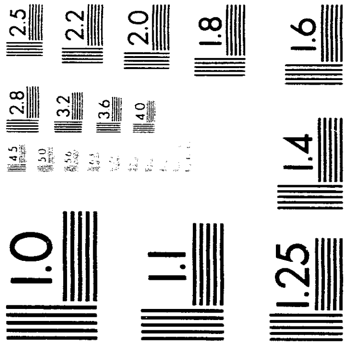
2.5
2.2
2.0
1.8
1.6
2.8
3.2
3.6
4.0
1.4
1.0
1.1
1.25

1 of 1

Conf-9308188--1

LA-UR-93- 3550

TITLE: DENSITY ESTIMATION BY MAXIMUM QUANTUM ENTROPY

AUTHOR(S): R. N. Silver, T-11
T. Wallstrom, T-13
H. F. Martz, A-1

SUBMITTED TO: Maximum Entropy and Bayesian Methods, 93

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

878

DENSITY ESTIMATION BY MAXIMUM QUANTUM ENTROPY

R. N. Silver, T. Wallstrom, H. F. Martz
MS B262
Los Alamos National Laboratory
Los Alamos, NM 87545

ABSTRACT. A new Bayesian method for non-parametric density estimation is proposed, based on a mathematical analogy to quantum statistical physics. The mathematical procedure is related to maximum entropy methods for inverse problems and image reconstruction. The information divergence enforces global smoothing toward default models, convexity, positivity, extensivity and normalization. The novel feature is the replacement of classical entropy by quantum entropy, so that local smoothing is enforced by constraints on differential operators. The linear response of the estimate is proportional to the covariance. The hyperparameters are estimated by type-II maximum likelihood (evidence). The method is demonstrated on textbook data sets.

1. Introduction

Non-parametric density estimation has been studied extensively by statisticians. If a set of N_s observations, $\{x_i\}$, is identically and independently drawn from a probability density function $f(x)$, the problem is to estimate f when no parametric form is known. A variety of non-Bayesian methods, such as histograms and kernel density estimators, have been developed and applied to density estimation [for reviews, see Silverman, 1986; Izenman, 1991]. There has been comparatively little work on maximum penalized likelihood methods [see, e.g., Good and Gaskins, 1980], despite their potential advantages such as a Bayesian interpretation, the ability to combine explicit prior knowledge with the data, the ability to combine data from different sources, etc. The situation in density estimation contrasts sharply with that of inverse problems and image reconstruction where maximum penalized likelihood methods are dominant [see, e.g., Titterington, 1985; Demoment, 1989].

In a maximum penalized likelihood (MPL) framework, the density estimate is determined from the maximum of

$$Q(f) \equiv \sum_{i=1}^{N_s} \ln\left(f(x_i)\right) - \alpha I(f; f_o, \beta) \tag{1}$$

as a functional of f. The first term in (1) is the log-likelihood function and the second term is the penalty function (alternative terms are *regularization functional*, or *information divergence*). Here f_o is a default model in the absence of data, $I(f; f_o, \beta)$ is zero when $f = f_o$ and monotonically increases as f diverges from f_o, α is a global smoothing hyperparameter (or *statistical regularization parameter*), and β is a local smoothing hyperparameter.

We propose a *Maximum Quantum Entropy* (MQE) method for density estimation, which corresponds to a choice for the penalty function, I_Q. The mathematical structures we use originated in *quantum statistical mechanics*; hence, an alternative name is *Quantum Statistical Inference - QSI* [Silver, 1993]. MQE is a variation upon the maximum entropy (ME) methods [Skilling and Gull, 1989] that have been applied extensively to inverse problems and image reconstruction. The penalty functions used in ME are various modifications of the Shannon entropy of information theory, which in fact originated in the 19th century development of classical statistical physics. The penalty function for MQE is the more general concept of *relative quantum entropy*, which was developed [von Neumann, 1927] for applications to quantum statistical physics.

Both ME and MQE enforce desirable properties of density estimators such as global smoothing toward a default model, positivity, normalization, extensivity, and convex optimization. But, in addition, MQE enforces local smoothing by constraining the expectation values of differential operators. The maximum local smoothing limit of MQE is traditional penalized likelihood [Good and Gaskins, 1980] which does not enforce extensivity. The zero local smoothing limit of MQE is classic ME. MQE was applied previously to inverse problems [Silver, 1993], where it was shown to improve upon ME wherever local smoothing is important. MQE may be compared to an alternative proposal [Skilling and Gull, 1989; Robinson, 1991] to smooth ME using 'intrinsic correlation functions' and 'hidden images', which does not incorporate local smoothing in the penalty function.

The purpose of the present paper is adapt MQE to density estimation. The theory will be developed within a Bayesian framework refering to [Silver and Martz, 1993] for mathematical details. The method is illustrated using textbook data sets [Scott, 1993].

2. MQE Density Functions

In MQE, the density function, the constraints and the entropy are all expressed in terms of a new concept in statistics, the *density matrix*. $D(x,x')$ is an $\infty \times \infty$ matrix which is real symmetric and positive semidefinite. The density function f is equal to the diagonal elements of $\mathbf{D}$,

$$f(x) = D(x,x) \quad . \tag{2}$$

$\mathbf{D}$ will be determined uniquely by the combination of constraints on f and a maximum entropy principle. Without loss of generality, we assume $0 \le x \le 1$ and impose appropriate boundary conditions.

