ARTICLES

Continuous probability distributions from finite data

David M. Schmidt
Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico 87545
(Received 11 August 1999)

Recent approaches to the problem of inferring a continuous probability distribution from a finite set of data have used a scalar field theory for the form of the prior probability distribution. This paper presents a more general form for the prior distribution that has a geometrical interpretation and can improve the specificity of likely solutions. It is also demonstrated that a numerical sampling of the posterior probability distribution can be used as an alternative to a histogram for visualization and to make probabilistic inferences from the data.

PACS number(s): 05.10.−a, 02.50.Wp

Inferring the continuous probability distribution from which a finite number of data samples were drawn is an example of an ill-posed inverse problem: there are many different distributions that could have produced the given finite data. Often one has prior information, separate from the data themselves, that could be used to reduce or weight the space of possible solutions. The ability to combine prior information with that from the data and to produce more informed results is one reason Bayesian approaches have been used to address a number of ill-posed inverse problems [1]. In the Bayesian approach a prior probability distribution over the space of possible solutions is constructed that reflects the prior information. This prior probability distribution is combined with the likelihood of the data given any particular possible solution, using Bayes's rule of probability, to produce a posterior probability distribution over the space of possible solutions. This posterior distribution encapsulates all the information available, both from the data and from the prior information, and can be used to make probabilistic inferences. Although prior information is sometimes considered subjective and its use may be controversial [2], prior information is essential in order to reduce the large range of likely solutions associated with ill-posed inverse problems. Indeed, one should tailor the prior distribution to incorporate all of the pertinent prior information available for each problem in order to maximize the specificity of the resulting posterior distribution. Even so, the posterior distribution may be broad and the most likely solution may not be representative of the full range of likely solutions. In such cases it is important to consider the full range of likely solutions when making inferences.

For the problem of inferring the continuous distribution from which a finite number of data have been drawn, a number of recent articles have described the utility of using a particular form for the prior distribution that may be viewed in field-theoretic terms and is designed to favor ''smooth'' distributions by penalizing large gradients [3–5]. The present paper describes a generalization of this prior distribution that can incorporate a wider range of prior information and has a geometrical interpretation that aids in understanding its properties. A number of examples are presented that illustrate how both of these features can be useful for generating more informed results. This paper concludes by demonstrating how a numerical sampling of the posterior can be used for displaying results (as an alternative to the histogram) and for making probabilistic inferences.

Let $P[Q|x_1,\ldots,x_N]$ denote the posterior probability that the target distribution $Q(x)$ describes the data $x_1,\ldots,x_N$. By Bayes's rule,

$$P[Q|x_1,\ldots,x_N]=\frac{P[x_1,\ldots,x_N|Q]P[Q]}{P[x_1,\ldots,x_N]} \tag{1}$$

$$=\frac{Q(x_1)\cdots Q(x_N)P[Q]}{\int \mathcal{D}Q\,Q(x_1)\cdots Q(x_N)P[Q]}, \tag{2}$$

where $P[Q]$ is the prior probability of the target distribution Q. By setting $Q(x)=\psi^2(x)$ [6], where ψ may take any value in $(-\infty,\infty)$, we may insure that Q is non-negative. ψ is referred to as the *amplitude* by analogy with quantum mechanics [4].

A particular form for $P[Q]$, or rather $P[\psi]$, that has been presented in order to, the authors say, incorporate a bias that Q be ''smooth'' is [4,3,5]

$$P[\psi]=\frac{1}{Z}\exp\left[-\int dx\,\frac{\ell^2}{2}(\partial_x\psi)^2\right]\delta\left(1-\int dx\,\psi^2\right), \tag{3}$$

where Z is the normalization factor and ℓ is a constant that controls the penalty applied to gradients. The δ function enforces normalization of the distribution Q. While Eq. (3) may be viewed as defining a scalar field theory [3] it may also be viewed in more traditional statistical terms. After integrating by parts and using standard quantum mechanics notation, Eq. (3) becomes

$$P[\psi]=\frac{1}{Z}\exp\left[-\frac{1}{2}\langle\psi|V^{-1}|\psi\rangle\right]\delta(1-\langle\psi|\psi\rangle), \tag{4}$$

where $V=-\ell^2\partial_x^2$ is a positive, symmetric (Hermitian) operator within the Hilbert space for ψ. This prior distribution

