14

Image Quality

In this chapter we consider the many practical issues one must wrestle with in the objective evaluation of imaging systems. Unlike Chap. 13, where knowledge of the relevant population statistics of the image classes is assumed, the emphasis here is on the practical issues that come to the fore when only a finite sample of images is available for determining the image statistics or the observer's performance, or both.

We begin in Sec. 14.1 with a description of various approaches to the assessment of image quality, including methods based on preference assessments, fidelity measures, and information-theoretic approaches. Then, in Sec. 14.1.5, we introduce the key elements that are required for the approach we advocate: the method must be objective, task-based, and account for the statistical properties of the relevant images and observers.

Properties of the human visual system and the determination of classification performance by human observers is the subject of Sec. 14.2, including the conduct of psychophysical experiments and the estimation of summary statistics for human performance. In Sec. 14.3 we turn to the subject of model or algorithmic observers for classification and estimation tasks. The approaches presented in Secs. 14.2 and 14.3 may make use of actual data sets derived from real imaging systems or, more often in research investigations, simulated images. Methods for image simulation are discussed in Sec. 14.4. As emphasized in that section, accurate models of the properties of the object and the physics of the image acquisition system are required if simulated images are to lead to accurate assessments of system performance.

14.1 SURVEY OF APPROACHES

14.1.1 Subjective assessment

The simplest approach to the assessment of image quality is to rely on a viewer's subjective assessment regarding how good an image looks. This approach can be as crass as the presentation of just a single pair of images, one processed by algorithm A and the other processed by contender B, with the developer of algorithm A drawing sweeping conclusions regarding the merits of A over B. A panel of experts might be used to make a stronger case regarding the merits of one algorithm over another, but here again the panel's decision is based on subjective preference rather than objective, task-based performance. There may be a place for beauty contests in the evaluation of imagery, such as when an individual selects a home-entertainment video system. We would argue that even then, most buyers base their subjective preference of one system over others by viewing a range of images; buyers usually take into account technical data across competing systems as well.

In an effort toward putting subjective preference methods on more solid footing, Zetzsche and Hauske (1989) developed a model based on the visual system with the goal of predicting subjective ratings of image quality. If this goal were met, the authors reasoned that they could determine image quality without the need for building physical prototypes of display devices. The predictions of the model were found to have correlations with mean subjective ratings ranging from 0.74 to 0.95 for images in which various artifacts were present.

Methods based on multidimensional scaling (MDS) have been applied to the analysis of subjective image quality ratings (Ahumada and Null, 1993). MDS methods incorporate various approaches for collecting numerical rating from multiple observers given the task of rating the quality of a set of images. Images can be presented in pairs, with the observer given the task of selecting the one with higher quality, or a set of images can be rank-ordered by quality. Normalizations can be done to account for differences in how observers scale the rating values; Thurstone scaling is a procedure that allows observers to use a rating scale nonlinearly (Torgerson, 1958). Once the rating data are in hand, MDS enables the dimensions of image quality to be extracted (Farrell *et al.*, 1991). Standard software packages are available for performing MDS. The difficulty with the MDS approach is that the labeling of the extracted dimensions, in terms of physical characteristics of the images or the image acquisition system, is left to the investigator (Shepard *et al.*, 1972). Moreover, the connection between an observer's rating of the quality of an image and the usefulness of the image for a specified task is never made.

Structured preference assessments formalize the subjective approach through the use of trained observers who perform a prescribed set of analyses. The well-known National Imagery Interpretability Rating Scale (NIIRS) system, which uses an interpretability rating scale for analyzing military reconnaissance images, is an example of a structured-preference approach. The NIIRS system was developed under the leadership of the U.S. Imagery Resolution Assessments and Reporting Standards (IRARS) Committee in the early 1970s. The first NIIRS system evaluated the visibility of military objects in images acquired in the visible spectrum. Later, the NIIRS system was extended to incorporate objects like buildings, roads, railroads and bridges, enabling the evaluation of images without military objects.

The NIIRS system is now able to handle data outside the visible spectrum, including thermal, radar, and multispectral imagery.

Models have been developed for predicting NIIRS ratings just as models have been developed for predicting subjective preference ratings. Given a set of input variables that can include the scene contrast, scene illumination, and imaging system characteristics, the models generate measures of image quality that can be related to the NIIRS scale. Another approach to the estimation of an image quality metric that correlates with the NIIRS scale is based on the power spectrum of the image to be rated, indicating that the measure is heavily influenced by the noise properties of the image.

The NIIRS approach is almost exclusively used for military applications; NIIRS refers to the value of an image for "intelligence purposes," rather than image quality *per se*. Furthermore, the NIIRS approach is not amenable to the analysis of the variation in true- and false-positive fractions [TPF and FPF, defined in (13.11)] of image interpretations as a function of the reader's mindset (see Fig. 13.5). Some argue that a preference-based approach is appropriate whenever the task is not well defined. We have not encountered an example where there truly is no specific task. There may be several tasks, in which case the system could be evaluated for each.

We regard preference assessments as useful in go/no-go decisions, giving information on the adequacy of images for further, more rigorous, testing. Indeed, rank-order studies of image quality have been proposed as a formal approach to determining whether the cost of a large-scale objective study is justified (Gur *et al.*, 1997; Rockette *et al.*, 1997; Good *et al.*, 1999; Towers *et al.*, 2000). These studies can make use of highly trained observers and specified tasks; their drawback is that they identify trends without providing an absolute measure of image quality. Statistical methods for planning and analyzing rank-order experiments have been introduced (Rockette *et al.*, 2001).

14.1.2 Fidelity measures

A common approach to image assessment is to assume that the goal in imaging is to reproduce a likeness of the object, leading to the conclusion that the best imaging system is the one that gives the smallest discrepancy between object and image. The most common measure of fidelity is the *mean-square error* (MSE) between object and image; some flavor of MSE is quoted in the majority of papers on image processing or image reconstruction. As we saw in Sec. 13.3.2, however, there are some arbitrary choices to be made in defining MSE, and different choices can lead to quite different conclusions about the quality of an imaging system or processing algorithm.

Problems with fidelity measures MSE and any other fidelity measure will be sensitive to many different properties of an image. If we rotate an image slightly with respect to the object or change the magnification, for example, we can produce a large discrepancy between the object and the image, even if they would otherwise be identical. Similarly, image distortion, such as barrel or pincushion effects, can lead to a large MSE. Finally, gray-scale errors such as nonlinear mapping of the image intensity or even an error in overall brightness can contribute heavily to any fidelity measure.

In many cases, these image modifications are trivial in the sense that they do not degrade the information we want to extract from the images. For example, a radiologist can interpret a chest radiograph just as well if it is rotated by a few degrees on a light box or displayed at a different magnification on a computer monitor.[1] MSE or other measures of fidelity would show that the rotated or scaled image was a poor representation of the object, but the user might not even notice the discrepancy.

Sometimes, however, apparently trivial modifications of an image are important. A cartographer wanting to derive accurate distances from an aerial photograph, for example, would worry a great deal about the magnification, and an astronomer wanting to track the angle between the two members of a binary star would worry about the rotation angle. In designing a lens system for photolithography, distortion might be critical, though for portrait photography it would be imperceptible. Even in these cases, however, a fidelity measure such as MSE is too blunt an instrument to say anything meaningful about the usefulness of an image.

Why not MSE? As delineated in Sec. 13.3.2, there are many arbitrary choices to be made in defining an MSE. For digital images, we must decide whether to discretize the assumed object for comparison with the digital output, to interpolate the digital image in order to get a continuous function to compare to the real object, or just to do a simulation and hope that the results will mean something. For each of these options, we must select a set of functions for discretization or interpolation, and we must select either a single object or a class of objects for comparison in some sense to the images. If the object contains null functions of either the system operator or the discretization operator, as any real object will, then any MSE will be very sensitive to the choice of object.

MSE measures can be very sensitive to relatively trivial image modifications such as magnification, rotation and gray-scale mappings, but they may be completely insensitive to small details that we really want to capture in the image. Furthermore, MSE measures make no distinction between blur and noise. It is easy to construct two very different images, one with high noise but good sharpness and a blurred one with low noise that have the same MSE. The main objection to any MSE metric, however, is that it has nothing to do with the intended use of the image..

14.1.3 JND models

There exists a school of thought in the field of image evaluation that the goal of an image processing or compression algorithm is to create an image that is perceptually equivalent to the original. This school measures image degradation in units of just-noticeable differences, or JNDs, between the original image and its processed counterpart. One JND unit corresponds to a fixed probability, say 50 or 75 percent, that an observer would detect the difference between two images or image regions (Lubin, 1993).

[1] There is anecdotal evidence that sometimes such image modifications can even aid the observer by changing the appearance of an image such that a previously-missed signal becomes visible.

The JND approach to image quality is rooted in the threshold theory of vision. Threshold theory states that signal detection occurs when a signal's perceptibility exceeds an observer's threshold; signal detection is a yes or no event. Furthermore, by the Weber-Fechner law, discussed in more detail in Sec. 14.2.1, the threshold for detecting an extended signal increases proportionally with background intensity. In the early days of vision science, much effort was expended on the measurement of the detection thresholds of various signals on different backgrounds. In the JND approach to image quality, the "signal" is a difference in a pair of images; if that difference is below threshold, the images are of equal quality.

All JND models are based on a model of the human visual system with the intent of predicting human performance in the ranking of image quality or the detection of image differences. The simplest approach is to weight image differences using a function that models the sensitivity of the human visual system to spatial frequency, referred to as a contrast sensitivity function (Daly, 1993). The JND model of Carlson and Cohen (1980) decomposes the input images into frequency bands. After the contents of the bands are processed nonlinearly, the outputs are compared to determine where image differences as seen through this simple model of the visual system are greatest. This model has been used to predict the detectability of edges and artifacts. Barten has also presented a model of the visual system that has been used to predict image quality (Barten, 1992, 1993). The Barten JND model utilizes a single integral over spatial frequencies rather than a decomposition into frequency bands, making use of an average contrast sensitivity function of the visual system. The Barten model has been shown to predict subjective image quality for several simple tasks and is the centerpiece of a recent National Electrical Manufacturers Association standard on display quality (NEMA, 2001).

More complex mechanistic models of the visual system have been developed for use in the prediction of visually perceptible differences in gray scale, color, and video imagery (Hultgren, 1990; Lubin, 1993; Daly, 1993). The models can account for such observation factors as viewing distance and light level (pupil diameter). The most comprehensive models include a nonlinearity representing the visual system's nonlinear response to luminance, a contrast sensitivity function, a bank of spatial-frequency and orientation-sensitive filters, and models of the chromatic and temporal properties of the visual system. The output is a JND map of the image differences, quantified per pixel, field, frame, or sequence.

One argument for the use of a JND metric is that the approach implies the matching of the processing algorithm with the visual system, similar to the way in which the information in a color television signal is matched to the human; because color resolution in the visual system is less than gray-scale resolution, the National Television Standards Commission (NTSC) represents color information more sparsely than luminance information.

Advocates of the JND approach argue that it is objective, it correlates with subjective assessments of image quality, and it predicts a large body of human data for both detection and discrimination tasks without the need to fit any free model parameters. The tasks have included disk detection, sine grating detection, checkerboard detection and edge-sharpness discrimination. The task can utilize real objects on real backgrounds; a recent comparison of image quality for the task of microcalcification detection in mammographic images showed a high correlation between JND measures of image quality and human observer performance (Krupinski *et al.*, 2003). Commercial JND-based image evaluation packages are readily available.

JND measures suffer from some of the same problems we have enumerated for fidelity measures, including the lack of distinction between blur and noise and the questionable definition of task. Both fidelity and JND measures quantify some form of image discrepancy: fidelity measures give all image differences equal importance, while JND measures weigh image differences according to their predicted manifestation at the output of the visual system. In order to calculate perceptual image differences, the JND approach requires twinned-noise image pairs, that is, two images in which the noise realization in each is the same. This paradigm is significantly different from the one underlying statistical decision theory, in which each image represents an independent sample from the signal, background, and noise distributions. It is not clear how the JND approach can be extended beyond simulated targets to real images with real signals because it is not possible to acquire real images that are identical except for the presence or absence of some target. An active area of current research is the usefulness of the JND approach for predicting the quality of an imaging system given random signals on random backgrounds in images with unpaired noise realizations.

Nevertheless, the JND community has much to offer the field of objective assessment of image quality. For example, we shall see that model observers play a significant role in the objective assessment of image quality; the sophisticated models of the visual system developed by the JND community may be of use in the development of predictive models of human task performance for more realistic tasks.

14.1.4 Information-theoretic assessment

In 1948, Claude Shannon published his now-famous theory of communication, in which he defined the information content of a message as a measure of the degree to which it is unexpected.[2] Shannon defined the information content of a single message state n as $I(n) = \log[1/\Pr(n)]$, where $\Pr(n)$ is the prior probability of occurrence of the n^{th} message. Messages with high probability carry little information; high information content is associated with messages that are least expected. By this definition, the mean information content of a message is

$$\overline{I} = \sum_{n=1}^{N} \Pr(n)\, I(n) = \sum_{n=1}^{N} \Pr(n) \log\left[\frac{1}{\Pr(n)}\right] = -\sum_{n=1}^{N} \Pr(n) \log[\Pr(n)]\,, \qquad (14.1)$$

which becomes

$$\overline{I} = \sum_{n=1}^{N} \frac{1}{n} \log[n] = -\sum_{n=1}^{N} \frac{1}{n} \log\left[\frac{1}{n}\right] \qquad (14.2)$$

when the messages are equally likely.

Shannon's model for a communications system was a nonimaging system comprised of a single source (the message), an encoder, a communications channel that transmitted the message, and a decoder. The purpose of the communications system was to provide the user with a reproduction of the message. Designers of

[2] In his book on the relationship between information theory and thermodynamic entropy, Brillouin (1956) points out that the theory developed by Shannon came to light earlier in Szilard's discussion of the Maxwell demon (1929). (We thank B. R. Frieden for this historical note.)

encoders, decoders, and transmitters were seeking to ensure that the user received the message that was sent. Not surprisingly, systems whose goal was to reproduce a transmitted message were most often evaluated using fidelity measures.

There is a large literature on the application of information theory to the evaluation of imaging systems. Fellgett and Linfoot (1955) and Linfoot (1955) considered a simplified model of an optical system in which the source is divided into small discrete elements, each capable of a finite number of discrete brightness levels. The information content of the values of the elements can then be defined in terms of their degree of unexpectedness. That is, the information carried by a particular object $\mathbf{f}$ is given by

$$I(\mathbf{f}) = \sum_{n=1}^{N} \mathrm{pr}(f_n) \log \left[\frac{1}{\mathrm{pr}(f_n)} \right] = - \sum_{n=1}^{N} \mathrm{pr}(f_n) \log \left[\mathrm{pr}(f_n) \right] , \qquad (14.3)$$

where f_n is the brightness of the n^{th} object element. In this simple model the object values are assumed to be independent, and we see that the entropy of the set of values becomes the measure of information content [*cf.* (15.158)].

Fellgett and Linfoot generalized this simple model to allow for a continuous distribution of object values and a division of object space into isoplanatic patches. With these additions to the model, Fourier methods can be used to describe the transfer characteristics of the imaging system. Felgett and Linfoot considered the assessment of an optical system for two tasks: the formation of an image that is similar to the object, and the production of an image that carries the most information about the object without regard to a specific inference or interpretation process. Assessment by similarity leads to fidelity measures; the same issues raised in the previous section on fidelity measures then apply, and Felgett and Linfoot point out many of these shortcomings as well. Thus Felgett and Linfoot turn to assessment by information content. Using the object's information measure as a starting point, an imaging system's ability to transfer information is computed and maximized. However, their resulting figure of merit is independent of the statistics of the object set and the measurement noise (film type, in those days). This is seen as a positive result by these authors, because it allows for optimization of optical systems without regard to the statistical properties of the object and the measurements, and no specific task must be considered.

More modern works have followed the approach of Fellgett and Linfoot, emphasizing the information rate of an imaging system (Huck *et al.*, 1997) and its correlation with the visual quality of the resulting images, where visual quality is measured in terms of image sharpness, clarity, and fidelity. Of course, all of these measures encounter the commensurability problem discussed in Sec. 13.3.2. Moreover, these measures are not uniquely related to the performance of a specified observer on a particular task.

Dainty and Shaw (1974) and Shaw (1978) related the information theory of Shannon to their noise-equivalent quanta (NEQ) approach to image assessment. According to these authors, an actual imaging system that degrades the information content of the input is associated with an NEQ relative to the real exposure quanta. As described in Sec. 13.2.13, this theory assumes a linear shift-invariant imaging system and stationary noise, leading to a Fourier-domain framework for describing the detection SNR as a function of spatial frequency. Spatial frequencies correspond to Shannon's channels in this approach (Wagner and Brown, 1985).

A broader view of information-theoretic image formation and assessment exists (O'Sullivan *et al.*, 1998). Object representation is achieved by combinations (not necessarily linear) of basis functions that may or may not be orthogonal; this approach does not automatically assume the object space is decomposed into pixels. The objects may be known exactly or random. The imaging system may be deterministic (low noise, nonrandom) or may be stochastic, and may be direct or indirect. This view of information-theoretic image formation is consistent with the framework shown in Fig. 7.14 for the imaging process. Moreover, in this treatment the task is more generally cast to include measures of optimality for detection, recognition (classification), parameter estimation, and scene estimation (image reconstruction). When the task is detection or classification, the overall performance of a system is measured by the performance of the recognition or detection function; performance measures for detection and recognition tasks include such familiar measures from Chap. 13 as the probability of detection and the probability of a false alarm. Optimal estimation for random objects is achieved using the familiar *maximum a posteriori* (MAP) procedure derived in Chap. 13 when a prior for the object exists; without a prior, maximum-likelihood methods result and are characterized by the Fisher information matrix and the Cramér-Rao bound.

Thus we see that the information-theoretic approach, when presented in this broad manner, is akin to the statistical-decision-theoretic approach presented in Chap. 13. In the information-theoretic approach, all performance metrics quantify the information provided by the measurements and the likelihood function plays a fundamental role in all cases. Similarly, we found in Chap. 13 that the likelihood ratio is central to all measures of task performance that characterize optimal decision/estimation strategies in statistical decision theory. The information-theoretic approach postulates that the user knows "everything except the decision" (O'Sullivan *et al.*, 1998). In other words, an ideal observer is assumed. Information measures are therefore useful for the assessment of raw data, but they are not necessarily good predictors of human performance. This point is particularly relevant to the use of information criteria in deriving optimal reconstruction algorithms. There is no guarantee that the resulting images are optimal when assessed in terms of human performance.

14.1.5 Objective assessment of image quality

For an image-assessment method to be acceptable, it must objectively quantify the usefulness of the images for performing a given task. Task-based measures of image quality have been advocated for many decades, starting with Harris (1964), and including Hanson (1977), Wagner (1978), Judy *et al.* (1981) and Myers *et al.* (1986). The resulting figure of merit must be computable and scalar, so that it can be used unambiguously in the optimization of imaging systems and the assessment of observer performance. Methods based on statistical decision theory satisfy these requirements.

Four key elements are essential in the objective assessment of image quality (Barrett, 1990):

1. Specification of a task;
2. Description of the object class(es) and imaging process, leading to a description of the data;

3. Delineation of the observer;

4. Figure of merit.

Let's consider each of these elements in more detail.

The task In Chap. 13 we considered two kinds of tasks in some detail. One kind of task is the detection of an object in the presence of a background or clutter. The object might have one or more random parameters and the background may or may not be random. A related task is the classification of an image into one of a finite number of alternative classes. A second type of task is the estimation of parameters describing the object or background or both. Chap. 13 gives many examples of detection, classification, and estimation tasks.

We have seen that many of the approaches described in earlier sections define the task as the reproduction of a single object. While object reproduction might be construed as an estimation task, there are several important differences between estimation and object reproduction. First, defining the task as object reproduction leads to the problem of commensurability delineated in Sec. 13.3.2: objects and images live in different spaces. No imaging system can exactly reproduce a continuous object. How then, to choose among systems that all fall short of this impossible goal? In addition, when the stated task is object reproduction, an assumption is being made that all object locations/elements/parameters are equally important; this is not the case in real situations. Finally, no imaging system will be utilized for a single object, so the task definition should encompass the use of the system over the expected range of objects.

Properties of objects and images From the preceding discussion we know that the evaluation of an imaging system should take into account the physical and statistical properties of the set of objects to be imaged. In a classification task, the objects are categorized into a finite set of classes. For example, the evaluation of mammographic imaging systems for the task of breast lesion detection requires the characterization of normal breast tissues and breast lesions in terms of the full probability density function of the objects under each class. While this is an impossible task, tremendous progress is being made toward the characterization of the mean and low-order joint densities of real tissues using ultra-high-resolution projection imaging and autoradiography, among other methods (Hoeschen *et al.*, 2000).

Another method for creating and characterizing a set of objects is through the use of simulations. The use of numerical algorithms to generate random objects gives the investigator the ability to characterize the deterministic and stochastic properties of the objects. Modern simulations are becoming increasingly realistic. Investigators have added simulated targets to real images (creating so-called hybrid images) with sufficient realism that in some cases human observers were unable to discriminate the artificial targets from real ones (Revesz *et al.*, 1974; Eckstein and Whiting, 1996). The future will bring even greater flexibility and realism to simulated images, with the entire anatomy and physiology of a human being modeled on a fine scale as a starting point toward the creation of simulated, highly realistic imagery of normal and abnormal states. Nonmedical imaging applications are following the same trajectory; in astronomy, acoustical imaging, radar, and so on, simulations of objects and imaging systems are vastly improving and leading to

new abilities to generate realistic data sets for image evaluation. Image simulation methods are described in some detail in Sec. 14.4.

The observer Given a task and a set of objects, the next requirement for the assessment of image quality is an observer or strategy for performing the task. The observer might be a human, such as a radiologist or an expert photointerpreter. Models of human observers can be used to predict human performance. Model observers make it possible to optimize imaging systems without the need for lengthy human-observer studies at every design stage. Human observers and their models are relevant to the assessment of images to be displayed for human consumption. For example, the assessment of display devices, reconstruction algorithms, and all manner of image-processing routines are evaluated appropriately using human observers or their surrogates.

The ideal observer is defined in Chap. 13 as the observer that makes optimal use of all available information to perform the specified task. Having no need for image reconstruction, the ideal observer is appropriate for the evaluation of the quality of the raw data for classification tasks.[3] Thus the ideal observer is the observer of choice for the assessment of imaging hardware. As detailed in Chap. 13, the ideal observer requires the complete PDF of the data under each hypothesis. In cases where this information is not available, the Hotelling observer can be a useful alternative, requiring only the first- and second-order statistics of the data.

The figure of merit Having specified the task, the objects, and the observer, all that is needed is some way of telling how well the observer performs. For classification tasks, useful figures of merit include the area under the receiver operating characteristic (ROC) curve (AUC), partial ROC areas, sensitivity/specificity pairs, the percent of correct decisions (PC), and the classification signal-to-noise ratio, or SNR. Those readers unfamiliar with the theory of ROC curves are referred to Chap. 13 for background material necessary for understanding the terminology here.

Possible figures of merit for estimation tasks include bias, variance, mean-square error (MSE), and ensemble mean-square error (EMSE). The MSE summarizes the performance of an estimation algorithm in determining the estimable parameters of a single object averaged over multiple data sets. In contrast, EMSE describes estimation performance averaged over both measurement noise and a distribution of objects, allowing for nonestimable parameters. Estimators can also be evaluated using bounds on their performance, the most notable being the Cramér-Rao bound for maximum-likelihood estimators. In Sec. 13.3 the reader can find a lengthier treatment of performance measures for estimation tasks.

Returning to the set of requirements listed at the beginning of this section, we can see that each of the methods described in the previous sections lacks one or more of these key elements. For example, JND methods measure image quality using a distance between two scenes without specification of a task or an object class. Thus, in the remainder of this chapter we shall rely on the approach to the objective assessment of image quality outlined in this section.

[3]Wagner, Brown, and Pastel suggested the division of imaging systems into detection and display components for assessment purposes as early as 1979.

14.2 HUMAN OBSERVERS AND CLASSIFICATION TASKS

A wide variety of imaging applications make use of a human as the observer or expert reader. The task is almost always classification, because humans are not as adept as machine algorithms at the absolute quantitation of parameters using images as input. The purpose of this section is to chronicle what is known regarding the perception of form by the human visual system, how we measure human performance on classification tasks, and what we have learned regarding human performance for various classification tasks. We shall focus on the perception of pattern and form, with the goal of connecting this to an understanding of human performance on single, static images. The extension to tasks involving temporal information, color, or stereo are beyond our scope, although in many cases the generalizations required to include this kind of information will be suggested.

14.2.1 Methods for investigating the visual system

Centuries ago, the human eye was assumed to work as a simple camera. This view was espoused by the famous astronomer Johannes Kepler as early as 1604. Not long after, Réné Descartes' famous treatise, *La Dioptrique* (1637), described an experiment in which an eye from an ox was used to "view" the image formed on the retina, which had been scraped away to make the eye translucent. The discovery that the image formed by the eye's lens was inverted was a source of much confusion, since none of us has the experience of seeing the world upside down. Since that time we have come to realize that we do not directly "see" the retinal image; what we perceive is a processed and interpreted version of the image formed at the back of the eye. The retina and the visual components of the eye-brain system are complex entities that have been the subject of amazing discovery since the time of Kepler.

The images formed by the eye's lens onto the retina stimulate the approximately 130 million photoreceptors we know as the rods and cones. These units stimulate bipolar cells that lead to the ganglion cells, whose axons form the optic nerve. The axons of the optic nerve terminate in the lateral geniculate nucleus (LGN) of the thalamus. The cells of the LGN relay signals to a region of the striate cortex called the primary visual cortex. The activity of a cortical cell is thus the result of millions of retinal inputs. Within the visual cortex further signal processing and feature extraction occurs, leading to our visual perception of the world around us.

Early discoveries of the visual system were anatomical, as described so graphically by Descartes. Anatomical studies tell us the spatial sampling of the rods and cones, the number of fibers making up the optic nerve, and the location of their termini. We need other means of determining how these entities function and interrelate.

One means of elucidating the functional properties of the elements of the visual system is through electrophysiological studies in animals. These studies involve the placement of electrodes into single cells in the visual pathway and the subsequent measurement of the cell's response to visual stimuli. In 1940, Hartline became the first to insert electrodes into a single ganglion cell in a vertebrate (a frog) and record axon potentials, following his earlier experiments in the horseshoe crab (1934). Hartline's work was the precursor to the acclaimed work of Hubel and

Wiesel (1962), who shared the Nobel Prize for their pioneering study of the visual system of the cat. Hubel and Wiesel studied the response of single cortical cells to visual patterns of specific orientation and location (bars, edges, and spots) and found that the cells demonstrate orientation selectivity and binocularity. They soon reported similar findings in monkeys (1968).

Many electrophysiological investigations in animal models have followed in the giant footsteps of Hartline, Hubel and Wiesel. For such studies to be relevant to the human visual system, the animal's characteristics must be able to be extrapolated to the human. Since the visual systems of all vertebrates are similar, these measurements provide especially valuable information regarding the behavior of the human visual system.

The functioning of the visual system can also be studied using *psychophysics*, the measurement of the reactions of observers to visual scenes and the development of quantitative relationships between response data and physical characteristics of the input images. The physical characteristics of the images include quantities such as the display luminance, the noise and resolution properties of the images, as well as parameters that specify the target and background. Observer performance is measured in terms of indices such as the area under the ROC curve or the percentage of correct detection or localization responses. Thus psychophysical experiments determine external measures of the visual-system function. Methods for the conduction of psychophysical studies using human observers are presented in Sec. 14.2.3.[4]

Modern imaging methods have brought new tools to the study of the function of the visual system. Using functional imaging methods such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET), investigators are determining areas of the brain involved in the performance of visual tasks. Imaging provides a noninvasive alternative to electrophysiological techniques with the ability to map both spatial and temporal response to stimuli.

In what follows we shall describe the more salient features of the visual system that are relevant to understanding human performance on classification tasks using images as inputs. These characteristics play a key role in the development of predictive models of the human observer.

Receptive fields A *receptive field* is an area on the retina that gives excitation or inhibition of a neuron's activity upon changes in illumination. Receptive fields can be defined for ganglion, geniculate, and cortical cells. The receptive field is evidence of a many-to-one relationship between photoreceptors in a region of the retina and the neural cell. In fact, there are about 1000 cortical neurons per retinal cone for visual information processing (Kronauer and Zeevi, 1985).

Receptive fields for the ganglia can be organized into two broad classes: those that have plain receptive fields, and those that have complex receptive fields. Plain receptive fields have a center-surround structure. When a spot of light illuminates their center, an increase in firing rate occurs (excitation); light on the surround region decreases the rate (inhibition). Diffuse light that illuminates both regions gives a cancellation of the signal, resulting in no response. Simple cells are often

[4]While it might be expected that psychophysics is exclusively applied to the study of human observers, psychophysical experiments using trained animals have been performed to elucidate properties of the cat and monkey visual system.

referred to as Kuffler cells, after the early investigator who mapped their behavior in the cat (Kuffler, 1953).

Complex cells, not surprisingly, are tuned to more complex retinal patterns, such as gratings. Enroth-Cugell and Robson (1966) were the first to measure the response of single ganglion cells to cosine gratings in the cat. Both even and odd receptive fields exist, giving the cat visual sensitivity to gratings and edges. At one time it was thought that complex cells were in series with simple cells, but we now know that simple and complex cells act in parallel. For some tasks the response of complex cells occurs earlier than that of simple cells, and for other tasks the opposite is the case (Hoffman and Stone, 1971).

The LGN and cortical neurons are also associated with receptive fields at the retina. Cortical cells have been found that are tuned to edges, lines, movement of lines and gratings at certain orientations, speeds and accelerations, and even angles between lines. While the responses of ganglion, LGN, and cortical neurons to stimuli have significant similarities, there are interesting differences across them as well. Maffei and Fiorentini (1973) compared the responses of these neurons to cosine gratings in the cat and found that the stages respond to different ranges of spatial frequencies. Moreover, the spatial frequency selectivity becomes narrower from the retina to the LGN to the cortex. DeValois *et al.* (1982) also found in the macaque that the sharpest tuning occurs in the cortex.

Lateral inhibition The output of a receptor's ganglion cell is not only impacted by multiple retinal inputs, but also by the behavior of nearby neurons. In studies of the horseshoe crab, Hartline and Ratliff (1957) were the first to show that the output of a ganglion cell can be inhibited when a nearby neuron is excited. This effect is referred to as *lateral inhibition*. The opposite can also occur, in which case the effect is termed *lateral summation*. Lateral inhibition and summation demonstrate one role of the synapse of the bipolar cells that communicate with the ganglion cells, enabling ganglion cells to interact.

Contrast sensitivity function By measuring the electrophysiological response of animals to patterns, the structure and function of multiple receptive fields have been elucidated. In humans, the ability to detect patterns is measured via psychophysics, giving a global response to the pattern rather than a response localized to a single neuron. The *contrast sensitivity function* (CSF) describes the overall sensitivity of the visual system to sinusoidal patterns as a function of pattern frequency. High sensitivity signifies that the pattern can be seen with little contrast; low sensitivity implies that a large contrast is required for the observer to detect the pattern. A great many psychophysical studies have been conducted to determine the sensitivity of human observers to grating patterns, starting with DePalma and Lowry in 1962. Robson (1966) measured both spatial and temporal CSFs in humans in the 1960s. Campbell and Robson (1968) measured the contrast sensitivity to single sinusoidal gratings over a broad range of spatial frequencies at fixed background luminances. By determining the just-visible contrast of sine-wave targets, these authors found that the CSF follows a band-pass shape with a pronounced maximum at 2 to 4 cycles per degree, falling off at both low and high spatial frequency. An idealized CSF is shown in Fig. 14.1.

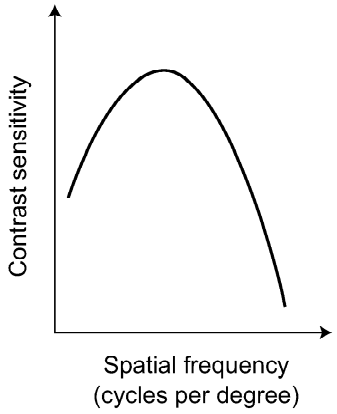

Fig. 14.1 Idealized example of a contrast sensitivity function.

