LINEAR DISCRIMINANTS AND IMAGE QUALITY

H H Barrett[1,2,3], T Gooley[3], K Girodias[2],
J Rolland[1,2,4], T White[1,2] and J Yao[1,2]

[1]Optical Sciences Center, University of Arizona, Tucson, AZ 85721
[2]Department of Radiology, University of Arizona, Tucson AZ, 85724
[3]Program in Applied Mathematics, University of Arizona, Tucson, AZ 85721
[4]Dept. of Computer Science, University of North Carolina, Chapel Hill, NC

Abstract

The use of linear discriminant functions, and particularly a discriminant function derived from the work of Harold Hotelling, as a means of assessing image quality is reviewed. The relevant theory of ideal or Bayesian observers is briefly reviewed, and the circumstances under which this observer reduces to a linear discriminant are discussed. The Hotelling oberver is suggested as a linear discriminant in more general circumstances where the ideal observer is nonlinear and ususally very difficult to calculate. Methods of calculation of the Hotelling discriminant and the associated figure of merit, the Hotelling trace, are discussed. Psychophysical studies carried out at the University of Arizona to test the predictive value of the Hotelling observer are reviewed, and it is concluded that the Hotelling model is quite useful as a predictive tool unless there are high-pass noise correlations introduced by post-processing of the images. In that case, we suggest that the Hotelling observer be modified to include spatial-frequency-selective channels analogous to those in the visual system.

Keywords

Image quality; medical imaging; linear discriminant functions; ideal observer; Hotelling trace.

1. INTRODUCTION

A general definition of image quality has proven to be an elusive goal. Indeed, in the image-processing literature, image assessment is most often purely subjective, and no objective definition of quality is even attempted. The radiology literature is somewhat more sophisticated in this respect; image quality is usually defined there in terms of how well some observer can perform some task of diagnostic interest. The difficulty in that case is in choosing a task and an observer.

By far the most common observer of real radiographic images is the physician, though there is also considerable interest in automated or machine observers. For the human observer, task performance can be measured by psychophysical studies. If the task is binary (i.e., the observer has only two possible choices), the results of such studies can be analyzed by use of ROC (receiver operating characteristic) curves. A common figure of merit for image quality is thus the area under the ROC curve

(AUC) or the associated detectability index d' or d_a.

Though psychophysical studies and ROC analysis satisfy our requirement for a rigorous definition of image quality, there are still many problems in practice. The studies are time consuming and expensive, especially if the observers are physicians or if real clinical images are used. Moreover, the results are too specific to answer many questions of practical importance. An ROC study can give a definitive comparison of two imaging systems for one particular disease entity and one set of engineering parameters for each system, but it says nothing about how either system would perform with other parameters or for other diseases.

For these reasons, there is considerable interest in the use of model observers for which the performance indices such as AUC can be calculated rather than measured. If we had a model observer whose performance correlated well with that of the human, we could use it to study the effects of variation of task or system parameters. Such a tool would be extremely valuable for optimizing and effectively using radiographic imaging systems.

The most widely investigated model observer is the ideal or Bayesian observer, defined as one who has full statistical knowledge of the task and who makes best use of that knowledge to minimize a suitably defined risk. The strategy of the ideal observer for a binary task is to calculate a test statistic called the likelihood ratio and to compare it to a threshold in order to decide between the two alternatives; this strategy maximizes the AUC. The performance of the ideal observer sets an upper limit to the performance obtainable by any observer, including the human, and it might be hoped that a system optimized for the ideal observer would also be optimized for the human.

Though this approach seems reasonable, significant problems are encountered in practice. Most importantly, the likelihood ratio is only rarely calculable. Indeed, almost all investigations of the ideal observer have concentrated on detection of an exactly specified signal (or perhaps discrimination of two exactly known signals) superimposed on an exactly known background. We refer to such situations as SKE/BKE (signal known exactly, background known exactly). The SKE/BKE paradigm is obviously quite different from clinical radiology where, even for simple lesion-detection tasks, the background is cluttered with normal anatomic structures and the lesion to be detected is highly variable in size, location, shape and contrast.

The reason for the concentration on SKE/BKE tasks is that the likelihood ratio in that case can be calculated by simple linear filtering. For detection of a known signal on a flat background, where the only randomness is measurement noise that can be modeled as a stationary, white, Gaussian random process, the likelihood ratio is the output of a matched filter. If the noise is stationary and Gaussian but not white, the likelihood ratio is calculated by a so-called prewhitening matched filter.

Even in the SKE/BKE case, the performance of an ideal observer can be very different from that of a human observer. For example, Myers et al. (1985) found that human performance relative to the ideal was dramatically degraded by certain kinds of noise correlations. One interpretation of this result, and of similar results by other authors, is that the human observer is incapable of performing the prewhitening operation. This interpretation has led to the suggestion that the correct model for predicting human performance is the quasi-ideal or non-prewhitening (NPW) ideal observer who uses a simple matched filter, even in the presence of colored noise, to derive a test statistic. Though this test statistic is inferior to the optimum test statistic (the likelihood ratio), it does have the virtue of correctly predicting human performance in a range of SKE/BKE tasks.

Unfortunately, as we shall see in Section IV, the NPW model can yield very poor correlation with the human if there is inherent randomness in the task. Furthermore, the ideal observer is usually not an option except for SKE/BKE since the likelihood ratio is impossible to calculate. We must therefore look for other observer models that remain calculable for a wide variety of realistic tasks yet correlate well

with the human observer.

These problems have led us to consider various linear discriminant functions, where the test statistic is a linear function of the data, as potential observer models. It is our hope that a suitable linear model will be found that will be computationally tractable for a wide variety of realistic tasks and will also be a good predictor of human performance as measured by ROC.

We have given particular attention to the optimum linear discriminant, which is often ascribed to Fisher (1936) but which had its origins in a classic paper by Hotelling (1931). We therefore refer to this observer model as the Hotelling observer.

It is the goal of this paper to survey efforts at the University of Arizona to determine the usefulness of linear discriminant models, and especially the Hotelling model, as tools for the assessment and optimization of imaging systems.

