In *Medical Imaging: Image Processing*, M.H. Loew, ed.
Proc. SPIE **1898**, pp. 716-731 (1993)

Introduction to Bayesian image analysis

K. M. Hanson*

Los Alamos National Laboratory, MS P940
Los Alamos, New Mexico 87545 USA
kmh@lanl.gov

ABSTRACT

The basic concepts in the application of Bayesian methods to image analysis are introduced. The Bayesian approach has benefits in image analysis and interpretation because it permits the use of prior knowledge concerning the situation under study. The fundamental ideas are illustrated with a number of examples ranging from a problem in one and two dimensions to large problems in image reconstruction that make use of sophisticated prior information.

1. INTRODUCTION

The Bayesian approach provides the means to incorporate prior knowledge in data analysis. Bayesian analysis revolves around the posterior probability, which summarizes the degree of one's certainty concerning a given situation. Bayes's law states that the posterior probability is proportional to the product of the likelihood and the prior probability. The likelihood encompasses the information contained in the new data. The prior expresses the degree of certainty concerning the situation before the data are taken.

Although the posterior probability completely describes the state of certainty about any possible image, it is often necessary to select a single image as the 'result' or reconstruction. A typical choice is that image that maximizes the posterior probability, which is called the MAP estimate. Other choices for the estimator may be more desirable, for example, the mean of the posterior density function.

In situations where only very limited data are available, the data alone may not be sufficient to specify a unique solution to the problem. The prior introduced with the Bayesian method can help guide the result toward a preferred solution. As the MAP solution differs from the maximum likelihood (ML) solution solely because of the prior, choosing the prior is one of the most critical aspects of Bayesian analysis. I will discuss a variety of possible priors appropriate to image analysis.

The first half of this paper is devoted to a brief introduction to Bayesian analysis illustrated with an interesting problem involving one and two variables. The second half comprises a discussion of the application of Bayesian methods to image analysis with reference to more sizable problems in image reconstruction involving thousands of variables and some decisive priors.

In this short paper I can not address many of the more detailed issues in Bayesian analysis. I hope that this review will encourage newcomers to investigate Bayesian methods in more depth and perhaps even stimulate veterans to probe the fundamental issues to deeper levels of understanding. A number of worthwhile textbooks can broaden the reader's understanding.[1–6] The papers by Gull and Skilling and their colleagues,[7–11] from which I have benefited enormously, provide a succinct guide to Bayesian methods. The series of proceedings of the workshop *Maximum Entropy and Bayesian Methods*, published mostly by Kluwer Academic (Dordrecht), presents an up-to-date overview of Bayesian thinking.

*Supported by the United States Department of Energy under contract number W-7405-ENG-36.

2. BAYESIAN BASICS

The revival of the Bayesian approach to analysis can be credited to Jaynes. His tutorial on the subject is recommended reading.[12] Gull's papers[7,8] on the use of Bayesian inference in image analysis and reconstruction provide a good technical review, from which this author has gained significant understanding. The Bayesian approach is based on probability theory, which makes it possible to rank a continuum of possibilities on the basis of their relative likelihood or preference and to conduct inference in a logically consistent way.

2.1. Elements of Probability Theory

As the essential ingredients of probability theory are only briefly summarized here, the reader may wish to refer to Refs.[1,3,4] for more thorough treatments.

The rules of probability have been shown by Cox[13] to provide a logically consistent means of inference, which establishes a firm foundation for Bayesian analysis. Because we are dealing mostly with continuous parameters, for example the pixel values of the image, probabilities are cast in the form of probability density functions, designated by a lowercase $p()$[†]. For heuristic purposes, the probability density function may be thought of as a limit of a frequency histogram. From an ensemble of N possible occurrences of the situation under study, the frequency of events falling in the interval $(x, x + \Delta x)$ is the fraction $F_{\Delta x}(x) = n_{\Delta x}/N$, where $n_{\Delta x}$ is the number of events in the interval. The probability density is the limit of the frequency density $p(x) = F_{\Delta x}(x)/\Delta x$, as $n_{\Delta x} \to \infty$ and $\Delta x \to 0$.

The normalization sets the scale for the probability density function:

$$\int p(x)\, dx = 1 \ , \tag{1}$$

where the integral is over all possible values of x. This normalization simply states that some value of x is certain to occur.

The transformation property of density functions

$$p(u) = p(x) \left| \frac{dx}{du} \right| \tag{2}$$

guarantees that the integrated probability over any interval in u is identical to that over the corresponding interval in x.

In problems involving two variables x and y, the joint probability density function $p(x, y)$ specifies the probability density for all x and y values. The normalization is again unity when the full range of possible values is included

$$\int p(x, y)\, dx\, dy = 1 \ . \tag{3}$$

The transformation rule (2) generalizes to multiple variables by replacing the derivative by the determinant of the Jacobian matrix $\mathbf{J}(x, y; u, v)$.

If x and y are statistically independent, $p(x, y) = p(x)\, p(y)$, that is, the joint probability density function is separable. Conversely, if this relation holds for all x and y, then the two variables are independent.

If one of the parameters is held fixed, the dependence on the other may be displayed by decomposing the joint probability density function

$$p(x, y) = p(y|x)\, p(x) \ , \tag{4}$$

[†]A more complete notation is often used in which the probability density is written as $p_x(\alpha)$, where x indicates the variable, and α its value. The abbreviated style is seldom ambiguous.

where $p(y|x)$ is called the conditional probability, which we read as the 'probability of y given x'. The variable on the right of the vertical bar is considered fixed (or given) and the one on the left is the free variable. To emphasize that x has a specific value, of say x_0, the conditional probability might be written as $p(y|x = x_0)$. Note that the independence of x and y implies for the conditional probability, $p(y|x) = p(y)$.

Often there are parameters that need to be included in the description of the physical situation that are otherwise unimportant for solving the stated problem. An example is the power spectrum of a time series, for which the phase of the Fourier transform is irrelevant.[14] The probability density function in y when x is irrelevant is found by integrating the joint probability over x

$$p(y) = \int p(x, y)\, dx \ . \tag{5}$$

This procedure for removing the 'nuisance variable' x is called marginalization.

2.2. Posterior Probability and Priors

The first aspect of Bayesian analysis involves the interpretation of Eq. (4). Suppose we wish to improve our knowledge concerning a parameter x. Our present state of certainty is characterized by the probability density function $p(x)$. We perform an experiment and take some data d. By decomposing the joint probability as in Eq. (4) both ways, and substituting d for y, we obtain Bayes's law:

$$p(x|d) = \frac{p(d|x)p(x)}{p(d)} \ . \tag{6}$$

We call $p(x|d)$ the posterior probability density function, or simply the posterior, because it effectively follows (temporally or logically) the experiment. It is the conditional probability of x given the new data d. The probability $p(x)$ is called the prior because it represents the state of knowledge before the experiment. The quantity $p(d|x)$ is the likelihood, which expresses the probability of the data d given any particular x. The likelihood is usually derived from a model for predicting the data, given x, as well as a probabilistic model for the noise. Bayes's law provides the means for updating our knowledge, expressed in terms of a probability density function, in light of some new information, similarly expressed.