We now know that the visual system is extraordinarily adaptive; the CSF is dependent on mean luminance level, noise level, color, accommodation, eccentricity, and image size (Kelly, 1977). While color CSFs have a shape similar to that shown in Fig. 14.1, high-frequency color patterns are less detectable than luminance patterns of the same frequency (Cornsweet, 1970). There is also significant variation in contrast sensitivity functions, as well as other parameters of the visual system, across human observers (Ginsburg *et al.*, 1982; Owsley *et al.*, 1983). Ginsburg and Evans (1984) measured the CSFs of a large population of observers and found that the peak value is dependent on the individual.

The CSF is often depicted as the envelope of multiple narrow spatial-frequency-selective responses internal to the visual system. This view stems from evidence that the bandwidth determined via psychophysical study in animals can be much greater than that determined via electrophysiological experimentation. For example, individual neurons in the cat have been found to respond to a narrower range of frequencies (Movshon *et al.*, 1978) than what is found externally via behavioral studies (Blake *et al.*, 1974).

Lateral inhibition reduces sensitivity to signals with large extent. That the CSF is very low at low frequency is consistent with the inhibitory behavior of receptive fields at low frequencies. Several studies have established that the human observer is unable to efficiently integrate information beyond a certain spatial extent (Blackwell, 1946; Burgess *et al.*, 1979; Boff *et al.*, 1986).

Masking Research has shown that the presence of one pattern can make another pattern less visible to an observer. This property is known as *masking*. The opposite of masking is *facilitation*, defined as the improved detection of a pattern in the presence of another. The pattern to be detected is referred to as the *signal*; the additional pattern is referred to as the *mask*. The mask is usually supra-threshold, meaning its contrast is above that required for detection. When the mask contrast becomes sufficiently low, the signal threshold is identical to the the signal threshold in the presence of a uniform background; that is, the signal threshold is what is expected based on the observer's CSF and no masking occurs.

Periodic patterns such as gratings or sinusoids have been shown to mask patterns with similar orientation or spatial frequency (Legge and Foley, 1980; Phillips and Wilson, 1984). This effect is known as *phase-coherent* masking. Another experimental paradigm is to use noise fields of different bandwidths as masks; the

effect is then called *phase-incoherent* masking (Pollehn and Roehrig, 1970; Pelli, 1981; Thomas, 1985). The presence of an aperiodic pattern such as an edge or a gradient can also mask a nearby feature (Fiorentini *et al.*, 1955). Masking demonstrates orientational selectivity as well as frequency-dependent behavior (Campbell and Kulikowski, 1966).

Diffuse light can mask signals. For this reason radiologists are trained to read images in a darkened viewing area after they have adapted to the ambient light level, to better detect low-contrast signals. Scattering in the lens and cornea of the eye can also mask low-contrast signals. This problem is known to worsen with age.

Channels *Channels* are independent processors tuned to different narrow ranges of spatial or temporal frequency. Channels were first hypothesized as visual scientists pondered data from studies using compound-frequency patterns such as sawtooth and rectangular gratings (Campbell and Robson, 1968). These data seemed to indicate that detection of the pattern occurs only when the most detectable component reaches its own threshold, independent of the presence of the other frequency components. Sachs *et al.* (1971) then carried out experiments using compound gratings consisting of just two frequency components. Whenever the second component differed in frequency from the first by more than a certain ratio, the data were consistent with the hypothesis that the two frequency components were being detected independently. Moreover, when two components with frequencies related by an even larger ratio were combined, the grating was no more detectable when the two components were phased so that their peaks added than when their peaks subtracted (Graham and Nachmias, 1971). The investigators concluded that different spatial-frequency components were detected by independent processors tuned to different narrow ranges of spatial frequencies. Detection of a stimulus occurs whenever the activity in one of these processors rises above a threshold. These processors were referred to as channels. Channels can be thought of as mosaics of receptive fields (Sachs *et al.*, 1971).

Many scientists have worked to corroborate the presence of frequency-selective channels in the visual system (Mostafavi and Sakrison, 1976) and to determine their properties in finer detail (Halter, 1976). The electrophysiological recordings of Hubel and Wiesel (1962) are construed by many as the first evidence for channels. Adaptation and masking experiments support the hypothesis that the channels are medium-bandwidth mechanisms (Blakemore and Campbell, 1969; Stromeyer and Julesz, 1972; Stromeyer and Klein, 1975; Legge and Foley, 1980). Narrow-bandwidth channels are suggested by the results of frequency-discrimination tasks (Campbell *et al.*, 1970). The entirety of the data suggests the presence of approximately octave bandwidth spatial-frequency channels over the entire visible range.

There is ample evidence, starting with the work of Hubel and Wiesel (1962), that the visual system also contains orientation-selective channels. DeValois *et al.* (1982) investigated simple cells in the macaque and found them to have an angular resolution of $\pm 20°$. These data are quite similar to the estimates of orientation selectivity in humans obtained using masking experiments (Campbell and Kulikowski, 1966; Phillips and Wilson, 1984). There are also channels tuned to object motion that have direction selectivity (Tolhurst, 1973), with a temporal two-octave bandwidth (Tolhurst, 1975; Watson and Robson, 1981).

Internal noise Human observers are noisy measurement devices. Thus, even if the images presented to a human observer were noise-free, the output of the human would have some variability. While it requires only one optical photon to excite a rod, the number necessary for "seeing" is larger (Hecht *et al.*, 1942). Barlow was the first to suggest that this discrepancy is the result of an internal noise mechanism (1956).

Burgess *et al.* (1981) compared human SKE (signal-known-exactly) detection performance in white noise to an ideal detector with an added internal noise contribution. While this modification to the ideal-observer model improved the model's agreement with the human data, it was suggested that some form of observer sampling inefficiency was also needed for the model to match the slope of the human data vs. noise spectral density. The authors further suggested that perhaps the observer noise might be a function of image noise. Data from subsequent classification experiments have borne out the suggestion that the visual system has two internal noise components (Burgess and Colborne, 1988). The first component is an additive noise term that is independent of the image luminance. This noise component may be the result of neural noise (Tolhurst *et al.*, 1983), as well as fluctuations in the observer's decision criterion (Eckstein *et al.*, 1997). The second component is an induced, or image-dependent, component. The induced internal noise has been shown to be proportional to the variance of the image noise (Burgess and Colborne, 1988).

Weber-Fechner law As stated earlier, diffuse light can mask low-contrast signals. As a result, objects on bright backgrounds are harder to detect than objects on dark ones (Cornsweet, 1970). The Weber-Fechner law states that the relative contrast of an object, given by $(L_{max} - L_{min})/L_{mean} = \Delta L/L$, is equal to a constant for a given probability of detection. By this law, the detection of a difference in luminance depends on the baseline, so that relative luminance is important, rather than absolute differences.

Evidence of behavior following the Weber-Fechner Law has been interpreted as a local gain mechanism or a saturating nonlinearity in the visual system, coupled with internal noise (Shapley and Enroth-Cugell, 1985). This law also plays a significant role in the approach used by many investigators in choosing the calibration method for their soft-copy display (Blume and Hemminger, 1997). Many investigators choose to use a *perceptually linearized* display, in which the output luminance at each digital driving level is set so that the step sizes between gray levels is higher at higher absolute luminance levels (Pizer, 1981).

Psychometric functions A *psychometric function* is a plot of the probability of a signal being detected as a function of signal contrast. For a signal of contrast c, the probability of detection is usually fit by a sigmoidal function of the form (Nachmias, 1981)

$$\Pr(D_2|c) = 1 - \exp[-(c/\alpha)^{\beta}], \tag{14.4}$$

where D_2 indicates that the observer chose in favor of the signal being present, β is a slope parameter, and α shifts the function relative to the signal contrast. Many experiments have been found to indicate approximately equal slope parameters (Mayer and Tyler, 1986).

14.2.2 Modified ideal-observer models

Given the vast array of anatomical, electrophysiological and psychophysical data now available to us, many researchers have worked to develop models for all or portions of the visual system. Some models are highly specialized, with the minimum number of components required to demonstrate the model's ability to predict data obtained in a narrow range of psychophysical experiments. Other models are extraordinarily complex, incorporating foveal sampling, a hierarchy of neural stages, and higher-level signal processing and decision making in an effort to replicate the entire visual system. We shall focus on models that have been developed for the specific purpose of objective evaluation of imaging systems for classification tasks.

In Chap. 13, the ideal observer was introduced as the optimal decision maker for classification tasks as determined by statistical decision theory. The ideal observer sets the upper bar for classification performance. Statistical-decision-theoretic models of the human observer thus use the ideal-observer model as a starting point. We do not need a model with millions of photoreceptors and receptive fields, so long as the model predicts human data on a range of tasks that are useful for image assessment. In fact, a simpler model facilitates imaging system evaluation and optimization over high-dimensional optimization spaces.

The modified-ideal-observer approach to modeling human performance is this: begin with the concept of the ideal observer; compare performance predictions with human performance on actual classification tasks; modify the model to better predict human performance. Modifications to the model should be grounded in the known features of the visual system described in the previous section.

We therefore require a rigorous basis for comparing observer performance. For this, we return to the concept of observer efficiency.

Observer efficiency In Chap. 12 we introduced the concept of detective quantum efficiency as a measure of the SNR transfer characteristics of a detector [*cf.* (12.23)]. In Chap. 13 we extended this concept to describe the efficiency of the Hotelling observer relative to the ideal observer [*cf.* (13.273)]. Analogously, we can define the statistical efficiency of the human observer relative to the ideal observer as

$$\eta_{human} = \frac{\mathrm{SNR}^2_{human}}{\mathrm{SNR}^2_{ideal}} \,. \tag{14.5}$$

The relative efficiency of any two observers can be similarly defined.[5]

When human and ideal performance are comparable, the efficiency approaches one and we conclude that the human observer is able to make almost complete use of the information in the data to perform the visual task. For efficiencies much less than one, we can conclude that the human observer is inefficient at extracting the relevant information in the image for performing the task. When this occurs, we look for features of the human visual system that might be the basis for the human observer's reduced performance.

[5]Some authors have defined observer efficiency as the ratio of SNRs required by the observers to perform the task. In this school, human efficiency equals the SNR required by the ideal observer divided by the SNR required by the human, where SNR is a physical quantity such as contrast; smaller SNRs denote better performance. We prefer the definition given in (14.5), where SNR quantifies task performance and high SNR is good!

Classification in uncorrelated noise As described in Sec. 13.2.13, the definition of observer efficiency given in (14.5) comes from the basic definition of DQE first given by Albert Rose (1948) as a means of comparing the noise level of an actual radiation detector with that of an ideal one. Rose compared the performance of the eye to an ideal picture pickup device and determined that the minimum contrast c_{min} required for detecting a uniform object on a flat background with quantum noise satisfies

$$c_{min}^2 NA = k\,, \tag{14.6}$$

where N is the photon density of the uniform background, A is the area of the object and k is a constant dependent on the observer; from experiments on human subjects, Rose determined that k is in the range of 3 to 7. A lower value of k implies a lower c_{min} and hence a more efficient observer.

Recall from Sec. 13.2.8 that the ideal observer takes on a special form when the task is the discrimination of two nonrandom signals in additive Gaussian noise. In this case the ideal observer is equivalent to a prewhitening matched filter (PWMF), which reduces to a simple matched filter when the noise is white. Lawson (1971) demonstrated that the Rose model of (14.6) is a special case of the PWMF for a pillbox signal in Poisson noise of sufficient count rate that the Poisson statistics can be approximated by Gaussian statistics.

The calculation of the ideal observer's SNR is straightforward for SKE/BKE (signal-known-exactly/background-known-exactly) tasks in Gaussian noise and can be done analytically. For this reason the first comparisons of human performance to ideal-observer performance were achieved in SKE/BKE tasks in white, or uncorrelated, Gaussian noise. Burgess *et al.* (1981) found human observers to be highly efficient (η of 0.5 to 0.8) for SKE/BKE detection and discrimination tasks in white noise. Human performance is well predicted by an ideal observer that positions a template over the location of the expected signal and performs a linear summation of the output. The fact that the efficiency is less than one can be explained by internal noise (Burgess and Colborne, 1988).

When the signal extent becomes sufficiently large, human detection efficiency in white noise declines (Burgess *et al.*, 1979). In effect, there is a spatial limit to the human's ability to perform the template-matching operation. We might have expected this from the shape of the CSF of the visual system. Other investigators have found that the human is unable to efficiently process "DC" information (Ratliff, 1965; Van Nes and Bouman, 1967). For this reason some investigators proposed that the PWMF model be modified by adding an "eye filter" (Loo *et al.*, 1984; Burgess, 1994).

Correlated noise Many experiments have been performed to investigate the impact of correlated noise on human discrimination performance (Judy, 1981; Guignard, 1982; Burgess, 1985b; Myers *et al.*, 1985; Blackwell, 1998). Of particular interest in the early 1980s was the character of the noise in computed tomography (CT) images and its impact on human perception. Raw CT data sets have Poisson noise, which is uncorrelated. When CT images are reconstructed from the raw data using the method of filtered backprojection (see Sec. 4.4.3), a filter with a ramp shape in the frequency domain is used, and the resulting images have a ramp-shaped power spectrum at low spatial frequency. Early on, Wagner (1978) hypothesized that human observers would be inefficient when faced with this noise-correlation structure, and suggested that a non-prewhitening matched filter model might be a good predictor

of human performance. Soon after, several studies found that human efficiency relative to the ideal observer is about 20% in CT noise, much less than the efficiencies found in white noise (Judy *et al.*, 1981; Burgess *et al.*, 1985b). Myers *et al.* (1985) investigated human performance for a family of noise power spectra of the form ρ^n, for $n = 1, 2, 3, 4$, where ρ is spatial frequency. Thus $n = 1$ corresponds to the CT case. These studies showed that human efficiency falls rapidly as n increases from 1 to 4.

A natural conclusion to draw from the reduced efficiency of the human observer in tasks limited by correlated noise is that the human observer is indeed unable to perform the prewhitening operation. For this reason the human observer was modeled by some investigators as a matched filter without the prewhitening operation. The efficiency of the human relative to this so-called non-prewhitening matched filter (NPWMF) was shown to be around 50% (Judy and Swensson, 1985), with the difference again explainable by internal noise.

Since the NPWMF equals the PWMF in white noise, the NPWMF model predicts human performance in both correlated and uncorrelated noise. Furthermore, by combining an eye filter and an internal noise mechanism with the NPWMF, an even larger body of human psychophysical data can be explained (Ishida *et al.*, 1984; Loo *et al.*, 1985; Ohara *et al.*, 1986; Giger and Doi, 1987, deBelder *et al.*, 1971; and Wolf, 1980). This observer is often called the NPWE in the literature, to denote the addition of an eye filter to the non-prewhitening matched filter; we shall use this same shorthand below.

The NPWE models the spatial-frequency response of the visual system with a single spatial-frequency filter. Given the experimental evidence that the human visual system has multiple narrow spatial-frequency channels, a preferred approach to modifying the ideal observer is to incorporate this recognized characteristic of the visual system.

Adding channels to the ideal observer The model of the human visual system as a matched filter is effectively a model with an infinite number of channels. Yet there is substantial evidence that the visual system processes images through a finite number of finite-width channels. Myers and Barrett (1987) introduced a handicapped ideal observer, constrained to process scenes through frequency-selective channels, and demonstrated that this modified Bayesian observer ably predicted human performance in correlated noise. They found that this model was robust to the choice of a channel width parameter. By requiring the lowest-frequency channel to have a finite turn-on frequency, this model also predicts the inefficient performance of human observers on tasks that have significant DC content.

Myers and Barrett found the performance predictions of the channelized ideal observer and the NPWMF to be indistinguishable for the problems they studied (stationary Gaussian noise, signal known exactly). They argued in favor of the channelized ideal observer because this model is consistent with a known mechanism of the visual system. Moreover, as we shall see in the following sections, this model has been found to be predictive of human performance over a much broader range of signal detection and discrimination tasks.

Random backgrounds The tasks described in the previous section were ones in which the background was known exactly; the only variation in the data was due to measurement noise. We now consider tasks in which the data are random due to

both background variability as well as measurement noise. The noise in the data is therefore said to have two components.

In Sec. 8.4 we described several approaches for generating random backgrounds, and in Chap. 13 we discussed model observers for tasks in which the background is random and known only in a statistical sense. Several investigators have made use of these methods to study the performance of human observers in random backgrounds and compare the results to model-observer predictions. Rolland and Barrett (1992) generated lumpy backgrounds by randomly superimposing Gaussian blobs on a uniform background according to the procedure described in Sec. 8.4.4. For the task of detecting Gaussian signals of known size and location on the lumpy backgrounds, Rolland and Barrett compared human performance to the performance of the Hotelling or optimal linear observer defined in Sec. 13.2.12 as well as the NPWMF. Rolland and Barrett found that the Hotelling observer was a good predictor of the human performance data. The performance of the NPWMF was not able to predict human performance over the range of system parameters investigated in the study.

Yao and Barrett (1992) combined the background model of Rolland and Barrett with power-law noise of the type investigated by Myers *et al.* (1987) and found that a channelized Hotelling observer was a good predictor of all the human data acquired in these experiments (Barrett *et al.*, 1993). Burgess *et al.* (1994, 1997, 1999) studied human performance in random lumpy backgrounds generated by filtering a Gaussian field. Their results were consistent with the findings of Rolland and Yao: a Hotelling observer constrained to process the frequency-selective channels is able to predict the data over the range of experimental parameters describing the signals and backgrounds. A NPWMF is not predictive, even when modified to include an eye filter. More recent experiments in power-law backgrounds generated by filtering a Gaussian random process were less conclusive; the most predictive model depended on the signal profile in a study by Burgess (2001).

Several studies have been performed to compare human performance to model observers using real images as backgrounds. In a study using backgrounds drawn from real x-ray coronary angiograms, Eckstein *et al.* (1999) found the channelized Hotelling model to be predictive of human performance in detecting simulated abnormalities. Bochud *et al.* (1995, 1999a, 1999b) studied human performance using simulated nodules in mammographic and angiographic backgrounds and compared their results to a non-prewhitening observer with and without an eye filter (a single channel). They found that, owing to the nonstationarity of the images, the models must be allowed to adapt to the statistics of the local background around the signal in order to better predict human performance. Interestingly, the data of Bochud *et al.* (1999b) suggest that the clinical backgrounds have higher-order statistical properties used by the human observer, although not by the Hotelling observer. Similarly, Caelli and Moraglia (1986) showed that a cross-correlator does not predict human performance when the background is a natural scene.

Signals of large spatial extent The inability of human observers to efficiently detect signals of large spatial extent described in Sec. 14.2.1 has direct ramifications on the task-based assessment of the quality of images derived from systems with significant artifact content. For example, the effective point response function (PRF) for images reconstructed from limited-angle tomographic data can be quite noncompact, yielding long-range streak artifacts. The images of compact objects are thus quite

extended and human efficiency for detecting such objects suffers a penalty (Wagner *et al.*, 1992; Myers *et al.*, 1993). These studies found that human performance is modeled quite well by an observer that performs only linear operations on the images. These studies involved signals at random locations, leading to location-dependent artifacts; the ideal observer is nonlinear in this case if the problem is cast as a binary signal-detection problem.

A long-tailed PRF can also arise when veiling glare is present in a display device or gamma rays penetrate the collimator in gamma-ray imaging. Rolland *et al.* (1989) has shown that human classification performance is inefficient for images formed by a system with a long-tailed PRF, consistent with the earlier literature on the inefficient spatial integration properties of the human. Rolland found that human performance is improved by linear filtering designed to narrow the overall system PRF, even though the ideal observer performance is unchanged by image processing (Sec. 13.2.6), as long as it is invertible.

Texture perception In some special circumstances the human can detect signals of large spatial extent quite efficiently. An example is the detection of a known grid of bright lattice points on a noisy background (Wagner *et al.*, 1990a). Another example is the detection of mirror symmetry patterns of dots (Barlow, 1978; Barlow and Reeves, 1979) buried in a background of random dots (Glass patterns). These results can be explained by an observer who uses the strategy of performing a series of local template-matching operations, skirting the need for integration over a large area (Wagner *et al.*, 1989).

A particular form of extended signal is a pattern of a different texture than the texture of the background in which the signal is embedded. In tasks where such an extended signal is to be detected, human efficiency can be extremely low. For example, the detection of a regular grid or lattice of objects, where some randomization of the object locations is involved, results in low human efficiency (Wagner *et al.*, 1990a). Similarly, the detection of random dot patterns (Maloney *et al.*, 1987; Tapiovaara, 1990) and the detection of diffuse liver disease (Garra *et al.*, 1989) can also be low-efficiency tasks. While many investigators have considered human performance in texture discrimination tasks (Julesz, 1981), these studies are rarely placed in the context of ideal-observer performance. Much more work is needed to understand human performance in textured tasks on an absolute scale.

Nonlinear tasks While the channelized Hotelling observer has been found to predict human performance over a wide range of experimental paradigms, that observer is constrained to perform linear operations on the data. In addition, as the previous section describes, there are ample examples of psychophysical studies showing low human efficiency relative to the ideal observer for nonlinear tasks. The question then arises, can the human do nonlinear operations?

There are many examples of tasks for which human efficiency is fairly high even though the optimal strategy is nonlinear. One example is the task of noise variance discrimination, wherein observers are asked to determine which of two scenes has higher pixel variance. The optimal discrimination strategy is quadratic in the data as seen in (13.163). In unpublished studies, we found that humans were able to perform this task quite efficiently. Does this mean the humans are able to do the computations of (13.163)? Maybe not. It can be shown (Wagner *et al.*, 1990b) that

a combination of linear and logic operations can approximate this ideal nonlinear strategy quite efficiently.

Similarly, the detection of 1 of M orthogonal signals in white noise is optimally performed with a nonlinear strategy [see (13.159)]. However, Nolte and Jaarsma (1967) showed that a series of linear operations, followed by the nonlinear operation of selecting the filter with the maximum output, approximates the ideal nonlinear strategy well over much of the signal parameter space of interest (the range of contrasts of use for psychophysical study). Other investigators have also shown that the "maximum-of" detector gives performance predictions very close to those of the optimal observer in the SNR ranges of experimental interest (Pelli, 1985; Wagner, 1990b).

Burgess and colleagues (Burgess and Ghandeharian, 1984a, 1984b; Burgess, 1985a) measured human efficiency in studies with signal uncertainty in white noise. To approximate ideal-observer performance, they computed the performance of an observer that compared the maximum of a series of matched-filter outputs to a threshold, following the theory of Nolte and Jaarsma (1967). Human observer performance was well predicted by this model observer, with an efficiency around 50%. Judy *et al.* (1997) found little degradation in human performance for the detection of sharp-edged disks and Gaussian signals when the disk diameter or Gaussian width was variable, relative to the SKE task.

Since selecting the maximum of a set of outputs from linear filters is a nonlinear or logical operation, we call this model a *linear+logic observer*. The closeness of the optimal observer to the linear+logic model may preclude one model being rejected in favor of the other using psychophysical data.

The field of neural networks sheds some light on the similarity of these models. A fully connected neural network can be shown to approximate ideal-observer performance. The neural network applies a series of filters in the form of weights to each input value, followed by a sigmoidal nonlinearity—a neural network is a linear+logic observer. As described earlier, there is good evidence that neural responses in the human visual system can be represented by a set of linear filtering operations followed by thresholding, and neural networks represent this scheme. Thus, when the human observer performs nonlinear tasks efficiently, we must be cautious before concluding the human can perform the optimal higher-order (in the data) nonlinear operations. It may well be that the human is smartly performing a series of linear operations, followed by a threshold nonlinearity, to obtain near-optimal performance.

Optimal processing of channelized data As demonstrated in Fig. 14.2, the addition of a channel mechanism to a model observer can be visualized by adding a block to the processing steps shown in Fig. 13.1. In the figure we suppose there are P channels, each represented by the column vector $\mathbf{u}_p$. The output of the channels is given by

$$\mathbf{v} = \mathbf{U}^t \mathbf{g}\,, \tag{14.7}$$

where $\mathbf{U}$ is the $M \times P$ matrix whose columns are the channel profiles $\mathbf{u}_p$, and $\mathbf{v}$ is the $P \times 1$ vector of channel outputs. The $\mathbf{u}_p$ represent the channel profiles, which we assume to be real. Processing the images through channels reduces the dimensionality of the data set from M to P. Possible choices for channel profiles are presented below. In some applications, P can be as small as 3.

Each channel output v_p is a random variable, made random by measurement noise and object variability — whatever sources of randomness are in the raw data. The probability density of each channel output is obtained using the methods for transforming random vectors presented in Sec. 8.1.5. Life is usually simpler, though, because in most models each channel output is the sum of multiple data values; the central-limit theorem tells us that the resulting random variable tends to be Gaussian distributed in that case.

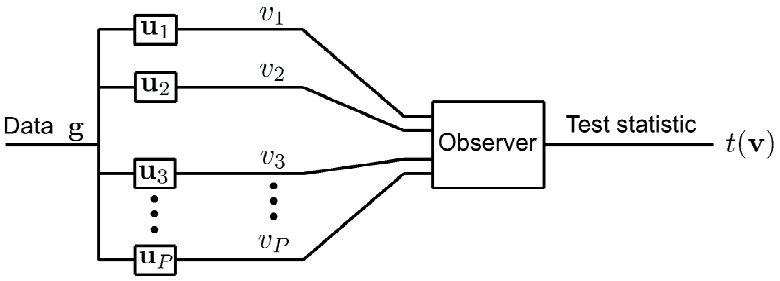

Fig. 14.2 Block diagram of a channelized observer.

Given the $\{v_p\}$, a strategy must be defined by which the channel outputs are combined to arrive at a decision that a stimulus either is or is not present. Possible options include adding the responses of the channels or using only the channel with the maximum response (Graham and Nachmias, 1971). Alternatively, the channel outputs may be combined via *probability summation* (Pirenne, 1943). Pirenne conjectured that binocular vision yields lower detection thresholds than monocular vision because the probabilities of detection from the left and right eyes are independent, and signals are detected if they are detected by the right eye, the left eye, or both. Formally, he suggested that the probability of detecting a signal using both eyes is

$$\Pr(D|L+R) = 1 - [1 - \Pr(D|L)][1 - \Pr(D|R)]\,, \tag{14.8}$$

where $\Pr(D|L)$ and $\Pr(D|R)$ are the probabilities of detecting the stimulus with the left and right eyes, respectively. While probability summation has been rejected as an explanation for the relative performance of binocular to monocular vision, it is encountered in some vision-system models as a means of combining the outputs of parallel channels (Daly, 1993). Combinations of differences in channels at each location/pixel have also been suggested (Lubin, 1993; Lloyd and Beaton, 1990; Zetzsche and Hauske, 1989). In some channel models, the sigmoidal form of (14.4) is imposed on the outputs of the frequency- and orientation-selective filters at each location (Legge and Foley, 1980) before the decision-making step.

Optimal methods for combining channel outputs The human observer can also be modeled as a quasi-ideal observer, that is, an observer who is constrained to process visual scenery through channels, but who is otherwise optimal in how the channel outputs are used to perform the task. If the human is modeled as a channelized ideal obsever, the model will achieve maximal AUC among all observers constrained to process data through the visual channels. A channelized ideal observer forms the

likelihood ratio of the channel outputs under each hypothesis, giving

$$\Lambda(\mathbf{v}) = \frac{\text{pr}(\mathbf{v}|H_2)}{\text{pr}(\mathbf{v}|H_1)}\,. \tag{14.9}$$

The model observer's decision strategy is to compare $\Lambda(\mathbf{v})$ to a threshold, choosing H_2 when $\Lambda(\mathbf{v})$ is greater than this value, and H_1 otherwise. As detailed in Chap. 13, the ROC curve and related performance measures for the channelized ideal observer can be determined using (14.9) as a starting point.

Alternatively, a channelized Hotelling observer (CHO) model might be invoked, thereby assuming that the human observer forms an optimal linear combination of the channel outputs. As described in Sec. 13.2.12, there is a well-established theory for determining the optimal linear combination of the channel outputs and the resulting CHO figure of merit using the statistical properties of the channel outputs. For a binary discrimination task, the Hotelling observer's template in the channel space is given by

$$\mathbf{w}_{Hot,\mathbf{v}} = \mathbf{S}_{2\mathbf{v}}^{-1}\Delta\mathbf{v}\,, \tag{14.10}$$

where $\mathbf{S}_{2\mathbf{v}}$ is the $P \times P$ intraclass scatter matrix of the channel outputs [*cf.* (13.187)] and $\Delta\mathbf{v}$ is the expected difference in the channel outputs under each hypothesis.

The separability of the data in channel space is written in terms of the interclass and intraclass scatter matrices for $\mathbf{v}$:

$$J_{\mathbf{v}} = \text{tr}[\mathbf{S}_{2\mathbf{v}}^{-1}\mathbf{S}_{1\mathbf{v}}] = \text{tr}[(\mathbf{U}^\dagger\mathbf{S}_{2\mathbf{g}}\mathbf{U})^{-1}(\mathbf{U}^\dagger\mathbf{S}_{1\mathbf{g}}\mathbf{U})]\,, \tag{14.11}$$

where $\mathbf{S}_{1\mathbf{v}}$ is the interclass scatter matrix of the channel outputs [*cf.* (13.186)]. While (14.7) has the form of the linear transformation given in (14.10), including a dimensionality reduction, these expressions differ significantly because transformation using the matrix of visual channel functions $\mathbf{U}$ may result in the separability of the channel outputs being less than the separability of the data, while the operation of (14.10) generates a test statistic that preserves the separability in the channel outputs.

When the channel outputs are Gaussian random variables with equal covariance under the hypotheses, the channelized Hotelling observer and the channelized ideal observer are equivalent. In Sec. 14.3 we shall discuss methods for computing performance measures for channelized model observers.

Channel choices The nature of the signal and the background play a significant role in determining an appropriate choice for the channel profiles $\{\mathbf{u}_p\}$ and the way they are imposed on the data. For example, in an SKE task the channels are centered at the known signal location and the sum represented by (14.7) is done. If, on the other hand, the signal can be located at N multiple orthogonal locations, the channels could be centered at each location to give an $N \times P$ vector of outputs for decision-making purposes. When the signals and background are rotationally symmetric, the channels do not require any angular dependence; orientation-dependent signals and backgrounds require channels with orientation-selective responses.

The channelized ideal-observer model of Myers and Barrett (1987) incorporated radially concentric channels to predict human performance in correlated noise. The model's predictive ability was found to be insensitive to channel width and low-frequency turn-on parameters for the tasks considered in that work. The simplicity of this channel structure was possible because the task was the detection of radially

symmetric signals at known locations. Because the signals were low contrast and the image noise was a filtered Gaussian random process, the model was equivalent to a channelized Hotelling observer.

More complex tasks involving asymmetric signals at varying locations may require more complex channel models. A variety of approaches for representation of channel mechanisms have been pursued by the developers of models of the visual system; these approaches can be incorporated into a CHO framework. Models based on Gabor functions (Daugman, 1988; Lloyd and Beaton, 1990; Watson, 1987) and wavelets (Daugman, 1985; Mallat, 1989; Marcelja, 1980; Watson, 1983) can be made to have both spatial and location specificity. Other options include ratio-of-Gaussian channels (Zetzsche and Hauske, 1989) and difference-of-Gaussian (DOG) models (Wilson and Bergen, 1979). Difference-of-mesa (DOM) filters can be used to model radial-frequency filters as well (Daly, 1993). "Mesa" is Spanish for table; a difference of two mesa functions gives a filter with a flat passband, a transition region, and a flat no-pass band. To give radial-frequency filters orientation selectivity, they can be multiplied by a set of functions tuned to orientation. For example, Daly uses what he calls *fan* filters to model the orientation-selective response. The product of the DOM and fan filters are termed cortex filters.

Once a selection has been made of the functions to be used to create a family of channels tuned to an array of orientations and frequencies, the next question is the number of such channels to include in the model. Many studies have indicated that only a fairly small number of channels is required for adequate modeling of human data. Myers and Barrett (1987) found good agreement between human data and CHO predictions with about 6 radial channels. The Daly visual-difference predictor model, designed to predict human performance for JND tasks (Sec. 14.1.3), uses only 6 DOM filters, combined with as few as 6 fan filters (30 degrees each), leading to 31 cortex filters in all [(# of fans) $\times$ (# of DOMs -1) + 1], since the lowest frequency filter has no orientation specificity). Wilson and Gelb (1984) also suggested the use of 6 spatial-frequency selective DOG filters, each with a range of orientations. There seems to be reasonable consensus that only about 6 channels are needed to cover frequency space; adding about 6 orientation-specific channels to each frequency-selective filter gives a complete model.