The density matrix, $\mathbf{D}$, can be diagonalized by a unitary transformation,

$$D(x,x') = \sum_{n=0}^{\infty} \psi_n(x) w_n \psi_n(x') \quad . \tag{3}$$

The ψ_n are orthonormal and complete forming a Hilbert space. The *weights* satisfy $w_n \ge 0$. Hence,

$$f(x) = \sum_{n=0}^{\infty} w_n \psi_n^2(x) \ge 0 \quad , \tag{4}$$

and

$$\sum_{n=0}^{\infty} w_n = 1 \quad . \tag{5}$$

(For practical calculations, we will show below that the ψ_n may be obtained as eigenfunctions of a linear differential operator, and the w_n are related to the eigenvalues.)

Linear Lagrange constraints on f may be written in terms of $\mathbf{D}$. Data constraints are

$$\int_0^1 U(x)f(x)dx = E(\mathbf{U}) = Tr\{\mathbf{UD}\} \quad , \tag{6}$$

where $(\mathbf{U})_{x,x'} = U(x)\delta(x - x')$. For example, if the constraints consist of a set of $E(\mathbf{O}_i) = \int_0^1 O_i(x)f(x)dx$, then $U(x) = \sum_i \lambda_i O_i(x)$ for Lagrange multipliers λ_i. The normalization constraint on f is

$$E(\mathbf{1}) = Tr\{\mathbf{D}\} = 1 \quad . \tag{7}$$

The key constraint is local smoothing, which is defined by the choice of an Hermitian differential operator $\mathbf{L}$ whose expectation value is the local smoothing constraint,

$$E(\mathbf{L}) = Tr\{\mathbf{LD}\} = \sum_{n=0}^{\infty} w_n \int_0^1 \psi_n(x)\mathbf{L}\psi_n(x)dx \quad . \tag{8}$$

We have used quadratic, $\mathbf{L}_2 \equiv -\partial^2/\partial x^2$, and quartic, $\mathbf{L}_4 \equiv \partial^4/\partial x^4$, differential operators. (We note that there are many other possible choices including x-dependent forms.) This explicit constraint applied to $\mathbf{D}$ is an implicit local smoothing constraint on f. In ME f will have the same singularity structure as U, whereas in MQE f will have smoother singularities than U depending on the choice of $\mathbf{L}$. The singularity structure of U is determined by the nature of the data analysis problem. For example, for inverse problems U consists of a sum of Lagrange multipliers times point spread functions which are most often already locally smooth. However, we shall see that for density estimation U consists of a sum of δ-functions. Then, ME produces an f with δ-function singularities, a MQE constraint on $\mathbf{L}_2$ requires f to be continuous, and a MQE constraint on $\mathbf{L}_4$ requires f to have continuous first derivatives. For a more comprehensive discussion see [Wallstrom, 1993].

These constraints are still not sufficient to uniquely specify $\mathbf{D}$, so now we invoke a maximum entropy principle. The *quantum entropy* of a density matrix is

$$S_Q \equiv -Tr\{\mathbf{D}\ln(\mathbf{D})\} = -\sum_{n=0}^{\infty} [w_n \ln(w_n)] \quad . \tag{9}$$

S_Q is invariant to unitary transformations of the Hilbert space. It is not a relative entropy, so that in the absence of constraints all eigenfunctions are equally likely. One can prove that S_Q is a concave function of $\mathbf{D}$ [Wehrl, 1978]. The maximum entropy principle is to maximize S_Q subject to the constraints of the problem. Using the method of Lagrange multipliers, maximize

$$Q(\mathbf{D}) \equiv S_Q - \beta E(\mathbf{L}) - E(\mathbf{U}) + (\mu + 1)E(\mathbf{1}) \quad , \tag{10}$$

where the Lagrange multipliers are chosen so that the constraints are satisfied. The local smoothing constraint on $E(\mathbf{L})$ has Lagrange multiplier β, the data constraint has Lagrange multiplier U, and the normalization constraint has Lagrange multiplier $\mu + 1$.

The maximum of (10) is found at

$$\mathbf{D} = \exp\left(-\mathbf{H} + \mu \mathbf{1}\right) \quad , \tag{11}$$

where

$$\mathbf{H} \equiv \beta \mathbf{L} + \mathbf{U} \quad . \tag{12}$$

This constitutes an exponential family of density matrıces parameterized by U, β, and μ. Within this family, there is a one-to-one correspondence between a choice of density function, f, and a corresponding density matrix, $\mathbf{D}$.

Diagonalizing $\mathbf{D}$, we find that the ψ_n in (3) are eigenfunctions of $\mathbf{H}$, i.e.

$$\mathbf{H}\psi_n(x) = \varepsilon_n \psi_n(x) \quad . \tag{13}$$

The weights are

$$w_n = \exp\left(-\varepsilon_n + \mu\right) \quad . \tag{14}$$

For example, for $\mathbf{L}_2$ (12) reads

$$-\beta \frac{\partial^2 \psi_n(x)}{\partial x^2} + U(x)\psi_n(x) = \varepsilon_n \psi_n(x) \quad , \tag{15}$$

which is analogous to the time-independent Schrödinger equation. Such eigenvalue equations may alternatively be derived from variational principles as developed in Sturm-Liouville theory.

The local smoothness of f is adjusted by tuning β. For reasonable choices of $\mathbf{L}$ (such as the quadratic and quartic), the ε_n increase monotonically with n and with β. The number of nodes in $\psi_n(x)$ also increase monotonically with n, so that small n corresponds to smoother $\psi_n^2(x)$. For $\beta = 0$ (ME) there is no local smoothing. As β is increased fewer eigenfunctions contribute to (4) resulting in smoother f.