The term in the denominator $p(d)$ may be considered necessary only for normalization purposes. As the normalization can be evaluated from the other terms, Bayes's law is often written as a proportionality, leaving out the denominator.

In one interpretation of Bayes's law, the prior $p(x)$ represents knowledge acquired in a previous experiment. In other words, it might be the posterior probability of the previous experiment. In this case, Bayes's law can be interpreted as the proper way to calculate the sum total of all the available experimental information.

In another interpretation of Bayes's law there may be no experimental data from a previous experiment, only general information about the parameter x. The prior might be viewed as a means to restrict x so that the posterior provides more information about x than the likelihood. In this situation, the prior is not necessarily restricted to what is known before the experiment. Indeed, many different priors might be employed in the Bayesian analysis to investigate the range of possible outcomes. As the proper choice for the prior should clearly depend on the domain of the problem, this topic will be addressed relative to image analysis in the next section.

However, a few uninformative priors are worth mentioning. Parameters can often be categorized as one of two types: location or scale. The values of location parameters can be positive or negative, and do not set the scale of things. Examples include position, velocity, and time. When x is a location parameter, there is no reason to prefer x over $x + \delta$. Thus, an uninformative prior for a location parameter is $p(x) =$

constant. On the other hand, scale parameters do set the scale of something and are constrained to be nonnegative. An example is the lifetime of a radioisotope. When x is a scale parameter, there is no reason to prefer a value of x over λx. Therefore, the appropriate uninformative prior for the logarithm of scale parameters is a constant, or, using (2), $p(x) \propto x^{-1}$. While both of these density functions cannot be integrated over their full range of $-\infty$ to $+\infty$ or 0 to $+\infty$, respectively, it is usually assumed that the normalization can be set by choosing appropriate generous limits to avoid divergence of the integrals.

2.3. Choice of Estimator

The second aspect of Bayesian analysis deals with the use of the posterior probability. While the posterior probability density function $p(x|d)$ fully expresses what we know about x, it may embody more information than is required. When attempting to summarize the results of an analysis, it may be necessary to represent $p(x|d)$ in more concise terms. For example, it might be desirable to quote a single value for x, which we call the estimate, designated as $\hat{x}$. Alternatively, it might be that a decision is required, perhaps in the form of a binary decision: yes, a certain object is present, or no, it is not. In this interpretation process some information concerning $p(x|d)$ is lost. The crucial point is that through cost analysis the Bayesian approach provides an optimal way to interpret the posterior probability. To achieve optimality, it is necessary to consider the costs (or risks) associated with making various kinds of errors in the interpretation process. The assignment of the proper cost function is typically considered to be a part of specifying the problem.

The choice of an estimation method to select the single value that is most representative of what we know about x should clearly depend on how one assigns significance to making errors of various magnitudes. As the posterior probability can be used to calculate the expected error, the choice of an estimation method can be based on the posterior probability density function itself.

A standard measure of the accuracy of a result is the variance (or mean square error). Given the posterior probability density function $p(x|d)$, the expected variance for an estimate $\hat{x}$ is

$$\int p(x|d)\,|x-\hat{x}|^2\,dx\ . \tag{7}$$

It is fairly easy to show that the estimator that minimizes (7) is the mean of the posterior probability density function

$$\hat{x} = \bar{x} = \int x\,p(x|d)\,dx\ . \tag{8}$$

An alternative cost function is the L1-norm:

$$\int p(x|d)\,|x-\hat{x}|\,dx\ , \tag{9}$$

for which the appropriate estimator is the median of the posterior density function

$$\int_{-\infty}^{\hat{x}} p(x|d)\,dx = \int_{\hat{x}}^{\infty} p(x|d)\,dx\ . \tag{10}$$

When any answer other than the correct one incurs the same increased cost, the obvious estimate is the value of x at the maximum of the posterior probability density function:

$$\hat{x} = \text{argmax}\ p(x|d) \tag{11}$$

This choice is the well-known maximum *a posteriori* (MAP) estimator.

For unimodal, symmetric density functions, which includes Gaussians, all the above estimators yield the same result. However, in many problems the posterior probability is asymmetric or has multiple peaks. Then these various estimators can yield quite different results, so the choice of an appropriate estimator becomes an issue. It is important to remember that each estimator minimizes a specific cost function.

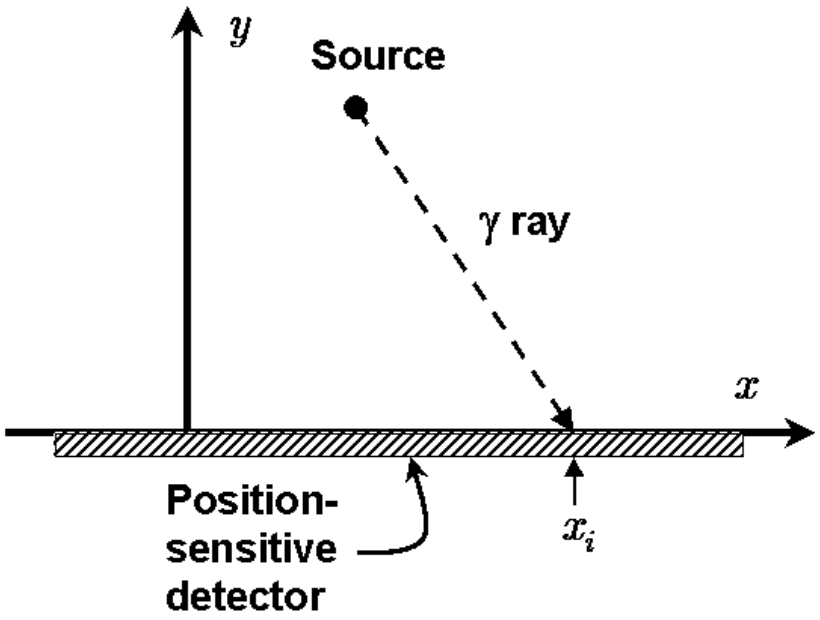

Figure 1. Schematic of a 2D problem. A radioactive source is placed at position (x, y). Gamma rays, emitted uniformly in angle, are detected by a linear detector that lies along the x axis. The x_i positions at which they hit the detector, denoted by dots on the x axis, are tallied. The problem is to estimate the location of the source.

The problem of making a decision based on data is very similar to that of estimating parameters. It involves interpretation of the posterior probability in terms of alternative models and deciding which is the 'best' match. The same considerations regarding the costs of making correct and incorrect decisions must be observed to achieve optimal results.

2.4. A Simple Example – the Infamous Cauchy Distribution

Suppose that a radioactive source is located at the position (x, y), as depicted in Fig. 1. A position-sensitive linear detector, colinear with the x axis, measures the position x_i that the ith gamma ray hits the detector. The measurements consist of the values x_i, $i = 1, ..., N$. Assume that the probability density function for the gamma rays is uniform in the angle θ at which they leave the source. From the relation $\tan(\theta) = -y/(x_i - x)$, which holds for $-\pi < \theta < 0$, the probability density function in x_i is obtained by using the transformation property (2) of density functions

$$p(x_i|x, y) = \frac{y}{\pi\,[y^2 + (x - x_i)^2]} \quad , \tag{12}$$

which is called the likelihood of event x_i. The coefficient in Eq. (12) provides the proper normalization (1). The likelihood for this problem is recognized to be a Cauchy distribution in x, notorious for the disagreeable attributes of possessing an undefined mean and an infinite variance.