Internal noise To achieve even better matching between CHO and human data, the internal noise of the visual system must also be addressed. One way to account for internal noise is to scale the detectability of the human observer to that of the model observer (Burgess *et al.*, 1997; Burgess, 1999), giving $\mathrm{SNR}_{human} = \kappa \mathrm{SNR}_{model}$, where κ represents the impact of internal noise on detectability. From (14.5) it can be seen that the scaling factor is related to observer efficiency according to $\eta = \kappa^2$.

Another approach is to add noise injectors to the channel model, as shown in Fig. 14.3, giving a modified definition of the channel outputs of (14.7):

$$v_p = \mathbf{u}_p^\dagger \, \mathbf{g} + \epsilon_p \,, \tag{14.12}$$

where ϵ_p is an additive Gaussian noise contribution in channel p (Legge and Foley, 1980).

For a channelized linear observer, (14.12) is equivalent to adding noise to the decision variable (Abbey and Bochud, 2000). In particular, the channelized Hotelling observer forms the scalar test statistic t according to

$$t = \mathbf{w}_{Hot}^t \mathbf{v} = \mathbf{w}_{Hot}^t [\mathbf{U}^t \mathbf{g} + \boldsymbol{\epsilon}] = \tilde{v} + \tilde{\epsilon} \,, \tag{14.13}$$

where the tildes represent the transformation to the decision-variable space. Both $\tilde{v}$ and $\tilde{\epsilon}$ are scalar random variables, and both can usually be assumed to be Gaussian. The point of (14.13) is that a single Gaussian-noise injector at the decision variable level can model additive internal noise.

Observer uncertainty acts as a source of internal noise. It might be expected that observer uncertainty about parameters of the signal would result in a signal-dependent noise contribution, or a channel-dependent noise contribution. For example, one might hypothesize that the magnitude of a matched-filter's template uncertainty would depend on actual signal size, which would result in signal-dependent internal noise. However, Eckstein *et al.* (2000a) note that signal uncertainty is most often modeled as statistically independent of the signal.

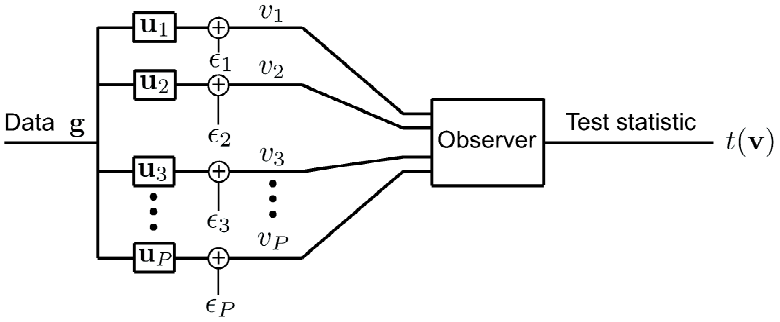

Fig. 14.3 Block diagram of a channelized observer including an internal noise mechanism.

Observer efficiency revisited The efficiency of the human observer relative to any model observer can be defined via (14.5). Choices for the denominator include the full ideal observer, the PWMF in SKE/BKE tasks in Gaussian noise, the NPWMF in SKE tasks, and the channelized Hotelling observer. A growing volume of psychophysical data, combined with model-observer calculations, is establishing the channelized Hotelling observer as an excellent predictor of human classification performance over a wide range of experimental paradigms, including ones in which the PWMF and NPWMF models are far less able to predict human performance. For this reason the remainder of this section will emphasize human efficiency relative to the CHO and review some of the many additional studies that have demonstrated the predictive capacity of the CHO model.

CHO success stories Many of the first comparisons of Hotelling and human performance involved the assessment of image quality in nuclear medicine. Fiete *et al.* (1987) investigated human performance in detecting simulated lesions in simulated liver scans and found excellent correlation with the Hotelling observer. Cargill (1989) used a more elaborate simulation for nuclear medicine, involving the detection of abnormalities in simulated images of a computer-generated 3D model of the liver with several possible disease states. She found excellent correlation between human performance and Hotelling predictions of image quality for 9 different collimator designs.

The CHO has also been shown to correlate well with human performance in the assessment of acquisition systems and reconstruction algorithms in tomographic imaging. Abbey and Barrett (1995) found good agreement between the human and

CHO across a range of linear iterative reconstruction algorithm parameters. Gifford *et al.* (1999, 2000a) used human and model observers to evaluate the impact of detector-response compensation on tumor detection in SPECT.

Acceptable levels of lossy image compression have been hotly debated for years. When the compressed images will be used by human observers, the evaluation of the compression algorithms must involve the assessment of the impact of the compression on human performance. Observer models can play a significant role in the evaluation of the large number of potential compression algorithms and the many parameters defined by each, provided the model observer predicts human performance. Eckstein *et al.* (1999) found the CHO and the NPWE to correlate well with human observer performance in the evaluation of image-compression algorithm settings. Based on this fact these investigators used a model observer to optimize the quantization parameters of the JPEG algorithm (Eckstein *et al.*, 2000b); the optimized parameter settings were then validated by psychophysical determination of improved human performance.

Human observers are known to be adaptive to noise level and image content, among other things. The CHO is also adaptive, with a decision strategy that changes when the signal or noise characteristics of the images are altered. Rolland and Barrett (1992) demonstrated that the adaptation of the human observer can be predicted by the CHO. In nuclear medicine, increasing exposure time shifts the dominant source of variability in the data from quantum noise toward the contribution from object variability. Rolland showed that human detection performance improves as exposure time increases, providing evidence of the human's ability to incorporate improved quantum statistics into its decision strategy. Similarly, the Hotelling observer's performance increases with increasing exposure. The correlation between the CHO predictions and the human data was extremely high, over several decades of observer performance. Conversely, the NPWMF strategy is not adaptive; the NPWMF applies a template determined by the difference of the signals under each hypothesis without regard for the character of the background. This observer's performance saturates as exposure time increases, failing to predict the performance of the human observer. No nonadaptive model could possibly predict human performance in this study.

A number of studies have extended the body of knowledge regarding CHO performance in random backgrounds. In addition to the work involving lumpy backgrounds of Rolland, Yao and Burgess discussed previously, it has been shown that the CHO correlates well with human performance in the presence of anatomical backgrounds (Eckstein and Whiting, 1995). Abbey and Barrett (2001) measured human-observer performance in several SKE tasks to investigate the effects of regularization and object variability in tomographic image reconstructions. Across a range of experiments that investigated parameters determining the signal profile, exposure time, and data covariance, the channelized-Hotelling observer was most able to predict the array of human data.

Abbey *et al.* (1999) give an elegant theoretical derivation of an unbiased procedure for determining the template of a linear observer for a detection task. The only inputs to the procedure are the images presented to the observer on each trial and the observer's decision as to which image was deemed "signal-present." The procedure requires the means and covariances of the data under each hypothesis. While the template-estimation procedure is applicable to any linear observer, Abbey *et al.* made use of the procedure for estimating the templates of human

observers and comparing them to the templates of model observers. Edwards *et al.* (2000) extended the template-estimation procedure to the case where the noise is a mixture distribution of Gaussians. Recently, Abbey and Eckstein (2001) suggested the use of Bayesian template-estimation methods; the reduction in variance obtained through these methods may outweigh the small bias that also results. These template-estimation methods are pointing the way toward a better understanding of the human-observer's decision strategy. Perhaps in the future they may even find use in the development of improved methods for computer-aided diagnosis (CAD).

14.2.3 Psychophysical methods for image evaluation

Psychophysical methods are used to measure human-observer performance and assess diagnostic accuracy. In this section we shall review the history of the development of ROC methodology as a tool for understanding the visual system and assessing imaging technologies. We shall then describe methods for the conduct of psychophysical studies.[6]

Early applications of ROC analysis ROC techniques were initially developed during World War II for analyzing the performance of radar systems for detecting aircraft. One of the earliest applications of psychophysical methods to a medical application was the work of Garland (1949), who investigated the diagnostic accuracy of roentgenographic and photofluorographic techniques and presented some of the earliest evidence of reader error and variability. The cross-fertilization that brought ROC methods to visual science was greatly facilitated when W. P. Tanner, a graduate student in psychology at the University of Michigan, was assigned a desk in the office of T. G. Birdsall, one of the early pioneers of ROC methods (Cohn, 1993). In 1954, Tanner and J. A. Swets, also of the University of Michigan, published a seminal paper in which statistical decision theory was first applied to the study of visual performance, even including a section entitled, "A new theory of visual detection." This paper demonstrated that the core principles of statistical decision theory were applicable to observer performance. Most notably, the mathematical model of Fig. 13.4 is applicable to human decision variables, and human observers can control their decision criterion and manipulate it in response to information regarding the prior probability of each hypothesis and the decision costs. The paper presented data collected by yes-no and forced-choice experiments and showed them to be consistent.

It took some time for the perception community to relinquish the theory of an absolute detection threshold for "seeing." In 1963, Nachmias and Steinman published an ROC study meant to determine whether humans have a decision criterion that could be altered by directives from the investigator. The paper concluded that the data supported the variable-criterion hypothesis, but did not rule out the absolute-threshold theory entirely. Finally, in 1969, Kratz published an analysis of the Nachmias and Steinman data that concluded that the absolute-threshold theory could be rejected.

[6]We gratefully acknowledge the presentation materials made available to us by Charles Metz for use in writing this section.

While the variable-threshold theory was being established, Swets and his colleagues were working with great gusto at extending the use of statistical decision theory to the study of decision processes in perception (Swets *et al.*, 1961; Swets, 1964; Green and Swets, 1966). Another mathematical psychologist at the University of Michigan, D. D. Dorfman, and his colleague E. Alf, Jr. published a maximum-likelihood method for estimating ROC curve parameters and determining confidence intervals (1968, 1969). In the same timeframe, L. Lusted became the first investigator to apply ROC methods to medicine in general and medical imaging in particular (1968, 1971). Also, in 1960, the First Freiburg Conference on the Neurophysiology and Psychophysics of the Visual System was held, creating a forum for the movement toward combining and correlating information about the visual system derived from electrophysiological investigation with that derived using ROC methods (Jung and Kornhumber, 1961). This was a time of tremendous growth in methodology and accumulation of data in visual science.

The next decade saw a shift in the center of the ROC universe from the University of Michigan to the University of Chicago, where ROC analysis was applied to a variety of problems in medical imaging. Metz *et al.* demonstrated the relationship between ROC analysis and Shannon's information theory (1973) and published a tutorial on the basic principles of ROC analysis for a medical imaging audience (1978). Goodenough (1975) made use of an L-alternative forced-choice paradigm. Starr *et al.* (1975) investigated the detectability of low-contrast disks and spheres on uniform backgrounds in radiography. The early work of Starr *et al.* was one of the first of a set of studies that together demonstrated the limitations of using single measures of imaging system performance like resolution as a measure of image quality. It was also one of the first investigations of the effect of search-region size on ROC curves.

In the 1980s, the advent of relatively inexpensive, fast computers enabled the development and dissemination of free software for curve-fitting of ROC data, making ROC analysis much more widely utilized for image evaluation.[7] Software for statistical testing also became available. There is now a wide variety of free packages available for the analysis of data acquired under a variety of experimental paradigms and providing an assortment of possible model fits, as well as the statistical comparison of results across imaging systems, observers, and tasks.

The last decade has seen continued development of numerical tools for the analysis of ROC data, the generalization of ROC methods to more complex and clinically relevant tasks, and a significant increase in the utilization of ROC-based methods for studies of image evaluation and observer performance. In the next sections we shall describe in more detail how ROC experiments are designed and performed and the methods for data analysis that are available to investigators today. Our purpose is not to provide a complete "how-to" manual, but rather to give an idea of the many options available to the investigator and the relevant literature where more specific experimental and analytical tools can be found.

The yes-no experiment A single point on an ROC curve can be determined for a given observer on a given binary-classification task using a simple "yes-no" experiment. In each experiment, an observer is presented with a set of images one at a

[7]Some twenty years later, the number of registered users of the free Metz software package is close to 4000!

time, and the observer responds either "yes – the signal is present" or "no – the signal is absent." (More generally, "yes – class 2 is true" or "no – class 2 is not true.") By tabulating the fraction of true and false responses at the end of the experiment, a single point on the ROC curve is determined. By instructing the observer to use a different mindset on each of a set of yes-no experiments, a set of points on the ROC curve is found, as depicted in Fig. 13.5. The finer the curve desired, the more yes-no experiments that must be performed.

Rating-scale approach Swets *et al.* (1961) showed that a single rating-scale experiment gives equivalent ROC estimates to that obtained via the inefficient process of repeated yes-no experiments. The rating-scale approach involves the presentation of single images to the observer at a time, with each image presentation referred to as a "trial." The data collected on each trial is the observer's certainty that the image belongs to class 2. Table 14.1 gives an example rating scale.

There are many variations on this theme. At one extreme, class 2 can be defined by the presence of an exactly-specified object at an exact location, while at the other extreme it can encompass the presence of any pathology of any kind, with the range of object variability in the middle ground. The number of rating levels can be as few as 5, although 6 or 7 is more commonly encountered, or the experiment can use a continuous rating scale. The use of a continuous rating scale, first advocated by Rockette *et al.* (1992) and validated by King *et al.* (1993), allows for finer distinctions of certainty levels by the observer, and a smaller chance of degenerate data (where the cells of the rating scale are not fairly evenly distributed with responses) in the analysis stage (Wagner *et al.*, 2001). However, some investigators shy away from this method because of a concern that diligent observers will find it difficult to report their rating so finely, and the concern that the intra-observer variability will be increased (the likelihood that the observer will rate the same case at the same level on two independent trials will be infinitely small).

A current controversy is the use of *action scales* like the BI-RADS scale (ACR, 1998) for classification of mammographic images. Action scales incorporate patient management as well as the reader's level of suspicion. Some investigators have recommended that a pure probability-of-disease rating be acquired in addition to an action rating to avoid the bias that can occur when using an action rating alone for ROC purposes.

Table 14.1 Example rating scale

Rating	Description of certainty level
1	Object is definitely a member of class 1
2	Object is likely to belong to class 1
3	Object is equally likely to belong to class 1 or class 2
4	Object is likely to belong to class 2
5	Object is definitely a member of class 2

Relationship to contrast-detail (CD) diagrams An early paradigm for image assessment was the *contrast-detail* approach. In this method the observer is shown an image containing multiple signals with a range of contrasts and sizes. The observer reports the smallest detectable signal at each contrast. A plot of the detection-contrast versus size (detail) is then generated. When a set of CD diagrams are plotted as a function of exposure or dose, it is termed a CDD diagram (Cohen *et al.*, 1981).

There are several difficulties with the CD-diagram approach. One is that the approach is subjective because it does not control for the observer's variable decision criterion; different observers can be lax or strict in their judgment and even the same observer's criterion for "seeing" the signal can vary. Also, there is no ability to correct for "wishful thinking" on the part of the observer, and without signal-absent locations there is no ability to determine the trade-off with false-positive responses. Thus, while the CD-diagram approach is routinely used as a quality-assurance protocol in many imaging applications, it is not recommended as a quantitative tool in the assessment of imaging systems unless the aforementioned concerns are addressed in the study.

The CD-diagram can be of use when machine observers are used in place of humans. Then the observer's threshold can be set to a fixed level, and the algorithm can be forced to evaluate both signal-present and signal-absent locations. Chakraborty and Eckert (1995) have developed a procedure for the machine evaluation of phantom images for use in the evaluation of image quality.

Forced-choice experiments We first encountered the forced-choice (FC) experimental paradigm in Sec. 13.2.5. In a forced-choice experiment, an observer is forced to make a decision in favor of one of the alternative hypotheses. In the binary-classification task, a pair of images is presented to the observer, either at the same time or sequentially, one from class 1 and the other from class 2. The order/placement of the images is randomized, and usually there is no restriction on viewing time. The observer must decide which alternative belongs to class 2. As derived in Sec. 13.2.5, the percentage of correct responses in a two-alternative forced choice (2AFC) experiment equals the area under the ROC curve. We shall have more to say on this when we discuss the analysis of forced-choice data in Sec. 14.2.4.

The generalization of the FC paradigm to the L-alternative task requires the observer to state which of L-alternative signals is present in a signal-present image, or which of L regions contains a specified signal, for example.

Generalized ROC methods In a classic ROC experiment designed to evaluate an SKE binary classification task, the observer records a single rating on each trial describing his or her certainty that the image belongs to class 1 or class 2 (see Table 14.1). The application of this experimental paradigm to more realistic problems in which the signal is not known exactly is problematic. When there are multiple possible signals, or multiple locations, a single probability score does not capture all the data available from the observer. Most notably, the observer may indicate a high certainty that a signal is present in a signal-present image, but in fact the observer may have missed the true signal and be responding to a noise-only location that is perceived to be signal. Without requiring the observer to provide location data along with the probability rankings, there is no ability to correct for this effect.

An alternative is to require the observer to point to the signal that is detected on an image, and rate the probability that it is there. A *localization ROC* (LROC) curve is a plot of the actually positive images detected with the lesion correctly localized vs. the fraction of actually negative images falsely called positive (Swensson, 1996). The x axis of an LROC curve is thus the same as in a conventional ROC plot. On each image there is either a single signal at an unknown location or there is no lesion. An example LROC curve is shown in Fig. 14.4.

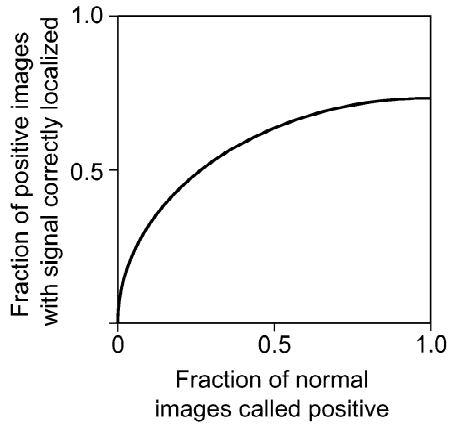

Fig. 14.4 An example LROC curve.

Free-response ROC curves (FROC) were introduced by Bunch *et al.* (1978) to enable the detection-and-localization analysis of images with an arbitrary number of signals. An FROC curve is a plot of the fraction of lesions detected vs. the average number of false-positive detections per image. FROC curves are often used in the assessment of CAD algorithms, where the number of false positives can be high. An example FROC curve is shown in Fig. 14.5

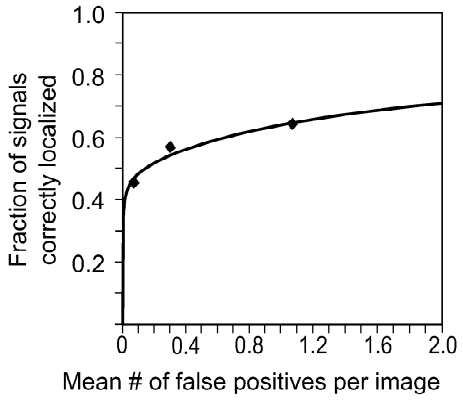

Fig. 14.5 An example FROC curve.

The *alternative free-response ROC* or AFROC curve is a plot of the fraction of lesions detected vs. the fraction of actually negative images falsely called positive (Chakraborty and Winter, 1990). An actually negative image is included in the

fraction of those called positive whenever one or more false-positive locations are identified on it. The x axis of an AFROC curve is similar to the x axis of ROC and LROC curves, only now the ability to mark more than one location on an image is allowed. An example AFROC curve is shown in Fig. 14.6.

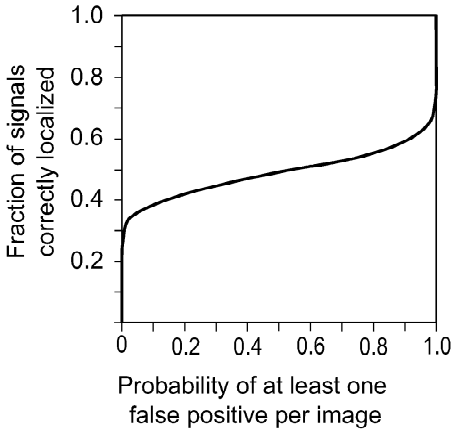

Fig. 14.6 An example AFROC curve.

14.2.4 Estimation of figures of merit

Once the experimental procedure has been run and the observer-response data have been collected, the question arises as to how best to analyze the data. In this section we shall describe methods for the analysis of ROC and related data, and the estimation of figures of merit for summarizing image quality for classification tasks.

Analysis of conventional ROC data The foundation for the analysis of ROC-like data is the analysis of conventional rating data. The simplest approach to the generation of an ROC curve from rating data is to determine the number of true- and false-positive responses associated with each rating level. For a rating scale with N levels, this will give a graph with $N-1$ TPF-FPF pairs, plus the $(0,0)$ and $(1,1)$ anchors for the plot. An *empirical* ROC curve is obtained by "connecting the dots" to generate a stairstep plot consisting of vertical and horizontal line segments obtained by adding true and positive responses to the curve as a threshold is swept across the response data. For continuous rating data, the AUC estimate obtained by integrating the area under an empirical ROC curve is a Wilcoxon statistic (Bamber, 1975).

When the number of rating levels is small, the empirical ROC curve will be jagged, but a fitting approach can yield a smooth estimate of the curve. However, the assumptions of conventional least-squares curve fitting are invalid, owing to the joint dependence of the ratings on the observer's mindset (Metz, 1986a). Hence a maximum-likelihood estimation procedure must be used instead. Dorfman and Alf (1968) published the first ML solution to the analysis of rating data.

The most widely used ML method assumes that the data underlying the ratings take on a parametric form with adjustable parameters under each hypothesis. Fig. 13.4 helps to make this concept more concrete: the fitting procedure assumes

the distributions for the decision variable conditioned on each hypothesis take on a particular form, most commonly a Gaussian, and the goal of the estimation procedure is to estimate the parameters of the two distributions given the rating data. The assumption that the two distributions are Gaussians is the so-called *binormal model* (Swets *et al.*, 1961). The binormal model does not limit the decision-variable data to Gaussian distributions; all that is required is that the data obtained under each hypothesis be transformable to Gaussian distributed random variables by the same unknown transformation.

The binormal model yields an estimated ROC curve with a straight-line plot on double-probability paper; the axes are given by the normal deviates z_{TPF} and z_{FPF} [*cf.* (13.15)]. The conventional ML procedure estimates the slope and intercept of this line, which can be related to the difference in means and the ratio of the variances of the two underlying distributions (Dorfman and Alf, 1969). Methods for obtaining ML binormal fits to ROC curves are also available for continuous rating data (Metz *et al.*, 1998b).

A large number of experiments have demonstrated the validity of the binormal model (Swets, 1986; Metz, 2000). Hanley (1988) has shown that ROC curves obtained from ML parameter fits to a variety of non-Gaussian distribution models, including binomial, Poisson, gamma, χ^2, rectangular, and triangular forms, are indistinguishable from the Gaussian-based curve, provided the number of data samples is large. Nevertheless, the standard binormal model can result in fitted ROC curves that cross the chance line and have a slope that does not decrease monotonically as the FPF increases (Berbaum *et al.*, 1990). These so-called "improper" ROC curves can be obtained from the standard binormal model when the number of cases is low, the data scale is discrete, or the operating points are not well distributed (degenerate data). To avoid this outcome, the "proper" ROC analysis was introduced by Dorfman *et al.* (1996). Proper ROC curves are constrained from crossing the chance line. Proper models based on bigamma distributions (Dorfman *et al.*, 1996) and binormal distributions (Pan and Metz, 1997; Metz and Pan, 1999) have been investigated. The proper binormal model transforms the data by forming the likelihood ratio associated with the two underlying normal distributions; the result is an ROC curve with a monotonically decreasing slope. An objection to this procedure is that the calculation of the likelihood ratio is not something that the actual observer under test is hypothesized to do. Rather, it is an additional transformation applied to the observer data and thus may not be representative of the observer to which the ROC curve applies.

An alternative fitting approach for ROC rating data is the "contaminated" binormal model (Dorfman *et al.*, 2000a). This model assumes the distribution of decision variables is the bimodal sum of two Gaussians under the signal-present alternative (Dorfman and Berbaum, 2000b). The model has been found to be useful in the analysis of data with small false-positive fractions and to give results very similar to those of the standard binormal fitting procedure for nondegenerate data (Dorfman and Berbaum, 2000c).

Analysis of forced-choice data We described the 2AFC experiment mathematically in Sec. 13.2.5 as one in which the observer is presented with two images $\mathbf{g}$ and $\mathbf{g}'$, where $\mathbf{g}$ is drawn from $\text{pr}(\mathbf{g}|H_1)$ and $\mathbf{g}'$ is drawn from $\text{pr}(\mathbf{g}|H_2)$. The images are presented simultaneously in different spatial locations or separately in time at the same location. The assignment of the two underlying sources of images to the two

presentation locations is randomized. The observer's task is to choose the image from class 2. To make the decision, the observer computes two test statistics $T(\mathbf{g})$ and $T(\mathbf{g}')$, and the data vector that gives the higher value is assigned to H_2. This assignment is correct if $T(\mathbf{g}') > T(\mathbf{g})$. By (13.39), the probability of a correct decision on any trial is AUC. Viewed this way, AUC is a criterion-free parameter-free distribution-independent figure of merit for a classification task (Massof and Emmel, 1987).

To estimate AUC from a forced-choice experiment, the percentage of correct decisions over a large number of trials is determined. To keep score of the number of correct responses, let n_i take on the value 1 for a correct response on trial i and 0 for an incorrect response. Mathematically, $n_i = \text{step}[T(\mathbf{g}_i') > T(\mathbf{g}_i)]$, where the subscript i denotes the i^{th} trial. Thus n_i is a Bernoulli random variable (see Sec. C.6.1). Over N trials, the AUC estimate is the proportion of correct responses (PC):

$$\widehat{\text{AUC}} = \text{PC} = \frac{1}{N}\sum_{i=1}^{N} n_i \,. \tag{14.14}$$

If we assume the response variables are independent from trial to trial, (14.14) is the sum of N i.i.d. Bernoulli random variables. From (C.159) we know that the summand must be a binomial random variable with parameters N and the true AUC. It is well known that (14.14) is an unbiased ML estimate of AUC.

The early work of Tanner and Swets (1954) demonstrated the consistency of data collected in yes-no and forced-choice experiments. While an FC experiment yields an estimate of AUC, it has the disadvantage of not providing any information regarding the shape of the underlying ROC curve. Burgess (1985b) compared ROC and FC experimental methods and concluded that ROC methods make more efficient use of the available images, giving AUC estimates with lower variance, while FC methods make more efficient use of observer time. In an effort to make more efficient use of the available images, some experimenters use a multiple-pass paradigm in which different images from each hypothesis are paired for presentation in each pass. It can be shown that the full ROC curve can be obtained in the limit of every image being paired with every other. Note that the response variables are no longer independent Bernoulli random variables in this case.

Analysis of generalized-ROC experiments The primary advantage of the generalized-ROC approaches described above is their applicability to tasks in which signal uncertainty, usually location uncertainty, plays a key role. Many advances have been made in the last decade toward the development of robust procedures for the analysis of generalized ROC data from LROC, FROC, and AFROC experiments.

Maximum-likelihood methods have been introduced for the fitting of LROC data (Swensson, 1996). An ROC curve can be obtained from LROC data. Swensson (1996) gives the following relationship between the area under the ROC curve and the area under the LROC curve: $A_{\text{LROC}} = 0.5(\text{AUC} + 1)$.

ML analysis tools for FROC data have been introduced by Chakraborty (1989). The procedure assumes that the underlying signal and noise distributions are Gaussians and the number of false-positive responses per image follows a Poisson distribution. The assumptions underlying the analysis of FROC and AFROC are detailed and their validity argued thoroughly in a recent book chapter by Chakraborty (2000), who also suggests that the FROC analysis gives estimates of system perfor-

mance with greater statistical power than those obtained using conventional ROC analysis.

While the use of localization brings a significant degree of reality to the task, compared to the classical ROC experimental design, there is also the added requirement for deciding what region around a signal will be considered a "true-positive" response in the analysis. The choice for this tolerance is arbitrary; yet it has ramifications on the results of the data analysis. There is yet no consensus on the appropriate localization tolerance to use in the analysis of LROC, FROC, and AFROC experiments.

There are few resources for the analysis of more complex experiments involving multiple hypotheses. Kijewski *et al.* (1989) developed an analysis procedure for determining the parameters specifying ROC curves between all pairs of classes in an L-class problem given ratings of the multiple alternatives. Still needed is a practical method for the analysis of multi-alternative tasks using ROC analysis.

Summary measures Once a satisfactory fit to the ROC rating data has been obtained, summary measures of performance can be derived. Global measures of system performance include the AUC, the detectability measure d_A obtained from AUC via (13.21), or the parameters of the binormal model. If it is known that a certain operating point on the ROC curve is more significant for the intended use of the system than others, a local measure of performance might be reported at that operating point; that is, the TPF at a given FPF or the FPF at a given TPF might be reported. Partial area measures giving either the area to the right or below the ROC curve from a specified operating point give a regional measure of performance (McClish, 1989). Jiang *et al.* (1996) provided an extension to partial-area index analyses for systems with high AUC. Finally, if sufficient information is available regarding the cost and benefit of decisions is known, these can be reported at the optimal operating point, or a full cost/benefit curve can be reported and summarized. There are many open questions regarding the best approach to summarizing performance. The AUC is the most widely used figure of merit today.

The ML theorem of (13.378) enables us to say something about the ML estimates of other performance measures based on the ML estimate of AUC. For example, an ML estimate of the observer SNR can be derived from $\widehat{\text{AUC}}_{\text{ML}}$ by inverting (13.20). An ML estimate of the observer's squared SNR can be obtained by similar reasoning and used in an ML estimation of observer efficiency.

When more than one observer has participated in an ROC study, there are two options for deriving an overall figure of merit. The first is to derive estimates of the binormal model parameters for each observer and average the parameters. The second is to pool the rating data and then perform the ML estimation procedure. Metz (1986b) has discussed the advantages and disadvantages of these alternatives. When multiple observers are used in the evaluation of multiple imaging systems, correlations in the data result. Tools for the analysis of multiple-reader, multiple-case experiments are discussed below.

Error analysis and the comparison of imaging performance When a measure of imaging performance is obtained, it is natural to ask how large the error bars are about that estimate. Moreover, when two imaging systems are being compared, we seek methods for determining the significance of the difference between figures of merit for the competing systems. In Sec. 13.1.1 we discussed a number of drawbacks to

the use of statistical tests of the null hypothesis. These same drawbacks are equally applicable to tests of the null hypothesis using estimated figures of merit for imaging systems.

In his 1920s work on estimation, R. A. Fisher (see Sec. 13.3.6) set the stage for randomized clinical trials by discussing randomized experiments in agricultural research. The analysis of independent imaging modalities was firmly established in the late 1970s, when the National Cancer Institute funded a contract to J. A. Swets and R. A. Pickett of Bolt, Beranek and Newman to develop methods for the assessment of diagnostic technologies. The outcome was a landmark text that presented computer code for the analysis of ROC data, including an analysis of the error in the the estimate of AUC for a single modality (1982).

Soon after, Metz and Kronman (1980) and Hanley and McNeil (1982) proposed methods for the comparison of ROC curves for which the data were assumed to be independent. In 1983, Hanley and McNeil extended their work to the situation where the data were obtained from the same set of patients. In 1984, Metz *et al.* provided a method for analyzing differences between ROC curves measured from correlated data. Differences could be given in terms of the difference in AUC, the TPF at a specified FPF, or the parameters of the standard binormal model. Nonparametric methods for comparing the areas under correlated ROC curves based on Wilcoxon statistics have been presented by DeLong *et al.* (1988) and Campbell *et al.* (1988). These early methods for estimating the uncertainty in AUC and comparing ROC curves took into account the variability in the data resulting from the measurement noise and the object variability sampled by the finite set of cases but did not describe or compensate for the contribution from observer variability.

