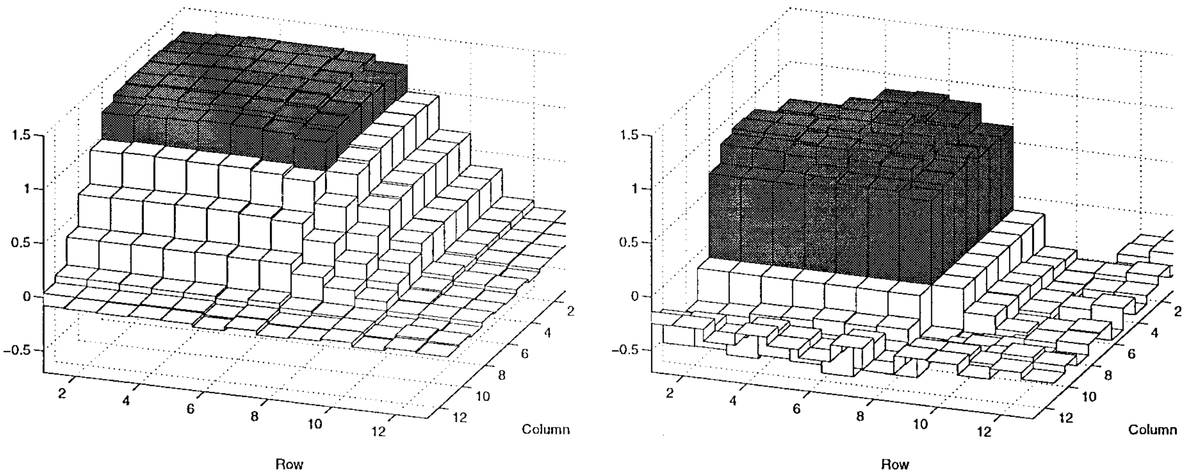
In this section, we present several types of examples. First, in Section V-A, we introduce a means of directly comparing the sensitivity of a coarse-scale likelihood statistic with that of an optimized statistic. In Section V-B we investigate the performance of the first, coarsest-stage of an MSHT algorithm in several cases of differing data quality for problems of sufficiently small size so that comparison to the optimal test is feasible. We focus on the coarsest scale because it is the scale at which the composite hypotheses are largest in cardinality and, therefore, most difficult to disambiguate [10]. That is, it is at this stage that the MSHT should have its greatest difficulty and, thus, is most in need of maximally sensitive statistics.

In Section V-C, we conclude with examples illustrating the performance of a "full" MSHT algorithm (one that continues past the coarsest scale and successively localizes the anomaly to the finest possible scale). This full algorithm includes three scales, the two coarsest scales are composite hypothesis tests. By the third scale, the anomaly has been sufficiently localized so that full enumeration and comparison of the remaining hypotheses is feasible. Hence, an optimal test on the remaining hypotheses is performed at the third scale of the full algorithm.

The data in all our examples are simulated based on the models presented in Section II-B and we include examples corresponding to two different field sizes. The examples in Sections V-A and V-B correspond to a 16 × 16 field while the examples in Section V-C correspond to a 32 × 32 field. The parameters of the problem addressed in each section are detailed in Table I. Throughout this section, a "false alarm" occurs when no anomaly exists but a "no anomaly" decision is not returned by the algorithm. A "detection" occurs when an anomaly exists and the true composite hypothesis (Section V-B) or hypothesis (Section V-C) is selected.

Before proceeding to examples, we present the definitions of *signal-to-noise ratio* (SNR) and *anomaly-to-background ratio*

[4] Under the assumptions stated in this section (i.e., ignoring overlap, etc.), the number of hypotheses considered in an MSHT is the number of hypotheses considered per scale (four) times the number of scales $(\log_2(N/s_{\max}))$.



(a) (b)

Fig. 6. $\mu_{1j}$ with a white background at SNR $= 1$. (a) For the coarse-scale likelihood statistic. (b) For the optimized statistic. Pixel $(m, n)$ corresponds to $\mathbf{f}_{\text{a}} = \mathbf{b}(4, m, n)$.

(ABR), as follows:

$$\text{SNR} \triangleq \frac{\text{trace}(\mathbf{T}\mathbf{\Lambda}\mathbf{T}^T)}{\text{trace}(\lambda\mathbf{I})} \tag{23}$$

$$\text{ABR} \triangleq \frac{\mathbf{f}_{\text{a}}^T\mathbf{f}_{\text{a}}}{\text{trace}(\mathbf{\Lambda})}. \tag{24}$$

The SNR measures the relative power between the projected background and the additive noise, while the ABR measures the relative power between the anomaly and background fields.

### *A. Direct Comparison of Statistics*

In this section, we illustrate in a direct way the superiority of the optimized statistics over the coarse-scale likelihood statistics. What we will show is that the optimized statistics are more sensitive and discriminating than the coarse-scale likelihood statistics. We illustrate this by examining the standard-deviation-normalized mean values of the two statistics conditioned on the hypotheses of a particular realization of the anomaly detection and localization problems. For the realization we focus on, summarized in Table I, we consider only the coarsest scale and we consider anomalies of size $s_{\min} = s_{\max} = 4$. The comparison between the statistics will be made for two types of background, fractal and white.