The normalization of f is maintained by choosing

$$\mu = -\ln\left(\sum_{n=0}^{\infty} e^{-\varepsilon_n}\right) \quad . \tag{16}$$

We are now ready to identify the penalty function, I_Q, in (1). The penalty function is a *relative quantum entropy*,

$$I_Q = Tr\{\mathbf{D}\ln(\mathbf{D}) - \mathbf{D}\ln(\mathbf{D}_o)\} \quad , \tag{17}$$

where $\mathbf{D}_o$ is the density matrix corresponding to the default model f_o. This may be regarded as a straightforward generalization of the Kullback-Liebler entropy used in ME methods from density functions to density matrices. In the limit of no local smoothing, $\beta \to 0$, MQE reduces to ME. Alternatively, let $Q_o(\mathbf{D})$ be the entropy variational functional similar to (10) whose maximum is at $\mathbf{D}_o$. Then

$$I_Q = Q_o(\mathbf{D}_o) - Q_o(\mathbf{D}) \quad . \tag{18}$$

It follows that $I_Q \geq 0$. We summarize the mathematical properties satisfied by I_Q which are critical to its relevance to statistics.

The concavity property of S_Q means that $\mathbf{G}$ defined by

$$\delta^2 S_Q = -\frac{1}{2}\int G(x,x')\delta f(x)\delta f(x')dxdx' \tag{19}$$

is positive semidefinite (no negative eigenvalues). The consequence is that one can prove *duality* properties between I_Q and its Legendre transform,

$$C_Q(U;U_o,\beta) \equiv I_Q(f;f_o,\beta) + \int f(x)U(x)dx \quad , \tag{20}$$

which is a cumulant generating functional. First order variations may be shown to be

$$\delta C_Q = \int f(x)\delta U(x)dx \qquad \delta I_Q = \int \left[-U(x) + U_o(x)\right]\delta f(x)dx \quad . \tag{21}$$

Second order variations are

$$\delta^2 I_Q = \frac{1}{2}\int G(x,x')\delta f(x)\delta f(x')dxdx' \quad \delta^2 C_Q = -\frac{1}{2}\int G^{-1}(x,x')\delta U(x)\delta U(x')dxdx' \quad . \tag{22}$$

Notice the dual symmetry between f and U in these relations, which is analogous to the dual symmetry between observables and Lagrange multipliers in traditional ME methods.

Legendre transform dual mathematical structures in statistics of this form may be given a differential geometry interpretation [Amari, 1985]. From (17) $I_Q(f_o;f_o,\beta)=0$, and from (21) $dI_Q(f;f_o,\beta)/df = 0$ at $f=f_o$. Hence, I_Q is an *information divergence*, and $\mathbf{G}$ is a *Riemann metric* in the manifold of f.

The concavity property ensures a dual (one-to-one) relation between conjugate variables, f and U,

$$\delta f(x) = -\int G^{-1}(x,x')\delta U(x')dx' \quad . \tag{23}$$

Because of this relation, $\mathbf{G}^{-1}$ may be termed a *linear response function*. For typical choices of local smoothing operator, $\mathbf{L}$, (including the quadratic and quartic) one can demonstrate that $G^{-1}(x,x')$ peaks at $x-x'=0$ and falls off faster than a power law as $\mid x-x' \mid$ increases, a property we term *locality*. The characteristic width of $G^{-1}(x,x')$ is termed the *correlation length*, γ. For $\mathbf{L}_2$, $\gamma \propto (\beta)^{1/2}$. For $\mathbf{L}_4$, $\gamma \propto (\beta)^{1/4}$. For example, let $\mathbf{G}_o^{-1}$ be the linear response function for no data constraints and a flat default model, i.e. $U=0$. Then for $\mathbf{L}_2$, one can prove $G_o^{-1}(x,x') \propto (1-erf(\mid x-x' \mid /\gamma))/\gamma$. Figure 1 illustrates the behavior of $\mathbf{G}_o^{-1}$ for quadratic and quartic local smoothing. Note that for quadratic smoothing $\mathbf{G}^{-1}$ is strictly positive, whereas for higher order smoothing $\mathbf{G}^{-1}$ can have negative components at large $\mid x-x' \mid$. The non-linearity of MQE guarantees that $f \geq 0$ regardless of the choice of local smoothing.

Readers familiar with density estimation may be tempted to identify $\mathbf{G}^{-1}$ with the kernel in a kernel density estimation procedure. Readers familiar with ME may be tempted to identify $\mathbf{G}^{-1}$ with the *intrinsic correlation function* used in the [Skilling and Gull, 1989] proposal to correct ME for local smoothing using *hidden ME images*. However, there

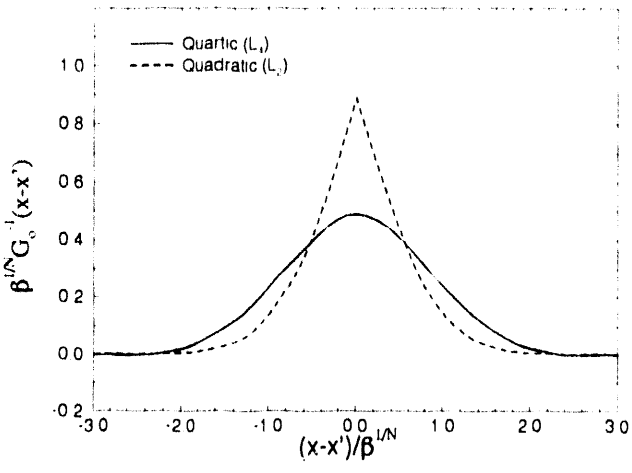

Fig. 1. Linear Response Functions - $\mathbf{G}_o^{-1}$ for local smoothing constraints of the form $\mathbf{L}_N = \partial^N/\partial x^N$, no data constraints, and a flat default model, i.e. $U = 0$. β is the Lagrange multiplier for the local smoothing constraint on the density matrix. Results are shown for quadratic (dashed) and quartic (solid) local smoothing.

are significant differences. For example, in both these methods no structure in f can be narrower than the width of the kernel or intrinsic correlation functions, whereas in MQE the non-linearity permits structure in f which is much narrower than the width of $\mathbf{G}^{-1}$.