The posterior probability density function for the x position of the source is given by Bayes's law

$$p(x|\{x_i\}, y) \propto p(\{x_i\}|x, y)\, p(x) \quad . \tag{13}$$

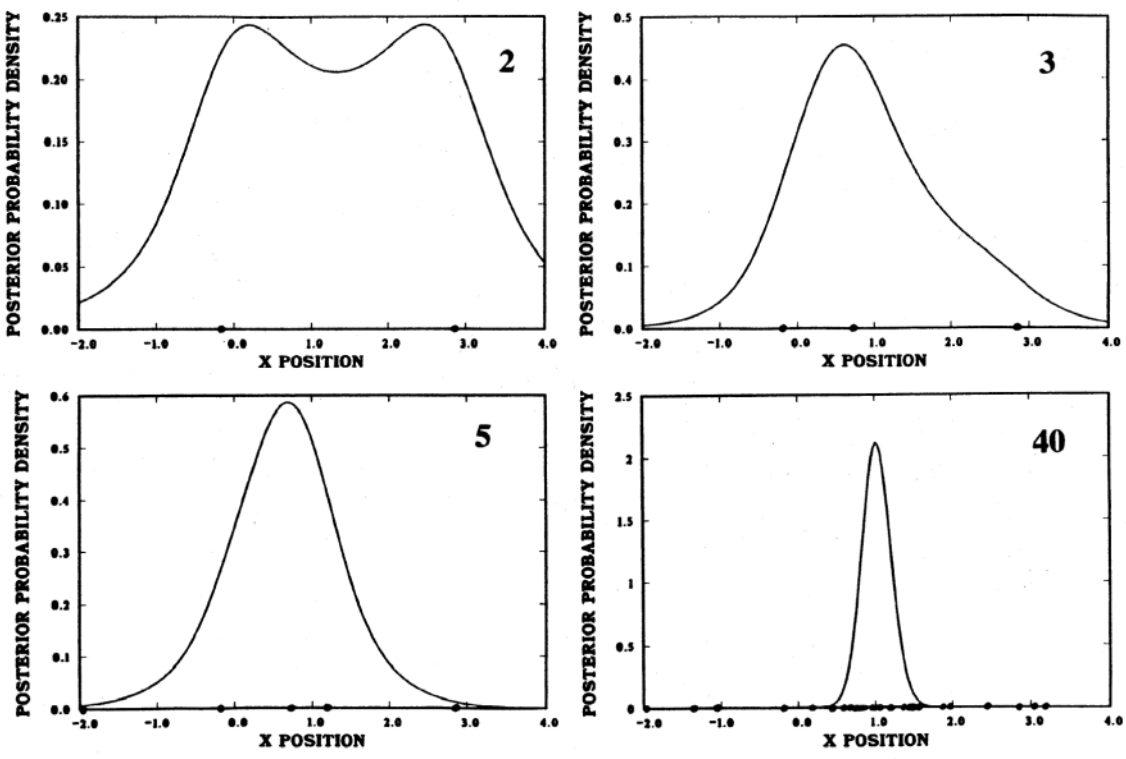

Figure 2. The posterior probability density function for the x position of the radioactive source assuming that the correct y is known. Each of these plots is shown for one simulated experiment. The number of x_i measurements is noted in the upper-right corner of each plot.

We do not have to keep track of multiplicative constants because the normalization determines these[‡]. Let us suppose we have no prior information about the location of the source. Then for the prior $p(x)$ we should use a constant over whatever sized region is required. Such a prior is noncommittal about the location of the source. The measured x_i clearly follow the likelihood (12), and, as the emission of one gamma ray can not effect the emission of another, the x_i are statistically independent. Thus the full posterior probability is

$$p(x|\{x_i\}, y) \propto \prod_i p(\{x_i\}|x, y) \propto \prod_i \left[\frac{y}{y^2 + (x - x_i)^2}\right] \tag{14}$$

Two fundamental characteristics of Bayesian analysis can already be identified. First, the calculation of the posterior probability is essentially a modeling effort based on the physical situation as we understand it. Second, the evaluation of the posterior probability proceeds through the model in the forward direction; one starts with assumed values of the parameters and predicts the measurements. Therefore, the numerical values of the posterior probability are often easy to calculate even when the model consists of a sequence of stages.

The posterior probability given by Eq. (14) is plotted in Fig. 2 for specific numbers of measurements. The plot for two measurements is bimodal, making it difficult to use the maximum posterior probability as an estimator for x. As the number of measurements increases, the width of the posterior density function

[‡]In careful analytical derivations it might be desirable to retain the constants to permit a final check of the results.

decreases, indicating less uncertainty in the knowledge of x. The broad tail of the Cauchy likelihood is suppressed as the number of measurements increases because the posterior probability involves a product of likelihoods.

We wish to estimate the x location of the source, assuming that y is known. We might consider using the average of the x_i measurements as an estimate of x:

$$\langle x_i \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i \ . \tag{15}$$

Alternatively, if we desired an estimate that minimizes the expected mean square error, we would use the mean of the posterior density function:

$$\bar{x} = \int x \, p(x|\{x_i\}, y) \, dx \ . \tag{16}$$

The performance of these two estimators was tested by simulating 1000 experiments, each involving a fixed number of measurements. The results are as follows:

Number of Measurements	rms Error in $\langle x_i \rangle$	rms Error in $\bar{x}$
1	112.3	112.3
2	56.80	56.80
3	42.57	1.73
5	28.21	1.14
10	28.40	0.52
20	18.48	0.35

We observe that the average value of the measurements performs terribly! This poor performance might have been anticipated, owing to the infinite variance of the Cauchy distribution. The variance in the average of N samples taken randomly from a density function is easily shown to be $1/N$ times the variance of the density function. Because the variance of the original density function is infinite, so will be the variance of the average of any finite number of samples. The reason that the rms error in $\langle x_i \rangle$ is not infinite is that only a finite number of trials were included. Because of the symmetry of the likelihood (12) for each event, the results for the posterior mean are identical to the sample average for one and two measurements. The posterior mean performs much better for three or more measurements.

We might want to include as part of the problem the determination of the y position. Then, instead of being a constant that is specified, y becomes an unknown variable. The full 2D posterior probability is easily written as $p(x, y|\{x_i\}) \propto p(x|\{x_i\}, y)p(y) \propto p(x|\{x_i\}, y)$, assuming the prior $p(y)$ is a constant. The 2D plots shown in Fig. 3 demonstrate how the lack of knowledge of y complicates the problem.

The prior is assumed to be a constant in the above example to reflect our complete ignorance of the source position. Thus the posterior probability is the same as the likelihood. The likelihood only states the relative probability of obtaining the specific set of measurements, given a particular x. Note that Bayes's law is necessary to gain information about x from the likelihood. If it were known that the source position was limited to a specific region, an appropriate prior would consist of a function that is a nonzero constant inside the region and zero outside. This prior would have the effect of eliminating the tails of the posterior probability outside the legitimate region.