To make the comparison, recall the definitions of $m_{ij}$ and $\sigma_i^2$. The former is the mean value of statistic $\tilde{\ell}_i$ associated with $\tilde{R}_i$, which is depicted in Fig. 2 conditioned on hypothesis $H_j$. The latter is the associated variance. We define the standard-deviation-normalized conditional mean as

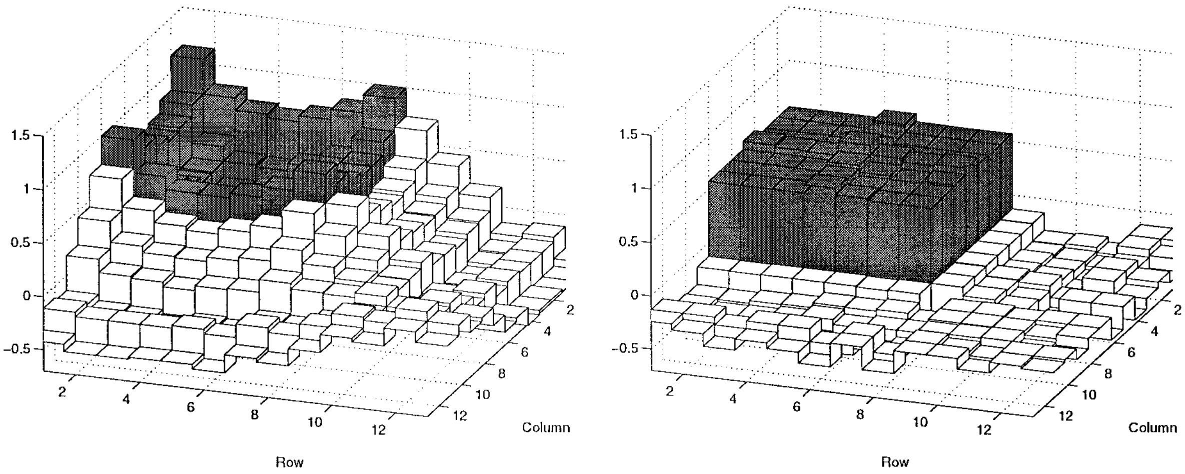
$$\mu_{ij} \triangleq \frac{m_{ij}}{\sigma_i}. \tag{25}$$

The value of $\mu_{ij}$ indicates how sensitive $\tilde{\ell}_i$ is to hypothesis $H_j$.

Figs. 6 and 7 illustrate values of $\mu_{1j}$ for the case of a white and fractal background respectively at $\text{SNR} = 1$. The 2-D bar at position $(m, n)$ in the plots of these figures corresponds to the hypothesis that $\mathbf{f}_{\text{a}} = \mathbf{b}(4, m, n)$. The shaded regions of these 2-D bar charts are the areas we wish to be large; that is, these correspond to values of $\mu_{1j}$ for hypotheses, $H_j$, that comprise the composite hypothesis $\tilde{H}_1$. The unshaded portions of these plots correspond to $H_j \notin \tilde{H}_1$, i.e., these are values of $\mu_{1j}$ we would like to be significantly smaller. In both figures, plot (a) corresponds to the coarse-scale likelihood statistic and plot (b) to the optimized statistic.

The shape of both plots in Fig. 6 exhibit precisely the type of behavior we want. The value of $\mu_{1j}$ is relatively high for $H_j \in \tilde{H}_1$ (shaded region) and relatively low for $H_j \notin \tilde{H}_1$ (unshaded region). Notice, however, that for the optimized statistic [Fig. 6(b)] there is a more abrupt transition between the shaded and unshaded regions as compared to the coarse-scale likelihood statistic [Fig. 6(a)]. Note that at this coarsest stage (and at any stage prior to the final one), our objective is to *discard* hypotheses and *not* to make an absolute decision about the presence or absence of an anomaly (this is considered at the final stage). As a consequence, it is the sharpness of the transition and the *relative* (not absolute) sizes of the $\mu_{1j}$ between the shaded and unshaded regions that are of importance. Thus, while it might appear that sensitivity is lost in Fig. 6(b) because the values in the shaded regions are somewhat lower than in Fig. 6(a), this is not the case. Indeed, because of the sharper transition in Fig. 6(b), the optimized statistic does a significantly better job at disambiguating composite hypotheses, as we will verify in the next section.