ME ($\beta = 0$) satisfies *local extensivity*, which means that the penalty function is an additive function of the $f(x)$ at each point. However, we often have prior knowledge or evidence in the data that f is locally smooth, which violates the local extensivity property. MQE relaxes this condition to *non-local extensivity*, defined as follows. Let δI_Q^i be a change in I_Q corresponding to a change δf^i in f. Let the δf^i have compact and disjoint supports separated by much more than γ. Then non-local extensivity means $\delta I_Q \simeq \sum_i \delta I_Q^i$ for $\delta f = \sum_i \delta f^i$. This may be shown by combining the locality properties of $\mathbf{G}^{-1}$ with (22). In comparison the MPL method of [Good and Gaskins, 1980] does not obey any form of extensivity, because it is equivalent to $\gamma \rightarrow \infty$.

These *convexity* and *non-local extensivity* properties of I_Q satisfy important desiderata for both image reconstruction and density estimation. In the latter case non-local extensivity is compromised only by the added constraint on the normalization of f. Many other mathematical properties of I_Q have been established in physics contexts [for reviews, see Wehrl, 1978; Balian, 1991].

3. Application to Density Estimation

We apply these properties of I_Q to the MPL problem defined by (1). From (21), the first order variation of $Q(f)$ requires that the MPL estimate satisfies

$$\sum_{i=1}^{N_s} \frac{\delta(x - x_i)}{f(x)} + \alpha(U(x) - U_o(x)) = 0 \quad . \tag{24}$$

From (22) the second order variation (Hessian matrix) is positive semi-definite, so that solution of (24) is a problem for convex non-linear optimization methods [Skilling, 1993].

A variety of interpretations of MPL methods exist including ways to estimate hyperparameters and quantify error estimates for any choice of penalty function [Thompson, 1991]. We specialize to the Bayesian interpretation of MQE. Bayes theorem is

$$P[f \mid \{x_i\}; f_o, \alpha, \beta] \times P[\{x_i\}; f_o, \alpha, \beta] = P[\{x_i\} \mid f] \times P[f; f_o, \alpha, \beta] \quad . \tag{25}$$

The *likelihood function* is

$$P[\{x_i\} \mid f] = \prod_{i=1}^{N_s} f(x_i) \quad . \tag{26}$$

The *prior probability* for f is taken to be

$$P[f; f_o, \alpha, \beta] \propto \exp\left[-\alpha I_Q(f; f_o, \beta)\right] \quad . \tag{27}$$

Then $P[f \mid \{x_i\}; f_o, \alpha, \beta]$ is the *posterior probability* for f, and $P[\{x_i\}; f_o, \alpha, \beta]$ is the *marginal likelihood*, or *evidence*. We take the best estimate, $\hat{f}$, from the maximum of the posterior probability which is equivalent to maximizing (1). Thus, the MPL estimate is equivalent to a Maximum A Posteriori (MAP) estimate in the Bayesian interpretation.

The hyperparameters α and β are estimated from the maximum of the evidence. This method is termed *type-II maximum likelihood* (ML-II) in the statistics literature [Good, 1983; Berger, 1985], and the *evidence procedure* in the ME literature. The marginal likelihood is obtained by integrating Bayes theorem (25) over f. A metric must be used in this integration over f in order to enforce invariance to coordinate transformations. The appropriate choice is the Jeffrey's prior $\sqrt{\det(\alpha\mathbf{G})}$, which is equivalent to a *Riemann volume* factor for the f-manifold in differential geometry. We evaluate the integral in a Gaussian approximation to the expansion of $Q(f)$ in $\ln(f/\hat{f})$ about $Q(\hat{f})$. The resulting marginal likelihood is

$$P[\{x_i\}; f_o, \alpha, \beta] \propto \frac{1}{\sqrt{\det\left(\mathbf{1} + \frac{\mathbf{M}}{\alpha}\right)}} \times \exp Q(\hat{f}) \quad , \tag{28}$$

where the $N_s \times N_s$ matrix $\mathbf{M}$ is

$$M_{ij} \equiv \frac{\hat{G}^{-1}(x_i, x_j)}{\hat{f}(x_i)\hat{f}(x_j)} \quad .$$

The first term on the r.h.s. of (28) favors the simpler f of large α and β, so that it may be termed an *Ockham factor*. The second term, $\exp Q(\hat{f})$, favors the more complicated f of small α and β, and it is termed the *data factor*. The balance between the Ockham factor and data factor determines the optimal hyperparameters, $\hat{\alpha}$ and $\hat{\beta}$. We find empirically that the ML-II optimization of hyperparameters is convex for all data sets studied so far.

The covariance of the MAP estimate can be calculated using the same Gaussian approximations employed in the calculation of the marginal likelihood. The result is

$$Cov[f(x), f(x')] = \frac{\hat{G}^{-1}(x, x')}{\hat{\alpha}} - \sum_{i,j=1}^{N_s} \frac{\hat{G}^{-1}(x, x_i)}{\hat{\alpha}\hat{f}(x_i)} \left(\mathbf{1} + \frac{\mathbf{M}}{\hat{\alpha}}\right)^{-1}_{ij} \frac{\hat{G}^{-1}(x_j, x')}{\hat{f}(x_j)\hat{\alpha}} \quad . \tag{29}$$

We interpret

$$N_d \equiv \alpha \int \frac{Cov[f(x), f(x)]}{\hat{f}(x)} dx \tag{30}$$

as the *number of degrees of freedom* in $\hat{f}$. One can prove $N_d \geq 0$. In the absence of data, the prior $N_d^o = Tr\{\mathbf{G}_o^{-1}\}$ is proportional to $1/\gamma$. This provides a simple interpretation of the local smoothing hyperparameter β, because it determines the correlation length scale γ which is inversely proportional to N_d^o. ME ($\beta = 0$) corresponds to an infinite N_d, which is why ME has infinite error bars on individual points of the MAP estimate, $\hat{f}$. MQE ($\beta \neq 0$) has a finite N_d and finite error bars on individual points.