Observer variability is a complex, multivariate phenomenon that was understood in principle as early as the text by Swets and Pickett (1982), which contains two chapters on the subject. Observers respond differently to different cases, and even the same observer's responses are not 100% correlated across repeated readings of the same data set. As we have seen, a reader's response depends on the latent decision criterion, but it also depends on the observer's training, experience, age, fatigue, and other factors. Readers have different skill levels, and some readers are better at some case sets or modalities than others. An excellent example of reader variability due to differences in decision criteria is contained in data published by Elmore *et al.* (1994) and the subsequent commentary by D'Orsi and Swets (1995). Beam *et al.* (1996) have published the largest study to date demonstrating radiologist variability in skill level and decision criterion in the case of mammographic interpretations.

The first practical multivariate method for the analysis of the variance in AUC estimates for correlated tests with the assumption that both observers and images (readers and cases in the medical literature) are random effects was the multireader, multi-case (MRMC) method of Dorfman, Berbaum and Metz (Dorfman *et al.*, 1992), now commonly referred to as the DBM MRMC method. The method makes use of a jackknife procedure to generate multiple estimates of AUC, each derived by leaving out one of the observations and analyzing those that remain. The results of each leave-one-out procedure are termed *pseudovalues*. An analysis of the variance in the pseudovalues gives an estimate of the variance in the estimate of AUC. By analyzing the statistics of pseudovalues, the contribution to the variance in the estimate of AUC from the cases or the readers can be obtained.

The DBM MRMC method was first developed for the analysis of discrete rating data. Roe and Metz (1997a, 1997b) further developed and validated the DBM method and made software freely available for either continuous or discrete rating data. An alternative, nonparametric method for analyzing the components of variance in ROC studies based on bootstrapping has been suggested by Beiden *et al.* (2000a). Gifford *et al.* (2001) have recently simulated the application of the DBM method to LROC studies and found it to be useful for studies with low numbers of readers and cases.

In some countries, double reading of certain clinical images is the standard, as a method for reducing the number of incorrect interpretations and reducing reader variability. In the U.S., several commercial CAD systems are now available for use as a second reader to the radiologist. The analysis of adjunctive systems requires careful consideration of the appropriate assumptions regarding the variability of the readers (for example, no threshold variability for a computer) and the means for combining their interpretations. The overall performance of the system will be dependent on these considerations. A method for analyzing the improvements in the accuracy of imaging studies derived by repeated observations was suggested by Metz and Shen (1992). Beiden *et al.* (2001a, 2001b) have presented a nonparametric estimate of the components of variance of AUC for comparing two modalities with different variance structures, for example, where one modality involves a CAD adjunctive device and the other does not.

The original MRMC method required every reader to interpret every image in each modality. Recently, the statistical analysis of "partially paired" data sets has been presented (Zhou and Gatsonis, 1996; Metz *et al.*, 1998a).

Ordinal regression Standard ROC methodology reports the performance of a particular observer on a particular task given a specified imaging system. The dependence of the performance measure on a parameter describing the object (size or amplitude, say) or observer (age or number of years of training) would require a series of studies across the range of the parameter of interest. Given the time and cost required for a single psychophysical study, the notion of performing repeated studies of this sort is daunting.

Tosteson and Begg (1988) proposed the use of ordinal-regression techniques for combining studies of multiple object and observer characteristics in a single study. Toledano and Gatsonis (1995, 1996, 1999) have further developed the method and provided extensions for handling incomplete data. The use of ordinal-regression methods in the optimization of imaging system parameters using realistic models for the imaging process deserves greater attention.

Sources of bias We have described methods for analyzing the uncertainty in estimated measures of system performance without mention of possible sources of error in the estimated mean system performance. There are many sources of bias that can creep into the evaluation of an imaging system (Begg and Greenes, 1983). Probably the most significant is the ground-truth problem, which we shall address again in Sec. 14.4.5. It is difficult in real imaging applications to know the true status of an object, be it an enemy aircraft in a reconnaissance image, a stellar object in astronomy, or an unknown feature in a medical image. Knowledge of ground truth can require expensive verification procedures like long-term follow-up, biopsy, or imaging using an alternative system. Thus, in order to know the truth status

required for scoring observer responses in an ROC study, the investigator might design the study with an absence of subtle objects or confounding cases (Rockette *et al.*, 1995, 1998). Without these cases in the study, the results of the study will not describe the performance of the system on these kinds of cases. Similarly, bias can also result from the skill of the observers involved in the study. In the design of the study, the investigator should carefully consider whether to use experts vs. nonexperts and the extent that they represent the intended use of the system.

In summary, the estimated AUC is a joint description of the performance of the imaging system and the population of images and observers used in the study.

Field tests vs. stress tests A *field test* samples the objects and observers as they are expected to be sampled in routine use. A *stress test* limits the objects or the observers (or both) in order to "challenge" the performance of systems where differences are expected. Studies over subpopulations of observers or objects can potentially enable significant differences in system performance to be demonstrated for those subpopulations. For example, it may be that expert and nonexpert radiologists utilize the output of a CAD algorithm differently, but a study that averages over the two sets of readers would possibly miss this important finding. In another example, the fraction of women with dense/heterogeneous breasts is small; a study comparing film-screen to digital mammography using a broad sample of patients might not uncover a significant advantage of one system over the other for that subpopulation of women.

As described in Sec. 13.1.1, the diligent investigator can always increase the number of cases in a study until a statistically significant result is obtained. However, a judicious selection of the cases used in the study can sometimes reduce the number of images required to show significant differences in system performance, by taking into account known differences in the physical performance characteristics of the imaging systems under comparison.

Summary of process Although there are many open areas of research, methods based on ROC analysis are still the best approach for evaluating classification tasks performed by human and model observers. Before beginning a psychophysical investigation, a few questions should be considered. The first is the nature of the classification task — will standard ROC methods suffice, or is a generalized method that incorporates localization/search needed? Consideration should be given to the need for realism and the adequacy of the information that will be gained, the available methods for data analysis as well as methods for statistical analysis.

Careful consideration should be given to sampling issues for images and readers, recognizing the impact these will have on the conclusions that can be drawn from the study.

The specific viewing conditions should be considered, including the degree of observer adaptation, the display settings, the observation distance, and so on. The human-machine interface is critical; small numbers of observer mistakes due to a poor interface can impact the data appreciably.

It is recommended that a block design be used to avoid image-order effects. Observers can read a subset of the images representing one imaging system, then another, then back to the first, until the entire set under all conditions has been read. Randomize the ordering across readers. Do not expect observation sessions to last more than about an hour, or fatigue can degrade observer performance.

Pilot studies can be used to determine the imaging conditions that will yield the best study power and highest efficiency of observer effort. Staircase methods have been described for determining the object contrast that will give high statistical power (Watson and Pelli, 1983; Watson and Fitzhugh, 1990). These methods adjust the object contrast iteratively over a sequence of 2AFC trials to find the signal contrast that yields a SNR_{human} of ~ 0.75 by decreasing the signal amplitude when the observer is correct, and increasing it after decision errors. Once the contrast has been determined that corresponds to that level of performance, the final study can make use of a fixed signal at that contrast — the method of constant stimulus — to give a more precise measure of system performance for that stimulus. Alternatively, the method of ordinal regression allows the evaluation of system performance across a range of stimuli.

Once the data are collected, they can be analyzed by the chosen fitting method. Free software packages are readily available on the worldwide web for this purpose. Then the figure of merit can be estimated along with its confidence interval. Tests for the differences between estimates of figures of merit are also included in several of the freeware packages.

14.3 MODEL OBSERVERS

Model observers serve many purposes. They can be used as tools in the study of the human visual system; by comparing the results of psychophysical studies to model-observer performance measures, researchers gain insights into human perception that can lead to improved models of the human visual system. Such studies give information regarding what tasks the human performs well and what image characteristics impact the human most significantly, potentially leading to improved imaging system designs for generating images for human interpretation. Model observers can also act in place of humans, or in concert with them, in which case we refer to the model as a computer-aided diagnosis (CAD) system.

Above and beyond the use of model observers as tools for understanding the visual system, model observers are an extremely valuable tool in the objective assessment of image quality. Model observers that operate on raw images or detected data enable the objective evaluation and optimization of image acquisition systems. Model observers designed to operate on reconstructed or processed images are useful for the assessment of image-processing algorithms without lengthy human-observer experiments. Because our emphasis in this text is on the design and evaluation of imaging systems, and not on the development of a better understanding of human perception, we shall focus on the use of model observers for the purpose of OAIQ — the objective assessment of image quality — in what follows.

Many of the same statistical methods used to evaluate imaging systems with human observers are applicable to the evaluation and comparison of model observers. The goal of this section is to present methods for determining the performance of a model observer for a given study. As we shall see, the particular model observer chosen and the method of determining its figure of merit will depend on the task as well as the extent to which we know or can characterize the statistics of the data.

We shall begin in Sec. 14.3.1 with a brief review of selected model observers for classification tasks and the requirements for determining each model observer's

performance. We shall then discuss how one chooses a model observer for system evaluation based on classification performance. Sec. 14.3.2 deals with the particular issues involved in the determination of classification performance by linear model observers. The determination of performance measures for ideal observers is the subject of Sec. 14.3.3. Finally, in Sec. 14.3.4, we shall discuss the use of estimation tasks in the objective assessment of image quality.

14.3.1 General considerations

Structure of observer models All model observers used in the objective assessment of imaging systems have a similar structure, illustrated in Fig. 13.1. As described in Sec. 13.2, for classification tasks every model observer computes a scalar test statistic t of the form

$$t = T(\mathbf{g}) \,, \tag{14.15}$$

where $\mathbf{g}$ might be either the raw data or a processed image and $T(\mathbf{g})$ is the observer's discriminant function. A decision is made in favor of hypothesis H_2 if t is greater than some threshold; otherwise H_1 is selected. By determining the number of images classified correctly for all threshold settings, an ROC curve can be generated.

The performance of the model observer can then be summarized using some metric related to the ROC curve. The area under the ROC curve (13.18) and the detectability d_A derived from AUC via (13.21) are common figures of merit. Alternatively, the SNR associated with the test statistic (13.19) can be determined from the first- and second-order statistics of t as a measure of the separability of the data from the two classes. The SNR and the detectability are the same when the test statistic is Gaussian under the two classes.

Categories of observer models for classification Observers can be classified according to whether $T(\mathbf{g})$ is optimal or suboptimal and whether it is a linear or nonlinear function of $\mathbf{g}$. By definition, optimal observers are the best possible in some sense. The Bayesian or ideal observer makes optimal use of all available information in the data and any additional nonimage information to achieve the highest AUC attainable. The ideal observer's test statistic is the likelihood ratio. In general, the ideal observer's discriminant function is a nonlinear function of the input data.

The Hotelling observer is the ideal linear observer; this observer's discriminant function is optimal in the sense that it achieves maximum SNR amongst all linear observers. The AUC-optimal linear observer is another privileged linear observer; as the name suggests, this observer employs the linear discriminant that achieves the highest possible AUC of all linear discriminants for the task.

Because of the large dimensionality of modern images, it may be necessary to make use of "efficient" features, or channels, that preserve the information in the data while enabling the determination of observer performance. Sec. 13.2.12 describes a method for deriving information-preserving linear features from an analysis of the known first- and second-order statistics of the data. Thus we can design a channelized Hotelling observer (CHO) such that it is still the optimal linear observer in spite of the reduced dimensionality.

The objective assessment of image quality may involve suboptimal model observers, particularly when the goal is to predict human performance. From Sec. 14.2 we know that the human observer has been modelled as an observer that processes images through frequency-selective and orientation-selective channels. The chan-

nelized Hotelling observer has been shown to be a useful predictor of the human observer for a variety of tasks, where the channels in this case are not efficient, but are instead chosen to predict human performance. Alternatively, more mechanistic models of the human visual system might be employed as surrogates for the human. These "anthropomorphic" models can incorporate highly nonlinear building blocks such as adaptive gain and contrast nonlinearity. Such models reduce the dimensionality of the data and incur an information loss as well.

Table 14.2 summarizes the types of model observers that can be employed in the objective assessment of image quality.

Table 14.2 Classification of observer models used in OAIQ

	Optimal	Suboptimal
Nonlinear	Ideal observer	Nonlinear model of human
Linear	Hotelling (max-SNR) CHO (efficient) AUC-optimal linear	CHO (visual channels, internal noise)

Computation vs. estimation As noted in Sec. 14.2, the goal of a psychophysical experiment is the *estimation* of human performance from a finite sample of images. This is in contrast to the methods presented in Chap. 13, which addressed the *computation* of ensemble performance measures for model observers. In this chapter we are concerned with the issues that arise when limited data are available for the estimation of observer performance. As we shall see, we might estimate the model observer's decision function from finite samples, and use that function to estimate the model's performance from the same or another set of finite data. Alternatively, a finite data set might be utilized to estimate the statistics of the data under competing hypotheses, with this information then used to estimate a figure of merit for the model observer's performance directly.

Why OAIQ is easier than pattern recognition While the objective assessment of image quality has striking similarities to classical pattern recognition, the two problems are significantly different. Whenever we evaluate an imaging system we do so in terms of a particular task and a specific observer performing the task; thus we have considerable prior information regarding the objects to be classified and the discriminant function to be utilized. In many circumstances we can make use of a signal-known-exactly task, where the background might be simulated or might be a real clinical background. In contrast to most pattern recognition problems, we also have tremendous knowledge of the physics and statistics of the imaging system under evaluation that we can exploit to simulate noise-free training images. Thus the mean data under each hypothesis is fairly easily determined. Furthermore, the noise PDF $\text{pr}(\mathbf{g}|\mathbf{f})$ is usually known from the physics; hence the noise covariance matrix is also known. With this information we are well positioned for determining a linear observer's discriminant function. Note also that we can avoid the gold-standard problem to be discussed in Sec. 14.4.5 by using simulated images; then we always know the underlying truth status of each image.

While we might estimate the model observer's template and evaluate the model observer's performance from finite data, the feature-extraction step is not ad hoc. It is dictated by the statistics of the data. If the purpose of the study is the prediction

of human performance, the features are further dictated by physiology — a channel model representing the visual system is then used as well.

In OAIQ the amount of prior information we bring to bear on the problem is tremendous relative to various approaches found in pattern recognition and data mining, where the statistics of the data may be completely unknown, the features are unspecified, and even the number of classes is uncertain. Moreover, OAIQ often makes use of simulated images, so there is no limitation to the number of images available, and there is no issue about their true classification.

Basic equations describing the ideal observer As derived in Sec. 13.2.6, the ideal observer achieves maximum AUC, maximum TPF at any FPF, and minimum Bayes risk. The ideal observer's test statistic is the likelihood ratio, given by

$$\Lambda(\mathbf{g}) \equiv \frac{\mathrm{pr}(\mathbf{g}|H_2)}{\mathrm{pr}(\mathbf{g}|H_1)}\,. \tag{14.16}$$

To classify a data set, the ideal observer compares $\Lambda(\mathbf{g})$ to a threshold.

Alternatively, the ideal observer forms the log-likelihood ratio, given by

$$\lambda(\mathbf{g}) \equiv \ln[\Lambda(\mathbf{g})] = \ln\left[\frac{\mathrm{pr}(\mathbf{g}|H_2)}{\mathrm{pr}(\mathbf{g}|H_1)}\right]\,, \tag{14.17}$$

which is then compared to a threshold to classify an image. Because the log-likelihood ratio is a monotonic function of the likelihood ratio, the AUC of the ideal observer is unchanged by this transformation.

Data needed for ideal-observer studies We see from (14.16) or (14.17) that the computation of the ideal observer's performance requires full knowledge of the probability density function for the data under the competing hypotheses. In general, these are high-dimensional functions, describing the full joint statistical behavior of M data values. There are well-known examples for which the ideal-observer's performance is calculable, most notably the SKE case in Gaussian noise (Sec. 13.2.8) and some non-Gaussian noise models as well (Sec. 13.2.9). However, for random signals and backgrounds, (see Secs. 13.2.10 and 13.2.11), the ideal observer's decision variable takes the form of an integral of huge dimensionality over the posterior density of the data conditioned on known signals and backgrounds. In Sec. 14.3.3 we shall consider various techniques for estimation of the ideal observer's performance.

Basic equations describing linear observers We may not be able to evaluate ideal-observer performance because of the computational complexity or because we simply do not have the statistical information required to use those tools. Or, we may not want to estimate the performance of the ideal observer because the goal of the assessment process is the prediction of human, rather than ideal, performance. For these reasons the assessment effort may focus on the estimation of the performance of linear model observers.

In a binary classification problem, an arbitrary linear discriminant computes a scalar test statistic t from the $M \times 1$ data vector $\mathbf{g}$ using a transformation of the form

$$t = \mathbf{w}^t\mathbf{g}\,, \tag{14.18}$$

where $\mathbf{w}$ is an $M \times 1$ template. The observer classifies each data set by comparing the value of t to a threshold. The statistics of t determine the performance of the

observer, as measured by AUC or SNR_t. When t is Gaussian-distributed, AUC and SNR_t are related according to (13.20). Given that the linear observer's test statistic is a linear weighted sum of many random variables, the Gaussian assumption for the PDF of t is usually valid as a result of the central-limit theorem.

Optimal linear (Hotelling) observer When the ensemble mean and covariance for $\mathbf{g}$ are known, the observer that maximizes SNR can be derived according to the procedure presented in Sec. 13.2.12. By (13.177) the Hotelling observer's template is known to be

$$\mathbf{w}_{Hot} = \mathbf{K}_\mathbf{g}^{-1} \Delta\overline{\mathbf{g}}\,, \tag{14.19}$$

where $\mathbf{K}_\mathbf{g}$ is the ensemble data covariance, assumed to be the same under each hypothesis, and $\Delta\overline{\mathbf{g}}$ is the difference in the mean data vector under the two hypotheses. The assumption of equal data covariance under each hypothesis is a reasonable approximation for weak signals, even though the signals may be random under each hypothesis. The subscript $\mathbf{g}$ on the covariance matrix, which refers to the raw data, is required because we shall later encounter covariance matrices that describe channel outputs, which will be subscripted accordingly. It can be seen that the Hotelling test statistic is the output of a prewhitening matched filter operation that attempts to compensate for all contributions to the correlations in the data.

The performance of the Hotelling observer is given by [*cf.* (13.178)]

$$\text{SNR}_{Hot}^2 = \Delta\overline{\mathbf{g}}^t \mathbf{K}_\mathbf{g}^{-1} \Delta\overline{\mathbf{g}} = \text{tr}\left[\mathbf{K}_\mathbf{g}^{-1} \Delta\overline{\mathbf{g}}\, \Delta\overline{\mathbf{g}}^t\right]\,. \tag{14.20}$$

In the SKE detection problem this expression simplifies to

$$\text{SNR}_{Hot}^2 = \mathbf{s}^t \mathbf{K}_\mathbf{g}^{-1} \mathbf{s}\,, \tag{14.21}$$

where we denote the signal to be detected by $\mathbf{s}$ in the data domain. Note that the data covariance matrix is assumed to be the same under each hypothesis in (14.20) and (14.21) because the contributions from background variations and measurement noise are assumed to dominate contributions from signal variability in the random-signal case.

The Hotelling observer achieves maximum SNR of all linear observers. An alternative approach is to determine the template $\mathbf{w}$ that gives maximum AUC of all linear observers. In Sec. 13.2.12 we presented the problem of classification in Poisson noise as an example for which the ideal observer is linear (without actually imposing a linearity requirement); this is the observer that achieves maximum AUC as discussed in the previous section. However the ideal observer is not the linear observer that achieves maximum SNR for this task. There is a much smaller literature on the AUC-optimal linear observer relative to the large literature on the Hotelling or max-SNR observer. In the case of a normally distributed test statistic, these two observers coincide.

Data needed for Hotelling-observer studies We see from (14.20) and (14.21) that computation of the performance of the Hotelling observer requires knowledge of the ensemble first- and second-order statistics of the data under each hypothesis. When information regarding the mean and covariance of $\mathbf{g}$ is unavailable, we must resort to procedures for estimating the performance of the optimal linear observer from samples.

The difference in the class means under each hypothesis, $\Delta\overline{\mathbf{g}}$, is an $M \times 1$ vector, where each element $\Delta\overline{g}_m = \overline{g}_{2m} - \overline{g}_{1m}$ is the difference in the average value in the m^{th} pixel in the image or data set under the two hypotheses. Its estimate can be obtained by determining the sample mean from sets of images known to be from each class; the behavior of the sample mean as an estimator is well-understood. Moreover, in many studies the signal is simulated and nonrandom, so that (14.21) is relevant and no estimation of the mean is required. Thus the determination of the mean data under each hypothesis is not a major stumbling block in most applications.

The most daunting issue in imaging applications is the determination of an estimate of $\mathbf{K_g}$, which we shall denote $\widehat{\mathbf{K}}_\mathbf{g}$. A natural inclination is to assume that $\widehat{\mathbf{K}}_\mathbf{g}$ is the sample covariance matrix, but the reader is cautioned against acting on this impulse. If the number of image samples, N_s, is less than the number of pixels in each image, M will be singular and noninvertible. This option therefore requires $N_s \geq M$.

Consider the number of elements of a covariance matrix to be estimated in typical imaging scenarios. A flat-panel digital x-ray imager can have 1024×1024 elements. A SPECT system with a 128×128 detector that collects data over 64 projection angles has the same number of elements. Thus these systems have a data vector with $\sim 10^6$ data elements, so $\mathbf{K_g}$ is a $10^6 \times 10^6$ matrix with about a trillion elements. The symmetry of this matrix allows us to reduce the number of elements to be estimated by about a factor of 2, but a half trillion is still a large number.

The linear discriminant based on sample means and covariances for the pixels in the raw data set is the approach commonly referred to as the Fisher discriminant. Because the number of values to be estimated to form the sample covariance matrix is almost always far greater than the number of samples available for the estimation procedure, the Fisher discriminant is rarely a useful estimate of the optimal linear discriminant in imaging applications.

We shall discuss several alternative approaches to the estimation of the Hotelling observer's performance in Sec. 14.3.2.

Basic equations describing channelized linear observers Any linear channel model can be represented by a matrix-vector multiplication like the one given in (14.7):

$$\mathbf{v} = \mathbf{U}^t\mathbf{g}\,, \tag{14.22}$$

where $\mathbf{U}$ is an $M \times P$ matrix whose columns are the channel profiles $\mathbf{u}_p$, and $\mathbf{v}$ is the $P \times 1$ vector of channel outputs. The $\mathbf{u}_p$ represent the channel profiles, which we have assumed to be real. Each channel output v_p is a number.

While both (14.7) and (14.22) represent a reduction of the dimensionality of the data set, the critical difference is that we are free to choose the channel profiles in (14.22) to suit our purpose. The channels could be designed to be efficient, giving minimal loss of detectability and thereby providing an estimate of the separability inherent in the data. Alternatively, the channels could be designed to estimate the separability of the data after processing through visual-system channels to predict human performance, which may or may not be efficient depending on the task. A number of possible channel profiles used in the vision literature are described in Sec. 14.2.

The performance of a channelized observer is given by the SNR associated with the channel outputs under each hypothesis:

$$\mathrm{SNR}_{\mathbf{v}}^2 = \Delta\overline{\mathbf{v}}^t \mathbf{K}_{\mathbf{v}}^{-1} \Delta\overline{\mathbf{v}} = \Delta\overline{\mathbf{g}}^t \mathbf{U}[\mathbf{U}^t \mathbf{K}_{\mathbf{g}} \mathbf{U}]^{-1} \mathbf{U}^t \Delta\overline{\mathbf{g}}\,. \tag{14.23}$$

Data needed for channelized-observer studies The information required to evaluate the performance of a channelized observer is the first- and second-order statistics of the data *as seen through the channels*. We see immediately from (14.23) that the channel covariance matrix to be inverted is much smaller than the data covariance matrix. If the number of channels is P, then $\mathbf{K}_{\mathbf{v}}$ is a $P \times P$ matrix, where P can be as small as 3 to 6. Even if P is 30 to 50, the matrix to be inverted is still a reasonably manageable size.

The second advantage to the use of a channelized model is the flexibility we have in choosing the channel profiles. As we shall see, prior knowledge of the characteristics of the signal and background can suggest particular forms for efficient channels. Alternatively, the channels can be chosen to model human performance. Given the nontrivial time required to perform psychophysical evaluations, the ability to evaluate a set of imaging system parameters using a model that predicts human performance can offer significant advantages.

Which model observer? The question of which model observer to employ is answered by the objective of the evaluation study. If the goal is to evaluate or optimize the hardware of the data acquisition system, then the ideal observer is the model of choice. Optimization with this observer will result in a system with the maximum information in the raw data in the sense of being able to perform the specified task. If it is not possible to compute ideal-observer performance because the calculation of the likelihood of the data under each hypothesis is not tractable, then the ideal linear observer is a useful alternative for use in hardware evaluation and optimization.

When the task is the evaluation of image-processing algorithms, ideal observers are of no use, because they are invariant to invertible image processing (see Sec. 13.2.6). Image processing algorithms, reconstruction methods and display devices exist for presenting images to human observers; thus the appropriate model should be one that predicts human performance. The model might be a highly detailed, mechanistic model of the visual system or a simpler linear channel model like the CHO.

In the next subsections we discuss in more detail each of these model observers and methods for estimating their classification performance.

14.3.2 Linear observers

In this subsection we shall present a number of approaches for determining the performance of linear model observers from finite data sets. We shall first consider the Hotelling observer that makes use of the raw data and describe several methods for estimating the SNR of this observer. As suggested by (14.20) and the discussion that followed, the estimation of this Hotelling observer's SNR must involve some method for dealing with (or circumventing) the need to estimate the inverse of $\mathbf{K}_{\mathbf{g}}$. Once we have exhausted our list of possible approaches to this problem, we shall explore methods that invoke dimensionality-reducing linear channels.

In many instances, image quality can be ascertained through a classification task involving nonrandom signals that are added to real or simulated backgrounds. Thus we shall first assume that the problem is the detection of a known signal, the so-called SKE problem, while allowing for a random background. In this case there is no need to estimate $\Delta\overline{\mathbf{g}}$ in (14.20); it is known, and our goal is to find an estimate of the SNR given in (14.21). This objective is only hampered by the fact that $\mathbf{K_g}$ is unknown. Subsequently, we shall consider methods for estimating linear-observer performance for random signals.

We shall then briefly discuss the characteristics of the estimated figures of merit. Finally, the subsection concludes with a short discussion of methods for determining the AUC-optimal linear observer. Throughout this subsection we make the assumption that the truth status of each image sample is known; methods for dealing with the no-gold-standard problem are presented in Sec. 14.4.5.

Nonrandom signals We consider the object to be the sum of a known signal and a random background according to the decomposition introduced in (8.306):

$$\mathbf{f} = \mathbf{f}_s + \mathbf{f}_b \,. \tag{14.24}$$

In the detection task, $\mathbf{f}_s$ is zero under H_1. The backgrounds are assumed to be random and drawn from the same ensemble under each hypothesis.

From (14.24), the mean data for a fixed object and a linear imaging operator $\mathcal{H}$ can be written as a linear superposition of signal and background [*cf.* (8.352)]

$$\overline{\mathbf{g}}(\mathbf{f}) = \mathcal{H}\mathbf{f}_s + \mathcal{H}\mathbf{f}_b \equiv \mathbf{s} + \mathbf{b} \,, \tag{14.25}$$

where $\mathbf{b}$ is the image of the particular background realization.

Without signal variability, the covariance matrix $\mathbf{K_g}$ describes the randomness in the data due to background variability and measurement noise. It can be written formally in terms of an expectation of the covariance of the data about the mean taken first over the noise for a single background, followed by an average over all backgrounds:

$$\mathbf{K_g} = \langle\langle(\mathbf{g} - \overline{\mathbf{g}})(\mathbf{g} - \overline{\mathbf{g}})^t\rangle_{\mathbf{n}|\mathbf{b}}\rangle_{\mathbf{b}} \,. \tag{14.26}$$

In the absence of object variability the data covariance matrix reduces to the noise covariance matrix, an entity that is usually known or computable through our knowledge of the image-formation process. Nonrandom backgrounds can be very useful in the validation of software intended to simulate realistic noise properties of an imaging system. However, the objective evaluation of imaging systems in the absence of object variability can yield misleading conclusions; thus image evaluation should employ a random background model if at all possible. Sec. 14.4 describes a number of approaches for simulation of random objects and images.

We have cautioned against the use of sampling methods to directly estimate the sample covariance matrix, and the use of exactly-specified backgrounds in the objective assessment of image quality. How, then, to simplify the calculation of Hotelling SNR in the presence of a random background? One assumption that is often made is that the background is stationary.

Stationarity? A stationarity assumption is attractive because the covariance matrix is then diagonalized by an appropriate Fourier transformation. For example, we

know from Sec. 7.4.4 that a circulant covariance matrix that satisfies $K_{\mathbf{m}\mathbf{m}'} = K_{[\mathbf{m}-\mathbf{m}']_M}$ (where the subscript indicates modulo-M arithmetic in both components of the multi-index) is diagonalized by a discrete Fourier transform. And from Sec. 8.2.8 we know that an infinite covariance matrix that satisfies $K_{\mathbf{m}\mathbf{m}'} = K_{\mathbf{m}-\mathbf{m}'}$ for all $\mathbf{m}$ and $\mathbf{m}'$ is diagonalized by a discrete-space Fourier transform. Following diagonalization by Fourier methods, $\mathbf{K}_\mathbf{g}^{-1}$ can be found by taking the reciprocal of each diagonal element.

While Fourier methods based on stationarity assumptions may seem attractive, this approach is fraught with problems. Real covariance matrices are neither infinite nor circulant. The assumption that $\mathbf{K}_\mathbf{g}$ is circulant implies digital wrap-around, meaning the statistical correlation of two pixel values representing adjacent detector elements is assumed to be equal to the correlation of two elements on opposite sides of the detector, or even in different projections. In an investigation of image quality in digital radiography, Pineda and Barrett (2001) have shown that stationarity assumptions can give misleading results.

Local stationarity Requiring stationarity of any sort over the whole image field is not only unrealistic, it is also unnecessary if our goal is to compute the SNR of a spatially localized lesion. Since (14.21) is the norm of the vector $\mathbf{K}_\mathbf{g}^{-1/2}\mathbf{s}$, we can compute it by summing over only those pixels for which the vector is substantially different from zero. Typically, in direct imaging systems, those pixel elements correspond to a restricted region in data space. If so, we can express the SNR in terms of the Wigner distribution function computed over this region, as discussed in Sec. 13.2.13.

For indirect imaging systems, a spatially localized lesion can contribute to a very nonlocalized set of detector elements. In this situation it is unlikely that an assumption of approximate stationarity would hold over the entire region for which $\mathbf{K}_\mathbf{g}^{-1/2}\mathbf{s}$ is significantly greater than zero. Thus for tomographic systems it is necessary to perform a reconstruction first to restore the local nature of the signal to be detected and allow the use of methods that invoke an assumption of approximate stationarity. The argument of the previous paragraph holds if we let $\mathbf{g}$ be the reconstruction and $\mathbf{s}$ be the reconstructed signal.

If $\mathbf{K}_\mathbf{g}$ were diagonal (in the multi-indices[8]), the region where approximate stationarity is required would be the same as the subset of pixels for which $\mathbf{s}$ is nonzero, but a nondiagonal covariance means that some elements of $\mathbf{K}_\mathbf{g}^{-1/2}\mathbf{s}$ are nonzero even if the corresponding elements of $\mathbf{s}$ are zero. Moreover, the range of the correlations is only a rough guide to selecting the correct subset of pixels; the matrix $\mathbf{K}_\mathbf{g}^{-1/2}$ can occupy a substantially larger band around the diagonal than $\mathbf{K}_\mathbf{g}$.

We do not know the width of this band if we cannot compute $\mathbf{K}_\mathbf{g}^{-1/2}$, but we can proceed experimentally. If we start with a measured covariance matrix, or one computed on a realistic nonstationary model, we can select an $L \times L$ subset of it centered on the signal location. Calling this matrix $\mathbf{K}_L$, we can compute $\mathbf{s}^t\mathbf{K}_L^{-1}\mathbf{s}$, which would be an estimate of the Hotelling SNR without any stationarity assumption if we were given only this subset of the data. We can then vary L and observe the behavior of this SNR; when it no longer changes, we can assume that we have

[8]See Sec. 8.2.8 for a discussion of discrete random processes and diagonality in multi-index notation.

found the band containing the nonzero elements of $\mathbf{K}_\mathbf{g}^{-1/2}$, and we can compare the resulting SNR to that computed with the Wigner distribution function. If agreement is good, we can use the Wigner expression to compute SNR for a variety of signals and all positions in the field and to define local NEQ and DQE as functions of spatial frequency and signal location (see Sec. 13.2.13). This approach may result in a number of nonzero elements in need of estimation that is small enough that the finite number of image samples can support their estimation.