2. MATHEMATICAL BACKGROUND

2.1. Problem Statement

A digital image consisting of M pixels can be represented as an Mx1 column vector **g**. This vector is related to the object being imaged, denoted **f**, by a relation of the form

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \mathbf{n} , \tag{1}$$

where **n** is a vector representing the measurement noise and **H** is an operator representing the imaging system, including any processing or reconstruction steps. If we consider only linear imaging systems and represent the object **f** in discrete form as an Nx1 column vector, then **H** is an MxN matrix. More generally, however, **H** can be a nonlinear operator, especially if a reconstruction algorithm is included, and **f** can represent a continuous object. There is no loss of generality in writing the noise as an additive term, even in the nonlinear case, provided the statistics of **n** take into account the statistics of **f** as well as the nature of **H**.

Note that **g** is a random vector, both because of the measurement noise and also because many different objects **f** will be imaged. We shall consider both sources of randomness in what follows, though only the measurement noise is present in SKE/BKE problems.

We assume that the task of interest is to observe a particular image **g** and use it to classify the corresponding **f** that produced the image into one of K classes. The simplest case is the binary task where K = 2 (e.g. normal vs. abnormal or lesion-present vs. lesion-absent). A general discriminant function for this binary task is a scalar test statistic $\lambda(\mathbf{g})$. The classification is performed by comparing this test statistic to a threshold λ_t; if $\lambda(\mathbf{g}) > \lambda_t$, **f** is said to belong to class 1, while otherwise it is classified into class 2. The theory of discriminant functions is concerned with finding the best functional form for $\lambda(\mathbf{g})$ and with assessing the accuracy of the classification procedure. If $\lambda(\mathbf{g})$ is a linear function of **g**, it is referred to as a linear discriminant.

2.2. Ideal Observers and Matched Filters

One theoretical route that yields a linear discriminant is to assume an ideal observer and to model the noise **n** as a Gaussian random process. In general, the test statistic used by the ideal observer is the likelihood ratio, defined by

$$\lambda_{ideal} = \frac{p(\mathbf{g}|\,1)}{p(\mathbf{g}|\,2)} , \tag{2}$$

where $p(\mathbf{g}|\,k)$ is the probability density of **g** given that it was produced by an object in

class k (k = 1 or 2). This test statistic is usually very difficult to determine and a highly nonlinear function of g. In the special case where p(g| k) is a multivariate normal probability density function, with the same covariance matrix K for both classes, λ_{ideal} is given by the so-called prewhitening matched filter:

$$\lambda_{PW}(g) = [\bar{g}_2 - \bar{g}_1]^t \mathbf{K}^{-1} g \ , \tag{3}$$

where $\bar{g}_k$ is the mean image for class k, and the superscript t denotes a matrix transpose. (Hence, if a and b are two vectors, $a^t b$ denotes their scalar product.) This test statistic λ_{PW}, which is clearly a linear function of g, can be calculated for each image if $\bar{g}_1$ and $\bar{g}_2$ and the common covariance matrix K are known. As an aside, if we model the noise as a Gaussian random process but with different covariance matrices for the two classes, the likelihood ratio turns out to be a quadratic function of g.

The model that led to Eq. (3) is very restrictive, and it is not obvious whether it is applicable to real radiographic images. One situation in which it is applicable is simple signal detection where the object consists of a flat background on which some weak signal can be superimposed. If the object is a gamma-ray emitter and we image it through a pinhole or collimator, the measurement noise is rigorously Poisson, but we can usually approximate the Poisson law by a Gaussian. If there is some linear post-processing filter, the Gaussian noise remains Gaussian but it becomes correlated. This is precisely the situation for which Eq. (3) describes the likelihood ratio, with $\bar{g}_2 - \bar{g}_1$ being the image of the signal to be detected, so in this case the prewhitening matched filter is indeed ideal.

Unfortunately, even in this simple situation, the prewhitening matched filter is not a good model for the human. There is considerable evidence, reviewed in Section IV, that shows that the human cannot perform the prewhitening operation. A better model for SKE/BKE signal-detection problems might be the non-prewhitening (NPW) matched filter where the test statistic is given by

$$\lambda_{NPW}(g) = [\bar{g}_2 - \bar{g}_1]^t g \ . \tag{4}$$

This form, which differs from Eq. (3) only by deletion of $\mathbf{K}^{-1}$, has a simple physical interpretation. For signal detection, the expected difference signal $\bar{g}_2 - \bar{g}_1$ is an image of the signal to be detected, so the observer simply lays a template of this signal image over the image g and integrates; the integral of the product of $\bar{g}_2 - \bar{g}_1$ and g is then the test statistic.

2.3. Scatter Matrices

If the object f is regarded as a random variable, the probability densities p(g| k) that enter into the likelihood ratio must take into account the variability of f as well as the measurement noise n. Even though Poisson noise can be modelled as Gaussian, it is highly unlikely that a multivariate Gaussian law would adequately describe realistic medical objects f. The true densities are very difficult to determine, and even if they could be found, the likelihood ratio would be a nonlinear function of g, making the performance of the ideal observer difficult to analyze.

For these reasons, we consider test statistics $\lambda(g)$ that are constrained from the outset to be linear. To define these test statistics and analyze their performance, we describe the first- and second-order statistics of g by use of two "scatter matrices" $\mathbf{S}_1$ and $\mathbf{S}_2$. The interclass scatter matrix $\mathbf{S}_1$, which measures how far the class means for the data values deviate from their grand mean $\bar{g}$, is defined as

$$S_1 \equiv \sum_{k=1}^{K} P_k(\bar{g}-\bar{g}_k)(\bar{g}-\bar{g}_k)^t \ , \tag{5}$$

where K is the number of classes (two in the binary problem), P_k is the probability of occurrence of class k, $\bar{g}_k$ is the class mean for the k^{th} class (where the overbar denotes an ensemble average that accounts for the variability in f as well as n), and the grand mean is given by

$$\bar{g} = \sum_{k=1}^{K} P_k \bar{g}_k \ . \tag{6}$$

The intraclass scatter matrix S_2 is the average covariance matrix, given by

$$S_2 \equiv \sum_{k=1}^{K} P_k K_k \ , \tag{7}$$

where the k^{th} class covariance matrix is given by

$$K_k \equiv (g-\bar{g}_k)(g-\bar{g}_k)^t{}_k \ , \tag{8}$$

where the angular brackets have the same meaning as the overbar, a full ensemble average over all objects f in class k and all realizations of n.