While Fig. 6 shows that there is some enhancement using the optimized statistic for the case of a white background, this point becomes much stronger if we consider a correlated background field as we do in Fig. 7. In particular, in Fig. 7 the background has a fractal covariance structure. In the case of Fig. 7(a), we see that $\tilde{\ell}_1$ is sensitive to the wrong hypotheses. That is, there exist $H_k \notin \tilde{H}_1$ and $H_j \in \tilde{H}_1$ for which $\mu_{1k} > \mu_{1j}$. Moreover, there is no clear, sharp transition between the shaded and unshaded regions in Fig. 7(a). Therefore, in the case of the fractal background, the coarse-scale likelihood statistic is ineffective. Comparing Fig. 7(a) with Fig. 7(b) we see that the optimized statistic is significantly better than the coarse-scale likelihood statistic—it is sensitive to the



(a)

(b)

Fig. 7. $\mu_{1j}$ with a fractal background at SNR $= 1$. (a) For the coarse-scale likelihood statistic. (b) For the optimized statistic. Pixel $(m, n)$ corresponds to $\mathbf{f}_{\rm a} = \mathbf{b}(4, m, n)$.

correct hypotheses and is highly discriminating between those hypotheses that belong to $\tilde{H}_1$ and those that do not. Though we shall not provide the corresponding figures for the three other coarse-scale composite hypotheses $(\tilde{H}_2,\ \tilde{H}_3,\ \tilde{H}_4)$, the same conclusions apply.

### *B. Coarse-Scale Performance*

In the previous section, we saw that the optimized statistics do a significantly better job at discriminating coarse-scale composite hypotheses as compared with the conventional log-likelihood statistics. In this section we continue to illustrate this point and also illustrate how much performance is lost by using a suboptimal algorithm as compared with the optimal (and computationally complex) GLRT. To make this latter point, we must focus on a realization of the problem, which is sufficiently small so that the optimal test can be conducted. Therefore, as described in Table I, we continue to consider a small field (16× 16) and only the coarsest scale. In contrast to the previous section, we now allow anomalies to be any size between $s_{\min} = 3$ and $s_{\max} = 4$, inclusive.[5]

In this section, we compare the performance of three algorithms for coarse-scale anomaly detection and localization. One algorithm is the GLRT. The other two are similar to the coarsest scale of a MSHT. For these, statistics associated with four coarse-scale regions (shown in Fig. 2) are computed. The largest is compared to a threshold. If it is above the threshold, its associated region is declared as containing the anomaly. Otherwise, a “no anomaly” decision is returned.

In Fig. 8, we illustrate receiver operator characteristic (ROC)[6] curves at different ABR’s for the three methods: GLRT (top curve in each plot); coarse-scale MSHT algorithm with optimized statistics (middle curve in each plot); and coarse-scale MSHT algorithm with coarse-scale likelihood statistics (bottom curve in each plot). The anomaly considered is $\mathbf{f}_{\rm a} = c\mathbf{b}(4, 6, 6)$ where $c$ varies with the ABR. The background is fractal and SNR $= 1$. Table II indicates the number of hypotheses formulated for each of the three methods considered here.

Fig. 8 unquestionably illustrates the superiority of the optimized statistic over the coarse-scale likelihood statistic. At all ABR’s, the coarse-scale MSHT performance using the optimized statistic outperforms that using the coarse-scale likelihood statistic by a wide margin. Indeed, in all but Fig. 8(a), the ROC’s for the coarse-scale MSHT using coarse-scale likelihood statistics do not stray far from a probability of detection of 0.2, the performance level that *blind guessing* achieves.[7] In fact, even in Fig. 8(a), at an ABR of 9.1, the coarse-scale likelihood statistics are not much better than blind guessing. These experimental results support the analytical ones of the previous section—the coarse-scale likelihood statistic has far lower discriminating power and sensitivity compared to the optimized statistic.