Decomposition of the covariance matrix Another approach is to make use of our knowledge of the physics of the imaging process, which often gives us powerful information regarding the distribution of data for a fixed object. Statistically speaking, we often know $\text{pr}(\mathbf{g}|\mathbf{f})$, from which we can determine the conditional mean $\overline{\mathbf{g}}(\mathbf{f})$ and the conditional covariance $\mathbf{K}_{\mathbf{n}|\mathbf{f}}$.

Key to making use of this prior information is a decomposition of the overall data covariance given in Sec. 8.5.3; we know from (8.347) that $\mathbf{K}_\mathbf{g}$ is the sum of two terms, written

$$\begin{aligned}\mathbf{K}_\mathbf{g} &= \left\langle \left\langle [\mathbf{g}-\overline{\mathbf{g}}(\mathbf{f})]\,[\mathbf{g}-\overline{\mathbf{g}}(\mathbf{f})]^t \right\rangle_{\mathbf{n}|\mathbf{f}} \right\rangle_\mathbf{f} + \left\langle \left[\overline{\mathbf{g}}(\mathbf{f})-\overline{\overline{\mathbf{g}}}\right]\left[\overline{\mathbf{g}}(\mathbf{f})-\overline{\overline{\mathbf{g}}}\right]^t \right\rangle_\mathbf{f} \\ &= \langle \mathbf{K}_{\mathbf{n}|\mathbf{f}} \rangle_\mathbf{f} + \mathbf{K}_{\overline{\mathbf{g}}} \equiv \overline{\mathbf{K}}_\mathbf{n} + \mathbf{K}_{\overline{\mathbf{g}}}\,, \end{aligned} \tag{14.27}$$

where $\overline{\mathbf{K}}_\mathbf{n}$ represents the noise covariance averaged over all objects. While both $\overline{\mathbf{K}}_\mathbf{n}$ and $\mathbf{K}_{\overline{\mathbf{g}}}$ are influenced by object variability, we emphasize that they are covariances for vectors in data space.

When the signal is random but statistically independent of the background, we can write the covariance matrix for $\overline{\mathbf{g}}$ as [see (8.359)]

$$\begin{aligned}\mathbf{K}_{\overline{\mathbf{g}}} &= \left\langle \left[\overline{\mathbf{g}}(\mathbf{f})-\overline{\overline{\mathbf{g}}}\right]\left[\overline{\mathbf{g}}(\mathbf{f})-\overline{\overline{\mathbf{g}}}\right]^t \right\rangle_\mathbf{f} = \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_\mathbf{f}\boldsymbol{\mathcal{H}}^\dagger \\ &= \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}_\mathbf{s}}\boldsymbol{\mathcal{H}}^\dagger + \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{K}}_{\mathbf{f}_\mathbf{b}}\boldsymbol{\mathcal{H}}^\dagger \equiv \mathbf{K}_\mathbf{s} + \mathbf{K}_\mathbf{b}\,, \end{aligned} \tag{14.28}$$

where $\mathbf{K}_\mathbf{s}$ and $\mathbf{K}_\mathbf{b}$ are the covariance of the data about the conditional mean resulting from signal and object variability, respectively. When the signal is nonrandom (14.28) simplifies to $\mathbf{K}_{\overline{\mathbf{g}}} = \mathbf{K}_\mathbf{b}$. Even in the random-signal case this simplification can be relevant; if the signal is of sufficiently low contrast, then $\mathbf{K}_{\overline{\mathbf{g}}} \approx \mathbf{K}_\mathbf{b}$ because the contribution due to the random background dominates.

Much of what follows on estimation of linear-observer performance is based on the decomposition of (14.27). Though we shall often use the approximation that $\mathbf{K}_{\overline{\mathbf{g}}} \approx \mathbf{K}_\mathbf{b}$, we note that (14.27) itself is exact; it requires no Gaussian assumptions regarding either the objects or the noise, and it does not assume that the noise is object-independent. Alternative forms for the object-variability term that make use of alternative ways of expressing the autocovariance of the object in object space are given in Sec. 8.5.3.

Role of the measurement noise To be more explicit about $\overline{\mathbf{K}}_\mathbf{n}$, we need to distinguish direct from indirect imaging and object-dependent from object-independent noise.

The simplest case is direct imaging with additive Gaussian measurement noise. As discussed in detail in Chap. 12, electronic noise in different detector elements is usually statistically independent and hence uncorrelated. If every detector element has the same noise variance σ^2, which is independent of the object $\mathbf{f}$, then

$$\overline{\mathbf{K}}_\mathbf{n} = \mathbf{K}_{\mathbf{n}|\mathbf{f}} = \sigma^2\mathbf{I}\,. \tag{14.29}$$

Thus $\overline{\mathbf{K}}_{\mathbf{n}}$ is a multiple of the unit matrix and hence full rank.

The situation is only slightly more complicated with Poisson noise. Since Poisson measurements are conditionally statistically independent with variance equal to the mean, we can write

$$\left[\mathbf{K}_{\mathbf{n}|\mathbf{f}}\right]_{mm'} = \overline{g}_m(\mathbf{f})\,\delta_{mm'} = \left[\mathcal{H}\mathbf{f}\right]_m \delta_{mm'}\,, \tag{14.30}$$

where the last form is for a linear digital imaging system characterized by the CD operator $\mathcal{H}$. Averaging over object variability is now straightforward:

$$\overline{\mathbf{K}}_{\mathbf{n}} = \langle \overline{g}_m(\mathbf{f})\rangle_{\mathbf{f}}\,\delta_{mm'} = \overline{\overline{g}}_m\,\delta_{mm'} = \left[\mathcal{H}\bar{\mathbf{f}}\right]_m \delta_{mm'}\,. \tag{14.31}$$

Thus the average noise covariance matrix is diagonal in spite of the object variability, though of course the overall covariance matrix $\mathbf{K}_{\mathbf{g}}$ is not diagonal.

It is not immediately obvious, however, that $\overline{\mathbf{K}}_{\mathbf{n}}$ is full rank. Indeed, the conditional noise covariance $\mathbf{K}_{\mathbf{n}|\mathbf{f}}$ is not full rank if any of the $\overline{g}_m$ is zero. Similarly, $\overline{\mathbf{K}}_{\mathbf{n}}$ is not full rank if any of the $\overline{\overline{g}}_m$ is zero, but this turns out to be of much less concern; the only way a particular $\overline{\overline{g}}_m$ could be zero is if the m^{th} detector element never receives radiation for any object in the ensemble, and in that case we might as well delete that detector element from the data set. Thus we can always assume that $\overline{\mathbf{K}}_{\mathbf{n}}$ is full rank for direct imaging, even with Poisson noise.

For indirect imaging the measurement noise is modified by the reconstruction algorithm. This issue will be discussed at length in the next chapter, but for now we note that analytic expressions for $\mathbf{K}_{\mathbf{n}|\mathbf{f}}$ can be developed, where $\mathbf{n}$ refers to the noise in an image reconstructed by a linear algorithm from either Gaussian or Poisson data (see Sec. 15.4.2). For nonlinear algorithms, analytic covariances are generally not possible, but practical computational methods are available for determining $\mathbf{K}_{\mathbf{n}|\mathbf{f}}$ numerically; for details, see Sec. 15.4.7. These numerical expressions can then be averaged over $\mathbf{f}$ to obtain $\overline{\mathbf{K}}_{\mathbf{n}}$.

Sample averages In Sec. 8.4 we discussed a variety of statistical models for objects and found that there were many circumstances where we could generate samples of $\mathbf{f}$; more discussion of methods for simulating random objects is also given in Sec. 14.4. However, it is usually not possible to determine $\text{pr}(\mathbf{f})$ from samples, and we almost always have to resort to the use of sample averages to determine the statistical properties of the data resulting from random objects.

Consider again the case of nonrandom signals (or where the signal is random but of low contrast), so that $\mathbf{K}_{\overline{\mathbf{g}}} = \mathbf{K}_{\mathbf{b}}$, and we want to estimate this covariance matrix. From Sec. 13.2.12 we know that the data covariance resulting from a general random background is given by

$$[\mathbf{K}_{\mathbf{b}}]_{mm'} = \left\langle (b_m - \overline{b}_m)(b_{m'} - \overline{b}_{m'})\right\rangle_{\mathbf{b}}\,, \tag{14.32}$$

where $\overline{b}_m$ is the mean contribution of the random background to detector element m. This expression describes the fluctuations in the data that would be observed over a large set of simulated or real noise-free images.

One approach to finding an estimate of $\mathbf{K}_{\mathbf{b}}$ is to use a theoretical object model such as a lumpy background (Sec. 8.4) for which the autocovariance function can be specified. This function is then mapped through the blur associated with the imaging system to produce the covariance matrix $\mathbf{K}_{\mathbf{b}}$. If we choose some functional

form (*e.g.*, fractal) for the autocorrelation, we can use sample images to estimate any unknown parameters in the function.

Another approach is to acquire a set of low-noise images and estimate the covariance matrix for the background (in data space) from them. If we do not want to make any assumptions about the form of the autocovariance, we can simply form the sample covariance matrix as a low-rank approximation to the desired ensemble covariance. The samples might be simulated noise-free backgrounds, or they might be experimental background images with low but nonzero noise, obtained with image-averaging or high-dose techniques. Methods for simulating noise-free backgrounds are discussed in Sec. 14.4.

Suppose we have a set of sample background images $\{\mathbf{g}_j, j = 1, ..., N_s\}$, which are either noise-free (simulated) or for which the noise is negligible compared to the effects of object variability (perhaps because the images were acquired with a long exposure time). We can array each of these images as $M \times 1$ column vectors. We can then subtract the sample mean from each image to form the set $\{\delta\mathbf{g}_j, j = 1, ..., N_s\}$, and the covariance matrix $\mathbf{K}_{\overline{\mathbf{g}}}$ can be estimated by

$$\widehat{\mathbf{K}}_{\overline{\mathbf{g}}} = \mathbf{W}\mathbf{W}^t \,, \tag{14.33}$$

where $\mathbf{W}$ is the $M \times N_s$ matrix with columns given by sample images:

$$\mathbf{W} = \frac{1}{\sqrt{N_s}} \left[\delta\mathbf{g}_1,\, \delta\mathbf{g}_2,\, ..., \delta\mathbf{g}_{N_s}\right] . \tag{14.34}$$

The sample covariance matrix of (14.33) is equally applicable when the set of sample images contains random signals of unknown statistical description as well as random backgrounds.

Once the background covariance matrix is estimated, the noise contribution can be determined (if it is not already known) to yield the full data covariance matrix. For example, we can make use of (14.31) to write the covariance of the data in the weak-signal approximation under Poisson measurement noise as

$$\begin{aligned} [\widehat{\mathbf{K}}_{\mathbf{g}}]_{mm'} &= \langle\langle (g_m - \hat{\bar{b}}_m)(g_{m'} - \hat{\bar{b}}_{m'})\rangle_{\mathbf{n}|\mathbf{b}}\rangle_{\mathbf{b}} \\ &= \hat{\bar{b}}_m\, \delta_{mm'} + [\widehat{\mathbf{K}}_{\mathbf{b}}]_{mm'} \,. \end{aligned} \tag{14.35}$$

The first term in the last line is $\widehat{\overline{\mathbf{K}}}_{\mathbf{n}}$, which in this case is a diagonal matrix with elements given by sample averages of the mean background. The second term $\widehat{\mathbf{K}}_{\mathbf{b}}$ is an estimate of the covariance $\mathbf{K}_{\overline{\mathbf{g}}}$ due to the random backgrounds.

Other non-Poisson forms of object-dependent measurement noise can be simulated to generate noisy images once the simulation of random objects and noise-free data sets is achieved satisfactorily. These images can be used to determine the first- and second-order statistics of the data necessary to determine the SNR of the linear observer, using methods described below.

Matrix-inversion tools Once we are assured that we have a covariance matrix with full rank, the next step is to compute the SNR. Given the size of $\widehat{\mathbf{K}}_{\mathbf{g}}$, direct inversion of the estimated covariance matrix to estimate the detectability via (14.20) or (14.21) is not feasible. We shall consider the following alternative approaches, none of which assumes stationarity in any sense:

1. Iterative computation of the template;
2. Neumann series;
3. Matrix-inversion lemma.

Iterative computation When the observer's template is known, it can be applied to a set of sample images (for which the ground truth is known) to yield a set of test statistics under each class that can be used to compute the observer's AUC or SNR. The optimal linear observer has a template given in (14.19); thus it would appear that the determination of $\mathbf{w}_{Hot}$ also requires the inversion of $\mathbf{K_g}$. Not so! Fiete *et al.* (1987) suggested that the Hotelling template could be calculated iteratively.

Finding the template amounts to solving the equation $\mathbf{K_g w} = \mathbf{s}$, where the signal $\mathbf{s}$ is assumed known and $\mathbf{K_g}$ is either known or estimated. This equation is analogous to the imaging equation $\mathbf{Hf} = \mathbf{g}$, where the unknown template takes the place of the unknown object and the covariance matrix plays the role of the imaging system. However, the covariance matrix is square, making it invertible in principle, unlike the system operator in most imaging problems.

The solution can be found by any of the iterative methods enumerated in Chap. 1 or by the regularized methods to be discussed in Chap. 15. One possible solution is given by the Landweber algorithm (1.231), which gives the following template estimates at each iteration:

$$\hat{\mathbf{w}}_{n+1} = \hat{\mathbf{w}}_n + \alpha[\widehat{\overline{\mathbf{K}}}_{\mathbf{n}}]^{-1}[\mathbf{s} - \widehat{\mathbf{K}}_{\mathbf{g}}\hat{\mathbf{w}}_n]\,, \tag{14.36}$$

where n denotes the iteration number and we have made use of the knowledge that the noise contribution to the covariance matrix is full rank. The beauty of this iterative approach is that no inversion of the full $\widehat{\mathbf{K}}_{\mathbf{g}}$ is required.

Once the template has been estimated, the SNR can be found by applying the template to a set of sample images, determining the mean and variance of the resulting scalar test statistic under each hypothesis, and computing the observer's performance via (13.19). Alternatively, we can directly estimate SNR^2 by (14.21) as $\mathbf{s}^t\hat{\mathbf{w}}$.

Neumann series The covariance matrix may not be diagonal in real situations, but it may be nearly diagonal (at least with the multi-index convention). For example, as we shall see in Chap. 16, for direct imaging applications using x rays the nondiagonal contributions to the data covariance are due to correlations in the object statistics and physical processes like escape of K x rays from the phosphor. When these contributions are not very long-range, the Neumann series approach can be advantageous.

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

$$\mathbf{K_g} = \sigma^2\mathbf{I} + \mathbf{A} = \sigma^2\left[\mathbf{I} + \frac{1}{\sigma^2}\mathbf{A}\right]\,, \tag{14.37}$$

where $\mathbf{A}$ describes the off-diagonal elements. Then we can use the Neumann series (A.59) to write the inverse covariance as

$$\mathbf{K}_\mathbf{g}^{-1} = \frac{1}{\sigma^2}\sum_{j=0}^{\infty}\left[-\frac{1}{\sigma^2}\mathbf{A}\right]^j = \frac{1}{\sigma^2}\mathbf{I} - \frac{1}{\sigma^4}\mathbf{A} + \frac{1}{\sigma^6}\mathbf{A}^2 - \ldots\,. \tag{14.38}$$

The Hotelling SNR then becomes

$$\mathrm{SNR}^2_{Hot} = \mathbf{s}^t\mathbf{K}_\mathbf{g}^{-1}\mathbf{s} = \frac{||\mathbf{s}||^2}{\sigma^2} - \frac{\mathbf{s}^t\mathbf{A}\mathbf{s}}{\sigma^4} + \frac{\mathbf{s}^t\mathbf{A}^2\mathbf{s}}{\sigma^6} - \ldots \,. \tag{14.39}$$

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 in (14.39) will converge if

$$\frac{\mathbf{s}^t\mathbf{A}^{n+1}\mathbf{s}}{\sigma^2\mathbf{s}^t\mathbf{A}^n\mathbf{s}} < 1 \tag{14.40}$$

for all n, and it may still converge (because of the alternating signs) even if (14.40) 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 $\mathbf{K_g}$ into a diagonal part $\mathbf{D}$ plus a matrix $\mathbf{A}$ with only off-diagonal terms. Assuming convergence, we then have

$$\mathbf{K_g} = \mathbf{D} + \mathbf{A} = \mathbf{D}\left[\mathbf{I} + \mathbf{D}^{-1}\mathbf{A}\right] \,; \tag{14.41}$$

$$\mathbf{K}_\mathbf{g}^{-1} = \left[\sum_{j=0}^{\infty}\left[-\mathbf{D}^{-1}\mathbf{A}\right]^j\right]\mathbf{D}^{-1} \,; \tag{14.42}$$

$$\mathrm{SNR}^2 = \mathbf{s}^t\mathbf{K}_\mathbf{g}^{-1}\mathbf{s} = \mathbf{s}^t\mathbf{D}^{-1}\mathbf{s} - \mathbf{s}^t\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{s} + \mathbf{s}^t\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{s} - \ldots \,. \tag{14.43}$$

The first term in this expansion, $\mathbf{s}^t\mathbf{D}^{-1}\mathbf{s}$, is what we computed above when we assumed there were no off-diagonal terms, and the remaining terms 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 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 $s^2_\mathbf{n}\,[\mathbf{K}_\mathbf{g}^{-1}]_\mathbf{nn}$, and the first correction term in (14.43) becomes

$$\mathbf{s}^t\,\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{s} = \sum_\mathbf{j}\sum_\mathbf{k} s_\mathbf{n}\,[\mathbf{D}^{-1}]_\mathbf{nj}\,[\mathbf{A}]_\mathbf{j\,k}\,[\mathbf{D}^{-1}]_\mathbf{kn}\,\mathbf{s_n} = \frac{s^2_\mathbf{n}}{D^2_\mathbf{nn}}A_\mathbf{nn} = 0 \tag{14.44}$$

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:

$$\mathbf{s}^t\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1}\mathbf{s} = \frac{s^2_\mathbf{n}}{D^2_\mathbf{nn}}[\mathbf{A}\mathbf{D}^{-1}\mathbf{A}]_\mathbf{nn} = \frac{s^2_\mathbf{n}}{D^2_\mathbf{nn}}\sum_\mathbf{k}\frac{[A_\mathbf{nk}]^2}{D_\mathbf{kk}} \,. \tag{14.45}$$

If we say that $\mathbf{A_{nk}} \simeq 0$ 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 $\mathbf{D_{nn}}$, 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 (14.40) must be satisfied.

Matrix-inversion lemma Suppose we want to invert an overall covariance matrix of the form

$$\mathbf{K_g} = \overline{\mathbf{K}}_\mathbf{n} + \widehat{\mathbf{K}}_{\overline{\mathbf{g}}} = \overline{\mathbf{K}}_\mathbf{n} + \mathbf{W}\mathbf{W}^t\,, \tag{14.46}$$

where we have assumed that $\widehat{\mathbf{K}}_{\overline{\mathbf{g}}}$ is given by (14.33). For electronic or Poisson noise $\overline{\mathbf{K}}_\mathbf{n}$ will be diagonal, but in some applications correlations will be introduced by the detector and $\overline{\mathbf{K}}_\mathbf{n}$ will be a nearly diagonal, banded matrix (see, for example, the discussion of x-ray detectors in Sec. 12.3.8).

By the matrix-inversion lemma (A.56a), we see that

$$\left[\overline{\mathbf{K}}_\mathbf{n} + \mathbf{W}\mathbf{W}^t\right]^{-1} = \overline{\mathbf{K}}_\mathbf{n}^{-1} - \overline{\mathbf{K}}_\mathbf{n}^{-1}\mathbf{W}\left[\mathbf{I} + \mathbf{W}^t\overline{\mathbf{K}}_\mathbf{n}^{-1}\mathbf{W}\right]^{-1}\mathbf{W}^t\overline{\mathbf{K}}_\mathbf{n}^{-1}\,. \tag{14.47}$$

The advantage of this form is that $[\mathbf{I} + \mathbf{W}^t\overline{\mathbf{K}}_\mathbf{n}^{-1}\mathbf{W}]$ is an $N_s \times N_s$ matrix, where N_s is a few hundred in practice, rather than an $M \times M$ matrix, where M may be 10^6. Moreover, since $\mathbf{W}^t\mathbf{K}_n^{-1}\mathbf{W}$ is positive-semidefinite, the inverse of the $N_s \times N_s$ matrix will always exist. Thus, if $\overline{\mathbf{K}}_\mathbf{n}$ 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.[9]

The matrix-inversion lemma reduces the size of the required inverse from $M \times M$ to $N_s \times N_s$ but it is not a dimensionality-reduction method in the sense that it does not entail potential information loss.

Dimensionality reduction using efficient channels The previous approaches depend upon writing the data covariance matrix as the sum of a full-rank, near-diagonal component representing the measurement noise and a low-rank contribution obtained from samples. When we do not have access to low-noise or noise-free samples from which to estimate the second term, an alternative approach is to make use of efficient channels that allow us to estimate the Hotelling observer's SNR in a lower-dimensional space.

In Sec. 13.2.12 we showed that a limited set of features, when properly chosen, preserves the information in the data in terms of yielding the same SNR for a linear observer. We found that an eigenanalysis of the inter- and intra-class scatter matrices results in full preservation of the separability using only $(L-1)$ features for optimal linear discrimination between L classes. In the binary classification problem, a single feature is all that is needed — quite a dimensionality reduction.

The requirement for finding that single privileged feature is that complete knowledge of the scatter matrices is available in order to do the eigenanalysis. Without such complete knowledge, we must judiciously apply whatever prior information we have regarding the signals to be discriminated and the background statistics to find channels that reduce the dimensionality of the problem with limited loss of detectability.

For a particular set of channel profiles, the Hotelling formalism can be applied in the channel space to determine the $\mathbf{w_v}$, which is the vector of optimal channel weights. By analogy with (14.19), we write the template in the channel space as

$$\mathbf{w_v} = \mathbf{K_v}^{-1}\Delta\overline{\mathbf{v}}\,, \tag{14.48}$$

[9]This idea was suggested to us by Brandon D. Gallas (see Barrett *et al.*, 2001).

where $\Delta\overline{\mathbf{v}}$ is the difference in channel outputs under the two hypotheses,

$$\Delta\overline{\mathbf{v}} = \mathbf{U}^t \Delta\overline{\mathbf{g}}\,, \tag{14.49}$$

and $\mathbf{K_v}$ is the $P \times P$ covariance matrix of the channel outputs:

$$\mathbf{K_v} = \mathbf{U}^t \mathbf{K_g} \mathbf{U}\,. \tag{14.50}$$

The SNR on the channel outputs is given by

$$\mathrm{SNR}_{\mathbf{v}}^2 = \Delta\overline{\mathbf{v}}^t \mathbf{K}_{\mathbf{v}}^{-1} \Delta\overline{\mathbf{v}}\,. \tag{14.51}$$

From Sec. 13.2.12 we know that efficient features are ones that preserve the separability of the data in a space of reduced dimensionality, achieving $\mathrm{SNR}_{\mathbf{v}}^2 = \mathrm{SNR}_{\mathbf{g}}^2$.

Laguerre-Gauss channels Consider the example of a detection task in which the detected signal is approximately radially symmetric, centrally peaked and smooth, and situated at a known location on a stationary background with a correlation that has no preferred orientation. With these assumptions it can be expected that the ideal linear template will be centered at the known position of the signal, rotationally symmetric and smooth before discretization to match the CD nature of the imaging system.[10] Laguerre-Gauss channel profiles have been proposed by Barrett *et al.* (1998c) for this task because they form a basis on the space of rotationally-symmetric square-integrable functions in $\mathbb{R}^2$.

The Laguerre polynomials are defined in (4.57) as

$$L_p(x) = \sum_{k=0}^{p} (-1)^p \binom{p}{k} \frac{x^k}{k!}\,. \tag{14.52}$$

The orthogonality relation for these polynomials is given by (4.58):

$$\int_0^\infty dx\; e^{-x} L_p(x)\, L_{p'}(x) = \delta_{pp'}\,. \tag{14.53}$$

We can transform this relationship to a two-dimensional form with the change of variables $x = 2\pi r^2/a_u^2$, where r is the radial distance and a_u plays the role of a scaling factor, giving

$$\frac{1}{2\pi} \int_0^{2\pi} d\theta \int_0^\infty \frac{4\pi r\, dr}{a_u^2} \exp\left(\frac{-2\pi r^2}{a_u^2}\right) L_p\left(\frac{2\pi r^2}{a_u^2}\right) L_{p'}\left(\frac{2\pi r^2}{a_u^2}\right) = \delta_{pp'}\,. \tag{14.54}$$

We see that the exponential factor of (14.53) has been transformed to a Gaussian factor in (14.54). From this equation we can define the Laguerre-Gauss (LG) functions as

$$u_p(r|a_u) = \frac{\sqrt{2}}{a_u} \exp\left(\frac{-\pi r^2}{a_u^2}\right) L_p\left(\frac{2\pi r^2}{a_u^2}\right)\,, \tag{14.55}$$

where the $\{u_p\}$ are orthogonal (without weighting factors) over $\mathbb{R}^2$ by (14.54).

Figure 14.7 shows radial dependencies of the first, third, and ninth LG functions, as well as their 2D forms. In order to apply these continuous functions to a

[10] Of course, a signal defined on a square pixel grid cannot be exactly rotationally symmetric, but we can ignore this problem if the template covers many pixels.

discrete data set, the functions must be sampled on the same grid used to discretize the data.

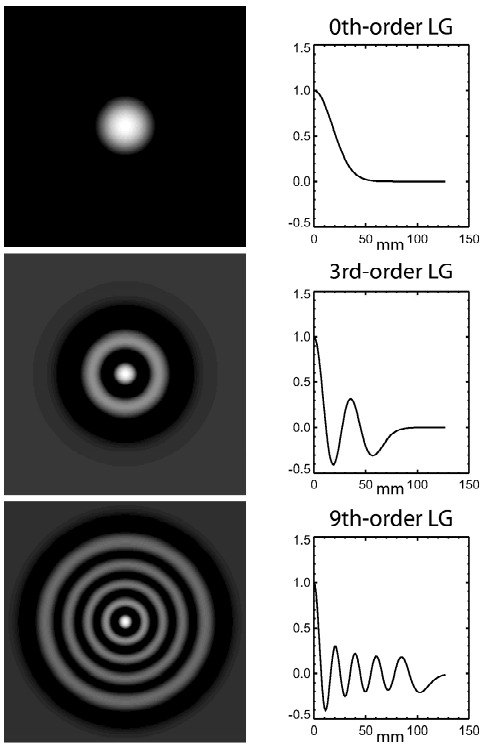

Fig. 14.7 The first, third and ninth Laguerre-Gauss functions: *Left*: 2D functions; *Right*: Radial forms. (Courtesy of Brandon Gallas.)

Because the LG functions form a basis for radially-symmetric functions in 2D, they can be used to exactly represent any rotationally symmetric function $f(r)$ by

$$f(r) = \frac{\sqrt{2}}{a_u} \exp\left(\frac{-\pi r^2}{a_u^2}\right) \sum_{p=0}^{\infty} \alpha_p \, L_p\left(\frac{2\pi r^2}{a_u^2}\right) , \tag{14.56}$$

where

$$\alpha_p = \int d^2 r \; u_p(r) \, f(r) \, . \tag{14.57}$$

Knowledge of the signal and background can be used to choose a_u and estimate the coefficients α_p for a finite set of channels. Alternatively, a range of values for a_u can be investigated, with the number of channels increased until the detectability reaches a maximum over a_u and P. This approach has been investigated extensively by Gallas and Barrett (2003) for an SKE task on a lumpy background with widely varying statistical parameters. These authors found excellent agreement between the channelized linear observer's performance and the ideal linear observer's performance with a small number of channels (5 – 30). The number of channels needed was found to depend on the complexity of the background statistics.

Channel models for predicting human performance While the previous paragraphs specifically address the considerations that come to the fore when using channels to estimate the performance of the optimal linear observer, the advantages offered by the dimensionality reduction of the channelized-Hotelling approach are common to all linear channel models. As described in Sec. 14.2.2, a variety of linear channel models have been proposed for use in the prediction of human performance. All such models have the similar characteristic that they result in a calculable figure of merit for model-observer performance based on dimensionality reduction.

All channels designed to model the human have another similarity: because the human visual system is insensitive to broad, structureless regions, channel models designed to predict human performance have zero response at zero spatial frequency. As stated in Sec. 4.1.4, Laguerre-Gauss functions are eigenfunctions of the 2D rotationally symmetric Fourier operator. Thus the LG channel profiles in the Fourier domain have the same form as the space-domain channels shown in Fig. 14.7. The LG channels are peaked at $\rho = 0$ in the Fourier domain, just as they are peaked at $r = 0$ in the space domain. The LG channels are therefore not recommended for use in modeling human performance.

The body of literature providing the range of applicability of the various candidate channel models for predicting human performance continues to grow. Whenever a given model is utilized, it is important to validate the performance predictions with psychophysical studies involving human observers if the task or the statistics of the data sets are outside the range of experimental conditions for which the model has previously been shown to be predictive of human performance.

Random signals We have described a variety of methods for estimating the Hotelling observer's performance for SKE tasks. Random signals present an additional level of complexity (and realism). Even so, the generalization of the Hotelling approach to random signals is often straightforward. In particular, when the signals are low contrast, we have already stressed that the data covariance is approximately equal to its composition in the SKE case. In that case the only new question that arises is the estimation of the mean data vector that appears in (14.20).

Estimation of the mean data vector If the task is the detection of a random signal, and there is no prior information regarding the signal distribution, it is straightforward to estimate the sample means for the two classes and subtract them to determine $\Delta\hat{\bar{\mathbf{g}}}$. The sample mean is the maximum-likelihood estimate of the true mean. The number of values to be estimated is the number of nonzero elements in the difference $(\hat{\bar{\mathbf{g}}}_2 - \hat{\bar{\mathbf{g}}}_1)$, which is determined by the extent of the signal as seen through the imaging system.

Prior information can be brought to bear on the estimation of the mean difference vector in a number of ways. If the signal is compact and there is prior information regarding its location, this information can be used to limit the number of values to be estimated to those within a certain region of the image. In the case of random signals of a specified shape, prior information regarding the signal's form can be used to reduce the number of parameters to be estimated to a small set, for example, signal amplitude, width, or location. Furthermore, prior information regarding the underlying distributions of the random parameters can be used to form Bayesian estimation procedures according to the theory presented in Chap. 13.

It should be noted that Hotelling SNR may be a poor indicator of system performance with large signal variability, as discussed in Sec. 13.2.12. If a signal 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. 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 nonoverlapping 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 allow signal location to be a parameter in the SNR and compute a detectability map as described next.

Signal known exactly, but variable Let us assume that the signal varies randomly but is known to the observer on each trial (the only uncertainty being whether it is present). This task is sometimes referred to as the signal-known-exactly-but-variable, or SKEV, task (Eckstein and Abbey, 2001; Eckstein *et al.*, 2002). Let the randomness in the signal be captured by a random parameter vector $\boldsymbol{\theta}$. For each value of $\boldsymbol{\theta}$, the optimum linear test statistic is given by [*cf.* (13.208)]

$$\hat{\mathbf{w}}(\boldsymbol{\theta}) = [\widehat{\mathbf{K}}_{\mathbf{g}}(\boldsymbol{\theta})]^{-1}\mathbf{s}(\boldsymbol{\theta})\,, \tag{14.58}$$

where the estimate of $\mathbf{K}_{\mathbf{g}}$ and its inverse must be determined using the methods described above. In particular, the method of template estimation given above may be used to estimate (14.58) without the need for finding an inverse of $\mathbf{K}_{\mathbf{g}}$ in some cases.