If there are M pixels in the image, both S_1 and S_2 are MxM matrices. Since S_2 represents an ensemble covariance matrix, it will usually also have rank M. The rank of S_1, on the other hand is much less than M, in fact just K-1, where K is the number of classes (Fiete et al., 1987). Thus, for a two-class problem, S_1 has rank one, and it can be written as a single outer product:

$$S_1 = P_1P_2(\bar{g}_2-\bar{g}_1)(\bar{g}_2-\bar{g}_1)^t = xx^t \ , \tag{9}$$

where

$$x \equiv \sqrt{P_1P_2}(\bar{g}_2-\bar{g}_1) \ . \tag{10}$$

2.4. Optimum Linear Discriminants

The first step in using the scatter matrices S_1 and S_2 to form a linear discriminant is to solve the eigenvalue problem:

$$S_2^{-1}S_1 \ u_p = \mu_p u_p \ , \quad p = 1....M, \tag{11}$$

where u_p is an Mx1 column eigenvector and μ_p is its associated eigenvalue. Since S_1 has rank K-1, as noted above, there are just K-1 nonzero values of μ_p. Each of the eigenvectors u_p corresponding to nonzero μ_p can then be used to form a linear feature given by

$$\lambda_p(g) = u_p{}^t g \ , \ p = 1...K-1 \tag{12}$$

The hyperplanes $\lambda_p(g) = C_p$, where the C_p are constants, then partition the g space into K disjoint regions corresponding to the K classes. One common choice of the constants leads to the following decision rule: Choose the class k for which

$$\sum_{p=1}^{K-1} [u_p^t(g - \bar{g}_k)]^2 = \min . \tag{13}$$

In other words, the class chosen by this rule is the one for which the image being tested is closest to the class mean, where distance is measured in the eigenvector basis as indicated in Eq. (13). Other choices for the constants C_p lead to other decision rules, but all of them are based on partitioning the g space with some set of hyperplanes, so in this sense all are linear discriminants.

The situation is much simpler in the two-class problem, for which there is only one $\lambda_p(g)$, and we therefore drop the index p. In that case, an explicit solution to the eigenvalue equation is given by

$$\mathbf{u} = S_2^{-1}\mathbf{x} \tag{14}$$

$$\mu = \mathbf{x}^t S_2^{-1}\mathbf{x} . \tag{15}$$

Direct substitution of Eqs. (14) and (15) into Eq. (11) will verify that the $\mathbf{u}$ and μ given by these equations are indeed an eigenvector and eigenvalue, respectively, of $S_2^{-1}S_1$. The decision rule is then to compute $\lambda(g) = \mathbf{u}^t\mathbf{g}$ and compare it to a threshold C; class 1 is chosen if $\lambda(g)>C$ and class 2 if $\lambda(g)<C$.

It is interesting to compare the Hotelling test statistic $\mathbf{x}^t S_2^{-1}\mathbf{g}$ to the prewhitening matched filter given in Eq. (3), which can be written as $\mathbf{x}^t\mathbf{K}^{-1}\mathbf{g}$. The difference is that the noise covariance matrix $\mathbf{K}$ has been replaced by the more general weighted covariance matrix S_2. Thus if the only source of variability in g is additive Gaussian noise, the Hotelling observer is the same as the ideal observer. More generally, the ideal observer uses a test statistic that is nonlinear in g, and the Hotelling test statistic is a linear approximation to it.

2.5. Performance Measures

A common and important measure of task performance for binary decisions is the detectability index d_a, defined by

$$d_a^2 = \frac{[E(\lambda(g)|\,2) - E(\lambda(g)|\,1)]^2}{P_1\,\mathrm{var}(\lambda(g)|\,1) + P_2\,\mathrm{var}(\lambda(g)|\,2)} , \tag{16}$$

where $E(\lambda(g)|\,k)$ is the conditional mean of the test statistic $\lambda(g)$ given that g comes from class k, while $\mathrm{var}(\lambda(g)|\,k)$ is the corresponding conditional variance. It is well known that, if $\lambda(g)$ is Gaussian, d_a is related to the area under the ROC curve by

$$\mathrm{AUC} = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left[\frac{d_a}{2}\right] , \tag{17}$$

where erf( ) is the error function.

Although d_a is an accepted index of performance in binary classification tasks, it is not readily extended to tasks with more than two alternatives. For such tasks a possible performance metric is the Hotelling trace J, defined by

$$J = \mathrm{tr}(S_2^{-1} S_1) \ , \tag{18}$$

where tr denotes the trace (sum of the diagonal elements) of the matrix. Since the trace is a scalar invariant, it can be calculated in a representation where the matrix is diagonal, and hence tr(A) is the sum of the eigenvalues of A. For the case at hand, we already know that $S_2^{-1}S_1$ has only K-1 nonzero eigenvalues, so the trace is a sum of K-1 terms. For K=2, $\mathrm{tr}(S_2^{-1}S_1)$ is just the μ given in Eq. (15).

The Hotelling trace is an intuitively appealing figure of merit for classification performance. It is a single scalar, so it can be used for system optimization. It increases if the system is modified in such a way that the class means become more widely separated, increasing the norm of the vectors that constitute S_1, and it also increases if the variability in the image due to noise or other factors is decreased, since that corresponds to reducing the covariance terms that go into S_2 and hence in a sense increasing S_2^{-1}.

Moreover, in the two-class problem with Gaussian assumptions, J is given by

$$J = P_1 P_2 [d_a(\mathrm{Hot})]^2 \ , \tag{19}$$

so optimizing a system for maximum J is equivalent to optimizing the d_a or AUC for an observer that uses the Hotelling feature operator u to form a test statistic.