The preceding example is a restatement of Gull's lighthouse problem.[7] Another more involved example of Bayesian analysis of a relatively 'simple' problem in 1D spectrum estimation can be found in Ref..[14]

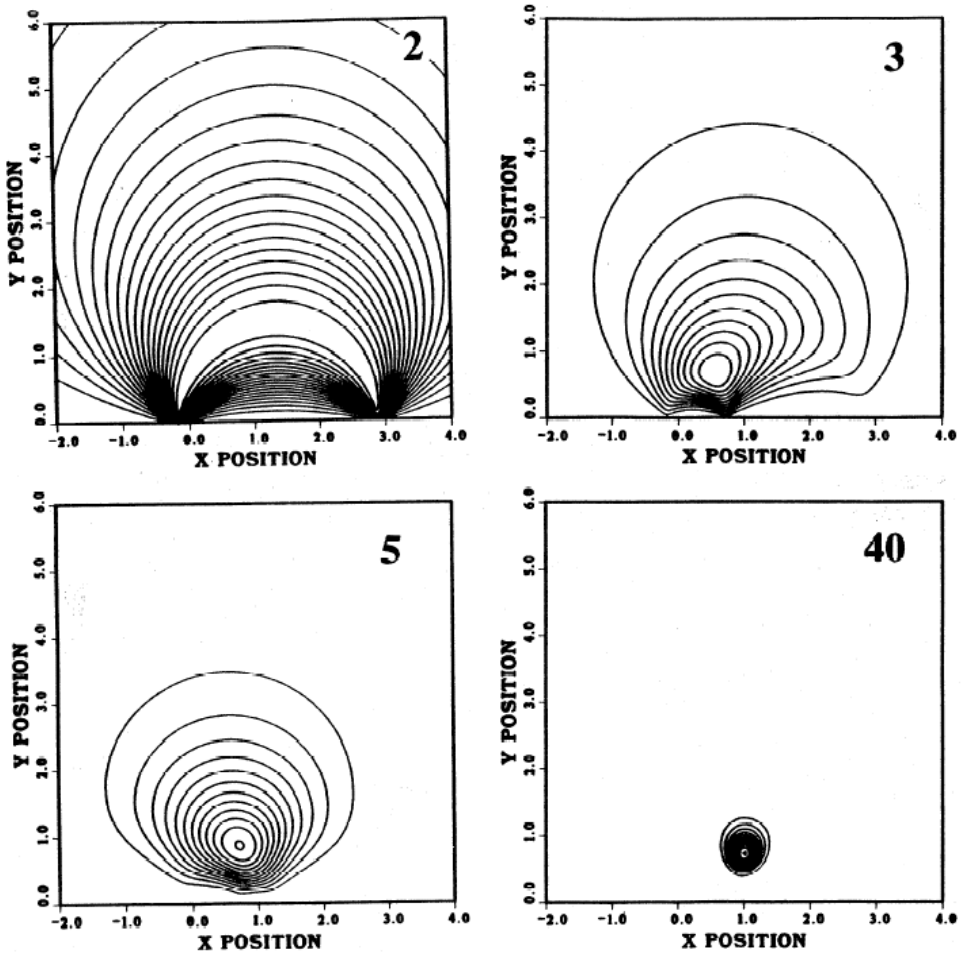

Figure 3. The joint posterior density function in both the x and y positions of the radioactive source.

3. BAYESIAN IMAGE ANALYSIS AND RECONSTRUCTION

We now move from problems involving a few variables into the realm of image analysis in which the number of variables can range from thousands to millions. The same Bayesian principles apply, but the computational burden is obviously magnified. The Bayesian method provides a way to solve image reconstruction problems that would otherwise be insoluble. The means is the prior, without which the MAP solution would collapse to the ML result. Thus the prior is indispensable. For an earlier review of Bayesian analysis applied to image recovery, see Ref..[15]

3.1. Posterior Probability

We follow the nomenclature of Hunt,[16] who introduced Bayesian methods to image reconstruction. An image is represented as an array of discrete pixels written as a vector $\mathbf{f}$ of length N (the number of pixels). We assume that there are M discrete measurements, written as a vector $\mathbf{g}$, which are linearly related to the amplitudes of the original image. We also assume that the measurements are degraded by additive, Gaussian noise, and for simplicity of presentation, that the noise is uncorrelated and stationary (rms noise is the same for all measurements). The measurements can be written as

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \mathbf{n} \ , \tag{17}$$

where $\mathbf{n}$ is the random noise vector, and $\mathbf{H}$ is the measurement matrix. In computed tomography (CT) the jth row of $\mathbf{H}$ describes the weight of the contribution of the image pixels to the jth projection measurement. Measurements that are nonlinearly related to the desired image can be easily dealt with in the Bayesian framework.[16–18]

Given the data $\mathbf{g}$, the posterior probability of any image $\mathbf{f}$ is given by Bayes's law (6) in terms of the proportionality

$$p(\mathbf{f}|\mathbf{g}) \ \propto \ p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f}) \ , \tag{18}$$

where $p(\mathbf{g}|\mathbf{f})$, the probability of the observed data given $\mathbf{f}$, is the likelihood and $p(\mathbf{f})$ is the prior probability of $\mathbf{f}$. The negative logarithm of the posterior probability density function is given by

$$-\log[p(\mathbf{f}|\mathbf{g})] = \phi(\mathbf{f}) = \Lambda(\mathbf{f}) + \Pi(\mathbf{f}) \ , \tag{19}$$

where the first term comes from the likelihood and the second term from the prior probability.

The likelihood is specified by the assumed probability density function of the fluctuations in the measurements about their predicted values (in the absence of noise). For the additive, uncorrelated Gaussian noise assumed, the negative log(likelihood) is just half of chi-squared

$$-\log[p(\mathbf{g}|\mathbf{f})] = \Lambda(\mathbf{f}) = \tfrac{1}{2}\chi^2 = \frac{1}{2\sigma_{\mathbf{n}}^2}|\mathbf{g} - \mathbf{H}\mathbf{f}|^2 \ , \tag{20}$$

which is quadratic in the residuals. The choice for the likelihood function should be based on the actual statistical characteristics of the measurement noise.

3.2. Priors

The second term $\Pi(\mathbf{f})$ in Eq. (19) comes from the prior probability density function. A Gaussian density function is often used for the prior, whose negative logarithm may be written in simplified form as

$$-\log[p(\mathbf{f})] = \Pi(\mathbf{f}) = \frac{1}{2\sigma_{\mathbf{f}}^2}|\mathbf{f} - \bar{\mathbf{f}}|^2 \ , \tag{21}$$

where $\bar{\mathbf{f}}$ is the mean and $\sigma_{\mathbf{f}}$ is the rms deviation of the prior probability density function. This oversimplified form excludes prior information about correlations and nonstationarities that might be essential to obtaining improved results,[19,15] but it suffices for now.