Convergence of density estimation can be monitored using N_d, because the effect of the data is to reduce it toward zero. Let f_t be the true density function. As N_s becomes large one may use the property

$$E\left(\sum_{i=1}^{N_s} O(x_i)\right) = N_s \int O(x) f_t(x) dx \quad , \tag{31}$$

to approximate the integral in (30). The result is

$$N_d \approx \frac{\alpha}{N_s} \sum_{i=1}^{N_s} \frac{\hat{f}(x_i)}{f_t(x_i)} \left(\mathbf{M}(\alpha\mathbf{1} + \mathbf{M})^{-1}\right)_{i,i} \quad . \tag{32}$$

In analogy with developments in ME [Skilling and Gull, 1989], we define

$$N_g \equiv Tr\{\mathbf{M}(\alpha\mathbf{1} + \mathbf{M})^{-1}\} \quad , \tag{33}$$

as the *number of good measurements*. Manifestly, $N_s \geq N_g \geq 0$. Then, to the extent that $\hat{f}$ has converged to f_t, (32) and (33) imply that $N_d \rightarrow \alpha N_g / N_s \leq \alpha$.

One can also derive a fundamental relation between the linear response of the MQE MAP estimate to perturbations and the covariance matrix,

$$\delta\hat{f}(x) = -\alpha \int Cov[f(x), f(x')] \delta U_p(x') dx' \quad . \tag{34}$$

Here δU_p is an infinitesimal perturbation in U which may be due to changes in the default model, changes in the data, changes in other constraints, etc. For example, an infinitesimal change in the default model corresponds to $\delta U_p(x) = -\int G_o(x, x') \delta f_o(x') dx'$. Putting (34) in words, the covariance matrix also describes the sensitivity of the MAP estimate to changes in prior knowledge or data. Large errors on the MAP estimate correspond to high sensitivity to input information, and small errors correspond to low sensitivity.

4. Examples

We apply MQE to three textbook examples of density estimation problems: the duration of eruptions of the Old Faithful Geyser; the amount of annual snowfall in Buffalo; and the Lawrence Radiation Lab (LRL) particle physics data. For each data set, we urge readers to

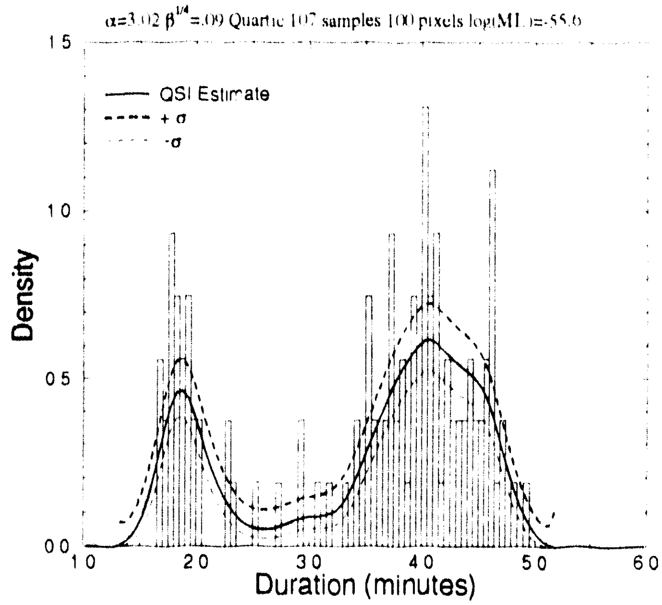

Fig. 2. Old Faithful Eruptions - 107 measurements of the duration of geyser eruptions are displayed as a histogram with 100 bins. The solid line is the optimal MQE estimate obtained with quartic local smoothing, $\mathbf{L}_4$. The dashed lines indicate ± one standard deviation errors on the MQE point estimate.

examine the corresponding sections of [Scott, 1993] to compare the performance of MQE with other approaches to density estimation.

To obtain the numerical results presented in this paper, we used Newton-Raphson for the non-linear optimization and matrix diagonalization of a discrete approximation to MQE to calculate f from knowledge of U. The number of pixels (bins) used is indicated directly on the figures for each data set. In other words, the raw data were histogrammed prior to applying MQE. We chose pixels widths which were much narrower than any structure in f, so the discretization should not significantly affect the estimate. All the MQE calculations used a flat default model, f_o, normalized to unit integral over the range of x. The values for the hyperparameters, α and β, are quoted for data scaled to the range $0 \leq x \leq 1$. The term, ***optimal estimate***, means that the hyperparameters were chosen to maximize the marginal likelihood.

Figures 2-4 show results for the duration of eruptions of the Old Faithful Geyser. The raw data from 107 eruptions are displayed as a histogram using 100 bins. Note that this histogram is not an optimal histogram estimate of f, which would use a much smaller number of bins. Rather, this histogram is simply a convenient way to display the raw data. In Fig. 2 the solid line is the optimal MQE estimate obtained for $\alpha = 3.02$ and $\beta^{1/4} = 0.09$ with quartic local smoothing. The dashed curve shows ± one standard deviation point estimates of errors on the MQE estimate, which are calculated from (29) according to $\sigma(x) = \sqrt{Cov[f(x), f(x)]}$. These provide only a partial representation of the full covariance matrix for the MQE estimate. The reader can be the judge of whether the optimal MQE estimate and errors are credible.

