Megalopinakophobia: Its Symptoms and Cures

Harrison H. Barrett[a,b,c,d], Kyle J. Myers[e], Brandon Gallas[a,b,d],
Eric Clarkson[a,c,d] and Hongbin Zhang[a,d,f]

[a]Department of Radiology, University of Arizona
[b]Program in Applied Mathematics, University of Arizona
[c]Optical Sciences Center, University of Arizona
[d]Center for Gamma-ray Imaging, University of Arizona
[e]Center for Devices and Radiological Health, Food and Drug Administration
[f]Department of Electrical and Computer Engineering, University of Arizona

ABSTRACT

This paper addresses issues in the calculation of a detectability measure for the ideal linear (Hotelling) observer performing a detection task on a digital radiograph. The main computational problem is that the inverse of a very large covariance matrix is required. The conventional approach is to assume some form of stationarity and argue that the matrix is diagonalized by discrete Fourier transformation, but there are many reasons why this assumption is unrealistic. After a brief review of the underlying mathematics, we present several practical algorithms for computing the detectability and some hints as to when each is applicable. The main conclusion is that large matrices should not be feared.

Keywords: Image quality, detectability, Hotelling, covariance matrix, digital radiology, NEQ, DQE

1. INTRODUCTION

Since the task in most radiological imaging procedures is the detection of a lesion, an important way of assessing image quality is in terms of some measure of detectability. Detectability for a human observer can be measured by psychophysical studies, but this approach is too time-consuming to use for system design or optimization. An alternative is to use the ideal observer, which sets an upper limit to the detectability achievable by any observer for a particular system. The ideal observer computes the likelihood ratio, which is the ratio of the probability density functions of the data with and without signal, but these densities may be unknown for complicated tasks with random signals or backgrounds. In these cases, a practical alternative is the ideal linear or Hotelling observer,[1–3] which requires knowledge of only the first- and second-order statistics of the data, not the full probability densities. The performance of this observer is specified by a signal-to-noise ratio (SNR), sometimes known as the Hotelling trace.

One difficulty in computing the SNR for the ideal linear observer is that it involves the inverse of a large covariance matrix (or autocovariance operator if we use continuous variables). Since brute-force inversion is precluded in practice, the common approach is to assume stationary noise, for which the covariance can be diagonalized by a Fourier transformation. The diagonal elements are commonly referred to as the noise power spectrum (NPS). If the covariance has been diagonalized, then inversion is simply a matter of taking the reciprocal of the diagonal elements, and the detectability for the ideal linear observer can be computed readily.

If we further assume shift invariance, then Fourier transformation can be used to simplify the signal part of the SNR as well as the noise part. With shift-invariance, the signal in the data can be expressed as the signal in the object convolved with a point spread function (PSF), and after Fourier transformation the convolution becomes a product. Thus the Fourier-domain signal in the data is the Fourier-domain signal in the object times the system transfer function (TF). When normalized to its value at zero spatial frequency, the TF is called the optical transfer function or OTF, and the modulus of this complex quantity is called the modulation transfer function or MTF.

Further author information: (Send correspondence to H.H.B.)
H.H.B.: Radiology Research Laboratory, Arizona Health Sciences Center, University of Arizona, Tucson, AZ 85724,
E-mail: barrett@radiology.arizona.edu

Use of both assumptions – stationarity of the noise and shift-invariance of the deterministic properties of the system – leads to the identification of a quantity known as NEQ (noise equivalent quanta), which is just the ratio of MTF^2 to NPS. The reciprocal of the NPS appears here as a surrogate for the inverse covariance.

A normalized version of NEQ, called DQE (detective quantum efficiency), can be defined as a putative measure of performance of the imaging system, independent of the particular signal to be detected. DQE was a popular metric for film-screen systems, and it is becoming a *de facto* standard for digital radiography. Unfortunately, the assumptions underlying the derivation of DQE are very questionable for film-screen systems, and several new unjustifiable steps are required to extend the concept to digital radiography.

The first problem is that the noise is never stationary for interesting radiographic images. An interesting image, virtually by definition, is one where the x-ray fluence conveys interesting information and hence varies spatially. The x rays constitute a random point process, and the mean and autocovariance function of this process are functions of position on the detector. For a stationary random process, by contrast, the mean would have to be constant, and the autocovariance function could depend only on the relative location of the two points to which it refers, not the absolute position. Thus the very fact that we are imaging an actual object, as opposed to just a uniform x-ray beam, immediately invalidates stationarity. Moreover, the simple fact that radiographs have finite size also spoils strict stationarity; the autocovariance between two image points cannot be invariant when one or both of the points falls off the detector.