For a moment consider the properties of the MAP solution for which

$$\nabla_{\mathbf{f}}\phi = 0 \ , \tag{22}$$

provided that there are no side constraints on $\mathbf{f}$. When the likelihood and the prior are based on Gaussian distributions as in Eqs. (21 and 20), the MAP estimate is the solution to the set of linear equations

$$\nabla_{\mathbf{f}}\phi = -\frac{1}{\sigma_{\mathbf{n}}^2}\mathbf{H}^{\mathrm{T}}(\mathbf{g} - \mathbf{H}\mathbf{f}) + \frac{1}{\sigma_{\mathbf{f}}^2}(\mathbf{f} - \bar{\mathbf{f}}) = 0 \ , \tag{23}$$

where $\mathbf{H}^{\mathrm{T}}$ is the transpose of $\mathbf{H}$, which is the familiar backprojection operation in CT. These equations can be easily rearranged to express the solution for $\mathbf{f}$ in terms of simple matrix operations, which is equivalent to a Wiener filter for deblurring problems involving stationary blur matrices.[20] The properties of this solution in the context of CT reconstruction shed light on the general characteristics of the MAP solution. The use of *a priori* known constraints, such as nonnegativity, with the Gaussian prior has been shown to provide bona fide benefits for reconstruction from limited data.[21–24]

It is clear from (23) that the effect of the prior is to pull the MAP solution toward $\bar{\mathbf{f}}$ and away from the ML solution, for which $|\mathbf{g} - \mathbf{H}\mathbf{f}|^2$ is minimized. The strength of the prior relative to the likelihood is proportional to $\sigma_{\mathbf{n}}^2/\sigma_{\mathbf{f}}^2$. The amount of pull by the prior is proportional to this factor. As the prior contribution to the posterior vanishes, the MAP result approaches the ML solution. Hence, the selection of this factor is critical.

A physical interpretation of Eq. (23) provides a useful way to think about the MAP solution. Equation (23) contains two terms, each linear in the difference between $\mathbf{f}$ and another quantity. The linearity prompts one to interpret this equation in terms of linear forces acting on $\mathbf{f}$. The prior contributes a force that pulls the solution toward the default solution $\bar{\mathbf{f}}$ in proportion to the difference $\mathbf{f} - \bar{\mathbf{f}}$. The likelihood similarly provides a force pulling in the direction of the ML solution. The MAP solution then can be viewed as a problem in static equilibrium. Note that through its pull on the solution, the prior introduces a bias in the solution. In this physical analog the negative logarithm of the posterior probability ϕ corresponds to a potential energy.

In accordance with this physical analogy, the negative logarithm of any prior $\Pi(\mathbf{f})$ may be interpreted as an energy. In other words, any prior probability can be expressed in the form of a Gibbs distribution

$$p(\mathbf{f}) = Z^{-1}(\beta)\, \exp[-\beta W(\mathbf{f})] \;, \tag{24}$$

where $W(\mathbf{f})$ is an energy function derived from $\mathbf{f}$, β is a parameter that determines the strength of the prior, and $Z(\beta)$, known as the partition function, provides the appropriate normalization: $Z(\beta) = \int \exp[-\beta W(\mathbf{f})]\, d\mathbf{f}$. From (24) the negative logarithm of the prior is $\Pi(\mathbf{f}) = \beta W(\mathbf{f}) + \log Z(\beta)$. This Gibbs prior makes increases in W less probable.

Another prior that is suited to positive, additive quantities[10] takes the entropic form

$$\Pi(\mathbf{f}) = -\alpha S(\mathbf{f}) = -\alpha \sum_i [\mathrm{f}_i - \tilde{\mathrm{f}}_i - \mathrm{f}_i \log(\mathrm{f}_i/\tilde{\mathrm{f}}_i)] \;, \tag{25}$$

where $\tilde{\mathrm{f}}_i$ is called the default value and $S(\mathbf{f})$ is the entropy function. The derivation of this entropic prior is based in part on the axiom that the probability density function is invariant under scaling (i. e. is a scale variable, defined in Sect. 2.2.) This prior virtually guarantees that the MAP solution is positive through the subtle property that its partial derivative with respect to f_i is $\log(\mathrm{f}_i/\tilde{\mathrm{f}}_i)$, which force becomes infinitely strong in the positive direction as $\mathrm{f}_i \to 0$. In the physical interpretation mentioned above, the entropic prior replaces the linear force law of the Gaussian prior with a nonlinear one. The entropic restoring force is stronger than the linear one when $\mathrm{f}_i < \tilde{\mathrm{f}}_i$ and weaker when $\mathrm{f}_i > \tilde{\mathrm{f}}_i$.

Many authors, noting that images tend to have large regions in which the image values are relatively smooth, have promoted priors based on the gradient of the image:

$$\Pi(\mathbf{f}) = |\nabla \mathbf{f}|^2 \text{ or } \Pi(\mathbf{f}) = |\nabla^2 \mathbf{f}|^2 \;, \tag{26}$$

where the gradient is with respect to position. These priors assign increased energy to spatial variations in $\mathbf{f}$ and thus tend to favor smooth results. The difficulty with priors of this type is that they smooth the reconstructions everywhere. Thus edges are blurred. As it seems reasonable that edges are the dominant source of visual information in an image, it would follow that the blurring of edges is not desirable. The introduction of line-processes by Geman and Geman[25] remedies this deficiency by effectively avoiding smoothing across boundaries.

In many situations the approximate shape of an object to be reconstructed is known beforehand. The author has introduced a prior with which the approximate structural characteristics of an object can be included in a Bayesian reconstruction.[26–28] This prior is based on a warpable coordinate transformation from the coordinate system of the approximate object into that of the reconstruction. The prior on the

warp is specified as a Gibbs distribution expressed in terms of the energy associated with the deformation of the coordinates. This type of prior is closely related to deformable models, which have been employed for several years in computer vision and are becoming a mainstay of that field.[29] Deformable models have recently been interpreted in probabilistic terms.[30,31]

3.3. Methods of Reconstruction

It is usually desired in image reconstruction that a single image be presented as the 'result'. We humans have difficulty visualizing the full multidimensional probability density function comprising the complete $p(\mathbf{f}|\mathbf{g})$. A typical choice for a single image called the reconstruction is that which maximizes the *a posteriori* probability, the MAP estimate. Then the problem reduces to one of finding the image that minimizes $\phi(\mathbf{f})$. Such an optimization problem is relatively straightforward to solve, particularly when it is unconstrained. The general approach is to start with a good guess and then use the gradient with respect to the parameters to determine the direction in which to adjust the trial solution. The gradient is followed to the minimum at which condition (22) holds. In truth, finding the MAP solution is just about the only thing we know how to do. The other Bayesian estimators, posterior mean or posterior median, are very difficult to implement because of the high dimensionality of typical imaging problems. Skilling's recent efforts,[32] discussed in the next section, may overcome this impasse.

With the introduction of nonconvex functionals for priors, the issue of the uniqueness of a solution must be considered. A pat answer to such concerns is that one can use a method like simulated annealing[25] to find the global minimum of $\phi(\mathbf{f})$. Unfortunately this type of algorithm requires a very large amount of computation compared to the method of steepest descents mentioned above. Furthermore, the question of the reliability or uncertainty in the Bayesian result becomes a very complex one to address.

3.4. Examples

Present space does not permit reproduction of many of the following examples. The reader is encouraged to obtain the citations, which are well worth reading in their own right.