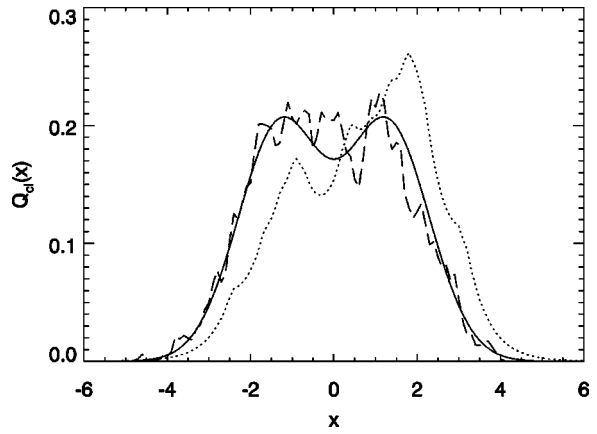

FIG. 1. The most likely distributions from an inverse Laplacian prior distribution with $\ell=6$ and from $N=20$ (dotted line) and $N=1000$ (dashed line) data drawn randomly from a target distribution consisting of the sum of two normal distributions (solid curve).

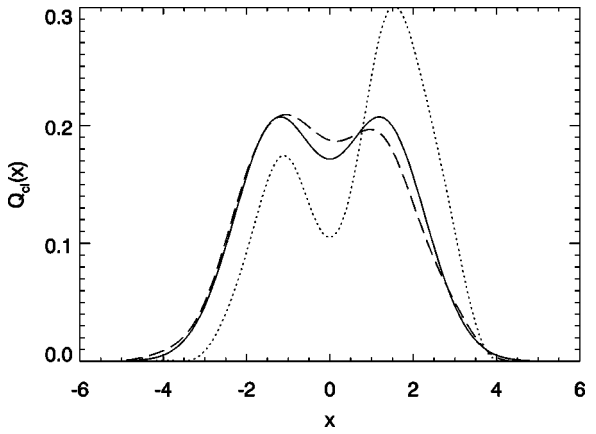

FIG. 2. The most likely distributions from a sinc function prior distribution with $k_0=3.33$ and from the same $N=20$ (dotted line) and $N=1000$ (dashed line) data used for the examples in Fig. 1, which were drawn randomly from a target distribution consisting of the sum of two normal distributions (solid curve).

is recognized as a generalization of a multidimensional Gaussian distribution with V acting as the covariance operator. Continuing this analogy, we write

$$V(\mathbf{x},\mathbf{y})=\sigma(\mathbf{x})\sigma(\mathbf{y})\rho(\mathbf{x},\mathbf{y}) \tag{5}$$

where $\sigma^2(\mathbf{x})$ is the variance at $\mathbf{x}$ and ρ is the correlation function. Information about smoothness is encoded in the correlation function. For example, if the distribution from which the $\{\mathbf{x}_i\}$ were drawn is expected to be smooth over distances smaller than a certain spatial scale then the correlation function should be near unity over distances smaller than this scale. The prior distribution used in [3,4] [Eq. (3)] uses one particular form for the covariance operator ($V=-\ell^2\partial_x^2$). Equation (4) allows for many other forms of the covariance operator and thus is a generalized form for the prior distribution that can be used to better encode whatever prior information is available for the particular problem at hand.

A useful feature of this prior probability distribution is that it can be viewed in geometrical terms. The eigenfunctions of the operator V form a basis for the space of ψ. The normalization constraint restricts ψ to lie on a hyperspherical surface of radius one. Those eigenfunctions with larger eigenvalues are more likely, *a priori*. If V has any eigenvalues that are zero then the corresponding eigenfunctions form a basis for a subspace that is, orthogonal to ψ; that is the prior distribution prevents ψ from having any components along these eigenfunctions. As will be demonstrated, this geometrical interpretation is useful for estimating the effects that different choices for V will have on the resulting solutions.