A more subtle reason for departure from stationarity is that objects are random. If we consider repeated images of the same object, then the x-ray patterns on the detector are sample functions of a Poisson process, but if we consider images of many different objects the Poisson character is lost, and the image statistics depend on the object statistics. We can avoid this issue by considering a BKE (background known exactly) problem, such as detection of a small lesion in an otherwise uniform beam, but in real clinical applications the random anatomical background may be more important than Poisson noise. Systems designed for BKE detection may perform poorly with realistic random backgrounds.[4,5] To use the SNR measure in these situations, therefore, the covariance should account for anatomical noise as well as x-ray noise, and there is no reason at all to think of anatomy as stationary.

The stationarity assumption runs into several additional difficulties for digital radiography. In this case the sampling of the x-ray beam by an array of detector elements converts the autocovariance function to a covariance matrix, and we need to assume that this matrix is stationary in some sense. One sense of stationarity in a discrete image is that the covariance matrix is *Toeplitz*, so that the covariance between the signals from two detector elements depends on their relative position in the array, not on absolute location. For an ideal detector that simply samples a continuous random process, the covariance matrix is Toeplitz if the random process is stationary, in the sense that the autocovariance function depends only on relative position. If we ignore the anatomical noise and consider only the Poisson noise of the x rays, then the covariance matrix is diagonal, and if we further assume that the x-ray fluence does not depend on position across the array, then the covariance matrix is a multiple of the unit matrix, hence Toeplitz.

Real digital x-ray detectors involve a gain mechanism where each x-ray photon is converted to many secondary optical photons or charge carriers, which are then detected. This gain process is random and hence a source of noise. Moreover, there is an electronic noise in the amplifiers associated with each detector element, and additional noise arises from production and possible escape of Compton-scattered or K x rays. Now there are altogether five contributions to the image statistics – Poisson noise in the x-ray beam, object randomness, the random gain mechanism, electronic noise and noise arising from K x rays and Compton scatter. All five of these effects must be stationary for the overall system to be stationary.

The Poisson and electronic contributions to the covariance are diagonal, so they are Toeplitz only if they are multiples of the unit matrix. That means that not only must the x-ray fluence be constant, but also the electronic noise variance must be the same in every element. The random gain mechanism and Compton and K x rays introduce short-range correlations between neighboring detector elements, and these effects must also be be invariant to absolute location for stationarity to hold. Again, this can happen only if the x-ray fluence is spatially uniform, but it also requires that the detector itself be uniform from element to element, which real detectors never are. Finally, as above, the inevitable departures from stationarity in the anatomy induce a departure from Toeplitz character in this term of the covariance.

The Toeplitz assumption, questionable though it may be, is still not strong enough for Fourier methods to be useful. To diagonalize a matrix by a discrete Fourier transform (DFT), we must assume that it is *circulant*, not

Toeplitz. For a circulant covariance matrix, two elements at opposite sides of the array have the same covariance as two adjacent elements in the center. This "digital wrap-around" has no physical basis, but it is what we must assume if we want to diagonalize the matrix with DFTs. Fourier aficionados skirt the Toeplitz-vs.-circulant issue by arguing that the correlations are short range, but this statement is not true for anatomical noise. Moreover, even if one considers only BKE problems without anatomical noise and does a simulation in which all of the other requisite assumptions are valid, substantial errors can still be made in the computation of SNR by approximating a Toeplitz matrix by a circulant one.[6]

It appears that the willingness – even eagerness – to embrace these questionable assumptions in digital radiography traces back to an inability to deal with the matrix inversion required in a rigorous calculation of SNR. The covariance matrix is enormous, an attempt to compute its inverse crashes the computer, so let's assume stationarity and do a Fourier transformation. To describe this syndrome, we have coined the term *megalopinakophobia*, from the Greek *pinakas*, matrix. It is the central tenet of this paper that megalopinakophobia is a curable disease.

2. MATHEMATICAL BACKGROUND

An $M \times M$ digital image can regarded as a vector $\mathbf{g}$ with M^2 random components. Most analyses reorder these components so that $\mathbf{g}$ can be regarded as an $M^2 \times 1$ column vector with a single scalar index j denoting position in the image, but we find it convenient to use a 2D vector index or multi-index $\mathbf{m}$ with components m_x and m_y, each taking on integer values in the range 1 to M. Thus a component of $\mathbf{g}$ is denoted $g_{\mathbf{m}}$.