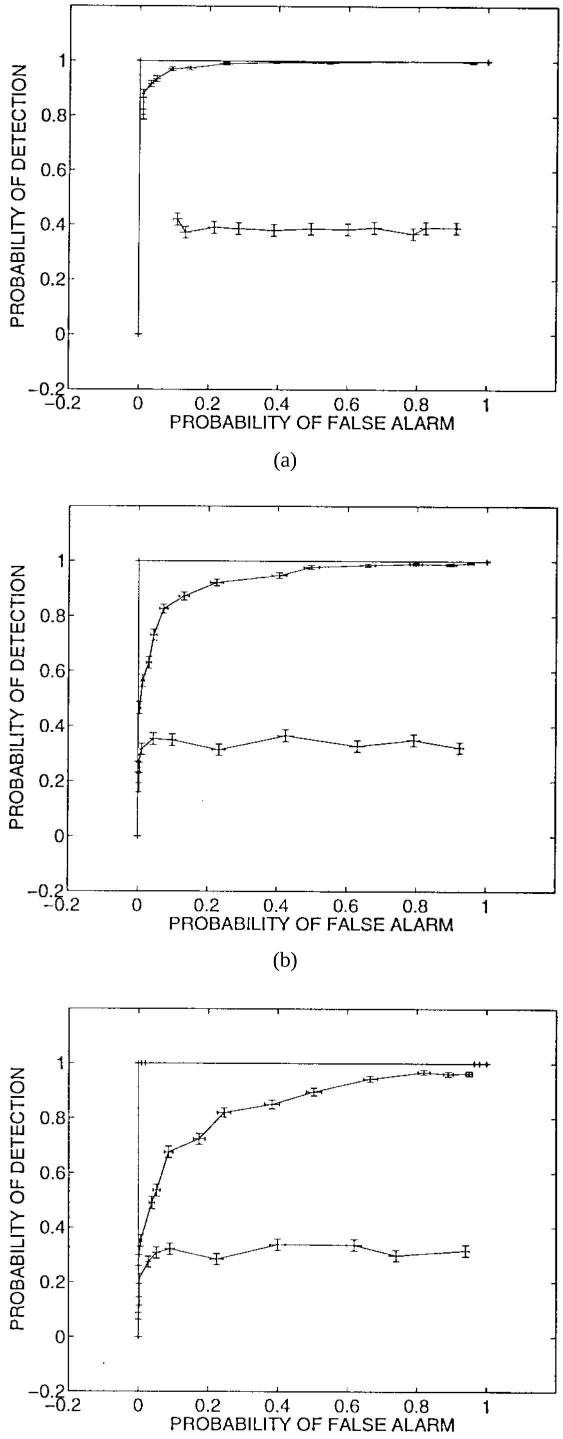
Comparing the performance of the coarse-scale MSHT with optimized statistics to the optimal GLRT provides us with a quantification of the performance loss due to the use of a computationally simpler decision rule. There are two pieces of information to extract from such results. The first is that it defines a range of ABR’s for which there is only minimal loss of performance as compared to the optimal test. For example, as Fig. 8(a) indicates, the performance loss in using the coarse-scale MSHT with affine statistics at ABR’s of about 9 or larger is quite small.

This comparison also identifies a range of ABR’s over which the constraint of using a single affine statistic for each composite hypothesis, $\tilde{H}_n$, is too severe, resulting in too great a loss of performance relative to the optimal test. For example, the performance in Fig. 8(c) and even in Fig. 8(b) suggest that

[5] This means, of course, that, in contrast to the previous section, the optimized statistics are now designed to be sensitive to all 3 × 3 and all 4 × 4 anomalies.

[6] Note that an ROC curve for an $M$-ary hypothesis test, unlike that for a binary hypothesis test, need not go through the point $(\Pr(\text{detection}), \Pr(\text{false alarm})) = (1, 1)$ because more is required for a detection in the $M$-ary hypothesis case. It is not enough that the null hypothesis is not chosen, the *correct* alternative hypothesis must be selected.

[7] Recall that there are five choices: “no anomaly” and the choice that an anomaly exists in region $\tilde{R}_n$ for $n = 1, 2, 3, 4$.



(a)

(b)

(c)

Fig. 8. Comparison of three algorithms at different ABR's. In each plot, ROC's for the GLRT (top curve), coarse-scale MSHT with optimized statistics (middle curve), and coarse-scale MSHT with coarse-scale likelihood statistics (bottom curve) are shown. Five-hundred Monte Carlo runs were conducted per data point and error bars are drawn plus and minus one standard deviation. The background is fractal and SNR $= 1$. (a) ABR $= 9.1$. (b) ABR $= 5.1$. (c) ABR $= 3.5$.

TABLE II
Number of Hypotheses Formulated by the Algorithms Considered in Section V-B

| *Algorithm* | *Number of Hypotheses Formulated* |
|---|---|
| GLRT | 365 |
| coarse-scale MSHT with coarse-scale likelihood statistics | 4 |
| coarse-scale MSHT with optimized statistics | 4 |

at this range of ABR's we need to consider more complex decision rules to achieve near-optimal performance. The GLRT and the optimization procedure we have described suggest a method for doing this. Specifically, the GLRT involves the calculation of *many* affine statistics for each $\tilde{H}_n$—one statistic matched to each individual hypothesis. In contrast, the coarse-scale MSHT algorithm using optimized statistics as we have described it, uses a single affine statistic. An obvious generalization is to use *several*—say, four—optimized statistics for each $\tilde{H}_n$, where each statistic is sensitive to a different subset of $\tilde{H}_n$. The resulting decision rule would then compute all 16 statistics and choose the $\tilde{H}_n$ corresponding to the largest one. Fig. 9 depicts the result of applying such a rule for the same cases shown in Fig. 8(b) and (c). We see from Fig. 9 that this rule, which has four times the complexity of the rule using a single statistic per $\tilde{H}_n$, has significantly better (and in the cases shown nearly perfect) performance.

### C. Full Algorithm Performance

In this section, we illustrate an example for a full MSHT algorithm. In contrast to the problems considered in the previous section, the one considered here is larger.[8] The field is 32 × 32 and the anomaly considered is $\mathbf{f}_\mathrm{a} = c\mathbf{b}(4, 3, 3)$ where $c$ varies with the ABR. One optimized statistic is used for each composite hypothesis in the tests at the first two scales of the algorithm. After the second scale, the anomaly has been localized to a 11 × 11 region, which is small enough that the optimal test can be performed over the hypotheses corresponding to anomalies with support in that region. Further details are provided in Table I. The background is fractal with SNR $= 1$.