With this form for the prior distribution [Eq. (4)] the probability $P[Q|x_1,\ldots,x_N]$ of a distribution Q given the data is

$$P[\psi|x_1,\ldots,x_N]\propto\psi^2(x_1)\cdots\psi^2(x_N)\times\exp\left[-\frac{1}{2}\langle\psi|V^{-1}|\psi\rangle\right]\delta(1-\langle\psi|\psi\rangle) \tag{6}$$

$$=e^{-S[\psi]}\delta(1-\langle\psi|\psi\rangle), \tag{7}$$

where the effective action S is

$$S[\psi]=\frac{1}{2}\langle\psi|V^{-1}|\psi\rangle-2\sum_i\ln(\langle x_i|\psi\rangle). \tag{8}$$

The most likely distribution given the data is that function $\psi_{\rm cl}$ which minimizes the effective action subject to the normalization constraint. To enforce this constraint a Lagrange multiplier term $\lambda(1-\langle\psi|\psi\rangle)/2$ is subtracted from the action. Variational methods then lead to the following equations for $\psi_{\rm cl}$ and λ:

$$|\psi_{\rm cl}\rangle=2\sum_i\frac{(V^{-1}+\lambda I)^{-1}|x_i\rangle}{\langle x_i|\psi_{\rm cl}\rangle}, \tag{9a}$$

$$\langle\psi_{\rm cl}|\psi_{\rm cl}\rangle=1. \tag{9b}$$

The solution to these equations may be written

$$|\psi_{\rm cl}\rangle=\sum_i a_iU(\lambda)|x_i\rangle, \tag{10}$$

where $U(\lambda)=(V^{-1}+\lambda I)^{-1}$. Equations (9) imply

$$a_i\sum_j a_j\langle x_i|U(\lambda)|x_j\rangle=2,\quad i=1,\ldots,N, \tag{11a}$$

$$\sum_{i,j}a_ia_j\langle x_i|U^2(\lambda)|x_j\rangle=1. \tag{11b}$$

These $N+1$ nonlinear equations determine λ and the a_i and may be solved using Newton's method [4].

The covariance operator V in the prior distribution should be chosen to reflect the prior information that one has available. A few examples with three different forms for V are described below in order to illustrate the effects that different choices of V can have on the resulting most likely distributions. First consider the case used in [4,3] in which the prior covariance operator is an inverse Laplacian in one dimension, $V^{-1}=-\ell^2\partial_x^2$. Here ℓ is assumed to be known. If its value is uncertain then in the Bayesian framework, a prior

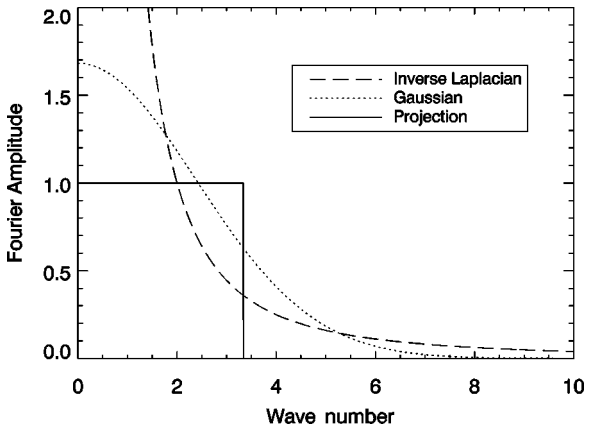

FIG. 3. The Fourier spectra of the three types of covariance operators shown in the legend. Free parameters in each case have been set to correspond roughly to a cutoff at wave number $k_0 = 3.33$.

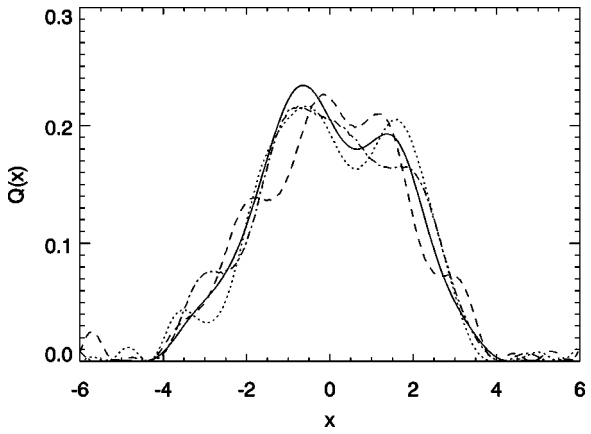

FIG. 4. Four different samples of $Q(x)$ drawn from the posterior distribution for the case in which 200 data points were drawn from the double normal target distribution used in the previous examples and the sinc function covariance operator was used for the prior distribution.