2.6. The NPW Observer

For later reference, we give the expression for the d_a index for the NPW observer in terms of S_1 and S_2. As shown by Barrett (1990), we have

$$[d_a(\mathrm{NPW})]^2 = \frac{1}{P_1 P_2} \frac{[\mathrm{tr}\,(S_1)]^2}{\mathrm{tr}(S_1 S_2)} \ . \tag{20}$$

3. COMPUTATIONAL METHODS

3.1. Image Modelling and Training Sets

One scheme we have used for implementing the Hotelling observer is to begin with a realistic three-dimensional mathematical model of some organ or organ system, allowing variability in both normal anatomy and in the nature and placement of lesions or other pathology (Cargill, 1989). This model is then used to create training sets of objects in two or more classes (normal and abnormal classes in the simplest case), and these simulated objects are used with an accurate model of the imaging system to create training sets of images. Ideally, these images would be indistinguishable from ones obtained with real clinical objects and physical imaging systems; the model developed by Cargill (1989) is very close to this ideal in the case of radiocolloid imaging of the reticuloendothelial system (liver, spleen and spinal bone marrow). Further work will be needed for other organ systems.

Given a training set of images created this way, a straightforward attempt to implement the Hotelling prescription would be to estimate the ensemble scatter matrices S_1 and S_2 from the training set, denoting the estimated matrices by $\mathcal{S}_1$ and $\mathcal{S}_2$, and then try to form $\mathcal{S}_2^{-1}\mathcal{S}_1$ by usual matrix manipulations; this method fails badly. One problem is that the matrices are huge. If the images are 64x64, then $\mathcal{S}_1$ and $\mathcal{S}_2$ are each 4096x4096. Furthermore, if the number of images in the training set is less than 4096, $\mathcal{S}_2^{-1}$ does not exist.

Fiete et al. (1987) have described a way to avoid the singularity of $\mathcal{S}_2$ in some cases. The trick is to take advantage of the fact that we can generate noise-free images in simulation studies. We can express S_2 rigorously as the sum of two matrices:

$$S_2 = S_2^{nf} + C_n \ , \tag{21}$$

where S_2^{nf} is the S_2 that would result from noise-free images, and C_n is the covariance matrix of $\mathbf{n}$. Using noise-free simulated images, we can get an estimate of S_2^{nf}, which we shall denote as $\mathcal{S}_2^{nf}$. Furthermore, C_n can often be modelled theoretically. For example, if we consider pinhole or collimator imaging, the noise is Poisson and C_n is a diagonal matrix with diagonal elements given by the mean values of each pixel in g. These mean values can be estimated rather accurately from the noise-free training set, so we also have an estimate $\mathcal{C}_n$ of C_n. Combining these two estimates, we get our final estimate for S_2, namely

$$\mathcal{S}_2 = \mathcal{S}_2^{nf} + \mathcal{C}_n \ . \tag{22}$$

Since the second term in $\mathcal{S}_2$ is full rank and both terms are non-negative definite, the inverse now exists, but it is still may not be practical to calculate it directly. Instead, we now take advantage of the fact that we do not need to know $\mathcal{S}_2^{-1}\mathcal{S}_1$ completely; rather, we need only to find its dominant eigenvectors and eigenvalues. Furthermore, $\mathcal{S}_1$, like S_1, has rank K-1, where K is the number of classes, so for a two-class problem we have to find only a single eigenvector and eigenvalue. The eigenvector is just a 64x64 image in our example, so now we have a relatively routine image reconstruction problem; an iterative algorithm for its solution is given by Fiete et al. (1987). Once the eigenvector is found, its scalar product with each image generates the scalar test statistic. The Hotelling trace can then be readily calculated, and its variation with any number of engineering parameters can be efficiently studied.

3.1. Region of Interest

The approach of estimating S_1 and S_2 from a noise-free training set requires that we have a theoretical model for C_n, but this is not always possible. If we have a nonlinear processing step in the imaging procedure, we cannot specify C_n, even though we may know a great deal about the statistics of the noise prior to the nonlinearity. Since nonlinear reconstruction algorithms are very valuable in tomography, this is a severe limitation.

One way to ensure that $\mathcal{S}_2$ is full rank is to have more images in the training set than pixels in one image. While this is difficult for even 64x64 images, it may be possible to restrict the task in such a way that the observer needs only a subset of the pixels to perform it. For example, if we wish to detect a small lesion in a fixed location, then S_1 corresponds to a template that covers only a few pixels. According to Eqs. (12) and (14), the template is applied not to g but rather to $S_2^{-1}g$, and the S_2^{-1} operation spreads information from the lesion location in g to surrounding pixels. The extent of this spread is difficult to predict, since it depends on the exact nature of S_2, but the spread should certainly be less than the full size of the image. Thus it may be possible to choose a small region that encompasses $S_2^{-1}g$ yet contains fewer pixels than the number of training images. An experimental approach to choosing the region size would be to start with the very small region defined by S_1 and to gradually increase it, estimating J at each region size. If the estimate of J approaches a constant before the number of pixels in the region exceeds the number of images, that constant value can reasonably be taken as a good estimate of the ensemble J.

3.3. Effects of Location Uncertainty

One objection to the method just described might be that the use of a fixed lesion location is unrealistic. A simple calculation will show, however, that this restriction is really not very severe. Suppose that the task is to detect a small lesion

that might be in one of L nonoverlapping locations. From Eq. (9), S_1 is (within a constant) just the outer product of $\bar{g}_2-\bar{g}_1$ with itself. Let s_ℓ denote the mean difference signal $\bar{g}_2-\bar{g}_1$ when the lesion is in the ℓ^{th} location. Since the lesion locations do not overlap, we can write

$$S_1 = \frac{P_1P_2}{L^2}\sum_{\ell=1}^{L} s_\ell s_\ell^t = \frac{1}{L^2}\sum_{\ell=1}^{L} S_{1\ell} , \tag{23}$$

where $S_{1\ell}$ is the S_1 matrix that would result from a lesion in the ℓ^{th} location alone. If we assume that the lesion is weak and does not significantly affect the noise level in the image, the class covariance matrices K_k are approximately equal, and $S_2 \cong K_1 \cong K_2$. Thus S_2 is approximately independent of the lesion location, and we have

$$J = \mathrm{tr}(S_2^{-1}S_1) = \frac{1}{L^2}\sum_{\ell=1}^{L} \mathrm{tr}(S_2^{-1}S_{1\ell}) , \tag{24}$$

which, if all locations are statistically equivalent, is just 1/L times the value of J for the fixed-location case. Thus, if we are comparing two imaging systems, the one that gives the higher J for detection of a lesion in a fixed location will also give the higher J for detection of the same lesion in one of L nonoverlapping locations; the absolute J value will be reduced by 1/L, but the rank ordering of the systems will be preserved.