The Hotelling SNR can be estimated for each value of the random parameter, following (13.209):

$$\widehat{\mathrm{SNR}}^2_{Hot}(\boldsymbol{\theta}) = \frac{\{[\hat{\mathbf{w}}(\boldsymbol{\theta})]^t\mathbf{s}(\boldsymbol{\theta})\}^2}{[\hat{\mathbf{w}}(\boldsymbol{\theta})]^t\widehat{\mathbf{K}}_{\mathbf{g}}(\boldsymbol{\theta})\,\hat{\mathbf{w}}(\boldsymbol{\theta})} = [\hat{\mathbf{w}}(\boldsymbol{\theta})]^t\mathbf{s}(\boldsymbol{\theta})\,, \tag{14.59}$$

where the second form follows from (14.58).

A summary measure of observer performance can be obtained by averaging (14.59) over $\boldsymbol{\theta}$ if $\mathrm{pr}(\boldsymbol{\theta})$ is known. Alternatively, a detectability map, which plots the SNR^2 as a function of $\boldsymbol{\theta}$, can be presented. Eckstein *et al.* (2002) have found that the optimal parameters for image compression are the same when evaluated using either an SKE or an SKEV paradigm.

AUC and the linear discriminant Thus far we have concentrated on the estimation of the SNR for the Hotelling observer. As discussed in Chap. 13, the Hotelling observer gives maximal SNR and maximal AUC when the data are Gaussian distributed. For non-Gaussian data, the Hotelling observer may not give the best AUC that can be achieved by a linear observer. It is therefore of interest to consider the behavior of AUC for an arbitrary linear discriminant and investigate methods for maximizing this alternative, and arguably superior, figure of merit.

It was shown in (13.44) that

$$\mathrm{AUC}_{lin} = \frac{1}{2} + \frac{1}{2\pi i}\mathcal{P}\int_{-\infty}^{\infty}\frac{d\xi}{\xi}\;\psi_{\mathbf{g}1}(\mathbf{w}\xi)\,\psi^*_{\mathbf{g}2}(\mathbf{w}\xi)\,, \tag{14.60}$$

where $\mathbf{w}$ is the arbitrary $M\times 1$ template of (14.18) that generates the test statistic t from each data vector $\mathbf{g}$, and $\psi_{\mathbf{g}j}(\cdot)$ is the characteristic function for the data under

hypothesis j. Limiting forms of (14.60) for nonrandom signals and for weak signals are given in Sec. 13.2.5. Note that (14.60) is just a 1D integral — only one line through the multivariate characteristic function under each hypothesis is needed once $\mathbf{w}$ is specified.

This formula for AUC is useful when we have analytic forms for the characteristic functions of the data under the two hypotheses. In background-known-exactly (BKE) problems, we might know the characteristic functions directly from the data statistics, but if the background is random we have to first characterize the object statistics and then propagate them into the data domain as discussed in Sec. 8.5.3. If the object is regarded as a continuous function, we need first to obtain an analytic expression for its characteristic functional, then apply (8.335) or (8.339) to obtain the characteristic functions for the data. For example, lumpy and clustered lumpy backgrounds were introduced in Sec. 8.4.4, and their characteristic functionals were derived in Sec. 11.3.10. Additional examples of analytic characteristic functions will be given in Chap. 18.

When the needed characteristic functions are available, an iterative search can be used to maximize the AUC given by (14.60); useful search algorithms are discussed in Sec. 15.4.3. Since the integral is one-dimensional, this search is not particularly computationally expensive.

A major advantage of the approach suggested by (14.60) is that no matrix inversion is required, unlike the determination of the full Hotelling SNR. While an iterative approach can be used to determine the Hotelling SNR when the noise contribution to the covariance matrix is known, it works by searching for the optimum linear template and indirectly obtaining the SNR. An iterative solution for (14.60) directly yields AUC.

The linear discriminant obtained by searching for the $\mathbf{w}$ that maximizes AUC may differ from the Hotelling observer, as discussed in Sec. 13.2.12. When this occurs, the linear discriminant that gives higher AUC is to be preferred whenever our goal is the linear approximation to the ideal observer.

Errors in estimates of SNR for linear observers It is natural to ask how close the estimated SNR is to the true SNR that would have been obtained with full knowledge of the ensemble statistics of the data. That is, we would like to know the bias and variance of the estimate. In this context, bias and variance refer to the first-and second-order statistics of the estimate when different finite sets of images are used. There are several methods, briefly surveyed below, to estimate the magnitude of the bias and variance from this source.

As with any real-world estimation problem, however, there can also be a systematic bias arising from invalid assumptions or modeling errors, and this kind of bias is much more difficult to assess. With computer-generated images, a major source of systematic bias is unrealistic or oversimplified simulation; with real images, a major problem is uncertainty in the true diagnosis. Both of these issues are discussed in Sec. 14.4; here we focus on statistical errors.

It is straightforward to estimate the variance of estimates of SNR or AUC when simulated images are used; all that is needed is to repeat the simulation several times with independent sets of images and compute the sample variance of the values obtained. More sophisticated resampling methods (see below) can also be used, but their only advantage is a saving in computer time, seldom a primary concern these days. In fact, with simulated images the variance and the statistical

bias can be made arbitrarily small simply by running the computer long enough. If real images are used, however, the number of images might be quite limited, and it becomes more critical to estimate the error associated with an estimate of SNR^2.

Errors in direct estimation of SNR in channel space One situation in which we can give not only the bias and variance but indeed the full probability density function of the estimated SNR^2 is when dimensionality reduction is performed with efficient or anthropomorphic channels and the resulting channel outputs are normally distributed. In that case we can estimate SNR^2 by

$$\widehat{\mathrm{SNR}^2} \equiv [\widehat{\Delta\overline{\mathbf{v}}}]^t \, \widehat{\mathbf{K}}_{\mathbf{v}}^{-1} [\widehat{\Delta\overline{\mathbf{v}}}] \,, \tag{14.61}$$

where the hats here denote estimates obtained by sample averages; it is assumed that the number of sample images is larger than the number of channels so that the sample covariance matrix is invertible.

The estimator defined in (14.61) is precisely the one studied by Hotelling in his classic 1931 paper, and it is often referred to as *Hotelling's T^2 statistic*. The PDF of T^2 is closely related to the F distribution; for details see Hotelling (1931) or Anderson (1971). The general behavior of the estimate is illustrated in Fig. 14.8, where it is seen that the estimate is highly biased unless the number of sample images is much larger than the number of channels (and of course it is not even defined if the number of sample images is less than the number of channels).

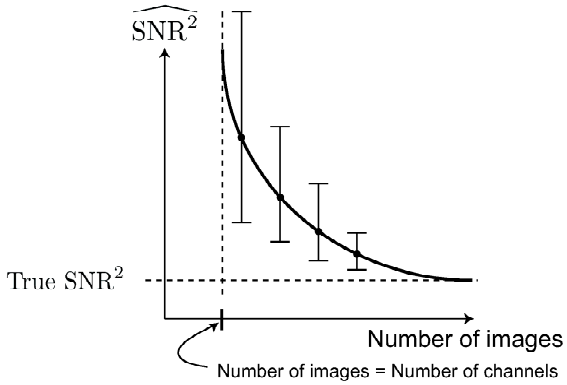

Fig. 14.8 Schematic behavior of the Hotelling T^2 estimate of SNR^2 as defined in (14.61). The dashed horizontal line indicates the true value of SNR^2 on the channel outputs, and the solid curve shows the mean of the estimate. The error bars are indicative of the variance. We thank Andy Alexander for suggesting this kind of plot.

The basic problem with the Hotelling T^2 estimate is that it makes no use of prior information about the quantity being estimated, namely the SNR^2 on the channel outputs. One key piece of prior information in many cases is knowledge of the mean difference signal in data space, $\overline{\mathbf{g}}$, from which we can determine the mean difference signal in channel space, $\Delta\overline{\mathbf{v}}$, by (14.49). If we regard $\Delta\overline{\mathbf{v}}$ as known and nonrandom, we can define a better estimate of SNR^2 by

$$\widehat{\mathrm{SNR}^2} \equiv [\Delta\overline{\mathbf{v}}]^t \, \widehat{\mathbf{K}}_{\mathbf{v}}^{-1} [\Delta\overline{\mathbf{v}}] \,. \tag{14.62}$$

Another key piece of prior information is the covariance decomposition (14.27). If we regard $\overline{\mathbf{K}}_{\mathbf{n}}$ as known and nonrandom and use (14.51), the estimate in (14.62) is modified to

$$\widehat{\mathrm{SNR}^2} \equiv [\Delta\overline{\mathbf{v}}]^t \left[\mathbf{U}^t \left(\overline{\mathbf{K}}_{\mathbf{n}} + \widehat{\mathbf{K}}_{\overline{\mathbf{g}}}\right) \mathbf{U}\right]^{-1} [\Delta\overline{\mathbf{v}}]\,. \tag{14.63}$$

Note that the hat, denoting sample estimates, now appears over only $\widehat{\mathbf{K}}_{\overline{\mathbf{g}}}$, so only that term contributes to the bias and variance of the estimate of SNR^2.

The statistical properties of (14.62) and (14.63) have not yet been derived, but they should offer substantially smaller bias and variance than the T^2 estimate of (14.61) simply because they use more prior information. All of these estimates, however, assume that the channel outputs are normally distributed; it is advisable to plot experimental histograms to check this assumption.

Training and testing An alternative to direct estimation of SNR^2 is first to estimate the template $\mathbf{w}$ and then to apply it to a set of sample images. When only a single, finite set of images is available, the experimenter must use the set of images for two purposes: training the observer (choosing the number of channels, their weights, and any parameters that characterize the channel profiles); and testing the observer (estimating its performance). This is the so-called "training-testing" paradigm. The training-testing label applies even without dimensionality-reducing feature extraction. When we estimate the template from samples by any method, we are training the observer.

There are two common ways to train and test an observer with a single set of sample images. The first option is to split the data into two independent sets, one set to be used to train the observer and the other to be used for testing the observer. The split does not need to be into subsets of equal size. This approach is sometimes referred to as the holdout method. A related method is the use of $N_s - 1$ images to train the observer, with the final sample used to test the observer. This method is known as the round-robin approach; by repeating the training/testing sequence N_s times, keeping score of the observer's decision variable each time, an estimate of the observer's performance is obtained over the entire data set. However, the round-robin method does not yield a single observer, but rather, each held-out image is tested on a different observer.

Gallas (2003) investigated various resampling approaches for determining the bias and variance of the performance estimate for the channelized linear observer trained and tested using variations on the hold-out method. Using a very large set of independent estimates of observer performance (the beauty of Monte Carlo image simulation), Gallas was able to determine the true performance of the channelized observer and thus calculate the bias as well as the variance of the finite-sample methods.

The second training-testing option is the resubstitution method, where the observer is trained and tested on the same set of images. The use of a single set of images to estimate the observer's template, followed by an estimation procedure that applies that template to the data to determine the first- and second-order statistics of t under each hypothesis to derive an SNR, will give an optimistic result (Wagner *et al.*, 1997). The resulting estimates of observer performance correspond to the results obtained via (14.61) and illustrated in Fig. 14.8.

14.3.3 Ideal observers

We learned in Chap. 13 that the ideal observer for binary classification tasks is one that bases its decision on the likelihood ratio. Many properties of the likelihood ratio and its logarithm, and of performance metrics derived from them, were given in Sec. 13.2. In this section we review a variety of approaches to using these often abstract mathematical concepts in the practical assessment of image quality.

Analogies with the Hotelling problem The basic challenge in computing the test statistic for both the Hotelling and the ideal observer is dimensionality. For the Hotelling observer, we need to construct and invert a huge covariance matrix; for the ideal observer, we need to form huge-dimensional multivariate probability density functions and take ratios of them. In neither case are brute-force methods likely to be fruitful; in both cases we must make use of prior information about the task and imaging system in order to make progress.

An important piece of prior information for the Hotelling problem is the conditional covariance $\mathbf{K}_{\mathbf{n}|\mathbf{f}}$, which is known from the physics of the measurement process. For example, $\mathbf{K}_{\mathbf{n}|\mathbf{f}}$ for raw, unprocessed data and Gaussian noise is given in (14.29), and for Poisson noise it is given by (14.30). The analogous prior information for the ideal observer is the conditional PDF $\text{pr}(\mathbf{g}|\mathbf{f})$, which is again known from the physics. Before processing, $\text{pr}(\mathbf{g}|\mathbf{f})$ is often multivariate Gaussian or multivariate Poisson, and in both cases the multivariate PDF can often be written as a product of univariate PDFs. The effect of processing is discussed in Secs. 15.2.6, 15.4.2 and 15.4.7.

In both Hotelling and ideal-observer studies, it is necessary to choose the object model carefully, allowing enough complexity and variability to capture the essence of real objects, yet retaining adequate mathematical tractability. In both cases, object models such as the lumpy and clustered lumpy backgrounds introduced in Sec. 8.4 are very useful.

The signal model, too, can be chosen to facilitate the computation. In particular, nonrandom signals are very attractive, though it remains an open question how well conclusions from SKE studies can be applied to more realistic tasks.

Decomposition of the PDFs The likelihood ratio is the ratio of two PDFs, each of which can be written somewhat abstractly as

$$\text{pr}(\mathbf{g}|H_j) = \int d\mathbf{f}\, \text{pr}(\mathbf{g}|\mathbf{f})\, \text{pr}(\mathbf{f}|H_j)\,, \qquad (j = 1, 2)\,. \tag{14.64}$$

The notation $\text{pr}(\mathbf{f})$ is explained in Sec. 8.2.2 [see especially (8.78) and (8.81)]. In brief, it denotes the density on the full (potentially infinite) set of parameters needed to specify the object as a random process $f(\mathbf{r})$.

The density on the data can also be written as

$$\text{pr}(\mathbf{g}|H_j) = \langle \text{pr}(\mathbf{g}|\mathbf{f}) \rangle_{\mathbf{f}|H_j}\,. \tag{14.65}$$

Numerous alternative forms of $\text{pr}(\mathbf{g}|H_j)$, with various assumptions about the object and the noise, are given in Sec. 8.5.4.

Thus, in order to determine the densities needed in the likelihood ratio, we need both the conditional density on the data for a given object, $\text{pr}(\mathbf{g}|\mathbf{f})$, and the densities $\text{pr}(\mathbf{f}|H_j)$ on the object under the two hypotheses. Note that $\text{pr}(\mathbf{g}|\mathbf{f})$ does

not depend directly on the hypothesis H_j; specifying the object specifies the mean of $\mathbf{g}$, and that in turn specifies the full density in most cases.[11] Note also that we do not refer to $\text{pr}(\mathbf{g}|\mathbf{f})$ as a likelihood since it is never our goal to estimate $\mathbf{f}$; it is not the goal in this section since we are discussing a classification problem, and it is not even the goal in image reconstruction (see Chap. 15).

Conditional PDFs To be more specific about $\text{pr}(\mathbf{g}|\mathbf{f})$, we need to distinguish direct from indirect imaging and object-dependent from object-independent noise, just as we did in Sec. 14.3.2 when we discussed $\mathbf{K}_{\mathbf{n}|\mathbf{f}}$ [see (14.29) and (14.30)].

Consider first the case of direct imaging with a detector array limited by Gaussian electronic noise. If we assume that all elements in the array are identical and that each generates its own noise independently of the other elements, then the probability density function of $\mathbf{n}$ is

$$\text{pr}_{\mathbf{n}}(\mathbf{n}) = (2\pi\sigma^2)^{-M/2} \prod_{m=1}^{M} \exp\left(-\frac{n_m^2}{2\sigma^2}\right) . \tag{14.66}$$

Since the electronic noise is independent of the mean detector output, the conditional density on the data is just a shifted version of the noise density:

$$\text{pr}(\mathbf{g}|\mathbf{f}) = \text{pr}_{\mathbf{n}}[\mathbf{g} - \overline{\mathbf{g}}(\mathbf{f})] = (2\pi\sigma^2)^{-M/2} \prod_{m=1}^{M} \exp\left\{-\frac{[g_m - \overline{g}_m(\mathbf{f})]^2}{2\sigma^2}\right\} . \tag{14.67}$$

For linear systems we can go a step further and write $\overline{g}_m(\mathbf{f}) = [\mathcal{H}\mathbf{f}]_m$.

Similarly, with raw Poisson measurements we have

$$\text{pr}(\mathbf{g}|\mathbf{f}) = \prod_{m=1}^{M} \exp\left[-\overline{g}_m(\mathbf{f})\right] \frac{[\overline{g}_m(\mathbf{f})]^{g_m}}{g_m!} . \tag{14.68}$$

Thus in both of these cases the multivariate density is a product of univariate densities.

The situation is more complicated if we regard $\mathbf{g}$ as the output of some data-processing or image-reconstruction step. Linear processing leaves Gaussian data Gaussian but introduces correlations. Nevertheless, it is straightforward to write down a multivariate expression for $\text{pr}(\mathbf{g}|\mathbf{f})$ since we know how to compute mean vectors and covariance matrices after linear operations, and a multivariate normal is fully specified by its mean and covariance. There is no simple way of expressing $\text{pr}(\mathbf{g}|\mathbf{f})$ after linear processing of Poisson data, but it may be valid to approximate it with a suitably correlated multivariate normal (see Sec. 15.2.6).

Noise on the output of iterative reconstruction algorithms is discussed in Secs. 15.4.2 and 15.4.7. If the algorithm is nonlinear and enforces a positivity constraint, then the noise cannot be Gaussian since negative values cannot occur. Specifically, with multiplicative algorithms such as MLEM (maximum-likelihood expectation-maximization), it often happens that the PDF on the reconstructed image is approximately a correlated log-normal (Wilson *et al.*, 1994; Barrett *et al.*, 1994).

[11] An exception to this statement will be given in Sec. 18.6.4 where we discuss speckle. As we shall see there, in some speckle problems the variance of the data is different for the signal-present and signal-absent hypotheses.

To summarize, with raw, unprocessed data, $\text{pr}(\mathbf{g}|\mathbf{f})$ usually has a simple analytic form (independent Gaussian or Poisson). With processing, the elements of $\mathbf{g}$ are no longer statistically independent, but it is usually possible to give at least an approximate form for the conditional density. In what follows we shall assume throughout that $\text{pr}(\mathbf{g}|\mathbf{f})$ is known analytically.

As a notational point, we see from (14.66) and (14.67) that the conditional density on unprocessed data $\mathbf{g}$ is completely determined by its mean with both the Gaussian and Poisson noise models, so $\text{pr}(\mathbf{g}|\mathbf{f}) = \text{pr}[\mathbf{g}|\overline{\mathbf{g}}(\mathbf{f})]$. The same is true after processing; if we know $\overline{\mathbf{g}}$, we can specify the density on $\mathbf{g}$, and we know $\overline{\mathbf{g}}$ if we know $\mathbf{f}$. We shall therefore write $\text{pr}(\mathbf{g}|\mathbf{f})$ and $\text{pr}(\mathbf{g}|\overline{\mathbf{g}})$ interchangeably, depending on which conditional variable we wish to emphasize.

Object statistics Statistical properties of objects were the subject of Sec. 8.4. The viewpoint adopted there regards the object as a random process for which each sample function is a vector in a Hilbert space. Since we are concerned only with the measurement component of the object, the Hilbert space of interest has a finite but huge dimensionality. We saw a few cases where the object statistics could be specified analytically, for example as a Gaussian random process or a Gaussian mixture, but in most cases analytic models are either unavailable or unrealistic. The two main options in those cases are to reduce the dimensionality of the statistical description of the object or to use a constructive model that allows us to simulate sample objects even if we cannot specify their statistics.

Dimensionality reduction rests on the assumption that somehow the essential features of a complicated random process can be captured with a relatively small number of parameters. As discussed in Sec. 8.4.1, approaches to finding this low-dimensional representation include principal components analysis (PCA) and independent components analysis (ICA). When ICA is applied to images, it is found that the independent components are the outputs of bandpass filters similar to wavelets or the channels in the human visual system; indeed, some have speculated that our visual system has evolved to extract approximately statistically independent components of natural scenes, thereby permitting efficient transformation of information to the brain.

We postulate that there exist similar low-dimensional representations of objects, as opposed to images, and that they again involve bandpass filters or channels. We know from the discussion in Sec. 8.4.3 that the univariate PDFs on the channel outputs have a long-tailed, kurtotic form. Sometimes they are described empirically in terms of the Lévy family, defined not by the density but by the characteristic function, which has the form $\psi(\xi) = \exp(-b|\xi|^q)$. If the channels are chosen so that the outputs are approximately statistically independent, the multivariate object statistics are described by a finite product of characteristic functions of this form.

The constructive models that have received the most attention in image-quality studies are lumpy and clustered lumpy backgrounds. As defined in (8.303), a sample function of a lumpy background is specified exactly by stating the lump positions $\{\mathbf{r}_n\}$ as well as the number of lumps N. The statistical properties are fully specified by giving the probability laws for $\mathbf{r}_n$ and N.

Alternatively, for many constructive models, the object statistics can be specified by giving an analytic form for the characteristic functional associated with the random field. This concept was introduced in Sec. 8.2.3, and the specific forms

for lumpy and clustered lumpy backgrounds were calculated in Sec. 11.3.9. Other constructive models that can be used to synthesize texture fields, and for which analytic characteristic functionals are available, will be introduced in Chap. 18.

To summarize, random objects may be specified by huge-dimensional PDFs, by lower-dimensional PDFs on channel outputs, by rules that let us construct sample functions, and/or by characteristic functionals. In what follows we shall see how each of these descriptions aids us in the computation of ideal-observer performance.

From object domain to data domain If we have either a statistical or a constructive specification of a random object, the next step is to transform it into the data domain. For constructive models, this step is straightforward in principle; one generates the random object and uses it to simulate the random image. Simulation methods are discussed in Sec. 14.4.

To discuss transformation of the PDF, we need to distinguish linear from nonlinear imaging systems. A general rule for nonlinear transformation of bivariate PDFs is given in (C.104), but it does not extend usefully to high-dimensional multivariate problems since the Jacobian cannot be evaluated. So far as the authors can see, there is no hope of transforming an object PDF through a nonlinear imaging system.

The transformation rules for linear systems are most easily expressed in terms of characteristic functions and functionals. If $\overline{\mathbf{g}}(\mathbf{f}) = \boldsymbol{\mathcal{H}}\mathbf{f}$, with $\boldsymbol{\mathcal{H}}$ a linear CD operator, then we know from (8.96) that the characteristic function for the random vector $\overline{\mathbf{g}}$ under hypothesis H_j is given by

$$\psi_{\overline{\mathbf{g}}|H_j}(\boldsymbol{\xi}) = \Psi_{\mathbf{f}|H_j}(\boldsymbol{\mathcal{H}}^\dagger\boldsymbol{\xi})\,, \tag{14.69}$$

where $\boldsymbol{\xi}$ is an $M \times 1$ vector, $\psi_{\overline{\mathbf{g}}|H_j}(\boldsymbol{\xi})$ is the characteristic function for $\overline{\mathbf{g}}$, and $\Psi_{\mathbf{f}|H_j}(\boldsymbol{\sigma})$ is the characteristic functional of the object $\mathbf{f}$, with $\boldsymbol{\sigma}$ being a vector in the same Hilbert space as $\mathbf{f}$, *e.g.*, $\boldsymbol{\sigma}$ corresponds to a function $\sigma(\mathbf{r})$.

If we write $\mathbf{g} = \boldsymbol{\mathcal{H}}\mathbf{f} + \mathbf{n}$ and assume that $\mathbf{n}$ is object-independent, then we know from (8.335) that

$$\psi_{\mathbf{g}|H_j}(\boldsymbol{\xi}) = \psi_{\mathbf{n}}(\boldsymbol{\xi})\,\psi_{\overline{\mathbf{g}}|H_j}(\boldsymbol{\xi}) = \psi_{\mathbf{n}}(\boldsymbol{\xi})\,\Psi_{\mathbf{f}|H_j}(\boldsymbol{\mathcal{H}}^\dagger\boldsymbol{\xi})\,. \tag{14.70}$$

For Poisson noise, (8.339) tells us that

$$\psi_{\mathbf{g}|H_j}(\boldsymbol{\xi}) = \Psi_{\mathbf{f}|H_j}[\boldsymbol{\mathcal{H}}^\dagger\,\boldsymbol{\Gamma}(\boldsymbol{\xi})]\,, \tag{14.71}$$

where

$$[\boldsymbol{\Gamma}(\boldsymbol{\xi})]_m = \frac{-1 + \exp(-2\pi i\,\xi_m)}{-2\pi i}\,. \tag{14.72}$$

Expressions for the data PDFs can be obtained by performing an inverse MD Fourier transform on each expression for $\psi_{\mathbf{g}|H_j}(\boldsymbol{\xi})$.

For signal-known-exactly tasks, it follows from the Fourier shift theorem that

$$\psi_{\mathbf{g}|H_2}(\boldsymbol{\xi}) = \exp(-2\pi i\boldsymbol{\xi}^t\mathbf{s})\,\psi_{\mathbf{g}|H_1}(\boldsymbol{\xi})\,, \tag{14.73}$$

where $\mathbf{s}$ is the nonrandom signal in data space. Thus it suffices to know the no-signal or background-only characteristic function in this case.

Estimation of object statistics In many cases we can express the object statistics in parametric form. For example, with stationary lumpy backgrounds the width of a single lump and the number of lumps per unit area (or volume) fully describe the random process. Similarly, if the object statistics are specified in terms of the outputs of bandpass channels, we could assume that the univariate characteristic function for the n^{th} channel has the Lévy form $\psi_n(\xi) = \exp(-b_n|\xi|^{q_n})$; if we assume further that the channel outputs are statistically independent, then the multivariate object statistics are specified by the sets $\{b_n\}$ and $\{q_n\}$.

We can use the freedom in choosing these parameters to create a wide variety of random object fields. Moreover, if we can estimate the parameters from a training set of real images, we can tailor the object description to a particular physical situation. The problem is that the training set will consist of *images*, and we want to find the parameters for describing *objects*, in spite of the blur and noise associated with whatever imaging system was used to form the images.

A way of estimating the object parameters from blurred, noisy images was devised by Kupinski *et al.* (2003a). They assumed that the object characteristic functional under the no-signal hypothesis was known except for some parameter vector $\boldsymbol{\alpha}$, so it could be written as $\Psi_{\mathbf{f}|H_1}(\mathbf{s};\boldsymbol{\alpha})$. The corresponding characteristic function in the data domain could then be obtained by one of the transformation rules given above; for example, (14.71) applies with Poisson noise, and

$$\psi_{\mathbf{g}|H_1}(\boldsymbol{\xi};\boldsymbol{\alpha}) = \Psi_{\mathbf{f}|H_1}\left[\boldsymbol{\mathcal{H}}^\dagger\boldsymbol{\Gamma}(\boldsymbol{\xi});\boldsymbol{\alpha}\right] . \tag{14.74}$$

Given a set of signal-absent training images $\{\mathbf{g}_n, n = 1, ..., N_s\}$, Kupinski *et al.* formed the *empirical characteristic function* for the data, which is basically a Monte Carlo estimate of $\psi_{\mathbf{g}|H_1}(\boldsymbol{\xi};\boldsymbol{\alpha})$, defined by

$$\widehat{\psi}(\boldsymbol{\xi}) \equiv \frac{1}{N_s}\sum_{n=1}^{N_s}\exp(-2\pi i\boldsymbol{\xi}^t\mathbf{g}_n) . \tag{14.75}$$

The estimation procedure was then basically minimization of the norm of the difference between the known $\widehat{\psi}(\boldsymbol{\xi})$ and the known analytic form $\psi_{\mathbf{g}|H_1}(\boldsymbol{\xi};\boldsymbol{\alpha})$ from (14.74), minimization being carried out by varying $\boldsymbol{\alpha}$. In practice a weighted least-squares norm was used, taking advantage of the fact that all characteristic functions are unity at $\boldsymbol{\xi} = 0$, so the variance of the estimate $\widehat{\psi}(\boldsymbol{\xi})$ is zero at that point. Moreover, a set of channels was applied to each $\mathbf{g}_n$ to reduce the dimensionality and ease the computational burden. For details, see Kupinski *et al.* (2003a).

The beauty of this procedure is that it gives a statistical description of the underlying objects, independent of the imaging system. Thus, even though a particular imaging system, say one described by an operator $\boldsymbol{\mathcal{H}}_0$, was used to obtain the training images, the characteristic function for another system, described by a general $\boldsymbol{\mathcal{H}}$, can be found from (14.74) once $\boldsymbol{\alpha}$ has been estimated. If we can devise a way of computing ideal-observer performance from this information, we can in principle vary $\boldsymbol{\mathcal{H}}$ and optimize the imaging system for the class of objects from which the training set was drawn.

Estimation of the likelihood ratio In an ideal-observer study, the basic quantity to be calculated is the likelihood ratio, defined by

$$\Lambda(\mathbf{g}) = \frac{\text{pr}(\mathbf{g}|H_2)}{\text{pr}(\mathbf{g}|H_1)} = \frac{\int d\mathbf{f}\ \text{pr}(\mathbf{g}|\mathbf{f})\ \text{pr}(\mathbf{f}|H_2)}{\int d\mathbf{f}\ \text{pr}(\mathbf{g}|\mathbf{f})\ \text{pr}(\mathbf{f}|H_1)} . \tag{14.76}$$

The integrals here are over a potentially infinite-dimensional Hilbert space, but they can be reduced to M dimensions (where M is the number of measurements) by observing that $\mathrm{pr}(\mathbf{g}|\mathbf{f}) = \mathrm{pr}[\mathbf{g}|\overline{\mathbf{g}}(\mathbf{f})]$. If we use (8.351) to decompose the object into background and signal parts,

$$\mathbf{f} = \mathbf{f}_b + \mathbf{f}_s \,, \tag{14.77}$$

and (for a linear system) transform the background and signal into data space as

$$\overline{\mathbf{g}} \equiv \mathbf{b} + \mathbf{s} \,, \qquad \mathbf{b} \equiv \boldsymbol{\mathcal{H}}\mathbf{f}_b \,, \qquad \mathbf{s} \equiv \boldsymbol{\mathcal{H}}\mathbf{f}_s \,, \tag{14.78}$$

then we can write the likelihood ratio as [*cf.* (13.166)]

$$\Lambda(\mathbf{g}) = \frac{\int_\infty d^M b \; \mathrm{pr}(\mathbf{g}|H_2, \mathbf{b}) \, \mathrm{pr}(\mathbf{b})}{\int_\infty d^M b \; \mathrm{pr}(\mathbf{g}|H_1, \mathbf{b}) \, \mathrm{pr}(\mathbf{b})} \,. \tag{14.79}$$

A useful alternative form of the likelihood ratio is given by (13.169) and (13.170) as

$$\Lambda(\mathbf{g}) = \langle \Lambda_{\mathrm{BKE}}(\mathbf{g}, \mathbf{b}) \rangle_{\mathbf{b}|\mathbf{g}, H_1} \,, \tag{14.80}$$

where the subscript BKE indicates background-known-exactly, and

$$\Lambda_{\mathrm{BKE}}(\mathbf{g}, \mathbf{b}) \equiv \frac{\mathrm{pr}(\mathbf{g}|H_2, \mathbf{b})}{\mathrm{pr}(\mathbf{g}|H_1, \mathbf{b})} \,. \tag{14.81}$$

The advantage of this form is that $\Lambda_{\mathrm{BKE}}(\mathbf{g}, \mathbf{b})$ is easy to calculate. In fact, for nonrandom signals it is just the ratio of two conditional densities like (14.67) or (14.68). Note, however, that the required average in (14.80) is with respect to the posterior density on the background, $\mathrm{pr}(\mathbf{b}|\mathbf{g}, H_1)$; we shall learn shortly how to do this average by Monte Carlo methods.

One way of evaluating the performance of the ideal observer on a signal-detection task is to generate sets of signal-present and signal-absent sample images, estimate the likelihood ratio of each image and form an ROC curve. Methods discussed in Sec. 14.2.4 can then be used to estimate the area under the curve or ideal-observer AUC.

A useful surrogate for ideal-observer AUC is the likelihood-generating function evaluated at the origin. We know from (13.97) that this quantity is given by

$$G(0) = -4 \ln \left\{ \int d^M g \; [\mathrm{pr}(\mathbf{g}|H_1) \; \mathrm{pr}(\mathbf{g}|H_2)]^{\frac{1}{2}} \right\} . \tag{14.82}$$

We can use $G(0)$ to estimate AUC by [*cf.* (13.20) and (13.96)]

$$\mathrm{AUC} \approx \tfrac{1}{2} + \tfrac{1}{2} \, \mathrm{erf} \left(\sqrt{\frac{G(0)}{2}} \right) . \tag{14.83}$$

If the log-likelihood ratio is normally distributed or $G(0)$ is large (which means that AUC approaches 1), then this result is exact. Clarkson and Barrett (2000) have found it to be an excellent approximation in a variety of cases with practical values of AUC.