In a detection task, we assume that $\mathbf{g}$ is drawn from one of two ensembles, one containing some sort of abnormality that we call a signal, and the other containing no signal. A linear observer is defined as one that performs this task by computing a linear discriminant or test statistic of the form:

$$\lambda = \mathbf{w}^t \mathbf{g} \,, \tag{1}$$

where $\mathbf{w}$ is another vector of the same size as $\mathbf{g}$, and superscript t denotes transpose. Thus $\mathbf{w}^t\mathbf{g}$ is the scalar product of $\mathbf{w}$ and $\mathbf{g}$, and the linear observer computes this scalar product and compares it to a threshold in order to decide whether the signal is present. We refer to $\mathbf{w}$ as the observer template.

The ideal linear observer uses a template given by

$$\mathbf{w} = \mathbf{K}^{-1}\mathbf{s} \,, \tag{2}$$

where $\mathbf{s}$ is the signal to be detected (or the mean signal if we are considering random signals), and $\mathbf{K}$ is the overall covariance matrix of the data.

The performance of the ideal linear observer is specified by the SNR, defined as

$$\mathrm{SNR}^2 = \mathbf{s}^t\mathbf{K}^{-1}\mathbf{s} = \mathbf{w}^t\mathbf{s} \,. \tag{3}$$

The SNR defined this way has several useful interpretations. For SKE/BKE (signal known exactly, background known exactly) tasks in Gaussian noise, the true ideal observer performs only linear operations on the data, so the ideal and Hotelling observers are identical, and SNR specifies the maximum achievable detectability. The area under the ROC (receiver operating characteristic) curve for the ideal observer can be computed directly from SNR in this case. This area, denoted AUC, is an alternative metric for detectability.

For SKE/BKE tasks in Poisson noise, the ideal observer is not precisely the same as the Hotelling observer, but in practice the difference is very small. In this case also, therefore, SNR as defined above is an excellent predictor of ideal-observer performance.

For random signals or backgrounds, SNR is not necessarily a predictor of ideal performance, but it still sets an upper limit to the performance of any linear observer, so it is a useful metric for design of imaging systems. Moreover, even if the noise is decidedly non-Gaussian (for example, anatomical noise), SNR is still an excellent predictor of the AUC for the Hotelling observer. Since linear test statistics are linear combinations of random components of $\mathbf{g}$ they tend to follow a univariate normal distribution as a result of the central limit theorem; for any univariate normal discriminant, AUC can be predicted from SNR.

Humans may be essentially linear observers, and there are many situations where the human performance can be predicted reasonably well by this SNR, even with random signals and backgrounds.[2] Human performance falls well below ideal in some kinds of high-pass noise,[7] but for low-pass noise of the kind found in digital radiography, SNR is a reasonable figure of merit for human observer performance.

3. COMPONENTS OF THE COVARIANCE

As noted in the introduction, there are five main contributions to the noise in digital radiography, and we can write $\mathbf{K}$ as

$$\mathbf{K} = \mathbf{K}^{(elec)} + \mathbf{K}^{(x)} + \mathbf{K}^{(gain)} + \mathbf{K}^{(Kx)} + \mathbf{K}^{(obj)} \,, \tag{4}$$

where the terms represent, respectively, the electronic noise; the Poisson statistics of the x rays as reflected through the gain mechanism; the excess noise of the gain mechanism; the effect of reabsorbed Compton-scattered and K x rays, and the effect of object randomness. For a detailed discussion of each term, see Barrett and Myers[8]; only a brief account will be given here.

For most readout circuits the electronic noise for different detector elements is uncorrelated, and we can write

$$[\mathbf{K}^{(elec)}]_{\mathbf{mm}'} = \sigma^2_{\mathbf{m}} \delta_{\mathbf{mm}'} \,. \tag{5}$$

This form is independent of the x-ray exposure, the object and the gain mechanism in the detector.

On the other hand, the second term, $\mathbf{K}^{(x)}$, does depend on the x-ray exposure, the object and the gains. It can be expressed as

$$[\mathbf{K}^{(x)}]_{\mathbf{mm}'} = \Gamma_{\mathbf{m}} \, \overline{g}_{\mathbf{m}} \, \delta_{\mathbf{mm}'} \,, \tag{6}$$

where $\overline{g}_{\mathbf{m}}$ is the mean number of x-ray photons absorbed in the detector element, and $\Gamma_{\mathbf{m}}$ is the overall gain (including gain in the detector and electronic gain) relating this number to output signal.