Even though we tend to think of reconstruction as the process of estimating a single image that best represents the collected data and prior information, in the Bayesian approach the posterior probability density function always exists. This point has been emphasized by the author[33,34] by displaying the posterior probability associated with a tomographic reconstruction as a function of two pixel values. Even when a pixel value is zero in a constrained reconstruction, the posterior probability for a positive value of that pixel can be readily calculated. The basic problem is that it is difficult to visualize the full information contained in the posterior probability regarding correlation between many pixels. Skilling, Robinson, and Gull[11] attempted to solve this problem by presenting a video that represents a 'random walk through the posterior landscape.' This effect is created by taking samples from the posterior probability in small random steps for each frame of the video. The viewer gets a feeling for the latitude in answers allowed by the posterior.

Skilling[32] has recently employed this random walk to implement a Monte Carlo evaluation of Bayesian estimators other than MAP, e. g. the posterior mean estimator. His results are encouraging.

The posterior probability density function may be used to interpret reconstructions. Binary decisions have been made using the posterior probability density functions associated with reconstructed images.[20,33–35] To date, this more complicated approach does not perform any better than basing the decision on the mean square difference between the reconstruction and the two test images.

In most clinical situations it is a human who interprets reconstructed images. Wagner, Myers, and Hanson[36–38] have compared how well humans can perform binary visual tasks relative to a variety of (algorithmic) machine observers. It is found that for the tasks investigated so far, human observers can

perform nearly as well as a computer. It is also concluded that the choice of the weight of the likelihood relative to the prior is important as it affects how well information is transferred to the decision maker.

Chen, Ouyang, Wong, and Hu[39] have used the Bayesian approach to interpret medical images for the purpose of improved display. They have adapted the prior of Geman and Geman[25] that smooths images inside identified regions but avoids smoothing across region boundaries. The boundaries are determined as an integral part of the Bayesian analysis. When applied to noisy bone-scan images, the results are very convincing. Boundaries are faithfully depicted and the noise is reduced without apparent loss in the detectability of very subtle features.

Spatial resolution is fundamentally limited in positron-emission tomography (PET). It has occurred to several investigators that the spatial resolution of PET reconstructions might be improved if high-resolution information could be incorporated. Application of the Geman and Geman prior[25] by Leahy and Yan[40] in a Bayesian approach seems to have yielded success. In their approach, the physiological structures in magnetic-resonance image (MRI) data are first delineated with line processes . The PET data are then reconstructed using these line processes, as well as freely adaptable ones. The final PET reconstructions show better separation between the various regions of activity. Chen *et al.*[41] have approached this problem in a similar manner and also obtained rewarding results. While both of these efforts yield quite reasonable PET reconstructions, final judgement on techniques such as these must await confirmation on the basis of subsidiary experiments that verify the physiological basis of the findings.

A final example due to the author[27,28] illustrates the use of prior structural information in a tomographic reconstruction. It is assumed that five parallel projections are available of the sausage-shaped object shown in Fig. 4. The projection angles are equally spaced over 180°. However, because the right-left symmetry of the object is assumed to be known, there are effectively only three distinct views. The 32×32 ART reconstruction presents only a fuzzy result. To exaggerate the extent to which flexibility can be incorporated in the reconstruction process, a circle is used for the prior geometric model. The important aspects to be included in the reconstruction are the expected sharp boundary of the object and its known constant amplitude. This structure is distorted to match the data through a coordinate transformation. A third-order polynomial is employed to permit a fair degree of flexibility. The assumed right-left symmetry reduces the number of warp coefficients from 20 to 11. The resulting MAP reconstruction (LR) reasonably matches the original, given that the available data consist of only three distinct views. Presumably a superior reconstruction would be obtained if the shape of the object were known better beforehand. This example illustrates the fact that this reconstruction comprises more than just the warped circular model; it also includes amplitude deviations from the model.

4. FURTHER ISSUES

The preceeding sections have provided only an overview of the use of Bayesian methods in the analysis of images. There are many other basic issues in Bayesian analysis. The existence of these issues does not necessarily detract from the Bayesian approach, but only indicates the depth of the subject that arises because of the introduction of the prior. As we have argued, the prior is the crucial additional feature of Bayesian analysis that facilitates solution of real-world problems.

The first and most important step in any analysis problem is to define the goal. This statement of purpose clarifies the direction in which to go and provides a means to determine whether there are any benefits to a proposed approach. The leading question in image analysis is whether the desired result should represent a final interpretation of the scene or whether the resulting image will be used as the basis of a subsequent interpretation. In the latter case, performance of the final interpretation may be adversely affected by too much use of prior information in the image processing stage. The image processing and final interpretation steps should be considered collectively as the ultimate performance of the system depends

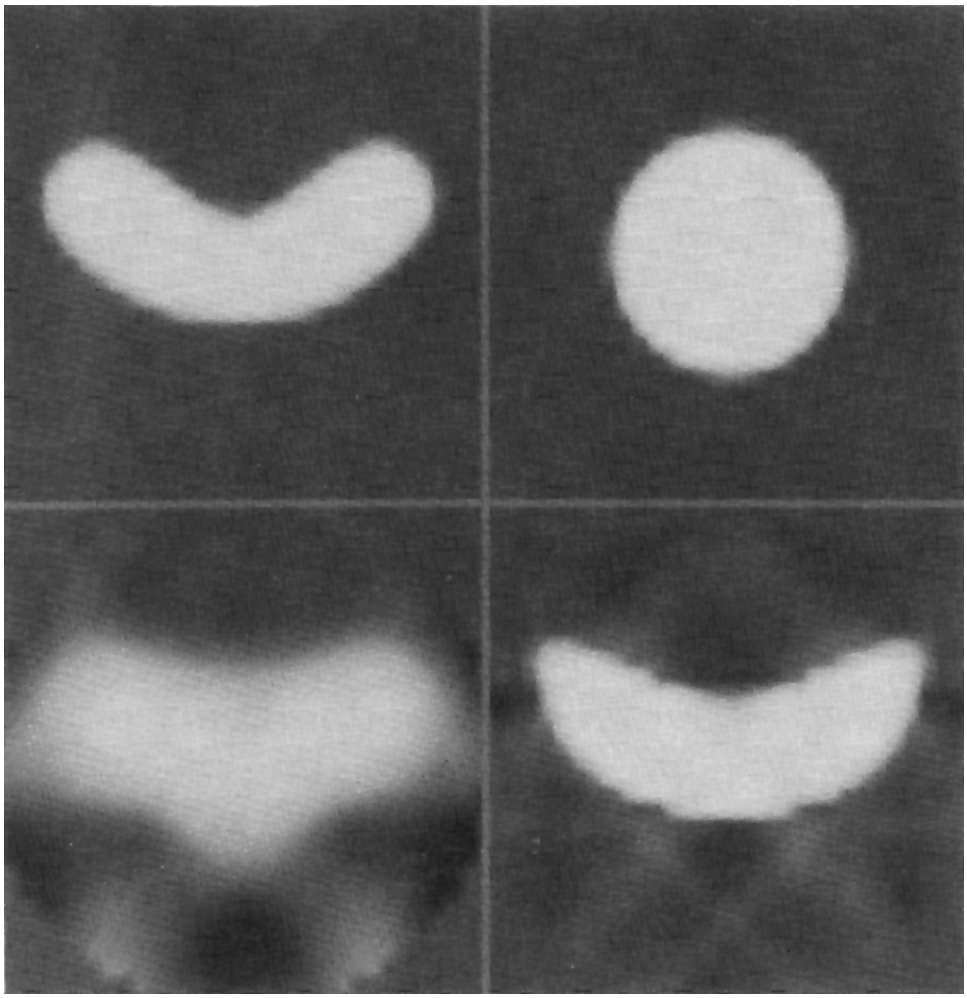

Figure 4. The ART reconstruction (lower-left) obtained from five noiseless parallel views poorly reproduces the original sausage-shaped object (upper-left). The MAP reconstruction (lower-right) obtained from the same data is based on the circular prior model (upper-right) subjected to a polynomial warp of third order, constrained to possess left-right symmetry.

on their cooperation. The resolution of most of the remaining issues revolves around this single, most critical point.