Figure 3 shows the effect of a different choice for the local smoothing constraint. The

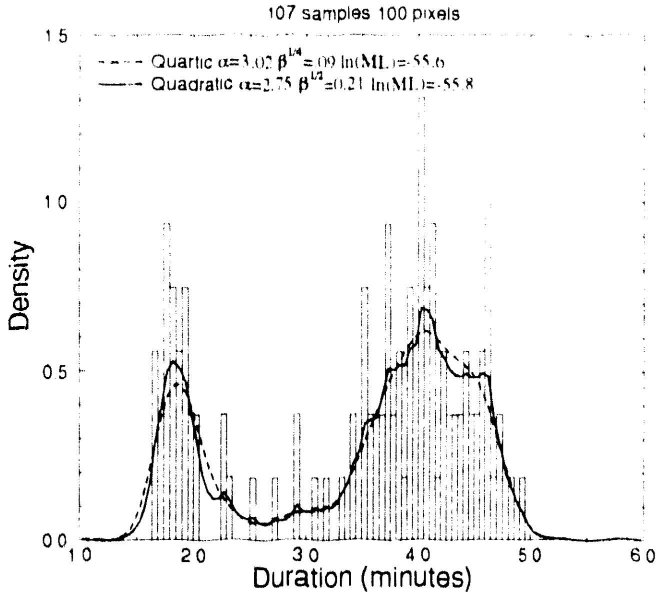

Fig. 3. Old Faithful Eruptions - Comparison of optimal MQE estimates for quadratic (solid) and quartic (dashed) local smoothing. The marginal likelihoods (ML) and correlation lengths are nearly identical. The quadratic estimate is unsatisfactory because it shows bumps at the positions of the data. The bumps are smaller than the error bars in Fig. 2 and not statistically significant. Nevertheless, the higher order smoothing of the quartic estimate is preferred.

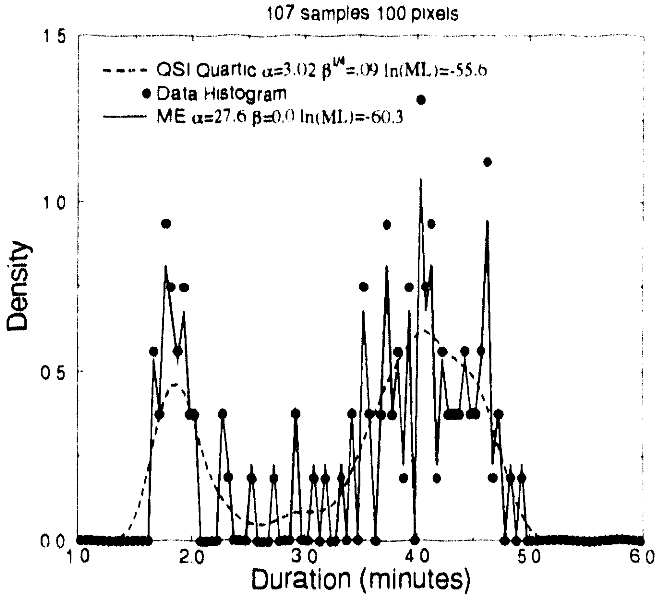

Fig. 4. Old Faithful Eruptions - Comparison of optimal MQE (dashed) with quartic smoothing and maximum entropy (solid) which has no local smoothing. Dots are the data histogram. The ratio of marginal likelihoods (ML) favoring MQE over ME is 110.

optimal MQE estimate obtained with quartic smoothing (dashed) is compared to the optimal MQE estimate obtained with quadratic smoothing (solid). The quadratic estimate appears to have bumps at the positions of the data, where the quartic estimate appears to be smooth. Therefore, the quadratic estimate is much less credible than the quartic, because the true density function should not depend on how the data were measured. However, one may argue that the apparent differences between quadratic and quartic are not significant. The MQE error bars are larger than the bumps. The derivative of the quartic estimate would also show bumps at the positions of the data. And the correlation lengths, γ, for the two estimates are nearly identical. (Let γ be defined as the half-width-half-maximum of $\mathbf{G}^{-1}$. Then from Fig. 1 and the values of β in Fig. 3 we find $\gamma_2 = .105$ and $\gamma_4 = .099$.) Indeed, there does not appear to be any Bayesian preference for the type of local smoothing, and the marginal likelihoods for the quadratic and quartic estimates are nearly identical. Nevertheless, we prefer, and we will use, quartic local smoothing for the rest of the figures in this paper. In Bayesian language, a strong *hyperprior* favors higher order smoothing.

Figure 4 compares the optimal MQE estimate (dashed) with the optimal ME estimate (solid) which has no local smoothing. The ME estimate consists of spikes at the positions of the data, and it is not credible. In this case there is a strong Bayesian preference; the marginal likelihood of the optimal MQE estimate is 110 times larger than the marginal likelihood of the ME estimate. This observation poses a question: Why does ME often work extremely well for inverse problems? As discussed earlier, the smoothness of f is determined by a combination of the smoothness of U and the local smoothing. The U's for inverse problems consist of a sum of Lagrange multipliers multiplying point spread functions (or kernels), whereas the U's for density estimation are sums of δ-functions. Typical point spread functions are already locally smooth, so that additional local smoothing is much less important. However, MQE would still be preferred over ME for most inverse problems because it provides point estimates of errors on f.