distribution for ℓ that reflects this uncertainty should be constructed and marginalized or integrated over when calculating the posterior distribution. With this form for V,

$$U(\lambda) = (-\ell^2 \partial_x^2 + \lambda I)^{-1}, \tag{12}$$

which is the Green's function of the modified Helmholtz equation. The solutions of this equation are well known, even for dimensions larger than one [7]. In particular, in one dimension the most likely solution $\psi_{\mathrm{cl}}(x)$ is, from Eq. (10),

$$\psi_{\mathrm{cl}}(x) = \sum_i a_i \frac{1}{2k\ell^2} \exp(-k|x - x_i|) \tag{13}$$

where $k = \sqrt{\lambda}/\ell$. Examples of the most likely probability distributions for this case with $\ell = 6$ are shown in Fig. 1. For these examples the data were drawn from a target distribution consisting of the sum of two normal distributions, shown as the solid curve in the figure. The most likely distributions are not very smooth, despite the fact that they result from a prior distribution that was used in order to produce smooth solutions [3,4]. This is a general feature that is not just peculiar to this particular example because, as can be seen by the form of Eq. (13), the most likely solutions from this prior distribution are constructed from functions that do not have continuous derivatives. The following examples demonstrate different choices of V that do a better job of encoding prior information about smoothness.

For the second example, consider the case in which the prior covariance operator has a correlation function which is a Gaussian,

$$V(\mathbf{x},\mathbf{y};r) = \sigma^2 \exp\left[\frac{-(\mathbf{x}-\mathbf{y})^2}{2r^2}\right]. \tag{14}$$

Here σ^2 is the prior variance for the magnitude of the target probability distribution and r is a correlation scale below which the target probability distribution is believed to be smooth. In this case it is useful to expand U in an operator product expansion in V,

$$U(\lambda) = V(1 - \lambda V + \lambda^2 V \cdot V - \lambda^3 V \cdot V \cdot V + \cdots). \tag{15}$$

Because $V^N \propto V(\mathbf{x},\mathbf{y};\sqrt{N}r)$ for this particular V, Eq. (15) generates a multiresolution expansion, analogous to a wavelet expansion, for U and therefore also for ψ_{cl}, consisting of Gaussians of ever increasing width, with the finest scale being represented by the original $V(\mathbf{x},\mathbf{y};r)$. This functional form for V therefore generates a most likely probability distribution that has finite derivatives to all orders and is generally more smooth than that from the first example.

For the final example, consider the case in which the prior covariance operator is a projection operator that projects onto the subspace formed by functions having only Fourier wave numbers smaller than a particular wave number k_0. In one dimension this covariance operator is the sinc function,

$$V(x,y;k_0) = \frac{\sin[k_0(x-y)]}{\pi(x-y)}. \tag{16}$$

Because this is a projection operator, $V \cdot V = V$ and from Eq. (15), U for this case is simply $U(\lambda) = V/(1+\lambda)$. The most likely amplitude therefore consists of sums of sinc functions centered at each data point. Examples of the most likely probability distribution using this prior distribution with $k_0 = 3.33$ are shown in Fig. 2. The same data used for the examples in Fig. 1 were used here. Even with only 20 data points the most likely solution indicates a doubly peaked distribution. Both of the examples here are more smooth than those generated by the prior distribution discussed above in the first example and shown in Fig. 1.

It is useful to examine the Fourier spectrum of the prior covariance operator in order to understand some of the properties of the resulting most likely distribution. The Fourier spectra of the three covariance operators considered in the above examples are shown in Fig. 3. Because of the geometrical interpretation of the prior distribution [Eq. (4)] it is clear that those wave numbers with larger Fourier amplitudes are more likely, *a priori*. However, in order to maximize the likelihood of the given data, the most likely amplitude will tend to consist of the largest possible wave number components. Because the sinc function covariance operator has the sharpest high wave number cutoff, it will tend to generate the smoothest most likely distribution. Conversely, the inverse