Fig. 10 illustrates ROC's at two ABR's. The top curve corresponds to ABR $= 5.1$ and the bottom to ABR $= 3.5$. At these ABR's, we see that performance is quite good, indicating that the MSHT is indeed an effective way to navigate a large hypothesis space to find the true hypothesis. At lower ABR's, however, where performance is significantly below the level we see here, additional statistics per composite hypothesis can be used to increase performance levels, as discussed in the previous section.

## VI. Discussion and Conclusion

We have presented the multiscale hypothesis test with optimized statistics as a new approach to the anomaly detection and localization problems from noisy tomographic projections. We have shown that, in certain data-quality regimes, this hierarchical, hypothesis zooming method can achieve good performance with great efficiency. The key to achieving high performance with low computational complexity is the design of highly selective statistics. We developed and solved an optimization problem for such statistics and, in several ways, quantified their superiority over conventional statistics.

[8] This larger size prohibits the computation of Monte Carlo results based on the optimal test.

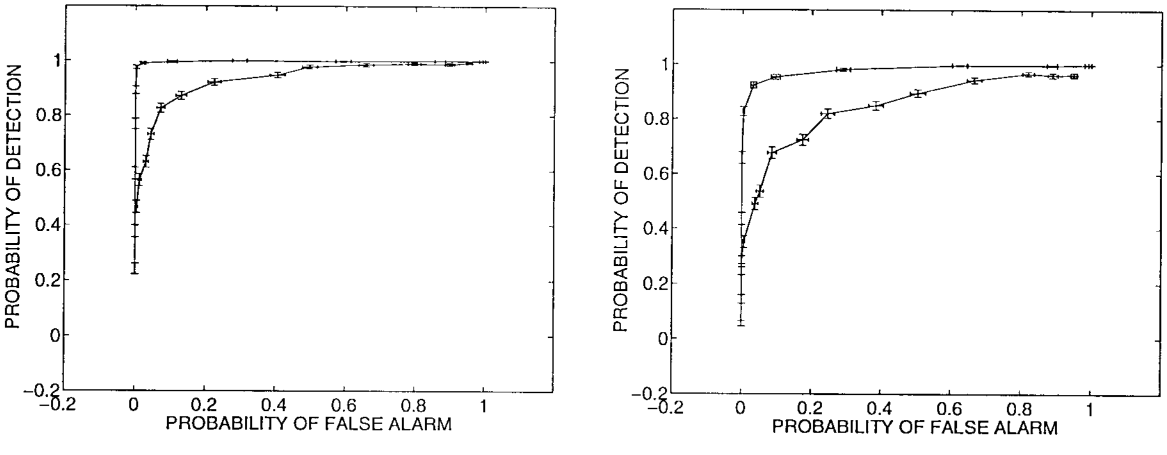


Fig. 9. ROC's for the coarse-scale MSHT. For the top curves in each plot, four affine statistics were computed per $\tilde{H}_n$. For the bottom curves, one affine statistic was computed. Five-hundred Monte Carlo runs were conducted per data point and error bars are drawn plus and minus one standard deviation. The background is fractal, SNR = 1. (a) ABR = $5.1$. (b) ABR = $3.5$.

While we have developed the MSHT framework in the context of anomaly detection and localization from tomographic data, we have touched on the fundamental issues relevant to a broad class of problems—those involving large hypothesis spaces. The key obstacles in dealing with large hypothesis spaces include how to organize the space for hypothesis zooming and how to process the data for efficient decision making. The MSHT framework as applied to the anomaly detection and localization problems addresses these challenges and provides a guiding philosophy for solutions to similar large-hypothesis-space problems.

In our development of the MSHT framework we have imposed certain conditions, the relaxation of which suggest ways to achieve additional performance gain using the methodology we have described. For example, as we described in Section V-B, it is possible to generalize our approach by designing several statistics per composite hypothesis. With this extension, our approach provides a set of MSHT algorithms ranging from the simplest (using one statistic per composite hypothesis), to the fully optimal GLRT (using as many statistics as there are individual hypotheses). This then provides a systematic framework for identifying the minimally complex MSHT as a function of data quality and performance level.