The data in Figure 5 are measurements of the annual snowfall in Buffalo over a period of 63 years. The data are displayed as a histogram with 100 bins. The optimal MQE estimate (solid) consists of a single bump. This data set has been studied using almost all available density estimation methods, and the results are displayed in [Scott, 1993]. Almost all methods, with the exception of a cross validation kernel method, produce density estimates showing three bumps. Figure 6 shows a non-optimal MQE estimate (dashed) with three bumps obtained by tuning the local smoothing hyperparameter down from large $\beta^{1/4}$ to $\beta^{1/4} = 0.1$. The parameter α is still adjusted to maximize the marginal likelihood. However, the optimal MQE estimate with one bump is 23 times more likely (judged by the ratio of marginal likelihoods) than the non-optimal MQE estimate with three bumps. And the error estimates are as large as the bumps, so they have no statistical significance.

The Buffalo snowfall is the only one of our three examples where optimal MQE agrees with the penalized likelihood method of Good and Gaskins using quartic smoothing. The equivalence means that Eq. (4) is dominated by the lowest ε_n eigenfunction. The only operative constraint is local smoothing and the quantum entropy is almost zero. This corresponds to a marginal likelihood which has a flat maximum for $0.4 \leq \beta \leq \infty$. For our other data sets, we find that this Good and Gaskins limit of MQE is not optimal and produces oversmoothed estimates. And for simulated f with a lot of sharp structure, the entropy constraint is dominant and local smoothing is unimportant.

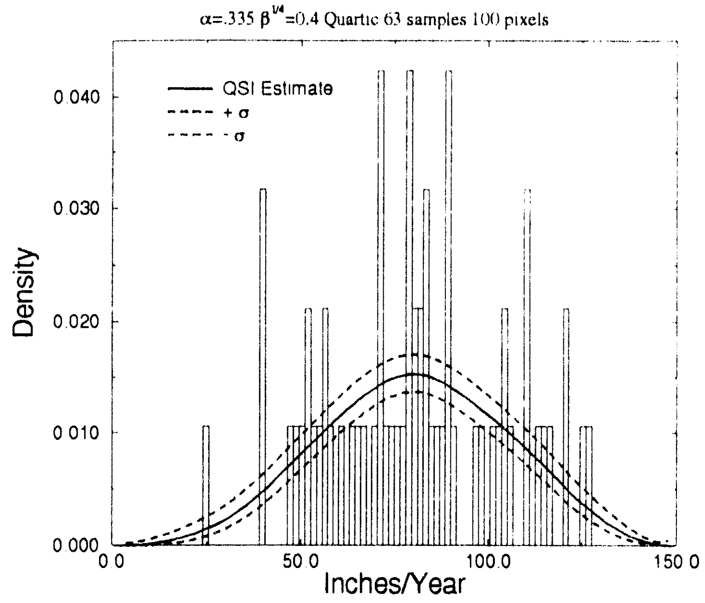

Fig. 5. Buffalo Snowfall - 63 measurements of the annual snowfall in Buffalo are displayed as a histogram with 100 bins. The solid line is the optimal MQE estimate obtained with quartic local smoothing. Dashed lines are the ± one standard deviation error bars on the MQE estimate.

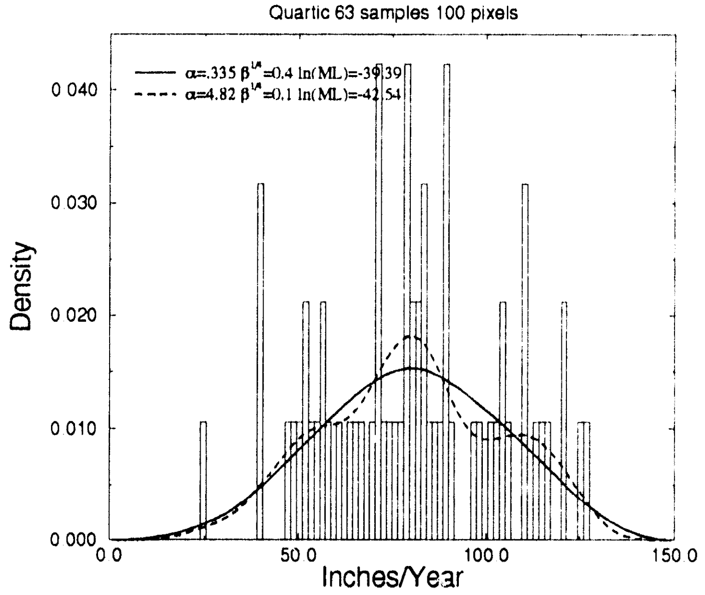

Fig. 6. Buffalo Snowfall - The solid line with a single bump is the optimal MQE estimate obtained at large β. The dashed line with three bumps is a non-optimal MQE estimate obtained by reducing $\beta^{1/4}$ to 0.1. The single bump estimate is 23 times more likely than the three bump estimate.

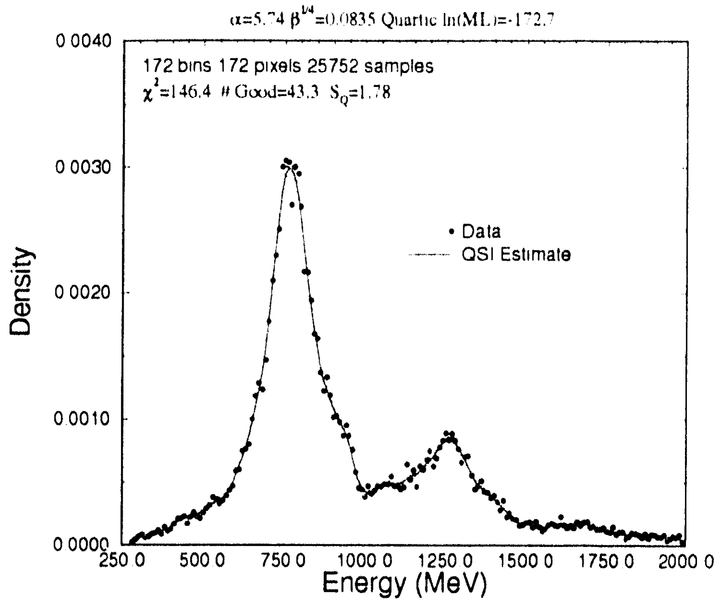

Fig. 7. LRL Particle Physics Data - Data consist of 25752 counts histogrammed into 172 10 MeV wide bins. The solid line is the optimal MQE estimate.