3.4. Pseudoinverses

While $\mathcal{S}_2$ may not have an inverse, it always has a unique Moore-Penrose pseudoinverse, which we shall denote by $\mathcal{S}_2^+$. Thus a rather obvious way to estimate J is by

$$\mathcal{J} = \mathrm{tr}[\mathcal{S}_2^+\mathcal{S}_1] . \tag{25}$$

This estimate of J has several nice properties, though space does not permit a full exposition here. First, note that $\mathcal{S}_2$, being a weighted sum of covariance matrices, may be defined in one of two ways. The unbiased estimate of the ensemble S_2 is given by

$$\mathcal{S}_2 = \frac{1}{N-1}\sum_{k=1}^{K} P_k \sum_{n=1}^{N} [g_{kn} - \bar{g}_k][g_{kn} - \bar{g}_k]^t , \tag{26}$$

where N is the number of training images per class, g_{kn} is the n^{th} training image from the k^{th} class, and $\bar{g}_k$ is the sample mean image for the k^{th} class. The alternative definition of $\mathcal{S}_2$ uses 1/N in place of 1/(N-1) in this equation, resulting in a small bias but yielding a maximum-likelihood estimate of S_2. It will be shown in a subsequent paper that $\mathcal{J}$ using the 1/N definition of S_2 is a maximum-likelihood estimate of J, while the 1/(N-1) definition leads to a minimum-variance estimate. In practice, N is relatively large, so these two estimates will not differ appreciably.

One might think that actual calculation of $\mathcal{S}_2^+$ would be an enormous computational task, since $\mathcal{S}_2$ is a 4096x4096 matrix for 64x64 images. In fact, however, it suffices to work with a KNxKN matrix. In most of our work, we take K=2 and N=32, so all that is required is a singular-value decomposition of a 64x64 matrix, an eminently feasible task. Again, details will be published separately.

3.5. Stationary Background Models

Another way to reduce the computational burden is to assume spatial stationarity for the image statistics. To see the advantage of this assumption, consider first one-dimensional images. In that case, the covariance matrices are Toeplitz and usually approximately circulant. This means that knowledge of one row or column of the matrix is sufficient to specify the full matrix, greatly reducing the number of independent parameters we must determine. Moreover, a circulant matrix can be diagonalized by means of a discrete Fourier transform, and the resulting diagonal elements are samples of the power spectral density. Thus knowledge of the power spectrum for each class yields the covariance matrices and hence S_2 and S_2^{-1}, with the latter guaranteed to exist unless all of the class power spectra vanish identically at some spatial frequency.

The situation in two spatial dimensions is more cumbersome, but the same basic conclusions hold. The covariance matrices are block-Toeplitz and approximately block-circulant, so they can be diagonalized by a 2D discrete Fourier transform. In both 1D and 2D, calculation of $\mathrm{tr}(S_2^{-1}S_1)$ reduces to performing an integral over the Fourier domain (Barrett et al., 1989; Myers et al., 1990).

4. EXPERIMENTAL RESULTS

4.1. Initial Studies Using Training Sets

As an initial test of the use of the Hotelling trace as a quality metric, we created a simple two-dimensional phantom of random, overlapping ellipses, roughly representing a liver (Fiete at al., 1987). The task was to detect a small cold lesion of random size, shape and contrast. The images were blurred with Gaussian blur functions of different widths, and Gaussian noise of various amplitudes was added; 32 normal and 32 abnormal images were generated for each combination of blur width and noise level. Each of these images was presented to 10 human observers, who were asked to use a six-point rating scale to specify how certain they were that a lesion was present. These data were analyzed to produce ROC curves from which values of the index d_a were computed for each value of noise and blur.

The Hotelling test statistic was calculated by use of Eq. (22), with $\mathcal{C}_n$ just being a constant times the unit matrix, and the Hotelling feature operator was found by an iterative search algorithm. This feature operator was then used to construct ROC curves for the Hotelling observer and to compute d_a. A remarkable correlation (r=0.99) between Hotelling and human d_a values was found.

4.2. Collimator Optimization

Later, a more realistic extension of this study was performed, with the objective of determining the optimal collimator to use in planar radiocolloid imaging of the liver (Fiete et al., 1987; White et al., 1989). In this study, three-dimensional mathematical phantoms (Cargill, 1989) of the reticuloendothelial system were generated to model a healthy class, while another group of mathematical phantoms with elliptical cold regions in the liver simulated a diseased class. Images of these objects through 24 parallel-hole collimators with various bore diameters (D_b = 1 mm to 7 mm) and bore lengths (L_b = 1 cm to 11 cm) were calculated, taking into account the attenuation and scatter in the body, spatial resolution due to the collimator and camera, and Poisson noise. The Hotelling trace J was calculated from these images for each collimator. As in the initial Fiete study, Eq. (22) formed the basis for the calculation, but this time $\mathcal{C}_n$ was derived from a Poisson noise model, so it was diagonal but not a multiple of the unit matrix.

Each liver was imaged through each collimator for two different lengths of time.

For a short imaging time, the best collimator had bore diameter of 1 mm and bore length of 1 cm. At a longer (and more reasonable) imaging time, the best collimator had a higher resolution and was slightly more efficient (D_b = 3 mm, L_b = 5 cm). Some typical images for the shorter time and the corresponding J values are given in Figure 1.

It is interesting to note that the ultra-high-resolution long-bore collimators performed poorest in both of these cases. An auxiliary study to chart the variations of J as a function of imaging time for a few of the collimators found that the long-bore collimators did indeed perform better than the others, but at such long imaging times (or high doses) as to be prohibitive for clinical studies. A psychophysical study to corroborate the conclusions of this theoretical investigation has also been performed, and again a good correlation between human and Hotelling performance was found (r=0.83).

4.3 Algorithm Optimization

Gooley has investigated a number of statistical image reconstruction algorithms for use with a coded-aperture cardiac SPECT imager. Among the methods studied were maximum likelihood techniques, including both the popular expectation-maximization (EM) algorithm and a Monte Carlo search routine. Other algorithms included various kinds of prior information through the specification of a prior probability density or hard constraints on the object (or class of objects) to be reconstructed.