The covariance $\mathbf{K}^{(gain)}$ can be thought of as the correlated part of the random gain process, but in fact it may also be a diagonal matrix, or nearly so. In a scintillator-photodiode detector, for example, the random process describing optical photons on the photodiode plane has a correlation, but the range of this correlation is approximately the range over which the optical photons spread in propagating to the output plane, which in turn is approximately the detector thickness. This range might be small compared to the size ϵ of a photodiode, and if so, then $\mathbf{K}^{(gain)}$ will be diagonal even though the optical random process is not delta-correlated. If the spread of the optical photons is comparable to ϵ, then a correlation will be induced in $\mathbf{g}$ between adjacent photodiodes but not between more distant ones.

The situation is slightly more complicated in semiconductor detectors where the charge carriers not only spread out as they propagate to the electrode plane, but they can also be trapped *en route*. Trapped carriers can induce correlated charges on neighboring electrodes, but there is no significant correlation from this effect if the electrodes are separated by several times the detector thickness.

Our use of the multi-index $\mathbf{m}$ to denote the detector elements makes it very simple to characterize the short-range correlations associated with $\mathbf{K}^{(gain)}$. We can say that

$$[\mathbf{K}^{(gain)}]_{\mathbf{mm}'} \simeq 0 \quad \text{if} \quad \epsilon \, |\mathbf{m} - \mathbf{m}'| > \delta_{gain} \,, \tag{7}$$

where δ_{gain} is the correlation length of the secondary process. If we had not used multi-indices, it would have been more complicated to state which elements of $\mathbf{K}^{(gain)}$ were zero; with multi-indices, we can think of this covariance matrix as confined to a narrow band around the diagonal, and we shall refer to it as a *banded matrix*.

The term $\mathbf{K}^{(Kx)}$ is the covariance associated with reabsorption of K x rays or Compton-scattered photons. If they don't escape from the detector material, the secondary x rays have a pathlength of order $1/\alpha_{tot}$, where α_{tot} is the total attenuation coefficient for the secondary photons in the detector material, and twice this pathlength is a reasonable estimate for the correlation range for this term. We can thus say, roughly, that

$$[\mathbf{K}^{(Kx)}]_{\mathbf{mm}'} \simeq 0 \quad \text{if} \quad \epsilon \, \alpha_{tot} \, |\mathbf{m} - \mathbf{m}'| > 2 \,. \tag{8}$$

Often this condition will lead to the conclusion that $\mathbf{K}^{(Kx)}$ is confined to a band one or two elements wide around the diagonal (in the multi-index notation).

Finally, the covariance $\mathbf{K}^{(obj)}$ arises because a random object creates a random x-ray fluence, which is then transformed through the amplification process to the output data. A key point about this term is that it varies quadratically with the x-ray exposure, so at large exposures it will be the dominant noise contribution. The range of the correlations associated with this term might be quite large since object structures can have large scales.

4. DETERMINATION OF THE COVARIANCE MATRIX

Before we can address computation of SNR, we must discuss how to determine the covariance matrix in the first place. Possible methods include model-based theoretical calculation, theory augmented by measurement, Monte Carlo simulation, and collection of sample images.

Analytic expressions for all five terms in Eq. (4) are given by Barrett and Myers.[8] To evaluate the expressions for $\mathbf{K}^{(x)}$ and $\mathbf{K}^{(gain)}$, we need to know the first two moments of the random number of secondaries as well as how the secondaries are distributed on the readout plane of the detector. These unknowns can be determined from theoretical models or Monte Carlo simulation, but simple measurements with tightly collimated x-ray beams can be used also.

Similarly, $\mathbf{K}^{(elec)}$ can be computed theoretically from knowledge of the circuit design and standard electronic simulation software. If we know that this component matrix is diagonal, as in Eq. (5), then the only unknowns are the variances, and we can also get those by direct measurement with no x-ray beam.

The contribution $\mathbf{K}^{(Kx)}$ is probably best determined by Monte Carlo simulation. Alternatively, the combination $\mathbf{K}^{(elec)} + \mathbf{K}^{(x)} + \mathbf{K}^{(gain)} + \mathbf{K}^{(Kx)}$ can be measured directly by using a uniform x-ray fluence and acquiring a large number of image frames. All of these terms are diagonal or banded around the diagonal, so only a very small subset of all possible elements in the covariance matrix needs to be computed; elements far from the diagonal are zero *a priori* by Eqs. (5) – (8) and do not need to be computed or measured. Once the diagonal and near-diagonal terms are established, we have a a full-rank estimate of the sum of the first four terms in Eq. (4). We denote that estimate as $\hat{\mathbf{K}}^{(noise)}$.