Further extensions of our approach are also possible. For example, in defining the MSHT, we have held the hierarchical hypothesis space decomposition fixed. For the problems of interest in this paper, the choice we have made (namely, defining $\tilde{H}_n$ to correspond to anomalies in region $\tilde{R}_n$) is natural. But, in more general problems, hypothesis space decomposition needs to be considered jointly with the design of decision statistics. In general, this is a prohibitively complex combinatorial problem. However, we expect that our framework can lead to feasible iterative algorithms for joint statistic design and hypothesis space decomposition. An iterative procedure might start with an initial composite hypothesis (call it $\tilde{H}$) and its associated optimized statistic. Then, using the type of information presented in Figs. 6 and 7, it can be determined which hypotheses are not placed in the correct composite hypothesis and ought to be redistributed. For example, the next suggested composite hypothesis would not include hypotheses which are in $\tilde{H}$ but have a relatively low average statistic value and would include hypotheses which are not in $\tilde{H}$ but have a relatively large statistic value. Iterating this process—i.e., redesigning the statistic to discriminate this new group of hypotheses from all others, identifying those that seem misplaced and redistributing them—provides a systematic procedure for using our methodology to build composite hypotheses that are maximally distinguished by an algorithm using optimized statistics.

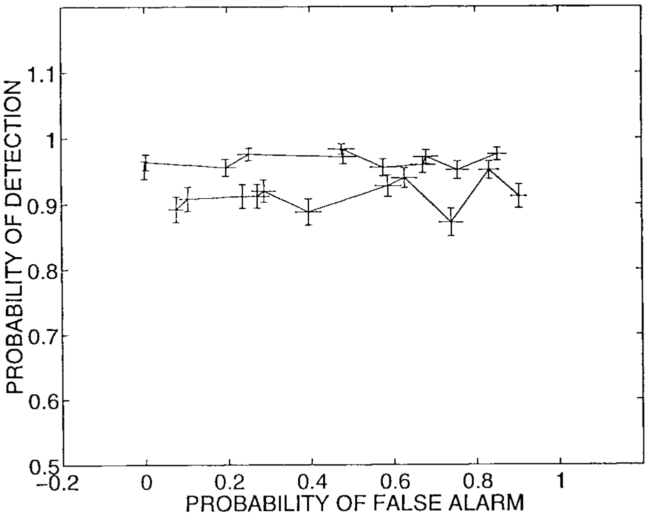


Fig. 10. ROC's for the MSHT at two ABR's. Top: ABR = $5.1$. Bottom: ABR = $3.5$. SNR = $1$, $\mathbf{f}_a = c\mathbf{b}(4,3,3)$ where $c$ varies with the ABR. The background is fractal, the number of Monte Carlo runs is 250 and error bars are drawn plus and minus one standard deviation.

Finally, there are a number of ways to broaden the range of problems for which our approach is applicable. One way is to relax the assumption that there is only one anomaly and consider multiple anomalies. One can apply our methodology to multiple anomaly problems in two ways. The simplest and crudest way is to use the statistics designed for single anomalies but to retain multiple regions at each scale. Exactly how to decide how many regions to retain is an open problem. Another, and more difficult approach, is to consider designing optimized statistics that are sensitive to different numbers of anomalies in different areas of the image. Without the assumption of some additional structure this latter approach seems infeasible due to the combinatorics of the problem.

While we have focussed on a particular 2-D linear/Gaussian problem from tomography, we feel that the methodology presented in this paper is applicable to a wider range of problems—both linear/Gaussian and also nonlinear/non-Gaussian. For example, one could consider applying our optimization procedure for statistics to nonlinear and/or non-Gaussian problems by considering only second order statistical characterizations. Three- and higher-dimensional problems could also be considered by appropriately changing the scale-recursive division of the hypothesis space. For example, in three dimensions, one might define eight overlapping cubes rather than four squares as shown in Fig. 2. The approach to finding optimized statistics would not change in higher-dimensional problems.

Although our approach may be applied to other problems, we have only explored those discussed in this paper. One feature of the problems explored here which we believe is relatively important is that tomographic data are *highly* nonlocal. It is this fact, we believe, that gives rise to the confusion of composite hypotheses and necessitates the search for better statistics. In other types of problems the conventional statistics may be adequate (see, for example, [10]). However, since our optimization procedure considers, as one possibility, the conventional statistics (and all other affine statistics), one can only do better by using optimized statistics.

## Appendix
## Quadratic Programming Formulation for Optimized Statistics

In this appendix, we show how to formulate the optimization problem for statistics posed in Section III-B as a quadratic programming problem. To do so we shall employ Lagrange duality theory which is a standard technique for recasting constrained optimization problems [2], [4]. The optimization problem we consider in Section III-B is

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}} \min_{(j,k)\in\mathcal{A}_i} \frac{\mathbf{a}^T\mathbf{T}\mathbf{b}_j - \mathbf{a}^T\mathbf{T}\mathbf{b}_k}{\sqrt{\mathbf{a}^T\mathbf{\Lambda}_{\mathbf{g}}\mathbf{a}}}. \tag{26}$$

It is sufficient to consider vectors $\mathbf{a}$ for which $\mathbf{a}^T\mathbf{\Lambda}_{\mathbf{g}}\mathbf{a} = 1$. Making a few additional notational changes, we rewrite the problem as