The costs of misinterpretation of the data, which represent a fundamental aspect of the Bayesian method, are often not given enough attention in image analysis. The cost (or risk) assessment can impact virtually every aspect of Bayesian analysis including determination of the strength of the prior, type of prior, type of estimator, etc.

If we are to present a single image as the reconstruction, which type of estimator should we use? Operationally the MAP estimator, which finds the maximum of the posterior, is usually the only one we know how to calculate. But, as discussed in Sect. 2.3, the mean or median of the posterior might be superior.

When our interest is restricted to a particular region of an image, should we marginalize over the

pixel values in the rest of the image? This issue has been investigated briefly[33,34] in regard to improving the performance of binary tasks without any firm conclusion. It is apparently not a large effect. There are potentially fundamental reasons to avoid marginalization, e. g. when there are many maxima in the posterior corresponding to disparate solutions.[2] As mentioned earlier, Skilling[32] may have found a way to implement the marginalization procedure.

Priors play a central role in Bayesian analysis. If the prior is too specific and too strong, the reconstruction can only closely resemble what is expected.[15] The subsequent interpreter will not know what is real and what is not. Furthermore, departures from what is anticipated may be missed. It seems reasonable to hypothesize that human observers make use of seemingly extraneous features in images, e. g. noise, to judge the reliability of the information contained in the image and thence to make decisions. Removal of all traces of such extraneous features might foil the human interpreter's ability to make reliable decisions. Therefore, one reason to avoid highly specific priors might be that they could interfere with the subsequent interpretation process. However, if the image analysis procedure is supposed to make the final interpretation, then a specific prior may be called for.

A methodology is needed for developing and refining priors for specific classes of problems in which very specific priors are desired. Consider as an example, radiographs taken of airplane parts at a fabrication plant. We might use a prior model based on the design of the 3D structure, together with a warpable coordinate system, to allow for deviations from the design.[26–28] How would we tailor the elastic moduli used to define the total strain energy of the warped coordinate system to incorporate accumulating information about which kinds of geometrical deviations are more likely to occur? And how would we include the relative costs of missing various kinds of deformations, some of which might be catastrophic if undetected?

The strength of the prior relative to the likelihood is a critical parameter in any Bayesian analysis. How should the prior's strength be determined? As the strength of the prior is increased, the MAP solution is biased more towards the default solution and blurred to an increased extent.[20,37,38] It seems that a relatively weak prior would be desirable to minimize these effects. But then the benefits of the prior might not be fully realized. Gull[8] has offered a means to determine the prior's strength based on the posterior, viewed as a function of α, the strength of the entropic prior (25). The value of α is arrived at by marginalization of the posterior with respect to the reconstruction values using a Gaussian approximation to the posterior at each stage of the search for the correct α. A number of experiments involving binary tasks performed by the author and his colleagues[36–38] have shown that this method yields reasonable results for both algorithmic and human observers. The strength of the prior determined by Gull's method falls within the rather broad region of optimal task performance of the defined visual tasks for underdetermined CT reconstructions. In these studies the old ad hoc method of selecting the strength of the prior so that χ^2 matches the number of measurements[16,8,42] is found to degrade observer performance. Gull's approach is based on a particular prior on α, namely, that it is constant in $\log(\alpha)$. In the final analysis the costs of making errors must play a significant role, which might have the effect of altering this prior.

As the Bayesian approach is typically based on a model, how do we choose which of many models to use? The historic answer to this question is to invoke the law of parsimony, or Occam's razor: use the simplest model that accounts for the data. The modern Bayesian answer[43,8,44] is to use the 'evidence', which is the probability of the model $\mathcal{M}$ given the data $p(\mathcal{M}|d)$. The evidence is introduced by broadening the hypothesis space to include the model as parameter. Taking the prior on the model to be constant and marginalizing over the parameters to be estimated (e. g. the pixel values), the correct model order is identified as a peak in the posterior. There are some concerns about the validity of the approximations used in this procedure.[45] However, the ability to determine models may become one of the most compelling reasons to use the Bayesian approach in image analysis.

At some point in every analysis problem it becomes necessary to determine how reliable the answer

is. In this regard Bayesian analysis is not much different than an analysis based on likelihood except that the prior can introduce a variety of effects that must be understood. The technical answer to specifying reliability[1] is that, for unbiased estimators the reciprocal of the Fisher information matrix, which is the expectation of the second derivative of the posterior, $-\langle \frac{\partial^2 p(\mathbf{f}|\mathbf{g})}{\partial \mathrm{f}_i \partial \mathrm{f}_j} \rangle$, sets the lower limit on the covariance between the i and j pixel values, f_i and f_j. This statement is the multivariate equivalent of the Cramér-Rao lower bound on the variance of a single variable (however Fisher wrote it down 23 years earlier). But this answer is not sufficient, because we are often interested in collective phenomena such as the presence/absence or accuracy of features that are spread over many pixels. Therefore, the correlations between pixels that are buried in the Fisher information matrix are crucial. Furthermore, we may wish to characterize the uncertainty by some measure other than the covariance, for example, in terms of confidence limits. We might want to know the probability that a specific object is present at a particular location in the image. Such questions will be difficult to answer in general terms when dealing with nonlinear, complex priors and complicated hypotheses. The answers will necessarily involve the numerical calculation of the posterior probability, and may even require a wholly new approach to man-computer interaction.

ACKNOWLEDGEMENTS

I have gained much from discussions with many, including Steve F. Gull, John Skilling, Harrison H. Barrett, Gregory S. Cunningham, James Gee, Kyle J. Myers, Richard N. Silver, Robert F. Wagner, David R. Wolf, and the Bayesian luncheon gang.