The most difficult term is $\mathbf{K}^{(obj)}$, which is simply neglected in many detectability studies. Any covariance matrix can be estimated from samples, but if we want the resulting sample covariance matrix to represent $\mathbf{K}^{(obj)}$, the other noise sources must be negligible. To achieve this goal, we can take advantage of the fact that $\mathbf{K}^{(obj)}$ is the only term that varies quadratically with the x-ray exposure. If clinical (or animal or cadaver) images can be taken at high enough exposure, the resulting sample covariance matrix on the images is directly an estimate of $\mathbf{K}^{(obj)}$. Alternatively, several images of each object can be taken at lower exposures, and the sample covariance matrix can be analyzed element by element to tease out $\mathbf{K}^{(obj)}$ separately. Another useful approach is to simulate realistic objects and then simulate their images without the other noise sources being present.

Suppose we have a set of low-noise or noise-free (simulated) sample images $\{\mathbf{g}_j, j = 1, ..., J\}$. We can subtract the sample mean from each image to form the set $\{\delta\mathbf{g}_j, j = 1, ..., J\}$, and the covariance matrix $\mathbf{K}^{(obj)}$ can be estimated by

$$\hat{\mathbf{K}}^{(obj)} = \mathbf{W}\mathbf{W}^t \,, \tag{9}$$

where $\mathbf{W}$ is the $M^2 \times J$ matrix with columns equal to sample images:

$$\mathbf{W} = \frac{1}{\sqrt{J}} \left[\delta\mathbf{g}_1, \, \delta\mathbf{g}_2, \, ..., \delta\mathbf{g}_J \right] . \tag{10}$$

Note that this sample covariance matrix is not sparse or banded, and its rank will be $J - 1$ for independent samples.

Given this estimate of $\mathbf{K}^{(obj)}$, we can write an estimate of the overall covariance matrix as

$$\hat{\mathbf{K}} = \hat{\mathbf{K}}^{(noise)} + \hat{\mathbf{K}}^{(obj)} . \tag{11}$$

Even though $\hat{\mathbf{K}}^{(obj)}$ is singular, the overall $\hat{\mathbf{K}}$ defined in this way will be nonsingular if the first term is nonsingular.

5. COMPUTATION OF THE DETECTABILITY

Given the estimate of the covariance matrix, what remains is to compute $\mathbf{s}^t\hat{\mathbf{K}}^{-1}\mathbf{s}$. Direct inversion loses its appeal when one contemplates the size of $\hat{\mathbf{K}}$; for a $1,000 \times 1,000$ detector, $\hat{\mathbf{K}}$ is $1,000,000 \times 1,000,000$! We shall present several approaches to estimating SNR^2 without actually doing such a huge inverse. None will assume stationarity in any sense, and all will take advantage of features specific to digital radiography. Most of the discussion will be for SKE tasks, but we shall show briefly how the formalism can be modified for random signals.

5.1. Iterative computation of the template

For an SKE problem, we can attempt to calculate the Hotelling template $\mathbf{w}$ rather than computing the SNR directly; the SNR can then be found by taking a scalar product as in Eq. (3). Finding the template amounts to solving the equation $\hat{\mathbf{K}}\mathbf{w} = \mathbf{s}$. This is a standard image-reconstruction problem, in which $\hat{\mathbf{K}}$ plays the role of the system matrix, $\mathbf{s}$ is the known data, and $\mathbf{w}$ is the unknown object. Unlike many image-reconstruction problems, however, the system matrix is square in this case, and in fact it is invertible as discussed above.

Thus we can be sure that $\hat{\mathbf{K}}\mathbf{w} = \mathbf{s}$ has a solution, and it can be found by means of the well-known Landweber algorithm, in which successive estimates of the template are formed according to

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \alpha[\hat{\mathbf{K}}^{(noise)}]^{-1}\left[\mathbf{s} - \hat{\mathbf{K}}\mathbf{w}_n\right] . \tag{12}$$

If this algorithm converges, it yields $\mathbf{s} = \hat{\mathbf{K}}\mathbf{w}_n$ as desired. For a discussion of conditions for convergence, see Barrett and Myers.[8]

5.2. The Neumann series

Another iterative algorithm is based on the Neumann series for the inverse of a matrix. Though formally identical to the Landweber algorithm, the Neumann algorithm can be much more efficient when the covariance matrix is nearly diagonal, as it is when we do not consider object variability.