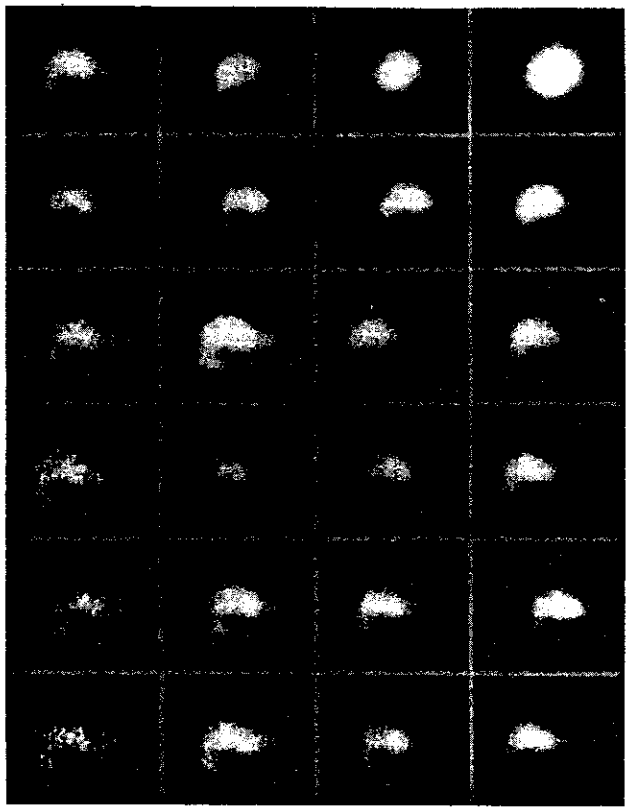
The objects used in this study were ellipses representing a left-ventricular cross-section at end systole. The abnormalities to be detected were akinetic wall segments represented as small protrusions of the upper surface of the ellipses. The objects, both normal and abnormal, were taken to be binary, i.e., each pixel was constrained to take on one of two distinct values. This assumption is reasonable in first-pass cardiac studies where a pixel either contains a uniform amount of the injected radiopharmaceutical or it contains none of the radiopharmaceutical. Some algorithms made use of this information while others did not.

Gooley has completed a psychophysical study of the images from a total of 16 algorithms and obtained d_a values for the human observer for each algorithm. Efforts to compare these psychophysical data with d_a(Hot) and d_a(NPW) are in progress at this writing. Preliminary results indicate that the psychophysical results do not correlate well with d_a(NPW).

4.4. Effects of Noise Correlation in an SKE/BKE Task

Myers et al. (1985) studied the performance of human and ideal observers for detection of a nonrandom disk signal superimposed on a spatially uniform, nonrandom background. The object was imaged through an aperture having a point spread function $p_1(r)$, and Poisson noise was added to the resulting blurred image. A deblurring filter with PSF $p_2(r)$ was then used to partially compensate the blur due to the aperture.

The main variable in this study was the aperture PSF $p_1(r)$. The PSF of the deblurring filter was adjusted so that all imaging systems considered had the same final point spread function, $p_1(r)*p_2(r)$ (where $*$ denotes convolution), regardless of the form of $p_1(r)$. Four different functional forms were used for $p_1(r)$; each was a low-pass filter but there were different rates of falloff in the frequency domain. Therefore, the corresponding deblurring filters $p_2(r)$, required to hold $p_1(r)*p_2(r)$ constant, had four different rates of high-frequency boost. Normally, increased high-frequency boost would increase the noise level in the images, but in this study the exposure time was also varied in such a way as to keep the performance of the ideal observer constant. Thus all images had the same overall point spread function and the same ideal-observer detectability performance; the only difference was in the noise correlation



BORE LENGTH (cm) \ BORE DIAMETER (cm)	0.1	0.3	0.5	0.7
1.0	Eff.=6.3×10^{-4} R=0.77 J=1.147	Eff.=5.6×10^{-3} R=2.34 J=0.608	Eff.=1.6×10^{-2} R=3.89 J=0.451	Eff.=3.1×10^{-2} R=5.45 J=0.316
3.0	Eff.=6.9×10^{-5} R=0.31 J=0.983	Eff.=6.3×10^{-4} R=0.94 J=1.081	Eff.=1.8×10^{-3} R=1.57 J=0.768	Eff.=3.4×10^{-3} R=2.19 J=0.547
5.0	Eff.=2.5×10^{-5} R=0.22 J=0.726	Eff.=2.3×10^{-4} R=0.66 J=1.034	Eff.=6.3×10^{-4} R=1.12 J=0.873	Eff.=1.2×10^{-3} R=1.54 J=0.767
7.0	Eff.=1.3×10^{-5} R=0.19 J=0.538	Eff.=1.2×10^{-4} R=0.54 J=0.927	Eff.=3.2×10^{-4} R=0.90 J=0.927	Eff.=6.3×10^{-4} R=1.26 J=0.825
9.0	Eff.=7.7×10^{-6} R=0.16 J=0.421	Eff.=6.9×10^{-5} R=0.48 J=0.812	Eff.=1.9×10^{-4} R=0.79 J=0.871	Eff.=3.8×10^{-4} R=1.11 J=0.813
11.0	Eff.=5.2×10^{-6} R=0.14 J=0.342	Eff.=4.7×10^{-5} R=0.43 J=0.714	Eff.=1.3×10^{-4} R=0.72 J=0.743	Eff.=2.5×10^{-4} R=1.01 J=0.343

Figure 1 Left: Simulated images of a 3D mathematical liver phantom through 24 different collimators. These images are representative of a large set consisting of a total of 1,536 images (64 phantoms, 24 collimators) that we have generated. Half of the phantoms, including the one used in this figure, contained a cold lesion. Right: Parameters of the collimators used to generate the images at the left, arranged in the same order as the images. Eff. denotes efficiency, R denotes resolution in cm at a distance of 8 cm, and J denotes the calculated Hotelling trace.

structure.

Since the task was SKE/BKE, the ideal observer and the Hotelling observer for this study were identical, and both were implemented by means of the prewhitening matched filter as specified in Eq. (3). The psychophysical study, on the other hand, revealed that the performance of the human observer was greatly degraded when there was a strong high-pass character to the noise. The ideal/Hotelling observer failed completely to predict the influence of noise correlations on human performance. The NPW model, on the other hand, accurately described the psychophysical data.