REFERENCES

1. H. L. Van Trees. *Detection, Estimation, and Modulation Theory - Part I.* John Wiley and Sons, New York, 1968.
2. G. E. P. Box and G. C. Tiao. *Inference in Statistical Analysis.* John Wiley and Sons, New York, 1973 (reprinted 1992).
3. A. P. Sage and J. L. Melsa. *Estimation Theory with Applications to Communications and Control.* Robert E. Krieger, Huntington, 1979.
4. B. R. Frieden. *Probability, Statistical Optics, and Data Testing: A Problem Solving Approach.* Springer-Verlag, New York, 1983.
5. J. O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.
6. J. C. Kiefer. *Introduction to Statistical Inference.* Springer-Verlag, New York, 1987.
7. S. F. Gull. Bayesian inductive inference and maximum entropy. In G. J. Erickson and C. R. Smith, editors, *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, pages 53–74. Kluwer Academic, Dordrecht, 1989.
8. S. F. Gull. Developments in maximum-entropy data analysis. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 53–71. Kluwer Academic, Dordrecht, 1989.
9. J. Skilling and R. K. Bryan. Maximum entropy image reconstruction: general algorithm. *Mon. Not. R. Ast. Soc.*, 211:111–124, 1984.
10. J. Skilling. Classic maximum entropy. In J. Skilling, editor, *Maximum Entropy and Bayesian Methods*, pages 45–52. Kluwer Academic, Dordrecht, 1989.
11. J. Skilling, D. R. T. Robinson, and S. F. Gull. Probabilistic displays. In Jr. W. T. Grandy and L. H. Shick, editors, *Maximum Entropy and Bayesian Methods*, pages 365–368. Kluwer Academic, Dordrecht, 1991.
12. E. T. Jaynes. Bayesian methods - An introductory tutorial. In J. H. Justice, editor, *Maximum Entropy and Bayesian Methods in Applied Statistics.* Cambridge University Press, Cambridge, 1986.
13. R. P. Cox. Probability, frequency and reasonable expectation. *Am. J. Phys.*, 14:1–13, 1946.
14. G. L. Bretthorst. Bayesian spectrum analysis and parameter estimation. In G. J. Erickson and C. R. Smith, editors, *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, pages 75–145. Kluwer Academic, Dordrecht, 1989.

15. K. M. Hanson. Bayesian and related methods in image reconstruction from incomplete data. In Henry Stark, editor, *Image Recovery: Theory and Application*, pages 79–125. Academic, Orlando, 1987.
16. B. R. Hunt. Bayesian methods in nonlinear digital image restoration. *IEEE Trans. Comp.*, C-26:219–229, 1977.
17. K. M. Hanson. Tomographic reconstruction of axially symmetric objects from a single radiograph. In *Proc. 16th Inter. Cong. on High Speed Photography and Photonics. (Proc. SPIE,* 491:180-187*)*, Strasbourg, 1984.
18. K. M. Hanson. A Bayesian approach to nonlinear inversion: Abel inversion from x-ray data. In P. Nelson, V. Fabor, D. L. Seth, and A. B. White, Jr., editors, *Transport Theory, Invariant Imbedding, and Integral Equations*, pages 363–368. Marcel Dekker, New York, 1987.
19. K. M. Hanson and G. W. Wecksung. Bayesian approach to limited-angle reconstruction in computed tomography. *J. Opt. Soc. Amer.*, 73:1501–1509, 1983.
20. K. M. Hanson. Object detection and amplitude estimation based on maximum *a posteriori* reconstructions. *Proc. SPIE*, 1231:164–175, 1990.
21. K. M. Hanson. Method to evaluate image-recovery algorithms based on task performance. *Proc. SPIE*, 914:336–343, 1988.
22. K. M. Hanson. Optimization for object localization of the constrained algebraic reconstruction technique. *Proc. SPIE*, 1090:146–153, 1989.
23. K. M. Hanson. Optimization of the constrained algebraic reconstruction technique for a variety of visual tasks. In D. A. Ortendahl and J. Llacer, editors, *Proc. Information Processing in Medical Imaging*, pages 45–57. Wiley-Liss, New York, 1990.
24. K. M. Hanson. Method to evaluate image-recovery algorithms based on task performance. *J. Opt. Soc. Amer.*, A7:45–57, 1990.
25. S. Geman and D. Geman. Stochastic relaxation, Gibb's distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-6:721–741, 1984.
26. K. M. Hanson. Reconstruction based on flexible prior models. *Proc. SPIE*, vol. 1652:183–191, 1992.
27. K. M. Hanson. Flexible prior models in Bayesian image analysis. In A. Mohammad-Djafari and G. Demoment, editors, *in Maximum Entropy and Bayesian Methods*. Kluwer Academic, Dordrecht, 1993.
28. K. M. Hanson. Bayesian reconstruction based on flexible prior models. To be published in *J. Opt. Soc. Amer.* **A**, 1993.
29. B. C. Vemuri, ed. *Geometric Methods in Computer Vision. Proc. SPIE*, 1570, 1991.
30. R. Szeliski. Probabilistic modeling of surfaces. *Proc. SPIE*, 1570:154–165, 1991.
31. R. Szeliski and D. Terzopoulos. Physically-based and probabilistic models for computer vision. *Proc. SPIE*, 1570:140–152, 1991.
32. J. Skilling. Bayesian computing: CLOUDS. In A. Mohammad-Djafari and G. Demoment, editors, *Maximum Entropy and Bayesian Methods*. Kluwer Academic, Dordrecht, 1993.
33. K. M. Hanson. Simultaneous object estimation and image reconstruction in a Bayesian setting. *Proc. SPIE*, 1452:180–191, 1991.
34. K. M. Hanson. Making binary decisions based on the posterior probability distribution associated with tomographic reconstructions. In G. J. Erickson, editor, *Maximum Entropy and Bayesian Methods*. Kluwer Academic, Dordrecht, 1992.
35. K. J. Myers and K. M. Hanson. Task performance based on the posterior probability of maximum-entropy reconstructions with MEMSYS 3. *Proc. SPIE*, 1443:172–182, 1991.
36. R. F. Wagner, K. J. Myers, and K. M. Hanson. Task performance on constrained reconstructions: Human observers compared with suboptimal Bayesian performance. *Proc. SPIE*, 1652:352–362, 1992.
37. K. J. Myers, R. F. Wagner, and K. M. Hanson. Binary task performance on images reconstructed using memsys 3: comparison of machine and human observers. In G. J. Erickson, editor, *Maximum Entropy and Bayesian Methods*. Kluwer Academic, Dordrecht, 1992.
38. R. F. Wagner, K. J. Myers, and K. M. Hanson. Rayleigh task performance in tomographic reconstructions: comparison of human and machine performance. to be published in *Proc. SPIE*, 1898, 1993.
39. C. Chen, X. Ouyang, W. H. Wong, and X. Hu. Improvement of medical images using Bayesian processing. *Proc. SPIE*, 1652:489–491, 1992.
40. R. Leahy and X. Yan. Incorporation of anatomical MR data for improved functional imaging with PET. In D. A. Ortendahl and J. Llacer, editors, *Info. Processing in Med. Imag.*, pages 105–120. Wiley-Liss, 1991.

41. X. Ouyang, W. H. Wong, V. E. Johnson, X. Hu, and C. Chen. Reconstruction of PET images by Bayesian data augmentation and Gibbs priors – Part II: Incoporation of prior boundary information. submitted to *IEEE Trans. Med. Imaging*, 1993.
42. E. Veklerov and J. Llacer. Stopping rule for the MLE algorithm based on statistical hypothesis testing. *IEEE Trans. Med. Imaging*, MI-6:313–319, 1987.
43. H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 1939.
44. D. J. C. MacKay. Bayesian interpolation. submitted to *Neural Comput.*, 1992.
45. C. E. M. Strauss, D. H. Wolpert, and D. R. Wolf. Alpha, evidence, and the entropic prior. In A. Mohammad-Djafari and G. Demoment, editors, *Maximum Entropy and Bayesian Methods*. Kluwer Academic, Dordrecht, 1993.