To see why the near-diagonal character of $\hat{\mathbf{K}}$ is useful, suppose initially that

$$\hat{\mathbf{K}} = \sigma^2\mathbf{I} + \mathbf{A} = \sigma^2\left[\mathbf{I} + \frac{1}{\sigma^2}\mathbf{A}\right] , \tag{13}$$

where $\mathbf{A}$ describes the off-diagonal elements. Then we can use the Neumann series to write the inverse covariance as

$$\hat{\mathbf{K}}^{-1} = \frac{1}{\sigma^2}\sum_{j=0}^{\infty}\left[-\frac{1}{\sigma^2}\mathbf{A}\right]^j = \frac{1}{\sigma^2}\mathbf{I} - \frac{1}{\sigma^4}\mathbf{A} + \frac{1}{\sigma^6}\mathbf{A}^2 + ... \tag{14}$$

The Hotelling SNR then becomes

$$\mathrm{SNR}^2 = \mathbf{s}^t\hat{\mathbf{K}}^{-1}\mathbf{s} = \frac{||\mathbf{s}||^2}{\sigma^2} - \frac{\mathbf{s}^t\mathbf{A}\mathbf{s}}{\sigma^4} + \frac{\mathbf{s}^t\mathbf{A}^2\mathbf{s}}{\sigma^6} + ... \tag{15}$$

Formally, the Neumann series will converge if $||\mathbf{A}||/\sigma^2 < 1$, but that requirement is too stringent for our purposes since it takes no account of the nature of the signal. By the ratio test, the series will converge if

$$\frac{\mathbf{s}^t\mathbf{A}^{n+1}\mathbf{s}}{\sigma^2\mathbf{s}^t\mathbf{A}^n\mathbf{s}} < 1 , \tag{16}$$

for all n, and it may still converge (because of the alternating signs) even if Eq. (16) is violated. In practice, convergence will be rapid if the correlations are weak and short-range and the signal is spatially compact.

More generally, we can always decompose $\hat{\mathbf{K}}$ into a diagonal part $\mathbf{D}$ plus a matrix $\mathbf{A}$ with only off-diagonal terms. Assuming convergence, we then have

$$\hat{\mathbf{K}} = \mathbf{D} + \mathbf{A} = \mathbf{D}\left[\mathbf{I} + \mathbf{D}^{-1}\mathbf{A}\right] ; \tag{17}$$

$$\hat{\mathbf{K}}^{-1} = \left[\sum_{j=0}^{\infty}\left[-\mathbf{D}^{-1}\mathbf{A}\right]^j\right]\mathbf{D}^{-1} ; \tag{18}$$

$$\mathrm{SNR}^2 = \mathbf{s}^t\hat{\mathbf{K}}^{-1}\mathbf{s} = \mathbf{s}^t\mathbf{D}^{-1}\mathbf{s} - \mathbf{s}^t\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{s} + \mathbf{s}^t\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{s} + ... \tag{19}$$

The first term in this expansion, $\mathbf{s}^t\mathbf{D}^{-1}\mathbf{s}$, is what we would get if there were no off-diagonal terms, and the remaining terms in Eq. (19) are the corrections arising from correlations induced by the detector. If these correlations are sufficiently weak, we may be able to truncate the series after a few terms, making the calculation of SNR very easy.

The banded character of the covariance is especially useful if we are trying to detect a spatially compact signal. At the extreme, suppose $\mathbf{s}$ is confined to a single detector element, say $\mathbf{m} = \mathbf{n}$. Then SNR^2 is simply $s_{\mathbf{n}}^2\,[\hat{\mathbf{K}}^{-1}]_{\mathbf{nn}}$, and the first correction term in Eq. (19) becomes

$$\mathbf{s}^t\,\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{s} = \sum_{\mathbf{j}}\sum_{\mathbf{k}} s_{\mathbf{n}}\,[\mathbf{D}^{-1}]_{\mathbf{nj}}\,[\mathbf{A}]_{\mathbf{jk}}\,[\mathbf{D}^{-1}]_{\mathbf{kn}}\,s_{\mathbf{n}} = \frac{s_{\mathbf{n}}^2}{D_{\mathbf{nn}}^2} A_{\mathbf{nn}} = 0\,. \tag{20}$$

since the diagonal elements of $\mathbf{A}$ are zero by definition.