Monte Carlo methods Perusal of expressions such as (14.79) or (14.82) shows that computation of ideal-observer performance in nontrivial cases requires evaluation of huge-dimensional integrals. In Sec. 10.4.5 we introduced the concept of Monte Carlo simulation and commented that it was useful in numerical evaluation of multidimensional integrals. As we shall show, Monte Carlo integration is a very valuable tool in ideal-observer evaluations, but in fact we need to move beyond the simple Monte Carlo methods of Sec. 10.4.5 to the more sophisticated and powerful approach of Markov-chain Monte Carlo (MCMC). Book-length treatments of MCMC are given by Robert and Casella (1999) and Gilks *et al.* (1996). We begin here, however, with simple Monte Carlo integration to illustrate the principles and problems.

To evaluate the numerator or denominator in the likelihood ratio as given in (14.76), we must in principle integrate over an infinite-dimensional space, though we could also use (14.79) to reduce it to M dimensions (which is of little consolation if M is of order 10^6). If, however, we can simulate a set of objects $\{\mathbf{f}_n\,, n = 1, ..., N_s\}$, then we can approximate those integrals by [*cf.* (10.300)]

$$\mathrm{pr}(\mathbf{g}|H_j) = \int d\mathbf{f}\ \mathrm{pr}(\mathbf{g}|\mathbf{f})\ \mathrm{pr}(\mathbf{f}|H_j) \approx \frac{1}{N_s}\sum_{n=1}^{N_s} \mathrm{pr}(\mathbf{g}|\mathbf{f}_n)\,, \tag{14.84}$$

where the sample must be drawn from $\mathrm{pr}(\mathbf{f}|H_j)$. That is, if H_2 denotes signal-present and H_1 denotes signal-absent, the simulations must include the signal and background for $j = 2$ but only the background for $j = 1$.

Recall that $\mathrm{pr}(\mathbf{g}|\mathbf{f}_n)$ in (14.84) is a known function, for example given by (14.67) or (14.68). In essence, the Monte Carlo integration associates this known function with every sample point $\boldsymbol{\mathcal{H}}\mathbf{f}_n$ in the data space. The method is thus reminiscent of *kernel estimation*, a technique often used to estimate probability densities from a discrete set of samples. The key difference is that choosing the kernel in kernel estimation is a black art. The kernel must be broad enough to fill in the gaps between samples, yet not so broad as to smooth out essential details in the density being estimated. No such issue arises with (14.84); the form of the kernel is dictated by the physics of the problem, and its width is dictated by the noise level.

This is not to say that (14.84) is a panacea. The kernel $\mathrm{pr}(\mathbf{g}|\mathbf{f}_n)$ falls off rapidly as $\boldsymbol{\mathcal{H}}\mathbf{f}_n$ gets farther from the particular $\mathbf{g}$ for which $\Lambda(\mathbf{g})$ is being calculated. If the noise level is small, most randomly chosen $\boldsymbol{\mathcal{H}}\mathbf{f}_n$ will be so far from $\mathbf{g}$ that $\mathrm{pr}(\mathbf{g}|\mathbf{f}_n)$ will be zero to computer precision, and few of the samples will make any contribution to the sum in (14.84). Even though the sum will asymptotically approach $\mathrm{pr}(\mathbf{g}|H_j)$ as N_s goes to infinity, and the estimator is unbiased for all N_s, the variance can be huge for practical finite values of N_s. The problem gets worse as M gets larger or as the noise level gets smaller.

One way to ameliorate this problem in some cases is by *importance sampling*. Suppose we know an analytic form for $\mathrm{pr}(\mathbf{f}|H_j)$, say as a Gaussian mixture or in terms of independent components. Then we are free to rewrite the data density as

$$\mathrm{pr}(\mathbf{g}|H_j) = \int d\mathbf{f}\ \frac{\mathrm{pr}(\mathbf{g}|\mathbf{f})\ \mathrm{pr}(\mathbf{f}|H_j)}{q(\mathbf{f})} q(\mathbf{f})\,, \tag{14.85}$$

where $q(\mathbf{f})$ is a probability density function (*i.e.*, a nonnegative function normalized to unity) with a support large enough that dividing by zero does not become an

issue. We can then approximate the data density as

$$\text{pr}(\mathbf{g}|H_j) \approx \frac{1}{N_s} \sum_{n=1}^{N_s} \frac{\text{pr}(\mathbf{g}|\mathbf{f}_n)\ \text{pr}(\mathbf{f}_n|H_j)}{q(\mathbf{f}_n)}\,, \tag{14.86}$$

where now the samples are drawn from $q(\mathbf{f})$. For this modification to be useful, we must choose $q(\mathbf{f})$ so that the samples in data space, $\boldsymbol{\mathcal{H}}\mathbf{f}_n$, are clustered near the actual $\mathbf{g}$. In simulation studies, we can do this by taking advantage of the knowledge of how we produced $\mathbf{g}$ in the first place. If we did so by simulating some particular object $\mathbf{f}_0$, then we know what this object was and can use this knowledge in computing the likelihood ratio. For example, if we describe objects by their independent components, with expansion coefficients $\{\alpha_k\}$, then the initial object $\mathbf{f}_0$ is described by $\{\alpha_{k0}\}$, and the importance sampler can generate random objects by random perturbations about $\{\alpha_{k0}\}$. So long as the perturbations are large enough to adequately sample the integrand, the sum in (14.86) is still an unbiased estimator of $\text{pr}(\mathbf{g}|H_j)$, and the variance is greatly reduced by using the prior knowledge of the point about which to take samples. A related approach, suggested by Zhang *et al.* (2001a), is to draw the samples from $\text{pr}(\mathbf{g}|\mathbf{f})$, renormalized as a density on $\mathbf{f}$.

Markov-chain Monte Carlo Direct Monte Carlo integration as sketched above has limited applicability because of the need for an analytic form for $\text{pr}(\mathbf{f}|H_j)$ in the importance sampler. A more general technique is MCMC, which will be discussed in the context of image reconstruction in Sec. 15.4.8. As we shall see there, the essence of MCMC is to propose random perturbations in the vector that is the variable of integration, and to accept or reject the proposed perturbations with a carefully chosen rule such that the sequence of accepted perturbations forms a Markov chain, and the equilibrium PDF for the chain is precisely the one from which we wish to sample.

For ideal-observer studies, MCMC is particularly applicable to the expression for the likelihood ratio given in (14.80). A Monte Carlo implementation of this formula is

$$\Lambda(\mathbf{g}) \approx \frac{1}{N_s} \sum_{n=1}^{N_s} \frac{\text{pr}(\mathbf{g}|H_2, \mathbf{b}_n)}{\text{pr}(\mathbf{g}|H_1, \mathbf{b}_n)}\,, \tag{14.87}$$

where the samples $\mathbf{b}_n$ are drawn from the posterior $\text{pr}(\mathbf{b}|\mathbf{g}, H_1)$.

To sample from the posterior, we can use a Metropolis-Hastings algorithm, which we shall discuss in more detail in Sec. 15.4.8. As applied to the present problem, the basic idea is to generate a sequence of samples of the background $\mathbf{b}$ in such a way that the samples are drawn from some target density $\pi(\mathbf{b})$ such as the posterior $\text{pr}(\mathbf{b}|\mathbf{g})$. If the current background in the sequence is $\mathbf{b}^{(k)}$, a new trial background $\mathbf{b}'$ is generated from a proposal density $q(\mathbf{b}'|\mathbf{b}^{(k)})$, which can depend on the current state. The probability of accepting this proposed change is [*cf.* (15.328)]

$$\Pr(acc) = \min\left\{1, \frac{\pi(\mathbf{b}')\, q(\mathbf{b}^{(k)}|\mathbf{b}')}{\pi(\mathbf{b}^{(k)})\, q(\mathbf{b}'|\mathbf{b}^{(k)})}\right\}. \tag{14.88}$$

If the change is accepted, we set $\mathbf{b}^{(k+1)} = \mathbf{b}'$; otherwise $\mathbf{b}^{(k+1)} = \mathbf{b}^{(k)}$. By a detailed-balance argument (see Sec. 15.4.8), it can be shown that the equilibrium distribution is indeed $\pi(\mathbf{b})$. Note that only ratios of target densities are required; if $\pi(\mathbf{b})$ is the posterior, then we can write it as $\text{pr}(\mathbf{b}|\mathbf{g}) \propto \text{pr}(\mathbf{g}|\mathbf{b})\ \text{pr}(\mathbf{b})$. The constant

of proportionality cancels out in (14.88), and we do not need to know the normalization of the posterior. We do, however, need to know the ratios $\mathrm{pr}(\mathbf{b}')/\,\mathrm{pr}(\mathbf{b}^{(k)})$.

Kupinski *et al.* (2003b) showed how this approach could be applied to likelihood-ratio calculations with a lumpy background model. Two types of perturbations to the background were allowed: changes in the location of a particular lump and changes in the number of lumps. This procedure was used to compare three rather stylized pinhole imaging systems in terms of ideal-observer AUC for an SKE task. By running the Markov chain multiple times, the variance in the estimate of the AUC was estimated. In subsequent work, Park *et al.* (2003) extended this method to random signals.

Channelized ideal observer We have mentioned low-dimensional representations of object statistics, but we can also consider dimensionality reduction in data space as a way of facilitating ideal-observer studies. Though dimensionality reduction would also be called feature extraction in pattern recognition, we have several advantages in assessment of image quality that we do not have in pattern recognition. As we discussed in the context of the channelized Hotelling observer in Sec. 14.3.2, we can consider SKE tasks where all details of the signal are known, we can construct backgrounds with known statistical properties, and we can simulate noise-free samples with these statistics.

Armed with this information, we can construct so-called efficient channels in such a way that the performance of the Hotelling observer operating on the channel outputs is a good approximation to that of the true Hotelling observer operating on the original data $\mathbf{g}$. For example, if we consider a rotationally symmetric signal in a known location in a statistically isotropic background, we can use rotationally symmetric channels defined by Laguerre-Gauss functions. Gallas and Barrett (2003) demonstrated that only 5–10 such channels were needed for good estimates of Hotelling-observer performance; we anticipate that a similar result will hold for ideal observers when we have a similar amount of prior information, but this hypothesis has not yet been confirmed.

Suppose we have a set of linear channels that we believe, based on our knowledge of the classification task, might be efficient with respect to the ideal observer. To check this possibility, we need to compute the likelihood ratio on the channel outputs for a training set of signal-present and signal-absent images and then create an ROC curve. Recent work by Subok Park and Matthew Kupinski offers some possible ways of computing the likelihood ratio. Though this work is unpublished at this writing, we sketch the main ideas here with the permission of the originators.

The approaches suggested by Park and Kupinski apply to situations where we have analytic expressions for the characteristic functions of the data $\mathbf{g}$ but no PDFs, yet still want to compute a likelihood ratio (LR). The basic idea is to reduce the dimensionality of the data by use of a set of P channels and then attempt to compute the LR on the channel outputs rather than on $\mathbf{g}$ itself.

As Park formulated the problem, the PD channel output vector is given by

$$\mathbf{v} = \mathbf{T}\mathbf{g} = \mathbf{T}\left\{\boldsymbol{\mathcal{H}}\mathbf{f} + \mathbf{n}\right\}, \tag{14.89}$$

and the characteristic function for $\mathbf{v}$ under hypothesis j (in the case of Poisson noise) is given by an extension of (14.71) as

$$\psi_{\mathbf{v}|H_j}(\boldsymbol{\omega}) = \Psi_{\mathbf{f}|H_j}\left[\boldsymbol{\mathcal{H}}^{\dagger}\boldsymbol{\Gamma}\left(\mathbf{T}^{\dagger}\boldsymbol{\omega}\right)\right], \tag{14.90}$$

where $\boldsymbol{\omega}$ is a $P \times 1$ vector. It is assumed that P is relatively small and that $\psi_{\mathbf{v}|H_j}(\boldsymbol{\omega})$ can be computed analytically. The likelihood ratio for a given $\mathbf{v}$ is[12]

$$\Lambda(\mathbf{v}) = \frac{\int d^P\omega \; \psi_{\mathbf{v}|H_2}(\boldsymbol{\omega}) \exp(2\pi i \boldsymbol{\omega}^\dagger \mathbf{v})}{\int d^P\omega \; \psi_{\mathbf{v}|H_1}(\boldsymbol{\omega}) \exp(2\pi i \boldsymbol{\omega}^\dagger \mathbf{v})} \,. \tag{14.91}$$

Since P is small, Park proposed doing these integrals as FFTs. Kupinski refined the idea by suggesting Monte Carlo integration with importance sampling:

$$\int d^P\omega \; \psi_{\mathbf{v}|H_j}(\boldsymbol{\omega}) \exp(2\pi i \boldsymbol{\omega}^\dagger \mathbf{v}) = \int d^P\omega \; \frac{\mathrm{pr}(\boldsymbol{\omega})}{\mathrm{pr}(\boldsymbol{\omega})} \psi_{\mathbf{v}|H_j}(\boldsymbol{\omega}) \exp(2\pi i \boldsymbol{\omega}^\dagger \mathbf{v})$$

$$\approx \frac{1}{N} \sum_{n=1}^{N} \frac{\psi_{\mathbf{v}|H_j}(\boldsymbol{\omega}_n) \exp(2\pi i \boldsymbol{\omega}_n^\dagger \mathbf{v})}{\mathrm{pr}(\boldsymbol{\omega}_n)} \,, \tag{14.92}$$

where the samples $\boldsymbol{\omega}_n$ are drawn from $\mathrm{pr}(\boldsymbol{\omega})$. Kupinski also suggested using a few sample images to determine the mean and covariance of $\mathbf{v}$ and then constructing a PD Gaussian with these estimated parameters to use as $\mathrm{pr}(\boldsymbol{\omega})$.

Much further work is needed to validate this approach and explore possible choices for the channels, but if efficient linear channels in the ideal-observer sense exist, it opens up many new avenues for evaluating imaging systems with the ideal observer and classification tasks.

Nonlinear features Several nonlinear approaches to dimensionality reduction have been suggested by Hongbin Zhang. Zhang *et al.* (2001a) discusses features derived from the ideal observer and based on $\Lambda_{\mathrm{BKE}}(\mathbf{g}, \mathbf{b})$ as defined in (14.81). Rather than attempt to average this expression over the posterior on the backgrounds, as required by (14.80), Zhang reasoned that a useful set of features for an SKE classification task could be defined as

$$\theta_p = \Lambda_{\mathrm{BKE}}(\mathbf{g}, \hat{\mathbf{b}}_p) \,, \qquad (p = 1, ..., P) \,, \tag{14.93}$$

where $\hat{\mathbf{b}}_p$ is some estimate of the background at the known signal location. Specifically, he argued that the background was likely to be slowly varying compared to a small signal, so he suggested that $\hat{\mathbf{b}}_p$ be taken as a smoothed version of $\mathbf{g}$, with different p corresponding to different widths of the smoothing filter. The resulting values of θ_p would not immediately be the ideal-observer discriminant function, but Zhang suggested that an artificial neural network might find a good approximation to the likelihood ratio in the PD space.

In related work, Zhang also suggested using a set of wavelets centered on the known signal location, followed by a nonlinear point transformation on each wavelet coefficient (Zhang *et al.*, 2001b). He suggested an iterative algorithm to train the nonlinear transformation so that the outputs would follow a PD multivariate normal law. Then, when a new image is passed through the same transformation, the likelihood ratio can readily be calculated (see Sec. 13.2.8).

[12]Even though the channels are linear, this likelihood ratio will usually be a nonlinear functional of the channelized data $\mathbf{v}$; it should not be confused with the AUC-optimal linear observer introduced in Sec. 13.2.12.

Checking the results We have sketched a number of approximate methods for estimating AUC for the ideal observer. How do we know if the results are correct? That is, how can we estimate the bias and variance of an estimate of ideal-observer AUC?

Variance in the estimate comes from two sources. First, as with any observer study, there is a variance arising from the random selection of images, or *cases* in medical parlance. Second, whenever the likelihood ratio is evaluated by using Monte Carlo or Markov-chain Monte Carlo methods to average over backgrounds, there is a variance associated with the random selection of backgrounds. This kind of variance is analogous to internal noise in the human observer; if the Monte Carlo calculation is repeated with the same image but a different random-number seed, it will not return the same value for the likelihood ratio.

Both kinds of variance can be estimated in simulation studies just by repeating the study many times, with different sets of images or with the same images but different random-number seeds. Alternatively, variance can be analyzed with MRMC methods as discussed in Sec. 14.2.4 or by using resampling methods as discussed in Sec. 14.3.2.

Bias is much more difficult to assess since we do not know what systematic errors we might be making in the likelihood-ratio calculation. To study the bias, Clarkson *et al.* (2003) proposed a set of consistency checks that must be satisfied in an ROC study if the test statistic is indeed a likelihood ratio. For example, we know from (13.85) that

$$\frac{\mathrm{pr}(\Lambda|H_2)}{\mathrm{pr}(\Lambda|H_1)} = \Lambda\,, \tag{14.94}$$

and in fact this relation holds if and only if Λ is a likelihood ratio (Clarkson and Barrett, 2000). It follows from (14.94) that

$$2\,(1 - \mathrm{AUC}_\Lambda) = \int_0^\infty d\Lambda_t\ [\mathrm{FPF}(\Lambda_t)]^2\,, \tag{14.95}$$

where $\mathrm{FPF}(\Lambda_t)$ is the false-positive fraction for threshold Λ_t. Again, this relation holds if and only if Λ is a likelihood ratio (Clarkson *et al.*, 2003).

Other useful relations are derived from moment-generating functions and from the likelihood-generating function. The moment-generating function for the log-likelihood ratio λ under hypothesis H_j is defined by (C.56) as

$$M_j(\beta) \equiv \langle \exp(\beta\lambda) \rangle_{\mathbf{g}|H_j} = \left\langle \Lambda^\beta \right\rangle_{\mathbf{g}|H_j}\,. \tag{14.96}$$

From (13.79) we know that $M_j(\beta)$, must satisfy

$$M_1(\beta + 1) = M_2(\beta)\,, \tag{14.97}$$

from which it follows that $M_1(1) = 1$. Moreover, a plot of $M_1(\beta)$ vs. β must be concave upward and pass through the points (0,1) and (1,1) as in Fig. 13.9. Once again, these properties are unique to the ideal observer.

Finally, a number of investigators have derived inequalities relating AUC to the likelihood-generating function (Clarkson, 2002; Clarkson and Barrett, 2000; Shapiro, 1999; Burnashev, 1998). One example is

$$\tfrac{1}{2}G(0) \le -\ln[2(1 - \mathrm{AUC}_\Lambda)] \le \tfrac{1}{2}G(0) + \sqrt{G(0) - \tfrac{1}{8}G''(0)}\,, \tag{14.98}$$

where $G(\beta)$ is the likelihood-generating function and primes denote derivatives.

If we have a set of simulated or real sample images and a way of estimating the likelihood ratio for each, we can check the validity of these relationships. For example, we can use no-signal images to estimate $M_1(\beta)$ directly from its definition (14.96) and see if it indeed passes through $(1, 1)$. Similar numerical methods can be devised for each of the relations that must be satisfied for the ideal observer.

If the relations are not verified, we must look for some error in our calculation of the likelihood ratio. If they are satisfied, we can have confidence that the test statistic we are calculating is *some* likelihood ratio, though not necessarily the likelihood ratio we think we are calculating, namely the one applicable to the image data. Clarkson *et al.* (2003) admit the possibility that the algorithm is finding a good estimate of some other likelihood ratio, but say they have a "natural tendency to regard (that) possibility as unlikely."

One case where it is quite likely, however, is when linear or nonlinear features have been extracted from the original image data for dimensionality reduction. Then the Markov chain or other algorithm applied to the features may indeed give a good estimate of the likelihood ratio on the features, and all of the consistency checks mentioned above will be passed, but there is no guarantee that this likelihood ratio will give the same performance as one calculated on the original image data; the consistency checks do not ensure that the features preserve the information content of the images.

14.3.4 Estimation tasks

Compared to the large literature on model observers for detection and classification tasks in image-quality assessment, much less attention has been given to computational methods for estimation tasks, and there is much less agreement about what one should be computing in the first place. Of course, there is a huge body of work on estimation of pixel values in image processing and reconstruction, but we have argued in Sec. 13.3.2 that there is no meaningful way of relating accuracy of the pixel values to image quality. We shall discuss image reconstruction further in the next chapter, but for now we concentrate on estimation problems other than reconstruction.

Thus, by "estimation task" we mean estimation of one or a few parameters characteristic of the object being imaged and (unlike pixel values) of direct relevance to the purpose for which the image was obtained. Our goal here is to survey some of the computational methods that can be used for assessment of performance on such tasks.

Since the parameter being estimated is determined by the object being imaged, we write it as $\boldsymbol{\Theta}(\mathbf{f})$. Boldface is used since the parameter will often be a vector, though almost always a low-dimensional one; when we intend a scalar parameter, we shall denote it as $\Theta(\mathbf{f})$. Upper case is used for $\boldsymbol{\Theta}(\mathbf{f})$ since we use $\boldsymbol{\theta}$ for several other things, including expansion coefficients (*e.g.*, pixel coefficients) in approximate object representations like (7.27), and that is definitely not what we mean here.

We emphasize here that we are viewing the parameters to be estimated as characteristics of the object. This is in contrast to the view of Sec. 13.3 where we were concerned with parameters characterizing the probability density function of the data. The relationship between the two viewpoints is subtle, yet critical for

assessing image quality on the basis of estimation tasks; we shall return to it at several points below.

Dichotomies Two useful dichotomies for the parameters are linear vs. nonlinear and estimable vs. nonestimable. The imaging systems that deliver the data from which the estimates are derived can also be categorized as linear or nonlinear. The estimators themselves can be linear or nonlinear functionals of the data, and they can be either biased or unbiased.

We encountered linear parameters in Sec. 7.1.4 when we discussed moment errors. In brief, a linear parameter is a linear functional of the object. If the components of $\boldsymbol{\Theta}(\mathbf{f})$ are derived linearly from the object, we know from (7.33) that they can be written as

$$\Theta_n(\mathbf{f}) = \int_{\infty} d^q r \; \chi_n^*(\mathbf{r}) \, f(\mathbf{r}) = \boldsymbol{\chi}_n^\dagger \mathbf{f} \, . \tag{14.99}$$

Equations of this form will be used in Chap. 15 for discussing image reconstruction, but here we should think of the components Θ_n merely as weighted integrals of the object. If the weighting function $\chi_n(\mathbf{r})$ is constant over some spatial region, we refer to $\Theta_n(\mathbf{f})$ as a region-of-interest integral, and its estimate $\hat{\Theta}_n(\mathbf{g})$ as a region-of-interest estimator. Of course, the estimate depends on the data $\mathbf{g}$ while the parameter itself depends on $\mathbf{f}$ but not on $\mathbf{g}$.

An important class of nonlinear parameters occurs in *mensuration tasks*, where the goal is to measure some physical dimension of a portion of the object. Examples include the area of an agricultural field in aerial photography, volume of the left ventricle in cardiology, and distance to a target in radar.

As we saw in Sec. 13.3.1, a parameter is said to be *estimable* or *identifiable* with respect to some data set if there is an estimator of it that is unbiased for all true values of the parameter. In terms of the likelihood $\text{pr}(\mathbf{g}|\Theta)$, a parameter is estimable if different values of the parameter lead to different likelihoods.

The imaging system that acquires the data $\mathbf{g}$ may be linear or nonlinear as defined in Chaps. 1 and 7. The distinction rests on the form of the *mean* data; the system is linear if $\overline{\mathbf{g}}$ is a linear functional of $\mathbf{f}$. We denote a general linear system by the operator $\boldsymbol{\mathcal{H}}$, so $\overline{\mathbf{g}} = \boldsymbol{\mathcal{H}}\mathbf{f}$.

For a linear system, we can be more precise about estimability, because in that case we can divide object space $\mathbb{U}$ into subspaces called measurement space and null space, and any object can be uniquely decomposed as

$$\mathbf{f} = \mathbf{f}_{meas} + \mathbf{f}_{null} \, . \tag{14.100}$$

We know from the discussion in Sec. 14.3.3 that the probability density function on the data in most cases is fully determined by the mean data, so for a linear system we have

$$\text{pr}(\mathbf{g}|\mathbf{f}) = \text{pr}[\mathbf{g}|\overline{\mathbf{g}}(\mathbf{f})] = \text{pr}(\mathbf{g}|\boldsymbol{\mathcal{H}}\mathbf{f}) = \text{pr}(\mathbf{g}|\mathbf{f}_{meas}) \, . \tag{14.101}$$

A general definition of estimability in this case is that $\boldsymbol{\Theta}(\mathbf{f})$ is estimable if and only if $\boldsymbol{\Theta}(\mathbf{f}) = \boldsymbol{\Theta}(\mathbf{f}_{meas})$ for all $\mathbf{f}$. If this condition is met, then a change in $\mathbf{f}_{meas}$ leads to a different $\boldsymbol{\Theta}(\mathbf{f})$ and a different likelihood $\text{pr}(\mathbf{g}|\boldsymbol{\Theta})$. Another definition of estimability is that $\boldsymbol{\Theta}(\mathbf{f})$ is estimable if and only if $\text{pr}(\mathbf{g}|\boldsymbol{\Theta}_1) = \text{pr}(\mathbf{g}|\boldsymbol{\Theta}_2)$ implies that $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta}_2$.

We can go a step further for a linear parameter. We can decompose the templates $\chi_n(\mathbf{r})$ into measurement and null components, and the n^{th} component of the parameter vector can be written as

$$\Theta_n(\mathbf{f}) = \boldsymbol{\chi}^{\dagger}_{n,meas}\mathbf{f}_{meas} + \boldsymbol{\chi}^{\dagger}_{n,null}\mathbf{f}_{null}\,. \tag{14.102}$$

Then $\boldsymbol{\Theta}(\mathbf{f}) = \boldsymbol{\Theta}(\mathbf{f}_{meas})$ for all $\mathbf{f}$ if and only if $\boldsymbol{\chi}_{n,null} = 0$ for all n. Otherwise a change in $\mathbf{f}_{null}$ would give a different value of the parameter but the same mean data and hence the same likelihood. Note that it is not necessary that the system have no null space, just that the templates have no components in that space. Since null components tend to involve high spatial frequencies, linear parameters derived from large, blobby templates are more likely to be estimable than ones derived from small or highly structured templates. In particular, as we shall discuss in more detail in the next chapter, integrals of the object over small pixels are almost never estimable.

The final dichotomies involve the estimator itself, which can be linear or nonlinear and biased or unbiased. Linear estimators were discussed briefly in Sec. 13.3, but considerable emphasis was placed there on maximum-likelihood (ML) estimators. Like the likelihood ratio used in ideal-observer classification problems, ML estimators are usually nonlinear functionals of the data. An exception in both cases occurs with Gaussian data. For Gaussian data with equal covariances under the two hypotheses, the ideal observer computes a test statistic (the log-likelihood ratio) that is linear in the data, and for Gaussian data and any linear parameter, the ML estimator is also linear in the data. In most interesting cases, however, neither the log-likelihood ratio nor the ML estimator is linear.

If the parameter is estimable, there exists an unbiased estimator, but we may not know it, or we may choose not to use it; Bayesian estimation, for example, deliberately introduces a bias toward the prior. Thus we must distinguish biased from unbiased estimators even for estimable parameters.

Performance metrics: MSE and EMSE From the discussion in Sec. 13.3.1, a natural choice for a figure of merit is the mean-square error or MSE, defined for a scalar parameter in (13.280) and for a vector in (13.286) or (13.287).

For estimable parameters, MSE has much to recommend it. It can be computed for any chosen object and estimator, it takes into account both bias and variance, and it is a scalar that can be used for system optimization. One drawback is that MSE is defined by averaging the error with respect to the density $\mathrm{pr}(\mathbf{g}|\boldsymbol{\Theta})$, so it will depend on the true value of $\boldsymbol{\Theta}$ in general. One solution to this problem is simply to plot $\mathrm{MSE}(\boldsymbol{\Theta})$ vs. $\boldsymbol{\Theta}$, much in the same manner that one can plot SKE detectability as a function of signal location or other parameters [see (13.209)].

With nonestimable parameters, MSE is more problematical. Since null components of the object influence $\boldsymbol{\Theta}(\mathbf{f})$ but not $\overline{\mathbf{g}}(\mathbf{f})$ in that case, many different objects can give the same mean data but different true values of $\boldsymbol{\Theta}$, and it is quite arbitrary which true value one associates with a given data set. Indeed, if there are no other constraints, it is usually possible to find an object so that *any* estimator of a nonestimable parameter is unbiased; whether that object is one that would ever be encountered is another matter. As we shall see in Sec. 15.1.4, positivity constraints limit the magnitude of null functions and alleviate issues of estimability, but they don't eliminate them.

Perhaps the best solution to defining a scalar figure of merit for estimates of nonestimable parameters[13] is to use the *ensemble mean-square error* or *EMSE* defined in (13.281) for scalars or (13.288) for vectors. The vector definition can be rewritten for our purposes as

$$\text{EMSE} = \left\langle \left\langle ||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||^2 \right\rangle_{\mathbf{g}|\boldsymbol{\Theta}} \right\rangle_{\boldsymbol{\Theta}} = \left\langle \left\langle ||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}(\mathbf{f})||^2 \right\rangle_{\mathbf{g}|\mathbf{f}} \right\rangle_{\mathbf{f}} . \tag{14.103}$$

In the last form, the average is over some ensemble of objects. For any particular object in the ensemble, a bias and hence an MSE can be defined, and the ensemble-average MSE is the quadratic error norm specific to the imaging system, the estimator *and* the chosen ensemble. Note that the use of an average over objects in the figure of merit does not imply that this same information was used in the estimator. The quantity $\hat{\boldsymbol{\Theta}}(\mathbf{g})$ might have been obtained by Bayesian methods, but it might also be an ML estimate or some other one that eschews prior information.

The question that remains is what ensemble to use in the averaging. The Bayesian answer would be to average over the prior, and indeed to use that same prior in the estimation process in order the minimize the EMSE. To a pragmatist, there are several difficulties with this approach. First, in practice we might not have enough verifiable prior information (as opposed to subjective or noninformative priors) that we would be willing to build it into the inference process. In practice, the only computationally tractable priors for Bayesian estimation might be some noninformative prior like entropy or simple analytic expressions such as conjugate priors[14] or the regularizing functions to be discussed in Sec. 15.3.3. Even if we were willing to use one of these analytic priors to do the estimation, there is no reason to think that samples drawn from it would bear any relation to the true distribution of $\boldsymbol{\Theta}(\mathbf{f})$ or $\mathbf{f}$, so it would be hard to have any confidence (belief) in the MSE computed from that prior.

What pragmatists can do well, however, is to perform realistic simulations (*i.e.*, ones consistent with a belief system honed in the field, laboratory or clinic), and these simulations can be used to compute sample approximations to the EMSE defined in (14.103). Specifically, if a set of sample objects $\{\mathbf{f}_n, n = 1, ..., N_s\}$ is generated, then we can approximate the EMSE by

$$\widehat{\text{EMSE}} = \frac{1}{N_s} \sum_{n=1}^{N_s} \left\langle ||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}(\mathbf{f}_n)||^2 \right\rangle_{\mathbf{g}|\mathbf{f}_n} . \tag{14.104}$$

The remaining average can be performed either analytically or by additional Monte Carlo simulations of $\mathbf{g}$ for a fixed $\mathbf{f}_n$.

Why ML? And how? As we saw in Secs. 13.3.4 – 13.3.6, ML estimators have many desirable properties. We know that ML estimators are efficient (*i.e.*, they achieve

[13] In spite of the terminology, nonestimable parameters can indeed be estimated. An estimate is merely a number associated with a data set. To be perverse, one could associate the number 3 with *any* data set. Then an estimate would be given for all $\mathbf{g}$ no matter whether the parameter was estimable, and in fact the variance of the estimate would be zero. The bias would, however, be completely meaningless.