A later theoretical investigation by Myers and Barrett (1987) found an alternative model that also explained the psychophysical data of Myers et al. (1985). These authors considered the effects of the spatial-frequency-selective channels in the visual system that have been found in a variety of psychophysical and neurophysiological experiments. The ideal-observer model was modified so that the observer had to preprocess the images through these channels, with a resulting loss of information, before calculating the likelihood ratio. The performance of this so-called channelized ideal observer was found by Myers and Barrett to be indistinguishable from that of the NPW observer for a wide range of tasks. The channels thus provide a plausible explanation of the human's inability to prewhiten.

4.5. Stationary Background Statistics

We have performed an extensive analysis of the effects of spatial inhomogeneity of the background on detection of a known lesion (Barrett et al. 1989; Rolland, 1990; Myers et al., 1990). The background was described as a stationary random process with a Gaussian autocorrelation function, and the signal to be detected was at a fixed location and had a Gaussian profile. The object, consisting of the background and, in half the images, the signal, was imaged through a pinhole aperture having either a hard-edged square profile or a smooth, Gaussian profile. Important variables included the width of the aperture (relative to the width of the signal) and the exposure time. The performance of three observers -- human, Hotelling and NPW -- were determined as a function of these variables.

The theoretical performance of the NPW and Hotelling observers was simple to determine since, as indicated in section 3.5, the S_2 matrix in the case of a stationary background is diagonalized by a Fourier transform, so calculation of the traces in Eqs. (18) and (20) reduces to integration in the Fourier domain. The integrals were performed numerically, resulting in plots of $[d_a(\text{Hot})]^2$ and $[d_a(\text{NPW})]^2$ as a function of aperture width and exposure time (Myers et al., 1990).

As shown in Figure 2, there are striking differences in these plots for the two observers. The performance of the Hotelling observer increases steadily with increasing exposure time, while the NPW observer shows a saturation at a very short time. Not surprisingly, the NPW observer, which does not take into account any statistical properties of the background, is far more sensitive to the inhomogeneity than is the Hotelling observer. Finally, the predictions of the two observer models for the optimum aperture size to use is somewhat different.

The psychophysical studies performed by Rolland (1989) allow an unambiguous choice between the two models. As seen from Figure 2, the dependence of the human d_a on aperture size, exposure time, and degree of background lumpiness is very well predicted by the Hotelling model and not at all by the NPW model.

4.6. Effect of Higher-order Statistics

Rolland's psychophysical study described above was later extended by Yao (unpublished). Yao considered two different ways of generating the stationary random process for the background. In one method, the background was obtained by spatial filtering of a white, Gaussian random process, so that the grey-level probability densi-

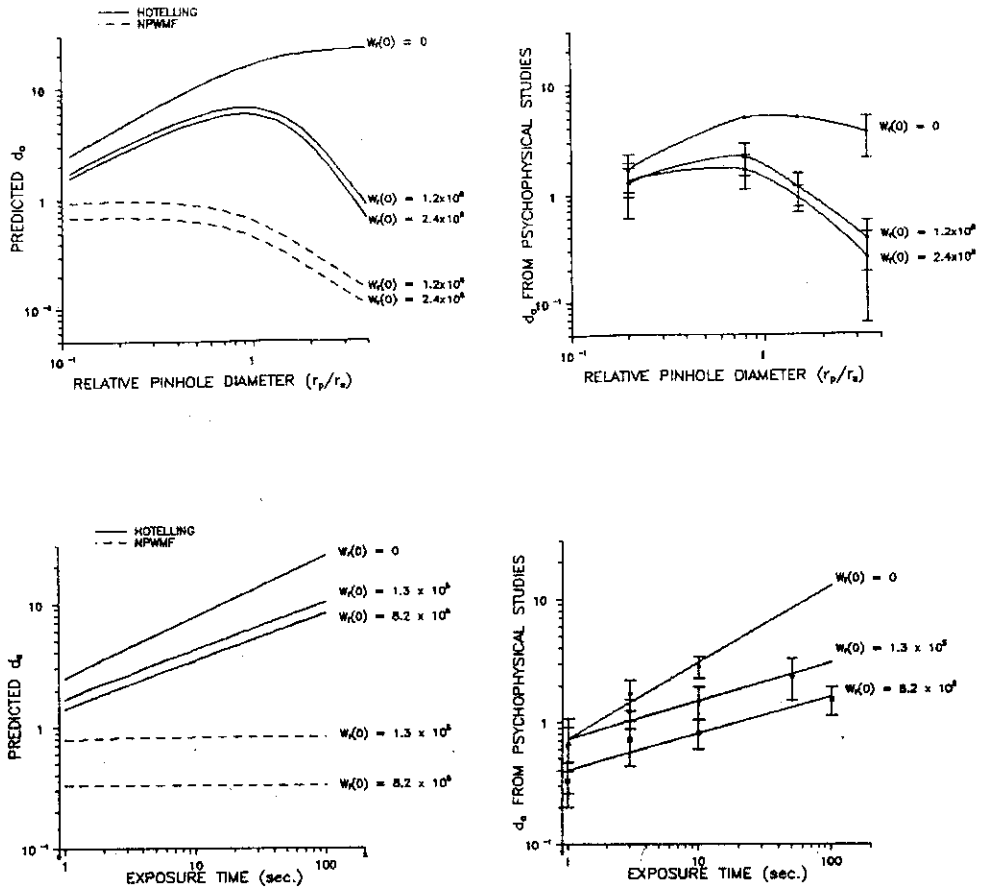


Figure 2 Results from Rolland (1990) on the detectability of a known lesion in an inhomogeneous background. Top left: Performance metrics (d_a) for the Hotelling and NPW observers as a function of aperture size. The parameter $W_f(0)$ specifies the degree of nonuniformity of the background. The upper curve, labelled $W_f(0) = 0$ is for the uniform background for which the Hotelling and NPW observers are identical. Top right: Performance of the human observer for the same parameters as in the theoretical curves at the left. Bottom left: Performance for the Hotelling and NPW observers as a function of exposure time. Bottom right: Performance of the human observer for the same parameters as in the theoretical curves at the left. Note that, except for a constant shift, the human curves agree very well with those for the Hotelling observer and not at all with those for the NPW observer.

ties remained Gaussian. The other method simply summed up N randomly placed Gaussian blobs. For large N this method also yields a Gaussian grey-level density, but for small N it is decidedly non-Gaussian.