The next term in the series is also simplified if we consider a signal confined to a single pixel:

$$\mathbf{s}^t\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{s} = \frac{s_{\mathbf{n}}^2}{D_{\mathbf{nn}}^2}[\mathbf{A}\mathbf{D}^{-1}\mathbf{A}]_{\mathbf{nn}} = \frac{s_{\mathbf{n}}^2}{D_{\mathbf{nn}}^2}\sum_{\mathbf{k}}\frac{[A_{\mathbf{nk}}]^2}{D_{\mathbf{kk}}}\,. \tag{21}$$

If we say that $\mathbf{A}_{\mathbf{nk}} \simeq 0$ when $|\mathbf{n}-\mathbf{k}|\epsilon > \delta$, then the number of terms we have to sum is of order $[\delta/\epsilon]^2$, which could be quite small. Moreover, if the elements of $\mathbf{A}$ are small compared to $\mathbf{D}_{\mathbf{nn}}$, then the correction terms are small and the series converges rapidly.

If the signal covers P pixels, the number of computations required is increased by a factor of P^2, and a convergence condition analogous to Eq. (16) must be satisfied.

5.3. The matrix-inversion lemma

Suppose we want to invert an overall covariance matrix of the form

$$\hat{\mathbf{K}} = \mathbf{B} + \mathbf{W}\mathbf{W}^t\,, \tag{22}$$

where $\mathbf{B}$ is a banded matrix representing all terms except $\hat{\mathbf{K}}^{(obj)}$ in the overall covariance. By the Woodbury matrix-inversion lemma,

$$\left[\mathbf{B} + \mathbf{W}\mathbf{W}^t\right]^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{W}\left[\mathbf{I} + \mathbf{W}^t\mathbf{B}^{-1}\mathbf{W}\right]^{-1}\mathbf{W}^t\mathbf{B}^{-1}\,. \tag{23}$$

The advantage of this form is that $[\mathbf{I} + \mathbf{W}^t\mathbf{B}^{-1}\mathbf{W}]$ is a $J \times J$ matrix, where J is a few hundred in practice, rather than $M^2 \times M^2$, where M^2 may be 10^6. Moreover, since $\mathbf{W}^t\mathbf{B}^{-1}\mathbf{W}$ is positive semidefinite, the inverse of the $J \times J$ matrix will always exist. Thus, if $\mathbf{B}$ can be inverted, either trivially because it is diagonal or by use of a rapidly convergent Neumann series, then it becomes feasible to add the sample covariance representing object variability.

5.4. Dimensionality reduction

The Landweber and matrix-inversion methods depend on writing $\hat{\mathbf{K}}$ as a full-rank, near-diagonal term plus a low-rank sample covariance matrix. The full rank of the first part arose because we assumed we had substantial prior knowledge of the form of the noise statistics. Sometimes we may have available only noisy sample images and may not want to postulate a full-rank term. That puts us in the standard dilemma of pattern recognition: the number of training images is necessarily much less than the dimensionality ($J << M^2$), and the sample covariance is far from invertible. The standard solution in pattern recognition is to reduce the dimensionality of the problem drastically by extracting features from the data. The problem that then arises is that there is no way to know how close the performance with these features comes to the performance with the optimal feature, the likelihood ratio.

We recently proposed[9] a solution to this problem specifically for the case of SKE detection. Again the trick is to use prior knowledge, this time about the form of the template. For SKE detection, we can assume that $\mathbf{w}$ is peaked around the signal location and that it is smoothly varying. Moreover, if we assume that the signal is rotationally symmetric and that there is no preferred direction for the background correlations, it is reasonable to look for a template that is rotationally symmetric, except of course that it must be sampled on the pixel grid.

To enforce this prior information, we suggested expanding $\mathbf{w}$ (before sampling) as a series of Laguerre-Gauss functions. The scalar product of the image $\mathbf{g}$ with a single sampled Laguerre-Gauss function was then the feature. The dimensionality of the feature space was the number of terms in the expansion, and a non-singular sample covariance matrix could be obtained if the number of sample images exceeded this dimensionality.

Gallas[10] has carried out an extensive validation of this method for SKE detection in lumpy backgrounds with widely varying statistical properties. He found excellent agreement with the true Hotelling SNR in every case with just 5-10 terms. Moreover, the estimated SNR was very robust to choice of the number of terms and to choice of the width of the Gaussian envelope of the Laguerre-Gauss functions.

5.5. Random signals

The Hotelling SNR applies to the case of random signals if we just replace $\mathbf{s}$ with its average $\bar{\mathbf{s}}$, but it may be a poor indicator of system performance with large signal variability. For example, if the lesion can be located anywhere within a wide field of view, the signal averaged over location is a broad, structureless function, and the detectability of the Hotelling observer, or any linear observer, becomes very small. The ideal observer is nonlinear and not well approximated by the Hotelling observer in this case.