[14] A conjugate prior is one chosen purely for mathematical convenience, to make the posterior have the same mathematical form as the prior. Unless one believes that nature is constructed for the convenience of statisticians, there is no reason to ascribe any degree of belief to conjugate priors.

the minimum possible variance as given by the Cramér-Rao bound) if any efficient estimator exists. Also, ML estimators are asymptotically efficient, asymptotically unbiased and asymptotically normally distributed. In the statistics literature, "asymptotic" refers to accumulating N i.i.d. data sets and letting $N \to \infty$, but it can have a broader meaning. All of the nice asymptotic properties of ML estimators apply if the variance of additive Gaussian noise goes to zero or if the number of counts in a photon-limited measurement gets large. Thus there is considerable motivation for using ML estimators, especially if we can get into one of these asymptotic regimes.

It is not obvious how we can perform ML estimation in general, since we seldom know the likelihood $\text{pr}(\mathbf{g}|\boldsymbol{\Theta})$ directly. Instead, as discussed in Sec. 14.3.3, we usually know the conditional density $\text{pr}(\mathbf{g}|\mathbf{f})$ or $\text{pr}[\mathbf{g}|\overline{\mathbf{g}}(\mathbf{f})]$; for direct imaging and Gaussian and Poisson noise, they are given by (14.67) and (14.68), respectively.

The general relation between the conditional densities on the data and the likelihood can be expressed either as an integral over the object space or an integral over data space:

$$\text{pr}(\mathbf{g}|\boldsymbol{\Theta}) = \int d\mathbf{f}\, \text{pr}(\mathbf{g}|\mathbf{f})\, \text{pr}(\mathbf{f}|\boldsymbol{\Theta}) = \int d^M\overline{g}\, \text{pr}(\mathbf{g}|\overline{\mathbf{g}})\, \text{pr}(\overline{\mathbf{g}}|\boldsymbol{\Theta})\,. \tag{14.105}$$

These forms are equivalent whenever the conditional probability on the data is determined solely by its mean, which is the case with our usual Gaussian or Poisson noise models, with or without post-acquisition data processing (see Sec. 14.3.3).

One situation where we can easily go from these conditional densities to the likelihood is in the estimation counterpart of the SKE/BKE problem. Suppose we decompose the object into background and signal as in (14.77), and we assume that the signal is known to be present but that it is characterized by some unknown parameter vector $\boldsymbol{\Theta}$. For a linear system, we can write the mean data for background and signal, respectively, as

$$\mathbf{b} \equiv \boldsymbol{\mathcal{H}}\mathbf{f}_b\,, \qquad \mathbf{s}(\boldsymbol{\Theta}) \equiv \boldsymbol{\mathcal{H}}\mathbf{f}_s(\boldsymbol{\Theta})\,. \tag{14.106}$$

For example, in medical imaging $\mathbf{f}_s(\boldsymbol{\Theta})$ might describe a spherical tumor with unknown center coordinates, gray level and diameter. In military reconnaissance, it might refer to a tank with unknown coordinates and heading.

If the background is known exactly and the signal is known except for these parameters, then

$$\text{pr}(\overline{\mathbf{g}}|\boldsymbol{\Theta}) = \delta[\overline{\mathbf{g}} - \mathbf{b} - \mathbf{s}(\boldsymbol{\Theta})]\,, \tag{14.107}$$

and the likelihood becomes

$$\text{pr}(\mathbf{g}|\boldsymbol{\Theta}) = \text{pr}(\mathbf{g}|\overline{\mathbf{g}})\Big|_{\overline{\mathbf{g}}=\mathbf{b}+\mathbf{s}(\Theta)}\,. \tag{14.108}$$

Explicit expressions for the likelihood ratio in the case of direct imaging can be found by substituting $\overline{\mathbf{g}} = \mathbf{b} + \mathbf{s}(\boldsymbol{\Theta})$ into (14.67) or (14.68). Since the number of parameters is small, there is no difficulty in maximizing the likelihood numerically.

Random backgrounds Just as in the signal-detection problem, the BKE assumption in estimation is oversimplified and can be misleading. It is much more realistic to consider random, cluttered backgrounds when we want to estimate signal parameters. We can regard the background components as a set of nuisance parameters,

in the sense that they do not enter into the overall cost or Bayes risk associated with the estimation problem. As we learned in Sec. 13.3.8, the optimal strategy for this problem is to marginalize over the nuisance parameters, at least if we have a believable way of generating or approximating the prior density or drawing realistic samples. The likelihood is then given by

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\Theta}) = \int d^M b \; \mathrm{pr}(\mathbf{g}|\boldsymbol{\Theta}, \mathbf{b}) \; \mathrm{pr}(\mathbf{b}) \,, \tag{14.109}$$

where $\mathrm{pr}(\mathbf{g}|\boldsymbol{\Theta}, \mathbf{b})$ is to be computed from (14.108). This form is quite similar to the likelihood expressions encountered in Sec. 14.3.3 [*cf.* (14.84) – (14.86)], and similar Monte Carlo and Markov-chain Monte Carlo methods can be devised to evaluate it (Kupinski *et al.*, 2003c). As in the detection case, direct sampling of backgrounds from $\mathrm{pr}(\mathbf{b})$ is unlikely to work well since a randomly chosen $\mathbf{b}$ will probably lead to a vanishingly small $\mathrm{pr}(\mathbf{g}|\boldsymbol{\Theta}, \mathbf{b})$, but importance sampling can be used as in (14.85). If an analytic form is known for $\mathrm{pr}(\mathbf{b})$, samples $\mathbf{b}_n$ can also be drawn from the BKE likelihood $\mathrm{pr}(\mathbf{g}|\boldsymbol{\Theta}, \mathbf{b})$, renormalized as a density on $\mathbf{b}$, and the likelihood estimate is proportional to $\frac{1}{N} \sum_{n=1}^{N} \mathrm{pr}(\mathbf{b}_n)$.

For a detailed survey of Monte Carlo methods in ML estimation, see Geyer and Thompson (1992).

PDFs of the estimates Monte Carlo methods can also be used to study the distribution of the estimates themselves. If we simulate multiple data sets with the same true value, say $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$, and compute $\hat{\boldsymbol{\Theta}}$ for each, then we have, in effect, drawn samples from $\mathrm{pr}(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_0)$. From these samples we can estimate the bias, variance, MSE and any other figure of merit we might devise.

In many problems, it is also possible to compute $\mathrm{pr}(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}_0)$ directly. Building on earlier work by Müller *et al.* (1990, 1995), Abbey *et al.* (1998) developed a method for approximating the density of maximum-likelihood and MAP estimates under a Gaussian noise model. They showed that the method was directly applicable to estimating parameters such as tumor volume from medical images, and they found that the predicted analytic PDFs were in good agreement with Monte Carlo simulation.

Rogala and Barrett (1997, 1998a, b, c) applied Abbey's method to a combination interferometer/ellipsometer where the goal was to estimate surface height and the real and imaginary parts of the refractive index at all points on a metal surface. Again, the analytic results were confirmed by Monte Carlo simulation.

Cramér-Rao bounds Rather than using the performance of a particular estimator as a figure of merit, it is also possible to use various performance bounds that might be easier to compute. In particular, the Cramér-Rao bound, introduced in Sec. 13.3.5, sets a lower limit to the variance of an unbiased estimator. For an unbiased estimator, the Cramér-Rao bound is given in (13.371) or (13.372), and for a biased estimator, the appropriate forms are (13.376) and (13.377). Both the biased and unbiased form are derived from the Fisher information matrix.

Kupinski *et al.* (2003c) developed MCMC methods to estimate the Fisher information matrix for the problem of estimating the position, width and amplitude of a Gaussian signal in a lumpy background. They did not assume that the signal was always present, so their treatment applied to a hybrid detection/estimation

problem, but the figure of merit was based only on the estimation performance, marginalized over the probability of detection.

Approaches based on the Cramér-Rao bound are attractive, but they have their limitations. For one thing, if more than one parameter is to be estimated, it is not clear how to combine the individual bounds into a single scalar figure of merit that can be used for system optimization. Second, in many problems no efficient estimator exists, and it is not clear in practice how far actual variance will be from the bound. Similarly, it is often the case that no unbiased estimator exists, so use of the unbiased form of the bound can be misleading; the biased form (13.376) is less useful since it requires knowledge of the bias gradient (derivative of the bias with respect to the parameter). Considerable work has been done at the University of Michigan on variance bounds in which the norm of the bias gradient is constrained, though mostly in the context of estimation of pixel values (Gorman and Hero, 1990; Hero and Fessler, 1994; Hero *et al.*, 1996).

14.4 SOURCES OF IMAGES

Simulated images play an important role in the practical assessment of image quality. They can be used to get a subjective impression of the effects of changing parameters of the imaging system, and they can serve as input for objective studies with either model observers or humans. If the simulations are realistic, they may even be preferable to real images since there is no question about the true state of the object. Most importantly, simulations can be used to assess imaging systems that do not exist, so they are essential to any program of systematic optimization.

Realistic simulations involve computer implementations of the object, the image-formation process and the detector, and they must accurately reflect both the deterministic and stochastic aspects of each of these components. The art of good simulation is thus necessarily specific to both the imaging system and the use to which the simulation will be put. Nevertheless, it is our goal in this section to give some general guidelines on the simulation process, with reference to specific systems only as examples. We shall refer to the methods for representing deterministic and random objects given in Chaps. 7 and 8, along with material on the simulation of image formation provided in Chap. 10.

In Secs. 14.4.1 and 14.4.2 we survey methods for deterministic and stochastic simulation of objects, and in Secs. 14.4.3 and 14.4.4 we treat deterministic and stochastic simulation of image formation. Finally, in Sec. 14.4.5 we discuss the gold-standard problem that arises when using real images instead of simulated ones.

14.4.1 Deterministic simulation of objects

In Sec. 7.1 we emphasized that real objects are functions, but we also acknowledged that numerical computations require approximate discrete representations. In all fields of image science, there is considerable emphasis on linear representations, and we know from (7.27) that the general form of such a representation is

$$f_a(\mathbf{r}) = \sum_{n=1}^{N} \theta_n \phi_n(\mathbf{r}) \,. \tag{14.110}$$

Thus object simulation involves two steps: choosing the expansion functions $\phi_n(\mathbf{r})$ and choosing the coefficients θ_n.

It is all but universal in simulation studies to choose the expansion functions as pixels or voxels, for two reasons. First, if we humans are simply inventing the simulated objects, it is easiest for us to think in terms of spatial variables. Pixels and voxels are discretizations of our natural visual domain, and it would be much harder for us to think in terms of, say, Fourier basis functions. Second, as we shall see below, we may also want to use images from some high-resolution imaging system as objects for another system of lower resolution. Since the first system is designed to present data to humans, it is likely to provide us with digital data in a pixel or voxel representation. Thus we have a ready-made discrete simulation if we stick with those expansion functions.

When the goal of the simulation is to evaluate imaging systems, it is not so much the simulated objects as the resulting simulated images that interest us. Our goal is to use the discrete representations of objects and systems to produce images that are as near as possible to those that would be obtained with actual continuous objects and continuous-to-discrete systems (see Sec. 7.4.3). That means that we should take N in (14.110) as large as possible. The only cost to increasing the number of pixels and voxels, in most cases, is increased computational time, and that commodity continues to plummet in price. In particular, we do not need to worry about whether the resulting system matrix is highly non-square and hence leads to an underdetermined inverse problem. In this section we are concerned only with accurate simulation of the forward problem; issues associated with choice of representation in inverse problems are discussed in detail in the next chapter.

Geometric objects The easiest way to get started in object simulation is to use superpositions of simple geometric shapes (circles, squares, ellipses...). In a pixel representation, the coefficients θ_n are assumed to have the same value for all pixels within one elemental shape, but generally different values within different elements. If the number of pixels N is large, we need not worry too much about pixels that straddle the border between elements. By using a range of sizes for the elemental shapes, we can get a simulated object that has small structures to challenge the spatial resolution of a simulated imaging system and large uniform structures with which to study system uniformity, radiometric accuracy and noise properties.

Such seemingly naive simulated objects have proven particularly valuable in tomographic imaging. They are known as *mathematical phantoms* in that field, and some have been so durable that they are commonly referred to by the name(s) of the investigators who devised them. Thus we have the Shepp-Logan phantom (an arrangement of ellipses somewhat resembling a 2D cross-section of the human brain; Shepp and Logan, 1974) and the Defrise phantom (a 3D set of thin parallel disks meant to challenge certain cone-beam tomographic systems; Defrise and Clack, 1994).

Geometrical shapes can also be manipulated to mimic much more complicated objects. For example, Tsui *et al.* (1993) devised a 3D representation of the human torso that includes a static model of the heart, and Pretorius *et al.* (1997) extended the work to a beating heart. This so-called MCAT (mathematical cardiac torso) phantom has become a *de facto* standard in simulation of nuclear-medicine cardiac studies.

The mathematical theory that treats efficient ways of representing and manipulating geometrical forms within the computer is called *computational geometry*. Two useful textbooks in this emerging field are O'Rourke (1998) at an undergraduate level and the more comprehensive graduate-level text by Preparata and Shamos (1985).

Digitized real objects Useful though these geometric objects may be, they do not capture the complexity of object variation from pixel to pixel within a given geometric element, and for this reason they may not give accurate results when used for the objective assessment of image quality. One way around this difficulty is stochastic simulation, discussed in Sec. 14.4.2, but another approach is use of real image data.

As mentioned above, we might have access to high-resolution images of objects that we also wish to image with a lower-resolution system. Often the high-resolution system will measure fundamentally different parameters of the object, or it might be that the higher-resolution system is more expensive or more invasive than the system under development. Under these circumstances, the higher-resolution system might not be one we would use in practice, but we can nevertheless use the images it produces to guide the development of the new system.

An example of considerable interest for medical imaging is the Visible Human Project. In this project a human cadaver was imaged with computed tomography at high spatial resolution and high (but irrelevant) radiation dose. High-resolution magnetic resonance imaging was also performed, and then literal tomograms[15] were obtained by slicing the cadaver into thin layers and photographing each.

The CT images obtained in this project have higher resolution and lower noise than any obtainable with living patients, so they can serve directly as objects for simulation studies of new CT systems. The MRI images are less useful for this purpose since the object in MRI is specified in a complicated way by three distinct scalar fields, the spin density and two relaxation times (see Prologue and Sec. 7.1.1). Any particular image represents some nonlinear combination of these three components and cannot be used to simulate objects for imaging systems that respond to other combinations. The optical images are useful mainly because they accurately delineate borders of the organs, so they provide an alternative to the stylized geometric shapes discussed above. The actual gray levels (or colors) do not, however, correspond to anything that would be seen with any real medical imaging system.

For many further details on the Visible Human images and their applications, the reader may consult the proceedings of conferences that have been held on the project (Banvard, 2000).

Similarly, Zubal *et al.* (1994) at Yale have developed torso and brain phantoms by starting with high-resolution CT images and painstakingly labelling different anatomical regions by hand. To simulate objects in lower-resolution nuclear-medicine simulations, these labelled regions can be assigned different gray levels, corresponding to uptakes of some radiopharmaceutical of interest.

Computer graphics Perhaps the greatest impetus to progress in image simulation today is computer games and the closely related field of virtual reality. Since

[15]Greek $\tau o \mu o \sigma$ = slice.

the everyday reality we see around us consists mainly of surface reflections from opaque objects, virtual reality and computer graphics are particularly useful for simulating such objects. Useful books in this area include works by Neelamkavil (1987), Anand (1993), Sillion and Puech (1994), Glassner (1995) and Rogers (1998). Graphics-related journals and magazines include: IEEE Transactions on Visualization and Computer Graphics, IEEE Computer Graphics and Applications and IEEE Multimedia; ACM Transactions on Modeling and Computer Simulation and ACM Transactions on Graphics, and Computer Vision, Graphics and Image Processing (CVGIP).

14.4.2 Stochastic simulation of objects

In Sec. 8.4 we discussed a wide variety of statistical models for objects. Each of these models provides a PDF that at least partially describes the random variation in objects, and stochastic simulation of objects amounts to drawing sample functions (or vectors) from those PDFs. Often the first thing we need to simulate is the overall shape of the object or of key components in the object; see Chap. 8 for a brief discussion of the statistical description of shape. Then we need to add in a random texture.

Random textures Methods of generating samples of texture fields with specified statistics were discussed in Sec. 8.4.4, and the literature on computer graphics can provide additional approaches. A common approach in image simulation is to assume that the texture is stationary within the boundaries of a single geometric element of the simulated object.

How accurately the texture needs to be simulated depends critically on the purpose of the simulation. If the task used to assess image quality is detection of a low-contrast lesion in a medical image, then, as we have noted earlier in this chapter, the texture results from anatomical variations that may, in fact, constitute the main noise source limiting task performance, so accurate modeling is essential. On the other hand, if measurement noise is high or if the task is estimation or mensuration, then task performance might be relatively insensitive to fine details of the object structure.

We urge the reader to be skeptical of simulations that omit texture modeling, particularly if the goal of the simulations is to provide input for image reconstruction. As we shall see in the next chapter, any reconstruction algorithm involves a choice of how much fine detail to attempt to reconstruct. Often this choice is made on the basis of claimed prior information, and the most common such claim amounts to saying that the object contains little or no fine detail. At the extreme, it may be asserted that the object is piecewise constant within boundaries of regions such as organs. Of course, it is easily possible to simulate objects and hence tomographic data consistent with this assertion, but the simulations then provide essentially no information about how the algorithms would perform on tasks that are sensitive to fine details.

Random signals In signal-detection studies it is useful to think of the object as a superposition of signal and background (see Sec. 8.4.5), and the signal component might be particularly amenable to simulation. In medical imaging, for example, a common task is tumor detection, and it might suffice to model the tumor as a small

sphere or ellipsoid of low contrast. When the task is discrimination between types of tumors or between benign and malignant lesions, however, it may be necessary to include other features such as spicules (needle-like protrusions from the body of the tumor), but these too can be incorporated in realistic simulations.

Simulated signals may be superimposed on simulated or real backgrounds. As we saw in Sec. 8.4.5, the signal can sometimes be regarded as simply added to background, and in those cases we can maintain separate files of simulated signals and real or simulated backgrounds, adding them together in various combinations as needed. Moreover, when we are dealing with linear systems, we can choose to add the images rather than the objects (see Sec. 8.5.4). This makes it possible to add simulated signals to actual images of normal (signal-absent) objects as seen through real imaging systems. Since normal images are much easier to acquire and verify than abnormal ones, this approach can be very beneficial in avoiding the gold-standard problem.

14.4.3 Deterministic simulation of image formation

Linear systems Once we have a discrete object representation, it is generally straightforward to compute its mean image through a linear imaging system by matrix multiplication. The only real difficulty is in formulating the matrix, and that problem is specific to the imaging modality. We shall give an example of how to construct the matrix for emission computed tomography in Sec. 17.2.6.

We emphasize again, however, that it is important in simulation studies to sample the object finely, especially in image-reconstruction problems. If the reconstruction algorithm assumes that the object consists of voxels of a certain size, and the data are generated on precisely this same assumption, then a false consistency may result. When the same matrix is used in data simulation and reconstruction, and the resulting images are good in some sense, all that has been proved is that the matrix is nonsingular; no useful conclusions can be drawn about the true CD system or about real data. Simulation studies that use the same matrix for both a forward problem and its inverse problem should be regarded with strong suspicion.

Sparseness of the H matrix Though the $\mathbf{H}$ matrix used in simulation may be huge, it is often very sparse, with most of its elements equal to or very near zero. In direct-imaging systems, for example, the point response function $h_m(\mathbf{r})$ will tend to be highly concentrated; for any chosen source point $\mathbf{r}$, only a small subset of the detector pixels will receive radiation. (Indeed, this is essentially a definition of direct imaging.) When such systems are represented by matrices, the same thing holds: for any chosen n, H_{mn} is nonzero for only a small subset of m. Put another way, each column of $\mathbf{H}$ is mostly zero, no matter how many columns we choose to use. The zero elements need not be stored, and of course there is never any point in multiplying by zero.

Indirect-imaging systems may also result in sparse matrices. In tomography, for example, radiation is received from points along or near a thin pencil through the object (at least when scatter is neglected). Conversely, for any chosen object point and any projection direction, only a small subset of the detector elements receive radiation. In this case elements in one column of $\mathbf{H}$ are indexed by both the detector index and the projection direction, but nevertheless only a small fraction

of the elements in each column are nonzero. For more discussion of this point in the context of emission computed tomography, see Sec. 17.2.6.

Shift-invariance Another structure that we might consider using is shift-invariance. As we have discussed in Sec. 7.2.3, it may be reasonable to describe certain CC systems with shift-invariant point spread functions. When the output of such a system is sampled with a regular detector array, and the object is represented by a regular grid of the same spacing, it is tempting to say that the system exhibits discrete shift-invariance and hence that the images are convolutions that can be computed efficiently with fast Fourier transforms or FFTs.

There are several problems with this approach. The first is that discrete convolutions (with N samples in 1D) are described by modulo-N arithmetic (see Sec. 3.6.2). The result is an entirely unphysical wrap-around such that images that disappear from one edge of the detector as the object point is shifted magically reappear on the other side. Various stratagems can be employed to minimize this effect, but their adequacy is seldom verified.

The second problem is that any real CC system must have some departures from strict shift-invariance. In a lens system with aberrations, for example, the form of the PSF varies with field angle. This problem is incompatible with any convolutional description. It can be minimized by restricting the object field and/or the image field, but then wrap-around effects may become more significant.

Finally, a great hazard of working with FFTs and discrete convolutions is that it entices the user to choose the number of samples in object space to be the same as the number of samples in image space. As we stressed above, accurate simulation of CD systems requires fine sampling of the object. The FFT approach to simulation requires sacrificing accuracy for speed; this tradeoff becomes increasingly difficult to justify as computers get faster, and it is especially questionable when only one or a few images are to be simulated.

These warnings do not imply that we should ignore approximate shift-invariance when constructing an $\mathbf{H}$ matrix or performing simulation studies. If neighboring columns of $\mathbf{H}$ are nearly equal but for a shift, we can take advantage of this structure and reduce the computation time needed to find the matrix and the memory needed to store it. For an example in the context of emission computed tomography, see Sec. 17.2.6.

Deterministic transport calculations In principle, the Boltzmann transport equation, discussed in detail in Secs. 10.3 and 10.4, allows us to compute the image of any object where the radiation can be considered particle-like, which for electromagnetic radiation means that interference and diffraction, polarization and quantum-mechanical effects such as squeezing can be neglected. To oversimplify, the domain of the Boltzmann equation is the same as that of geometric optics.

14.4.4 Stochastic simulation of image formation

Stochastic simulation was introduced in Sec. 10.4.5 as a broad class of methods in which some quantity is estimated by performing random experiments, either physically or in a computer. These methods can be applied to the generation of samples of noisy data for use in psychophysical experiments and model observer calculations.

Detectors and image noise So far we have discussed ways of computing or estimating the mean data $\{\overline{g}_m\}$, but for many purposes we need to simulate the actual noisy data $\{g_m\}$, so we must also be able to simulate the noise contributions $\{n_m\}$.

For simple noise processes, we can just call an appropriate random-number generator to generate a noisy image. For example, many detector arrays are dominated by electronic noise, which we know from the discussion in Chap. 12 to be usually well described by Gaussian probability laws. Moreover, we can often argue from physical grounds that the noise in different detector elements is statistically independent, so the noise can be simulated by calling an independent Gaussian random-number generator at each element. Similarly, if Poisson noise dominates, we can first calculate the mean number of counts at each element by deterministic methods and then call a Poisson random-number generator with this mean.

Some detectors generate excess noise as a result of a random amplification process (see Secs. 11.4), and the detector therefore introduces noise correlations. To accurately simulate a noisy image in this case, we must draw random vectors from a multivariate PDF, or we must simulate the amplification process itself.

In summary, for the results of an evaluation study to be valid, it is crucial that realistic object models and accurate models of the imaging system be employed. Particularly when the investigation involves an imaging system or a task for which model and human observer data have not been compared before, performance estimates based on simulated data sets and model observers should be verified using real data and human observers.

14.4.5 Gold standards

Conventional ROC analysis requires knowledge of the truth status of the images in order to score observer responses as correct or incorrect. Thus standard ROC methods are not directly applicable when the truth status of the images is unknown. The requirement that independent truth status be known can lead to case-selection bias that can favor one system over another.

Even when a method for establishing the truth status of the images exists, giving a so-called "gold standard," the method is more often a bronze standard rather than gold. New modalities are often evaluated with an older technology as the gold standard, even though the new modality may allow for the detection of subtle objects missed when using images from the older device. In medical applications, biopsy proof is the gold standard, but even biopsy is not perfect. Biopsy needles can miss their mark, and pathologists have been shown to make mistakes as well. Pathologist is another name for a human observer performing a classification task, so the process should be amenable to objective evaluation based on task performance. But what would be the gold standard?

Given the need to keep score of the observer's performance using a specified figure of merit, it is clear that simulations offer an added advantage — they solve the *ground truth* problem. For simulated images, the truth state of the objects are known because this information is in the hand of the investigator.

In this section we shall describe the effect of inaccurate gold standards on ROC methods and present approaches to the evaluation of imaging systems in the absence of ground truth. As we shall see, methods for the assessment of imaging systems in the absence of ground truth exist, but the uncertainty in the estimate of

the system's performance is much larger than what is achieved when ground truth is known.

Truth by expert panel One approach to the establishment of truth is the use of an expert panel of observers. This method raises a multitude of questions and concerns regarding the number of experts to be used, how they will be chosen, and how their responses will be combined to establish "truth." Revesz *et al.* (1983) showed that the ranking of 3 systems could be made to favor any one of the 3, depending on the way in which the expert opinions were used to determine truth.

Mixture-distribution analysis Mixture-distribution analysis is based on the assumption that experts are likely to be correct when they agree. Kundel and Polansky (1997) suggested the use of a mixture-distribution analysis as an alternative to ROC methods when ground truth is not available. The method is based on dichotomizing the images into groups on the basis of the extent of a set of observers' agreement on them. It is assumed that the image groups represent different levels of case difficulty; *i.e.*, lower agreement indicates harder cases while higher agreement indicates easy cases. The number of groups is arbitrary. Thus the underlying model is a mixture distribution with a user-defined number of groups. Given the observer's ratings, an expectation-maximization (EM) method can be used to estimate the proportion of images in each group and the probability of truth given a certain level of agreement. Having estimated the truth status, the reader ratings can be used to determine the ROC curve.

Kundel and Polansky (1997) have compared mixture-distribution analysis to the results of an ROC analysis, where the ROC method used a separate expert panel to determine truth. Both methods gave similar estimates for the percentage of correct diagnoses for the task of image interpretation in chest radiography. Kundel and Polansky (1998) have shown that the results are fairly robust to the number of groups used in the model. The method may be especially useful in the evaluation of CAD algorithms; Kundel *et al.* (2001) recently demonstrated the use of the mixture-distribution approach for the evaluation of CAD in mammography. Recent emphasis on lung cancer screening programs using high-resolution CT raises the spectre of a very large number of potential lesions in the images for each patient; biopsy proof is simply not viable. Mixture-distribution analysis may be useful for the assessment of adjunctive CAD algorithms for this application.

See Polansky (2000) for a tutorial on the mixture-distribution method and other agreement-based approaches.

ROC analysis without truth of diagnosis It is not possible to perform an ROC evaluation of a single imaging system in the absence of ground truth because the problem is underdetermined. However, if each object has associated with it ratings from images obtained on two or more modalities, Henkelman *et al.* (1990) demonstrated that an EM algorithm can be used to estimate the class prevalences and the model parameters of a mixture distribution for the underlying objects.

The EM model makes the assumption that there are two underlying distributions for the decision variables, one for each class, and the distributions are correlated by an unknown amount. The dimensionality of each distribution is equal to the number of modalities under test. The EM algorithm estimates the relative proportion of each distribution (the prevalences), the locations of the observer's thresholds

corresponding to each rating level along the decision axis for each modality, and the parameters specifying the distributions. For example, the use of a 5-point rating scale for 2 medical imaging modalities involves the estimation of 10 category boundaries, one disease prevalence and, in the case of a bivariate normal model for the distributions, the difference in means of the two distributions, their widths in two dimensions, and their correlations. Thus, given a sufficient number of images and observers, the estimation problem becomes tractable.

In a commentary on the Henkelman approach, Begg and Metz (1990) point out that the method breaks down if the imaging systems have low AUC (Henkelman *et al.* restricted their investigations to systems with an AUC ≥ 0.92). Begg and Metz suggest that each system must have an AUC of 0.80 or better for this technique to be applicable.

The work of Henkelman *et al.* has been extended by Beiden *et al.* (2000b), who performed Monte Carlo simulations to determine the uncertainties in the EM estimates of AUC obtained in the absence of ground truth. These authors found that many more patients were required in the truth-unknown case to yield estimates of AUC with standard deviations of those determined in the truth-known case, for the particular choice of true underlying distributions they investigated.

More investigation is required to better understand the usefulness of the EM approach to the no-gold-standard problem in ROC analysis in order to better understand the impact of the forms of the underlying distributions, the number of samples, the number of observers, the model assumptions made in the EM algorithm, and so on. Henkelman *et al.* suggest that the estimation problem may become better conditioned as the number of imaging modalities increases. More research is required to investigate this issue as well. Nonimaging diagnostic tests, including pathology readings, might also be included as additional modalities along with one or more imaging tests in the EM procedure.

Evaluation of estimation performance without a gold standard The issue of ground truth arises also in evaluating imaging systems on the basis of estimation tasks. For example, cardiac ejection fraction (the fraction of the blood expelled on each beat) can be measured by many different methods, including SPECT, planar nuclear medicine, MRI, ultrasound, CT and biplanar projection x rays. Each of these methods has significant errors, and none is universally accepted (except by its practitioners) as the "gold standard." When a new method is developed, it is customary (perhaps even mandatory) to publish a plot of ejection fractions obtained by the new method against ones obtained on the same patients with some older method. Ideally such plots would show a high correlation, with regression slopes near one and intercepts near zero. It is not uncommmon, however to find slopes around 0.6-0.8 and intercepts around 0.2-0.3. Something is wrong with one or both methods, but there appears to be no way of telling which without a gold standard.

It would be desirable to regress the estimates obtained from each modality against the true value of the parameter rather than against another estimate, and in fact it is possible to do so if each patient is studied on each of two or more modalities (Hoppin *et al.*, 2002; Kupinski *et al.*, 2002). The basic assumption is that there exists a linear relation between the mean value of the estimates and the true value for each patient (though nonlinear relations can also be used). If θ_{pm} is the estimate obtained from patient p on modality m, and Θ_p is the true value for

that patient, the assumed relation is

$$\theta_{pm} = a_m \Theta_p + b_m + n_{pm} \,, \tag{14.111}$$

where n_{pm} is a zero-mean random variable. For simplicity, Hoppin *et al.* assumed that n_{pm} was normally distributed, but this does not appear to be critical. It was also assumed that n_{pm} was statistically independent of $n_{p'm'}$ for $p \neq p'$ or $m \neq m'$, and that the random variables for different patients but the same modality had the same variance. If P patients are each studied on M modalities, there are a total of PM measurements and $3M$ unknowns, namely the M values of a_m and b_m as well as the variances of n_{pm} for each m.

The basic idea is to estimate the $3M$ unknowns from the PM measurements by maximum-likelihood methods. With the assumptions made about n_{pm} it is straightforward to write down a probability density function on the measurements conditional on the unknown parameters and on the true values Θ_p, but of course we don't know these true values. Therefore Hoppin *et al.* assumed that the Θ_p were drawn independently from some parametric density $\mathrm{pr}(\Theta_p|\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a vector of unknown parameters describing the density. For example, since ejection fraction is defined on 0-1, a natural choice for $\mathrm{pr}(\Theta_p|\boldsymbol{\alpha})$ is a beta distribution, which has two free parameters. These two parameters are of course unknown, so they are simply added to the list of parameters to be estimated. For example, with three modalities and the beta distribution, there are a total of $3M + 2 = 11$ unknowns, but if 100 patients are studied, there are 300 measurements.

This method has been well validated in simulation studies, and it has been placed on a firm theoretical footing by calculation of the Fisher information matrix. Not only does it give accurate estimates for the desired regression parameters, it also gives good values for the nuisance parameters contained in $\boldsymbol{\alpha}$.