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} z \tag{27}$$

$$\text{subject to} \begin{cases} \mathbf{P}\mathbf{x} - z\mathbf{e} \geq 0 \\ 1 - \mathbf{x}^T\mathbf{x} \geq 0 \end{cases} \tag{28}$$

where

$$\mathbf{x} \triangleq \mathbf{\Lambda}_{\mathbf{g}}^{1/2}\mathbf{a}, \quad \mathbf{P} \triangleq \mathbf{Q}\mathbf{\Lambda}_{\mathbf{g}}^{-1/2} \tag{29}$$

and $\mathbf{Q}$ and $\mathbf{e}$ are as defined in Section III-B. This is the *primal problem*. Notice that in the primal problem we want to maximize $z$ and therefore implicitly maximize $\mathbf{x}^T\mathbf{x}$. Hence, the constraint $\mathbf{x}^T\mathbf{x} \leq 1$ is equivalent to $\mathbf{x}^T\mathbf{x} = 1$.

Let us call the optimal cost to the primal problem $\hat{z}$. For simplicity, we assume that $\hat{\mathbf{x}}$ exists and that $\hat{z} < \infty$. Introducing Lagrange multipliers $\mathbf{y}$ and $\mu$, we define the Lagrangian cost function as

$$L(z, \mathbf{x}, \mu, \mathbf{y}) \triangleq z + \mathbf{y}^T(\mathbf{P}\mathbf{x} - z\mathbf{e}) + \mu(1 - \mathbf{x}^T\mathbf{x}). \tag{30}$$

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 $J(\mu, \mathbf{y}) \triangleq \max_{z,\mathbf{x}} L(z, \mathbf{x}, \mu, \mathbf{y})$. The function $J$ is the maximum of the Lagrangian cost as a function of the Lagrange multipliers. It is straightforward to show that in searching for the $\hat{\mathbf{y}}$ and $\hat{\mu}$ we must consider only nonnegative values. It is also clear that the weak duality relationship, $J(\mu, \mathbf{y}) \geq \hat{z}$, holds for all values of $\mathbf{y}$ and $\mu$. The dual problem attempts to find the smallest upper bound. In our case it is not hard to show strong duality, i.e., that the smallest upper bound is, in fact, tight $(\min_{\mu,\mathbf{y}} J(\mu, \mathbf{y}) = \hat{z})$.

The *dual problem*, therefore, is

$$[\hat{\mu} \quad \hat{\mathbf{y}}^T] = \arg\min_{\mu,\mathbf{y}} J(\mu, \mathbf{y}) \tag{31}$$

$$\text{subject to} \begin{cases} \mu \geq 0 \\ \mathbf{y} \geq 0 \end{cases}. \tag{32}$$

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 $\mathbf{y}^T\mathbf{e} = 1$ and $\hat{\mathbf{x}} = \frac{1}{2\mu}\mathbf{P}^T\mathbf{y}$. Plugging these conditions back into $L$ yields $J(\mu, \mathbf{y}) = \mu + \frac{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}{4\mu}$.

Having found a workable expression for $J$, the dual problem is to minimize it. Using the fact that $\mu$ cannot be zero (or else $\hat{\mathbf{x}}$ is unbounded) a necessary condition for $\mu$ at the minimum is $\hat{\mu} = \frac{1}{2}\sqrt{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}$. Having found the optimal $\mu$ we plug this into $J$ to get $J(\hat{\mu}, \mathbf{y}) = \sqrt{\mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y}}$. Putting all this together, the dual problem is

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y}} \mathbf{y}^T\mathbf{P}\mathbf{P}^T\mathbf{y} \tag{33}$$

$$\text{subject to} \begin{cases} \mathbf{e}^T\mathbf{y} = 1 \\ \mathbf{y} \geq 0 \end{cases}. \tag{34}$$

Recalling the fact that $\mathbf{P} \triangleq \mathbf{Q}\mathbf{\Lambda}_{\mathbf{g}}^{-1/2}$ and $\mathbf{x} \triangleq \mathbf{\Lambda}_{\mathbf{g}}^{1/2}\mathbf{a}$, we may rewrite the dual problem and the optimal primal solution as shown in Section III-B.

## Acknowledgment

The authors would like to thank Dr. S. P. Boyd and Dr. L. Vandenberghe, both at Information Systems Laboratory, Stanford University, for their assistance in reformulating the optimization problem posed in this paper. The reviewers are also acknowledged for their helpful comments.

## References

[1] J.-M. Beaulieu and M. Goldberg, "Hierarchy in picture segmentation: A stepwise optimization approach," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 150–163, Feb. 1989.

[2] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1995.

[3] M. Bhatia, "Wavelet transform-based multi-resolution techniques for tomographic reconstruction and detection," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, Aug. 1994.

[4] S. P. Boyd and L. Vandenberghe, private communication, Apr. 1996.

[5] A. B. Frakt, "Multiscale hypothesis testing with application to anomaly characterization from tomographic projections," Master's thesis, Mass. Inst. Technol., Cambridge, May 1996.

[6] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. New York: Academic, 1981.

[7] S. L. Horowitz and T. Pavlidis, "Picture segmentation by a tree traversal algorithm," *J. ACM*, vol. 23, pp. 368–388, Apr. 1976.