One way around this problem is to replace the original two-alternative detection problem with an $(L+1)$-alternative problem where the signal can be at one of L non-overlapping locations.The simple detection decision can then be made by choosing the location for which the response of the Hotelling observer is maximum, but we also get information on lesion location this way. Another possibility is to stick with the SKE problem but to allow signal location to be a parameter in the SNR. Then we can compute a detectability map, which is a plot of SNR vs. position of the lesion.

Variations in lesion size can be treated similarly. We can use the average lesion as $\bar{\mathbf{s}}$ in the Hotelling SNR, or we can compute the SKE SNR as a function of lesion size.

6. CONCLUSIONS

We have presented several different approaches to computation of SNR without unrealistic stationarity assumptions. All of them depend to some degree on prior knowledge. A key step in several methods is to decompose the covariance matrix into two terms, a term representing various noise processes in the detector plus a term with long-range correlations due to object variability. Since we know *a priori* that the first term is diagonal or near diagonal, we do not need to compute or estimate the elements far from the diagonal, and as a result we get a full-rank approximation to this term. That makes the overall covariance full-rank even if the object-variability term is a sample covariance matrix.

The assumption that the noise covariance is nearly diagonal is directly analogous to the assumption made with Fourier methods – it is just more defensible physically. In using a DFT, the implicit assumption is that the noise covariance is diagonalized, so there is no need to compute the off-diagonal elements. The assumption is invalidated if the noise is not circulant-stationary, and serious errors in the estimated SNR can then result. Similarly, if there should be long-range correlations in x-ray detectors (apart from those associated with object variability), then some of the methods proposed in this paper could also give errors. The only physical mechanism the authors can think of that might lead to such errors is veiling glare in image intensifiers.

We have also briefly described one method – the Laguerre-Gauss expansion – that is not based on dividing the covariance into a near-diagonal and long-range part. In that case the prior information is that we know the object location, that the template is smooth and peaked at this location, and that the noise statistics have no preferred direction.

More experience with each of these methods, both in our laboratories and in the digital-radiology community at large, would be highly desirable, but from what we have done so far, the techniques appear to be robust and to give results that do not differ much from method to method. We have, however, found substantial discrepancies between SNR and Fourier-based methods. Much further work is required, but we urge the community to use caution before adopting Fourier methods as a standard.

ACKNOWLEDGMENTS

We wish to express our thanks to Robert Wagner, Robert Gagne and Angel Pineda for stimulating discussions on the subject of this paper. We refrain from comment on their susceptibility to megalopinakophobia, but we thank Georgios Kastis for his assistance with Greek etymology.

This work was made possible by funding from NIH grants RO1 CA52643 and P41 RR14304.

REFERENCES

1. H. H. Barrett, "Objective assessment of image quality - effects of quantum noise and object variability," *J. Opt. Soc. Am. A.* **7**, 7, pp.1266-1278, 1990.
2. R. D. Fiete, H. H. Barrett, W. E. Smith and K. J. Myers, "The Hotelling trace criterion and its correlation with human observer performance," *J. Opt. Soc. Am. A.* **4**, pp. 945-953, 1987.
3. W. E. Smith and H. H. Barrett, "Hotelling trace criterion as a figure of merit for the optimization of imaging systems", J. Opt. Soc. Am. A, 3:717-725, 1986.
4. K. J. Myers, J. P. Rolland, H. H. Barrett and R. F. Wagner, "Aperture optimization for emission imaging: effect of a spatially varying background," *J. Opt. Soc. A.* **7**, pp. 1279-1293, 1990.
5. J. P. Rolland and H. H. Barrett, "Effect of random background inhomogeneity on observer detection performance", J. Opt. Soc. Am. A., 9(5):649-658, 1992.
6. A. Pineda and H. H. Barrett, "What does DQE say about detectability in digital radiography?", Proc. SPIE 4320, 2001.
7. K. J. Myers and H. H. Barrett "Addition of a channel mechanism to the ideal-observer model", J. Opt. Soc. Am. A., 4:2447-2457, 1987.
8. H. H. Barrett and K. J. Myers, "Foundations of Image Science", John Wiley, New York, to be published 2001.
9. H. H. Barrett, C. K. Abbey, B. Gallas and M. Eckstein, "Stabilized estimates of Hotelling-observer detection performance in patient-structured noise", Proc. SPIE 3340:27-43, 1998.
10. Brandon Gallas, to be published.