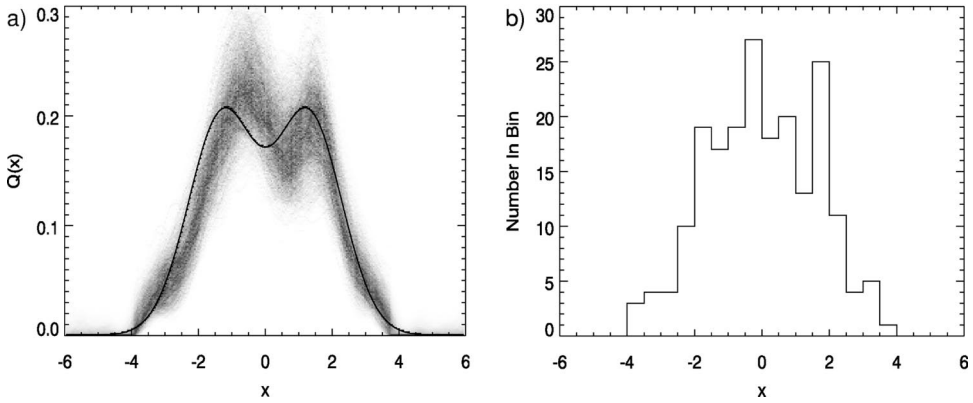

FIG. 5. Two ways of visualizing data drawn from a continuous distribution. (a) The Bayesian posterior distribution, represented by a two-dimensional histogram of the values of $Q(x)$ across the MCMC samples, shown in gray scale where darker shades indicate a larger number of occurrences. The solid line is the true or target distribution. (b) Conventional histogram of the same data.

Laplacian covariance operator will tend to produce the least smooth most likely distribution.

Although examining most likely solutions is useful for illuminating the effects that different prior distributions can have, it is the posterior probability distribution over the space of possible solutions that should be of the most interest. The posterior distribution summarizes all of the available information, both that from the data and that from the prior information, and is less sensitive to small differences in prior distributions than is the most likely solution. One way to investigate other likely solutions is to approximate the posterior distribution by a Gaussian distribution about the most likely solution, as discussed in [4]. Another way is to numerically generate samples from the posterior distribution, for example using the technique of Markov chain Monte Carlo (MCMC) [8]. As an illustration, 10^4 samples were generated, using a Metropolis MCMC algorithm [9], from the posterior distribution for the case in which 200 data points were drawn from the double normal target distribution used in the previous examples and the sinc function covariance operator was used for the prior distribution. Figure 4 shows a few of these sample distributions. The degree of variability among the samples is representative of the uncertainty in the solution, while features that are common across a large fraction of the samples are associated with a high posterior probability. Because these samples are distributed according to the posterior distribution, the posterior probability of any particular feature can be easily computed by calculating the fraction of samples possessing that feature. Finally, a convenient way to visualize the posterior distribution using these samples is to plot the two-dimensional histogram of the values of $Q(x)$ across the samples, as shown in Fig. 5a. This is a very effective way of conveying the general shape and uncertainty of $Q(x)$, and is more illuminating than the conventional histogram of the same data, shown in Fig. 5b.

I thank C.C. Wood for helpful discussions and for comments on the manuscript. This work was supported by Los Alamos National Laboratory, by NCRR/NIH Grant No. RR13630, and by NIDA/NIMH Grant No. DA/MH09972.

[1] Recent examples include K. Ertl, W. Vonderlinden, V. Dose, and A. Weller, Nucl. Fusion **36**, 1477 (1996); A. M. Thompson, J. C. Brown, I. J. D. Craig, and C. Fulber, Astron. Astrophys. **265**, 278 (1992); D. M. Schmidt, J. S. George, and C. C. Wood, Hum. Brain Mapp. **7**, 195 (1999).

[2] For a discussion of Bayesian and conventional statistics in the physical sciences, see, for example, B. Efron, Am. Stat. **40**, 1 (1986); and G. J. Feldman and R. D. Cousins, Phys. Rev. D **57**, 3873 (1998).

[3] W. Bialek, C. G. Callan, and S. P. Strong, Phys. Rev. Lett. **77**, 4693 (1996). These authors use a different, exponential relation between $Q(x)$ and $\psi(x)$ $(Q(x) \propto \exp[-\psi(x)])$ but the same prior distribution for $\psi(x)$ as in the current paper [i.e., Eq. (3)].

[4] T. E. Holy, Phys. Rev. Lett. **79**, 3545 (1997).

[5] For a reparametrization invariant geometrical formulation, see V. Periwal, Phys. Rev. Lett. **78**, 4671 (1997).

[6] I. J. Good and R. A. Gaskins, Biometrika **58**, 255 (1971).

[7] For example, see G. Arfken, *Mathematical Methods for Physicists* (Academic Press, Orlando, FL, 1985).

[8] For example, see W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman & Hall, London, 1996).

[9] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).