Yao performed a psychophysical study using both methods for generating the background, and with large and small N in the second method. The parameters of the background were adjusted so that all three approaches yielded backgrounds with exactly the same mean, variance and autocorrelation function, but with different higher-order statistics as indicated by the grey-level histograms. The result of the study was that the human observer had the same detection performance for objects located in the three backgrounds. Thus the higher-order statistics do not seem to play a role, at least for this task. This result lends further support for considering the Hotelling observer, who has knowledge of only first- and second-order statistics.

4.7. Estimation Tasks

The Hotelling and ideal observers are both based on the assumption that the task of the imaging system is either detection of some abnormality or classification of the objects into two or more classes (differential diagnosis). In nuclear medicine, on the other hand, the task is often to extract some quantitative information from the image. In the medical literature, this task is called quantitation, while in the statistics literature it is known as estimation. Important examples of quantities to be estimated in nuclear medicine include the cardiac ejection fraction or the concentration of some receptor-specific tracer in a region of the brain.

Smith and Barrett (1986) demonstrated that the Hotelling trace could be used to select an optimum coded aperture, and later these same authors (Smith and Barrett, 1988) showed that the performance of various aperture codes for a detection task, as measured by the Hotelling trace, correlated well with the performance on an estimation task as measured by a mean-square error. This prompted us to examine theoretically the relationship between these two apparently very different tasks. We were able to derive a quite general set of mathematical relations between performance metrics for detection and estimation (Barrett, 1990). For detection performance, we considered either the Hotelling trace or a non-prewhitening matched filter; as estimation metrics we considered either the ensemble mean-squared error or the relative variance in a region-of-interest estimate.

In each case we found that the detection metric can be rigorously written as the estimation metric times a product of four factors. The first factor just accounts for the lesion size and contrast as in the Rose model, so we call it the Rose factor. The second factor accounts for the bias in the estimator, the third factor accounts for the complexity of the scene (or, equivalently, the conspicuity of the lesion). while the final factor accounts for the effects of noise correlation. We call these last three factors the bias, conspicuity and correlation factors, respectively.

Taken together, these factors provide a comprehensive picture of how different characteristics of the imaging system or the object affect performance on different tasks. General matrix expressions for all four factors have been derived, and specific forms have been worked out for several radiographic imaging modalities. The details of this theory are found in Barrett (1990).

5. DISCUSSION

The experimental results presented above lend considerable credence to the use of the Hotelling observer to predict the performance of the human, at least for the purpose of evaluating and optimizing imaging systems. The psychophysical studies performed by Fiete, White, Rolland and Yao all show unequivocally that the human performance correlates well with that of the Hotelling observer for the tasks considered. It is hoped that the results obtained by Gooley on comparison of algorithms

will also fit this pattern, but computation of the Hotelling trace for these images is still in progress at this writing.

There is, however, one major study for which the Hotelling observer fails badly to predict human performance, and that is the work of Myers et al. on SKE/BKE detection in correlated noise. As noted above, the Hotelling and ideal observers are identical for this study, but the NPW observer is the one that predicts human performance. By contrast, in the Rolland study the Hotelling observer correctly predicted the effects of background inhomogeneity on human performance, while the NPW observer failed badly to do so.

One possible way to reconcile these apparently contradictory results is to include channels in the Hotelling model. We have already seen that this addition removes the discrepancy between the Hotelling (or ideal) model and the psychophysical results obtained by Myers. It might also be expected that a channelized Hotelling observer could also take proper account of background statistics and therefore correctly predict human performance in the Rolland study. Studies to examine this possibility are in progress.

Acknowledgements

The authors have benefitted greatly from discussions with many people, including Robert Wagner, Kyle Myers, Stephen Moore, Charles Byrne and John Denny. This work was supported by the National Cancer Institute under grants PO1 CA23417 and RO1 CA52643.

References

Barrett HH, Rolland JP, Wagner RF, and Myers KJ (1989). Detection of known signals in inhomogeneous, random backgrounds. Proc. SPIE, 1090:176-182.

Barrett HH (1990). Objective assessment of image quality: effect of object variability and quantum noise, J. Opt. Soc. Am. A. 7:1266-1278.

Cargill EB (1989). A mathematical liver model and its application to system optimization and texture analysis. Ph.D. Dissertation, University of Arizona.

Fiete RD, Barrett HH, Smith WE, and Myers KJ (1987). The Hotelling trace criterion and its correlation with human observer performance. J. Opt. Soc. Am. A, 4:945-953.

Fisher RA (1936). The use of multiple measurements in taxonomic problems. Ann. Eugenics 7 (part2):179-188.

Gooley TA (1990). Quantitative comparisons of statistical methods in image reconstruction. Ph. D. dissertation, University of Arizona.

Hotelling H (1931). The generalization of Student's ratio. Ann. Math. Stat. 2:360-378.

Myers KJ, Barrett HH, Borgstrom MC, Patton DD, and Seeley GW (1985). Effect of noise correlation on detectability of disk signals in medical imaging. J. Opt. Soc. Am. A, 2:1752-1759.

Myers KJ and Barrett HH (1987). Addition of a channel mechanism to the ideal-observer model. J. Opt. Soc. Am. A, 46:2447-2457.

Rolland JPY, (1990). Factors influencing lesion detection in medical imaging. Ph. D. dissertation, University of Arizona.

Smith WE and Barrett HH (1986). Hotelling trace criterion as a figure of merit for the optimization of imaging systems. J. Opt. Soc. Am. A, 3:717-725.

Smith WE and Barrett HH (1988). Linear estimation theory applied to the evaluation of *a priori* information and system optimization in coded-aperture imaging. J. Opt. Soc. Am. A,5:315-330.

White TA, Barrett HH, Cargill EB, Fiete RD, and Ker M (1989). The use of the Hotelling trace to optimize collimator performance. The Society of Nuclear Medicine 36th Annual Meeting, St. Louis, MO, (abstract).