[8] J. Le Moigne and J. C. Tilton, "Refining image segmentation by integration of edge and region data," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 605–615, May 1995.

[9] E. L. Miller and A. S. Willsky, "A multiscale, decision-theoretic algorithm for anomaly detection in images based upon scattered radiation," in *Proc. First Int. Conf. Image Processing*, Austin, TX, Nov. 1994.

[10] ——, "Multiscale, statistical anomaly detection analysis and algorithms for linearized inverse scattering problems," *Multidimens. Syst. Signal Process.*, vol. 8, Jan. 1997.

[11] L. M. Novak, G. J. Halversen, G. J. Owirka, and M. Hiett, "Effects of polarization and resolution on the performance of a SAR automatic target recognition system," *MIT Lincoln Lab. J.*, vol. 8, Spring–Summer 1995.

[12] J. B. Poline and B. M. Mazoyer, "Analysis of individual brain activation maps using hierarchical description and multiscale detection," *IEEE Trans. Med. Imag.*, vol. 13, pp. 702–710, Dec. 1994.

[13] 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 Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 886–906, 1984.

[14] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1968.

[15] B. J. West and A. L. Goldberger, "Physiology in fractal dimensions," *Amer. Scientist*, vol. 75, pp. 354–365, 1987.

**Austin B. Frakt** (S'96) received the B.S. degree in applied and engineering physics from Cornell University, Ithaca, NY, in 1994, and the M.S. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge, in 1996.

He currently holds a National Defense Science and Engineering Graduate Fellowship and is a member of the Stochastic Systems Group, Laboratory for Information and Decision Systems, MIT. He has worked at Hughes Aircraft Company, El Segundo, CA, and at Alphatech, Burlington, MA. His current research interests include stochastic signal and image processing with an emphasis on modeling and estimation of multiresolution signals and images including 1- and 2-D self-similar processes.

**W. Clem Karl** (M'91) received the S.M., E.E., and S.B. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, and the Ph.D. degree in electrical engineering and computer science from MIT in 1991.

He was Staff Research Scientist with the Brown–Harvard–MIT Center for Intelligent Control Systems and the MIT Laboratory for Information and Decision Systems from 1992 to 1994. He joined the faculty of Boston University, Boston, MA, in 1995, where he is currently Assistant Professor of Electrical, Computer, and Systems Engineering. Since January 1996, he has also held a joint appointment in the Department of Biomedical Engineering. His research interests are in the areas of multidimensional and multiscale signal and image processing and estimation, geometric estimation, and medical signal and image processing. He is a guest editor of the *International Journal of Pattern Recognition and Artificial Intelligence* (special issue on processing, analysis, and understanding of MR images of the human brain).

Dr. Karl is an associate editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He is a member of Sigma Xi.

**Alan S. Willsky** (S'70–M'73–SM'82–F'86) received the S.B. degree and the Ph.D. degrees from the Massachusetts Institute of Technology (MIT), Cambridge, in 1969 and 1973, respectively.

He joined the MIT faculty in 1973, and his present position is Professor of Electrical Engineering. From 1974 to 1981, he served as Assistant Director of the MIT Laboratory for Information and Decision Systems. He is also a founder and member of the board of directors of Alphatech, Inc., Burlington, MA. He has held visiting positions at Imperial College, London, L'Université de Paris-Sud, and the Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes, France. His research interests are in the development and application of advanced methods of estimation and statistical signal and image processing. Methods he has developed have been successfully applied in a wide variety of applications, including failure detection in high-performance aircraft, advanced surveillance and tracking systems, electrocardiogram analysis, computerized tomography, and remote sensing. He is the author of the research monograph *Digital Signal Processing and Control and Estimation* and is the co-author of the undergraduate text *Signals and Systems*.

Dr. Willsky was program chairman for the 17th IEEE Conference on Decision and Control, has been an associate editor of several journals, and special guest editor for several special issues, and has served as a member of the Board of Governors and Vice President for Technical Affairs of the IEEE Control Systems Society. In 1988, he was made a Distinguished Member of the IEEE Control Systems Society. He has given plenary and keynote lectures at a number of major scientific meetings, including the 20th IEEE Conference on Decision and Control, the 1991 IEEE International Conference on Systems Engineering, the 1991 SIAM Conference on Applied Linear Algebra, the 1992 Inaugural Workshop for the National Centre for Robust and Adaptive Systems, Canberra, Australia, the 1992 INRIA 25th Anniversary Symposium in Paris, the 1993 IEEE Symposium on Image and Multidimensional Signal Processing in Cannes, and the 1997 Wavelet Applications in Signal and Image Processing Conference. In 1975, he received the Donald P. Eckman Award from the American Automatic Control Council. He was awarded the 1979 Alfred Nobel Prize by the ASCE and the 1980 Browder J. Thompson Memorial Prize Award by the IEEE for a paper excerpted from his monograph.