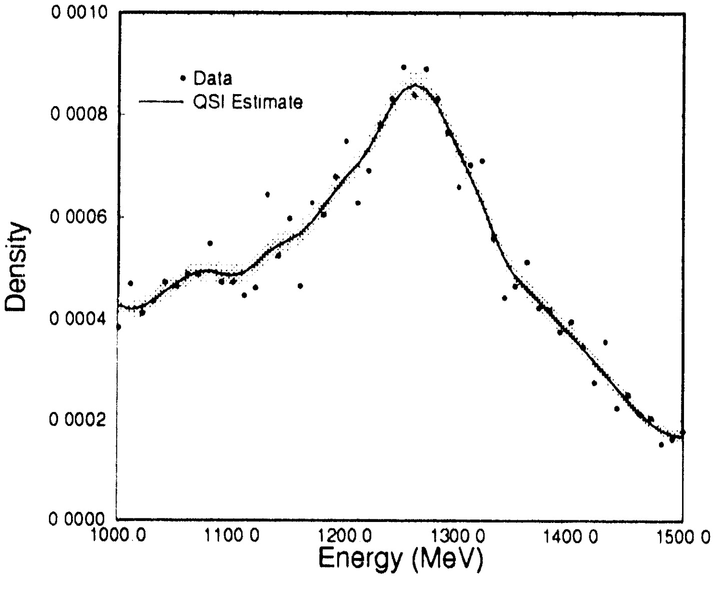

Fig. 8. LRL Particle Physics Data - Detail of Figure 7. The boundaries of the gray area are the ± one standard deviation errors on the MQE estimate.

Finally, Figs. 7 and 8 show MQE results (solid) for the LRL particle physics data. The data consist of 25752 counts histogrammed into 172 10 MeV wide bins. The gray area in Fig. 8 indicates the ± one standard deviation point errors on the MQE estimates. There are many counts in each bin, so the likelihood function can be approximately related to a

χ^2 statistic. We find for the optimal MQE estimate that $\chi^2 = 146.4$ and that $N_g = 43.3$, where N_g is the number of good measurements given by (31). This is in rough agreement with the relation, $\chi^2 + N_g \approx N_{bins}$, expected from an analysis of the ML-II procedure for inverse problems [Silver and Martz, 1993]. Note also that the quantum entropy, $S_Q = 1.78$, indicates that approximately six eigenfunctions are dominating the MQE estimate in (4).

We regard these maximum quantum entropy (or quantum statistical inference) results for density estimation as very encouraging. The introduction of quantum entropy dramatically expands the potential applications for maximum entropy methods. Considerable further testing and development will be needed to realize the full potential of quantum methods for statistics, inverse problems, and image reconstruction.

REFERENCES

Amari, S.: 1985, *Differential-geometrical Methods in Statistics*, Springer-Verlag, Berlin.

Balian, R.: 1991, *From Microphysics to Macrophysics*, Springer-Verlag, Berlin.

Berger, J. O.: 1985, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, Berlin.

Demoment, G.: 1989, 'Image Reconstruction and Restoration: Overview of Common Estimation Structures and Problems', *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**, 2024-2036.

Good, I. J.. 1985, *Good Thinking: The Foundations of Probability and Its Applications*, University of Minnesota Press, Minneapolis.

Good, I. J., and Gaskins, R. A.: 1980, 'Density Estimation and Bump Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data', *Journal of the American Statistical Association* **75**, 42-73.

Izenman, A. J.: 1991, 'Recent Developments in Nonparametric Density Estimation', *Journal of the American Statistical Association* **86**, 205-224.

Robinson, D. R. T.: 1991, 'Maximum Entropy with Poisson Statistics', *Maximum Entropy and Bayesian Methods* W. T. Grandy, L. H. Schick (eds.) Kluwer, Dordrecht, 337-341.

Scott, D. W.: 1993, *Multivariate Density Estimation*, John Wiley & Sons, Inc., New York.

Silver, R. N.: 1993, 'Quantum Statistical Inference', *Maximum Entropy and Bayesian Methods*, A. Djafari, G. Demoment, (eds), Kluwer Academic Publishers, Dordrecht, 167-182.

Silver, R. N. and Martz, H. F.: 1993, 'Quantum Statistical Inference for Inverse Problems', submitted to *Journal of the American Statistical Association*.

Silverman, B. W.: 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Skilling, J.: 1993, 'Bayesian Numerical Analysis', in *Physics & Probability*, W. T. Grandy, Jr., P. W. Milonni (eds), Cambridge University Press, Cambridge, p. 207-222.

Skilling, J. and Gull, S.: 1989, 'Classic MaxEnt', *Maximum Entropy and Bayesian Methods*, J. Skilling (ed.), Kluwer, Dordrecht, 45-71.

Titterington, D. M.: 1985, 'Common Structure of Smoothing Techniques in Statistics', *Int. Statist. Rev.* **53** , 141-170.

Thompson, A. M., Brown, J. C., Kay, J. W., Titterington, D. M.: 1991, 'A Study of Methods of Choosing the Smoothing Parameter in Image Restoration by Regularization', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 326-339.

von Neumann, J.: 1927, *Gött. Nachr.* **273**.

Wallstrom, T.: 1993, 'Generalized Quantum Statistical Inference', to be published.

Wehrl, A.: 1978, 'General Properties of Entropy', *Reviews of Modern Physics* **50**, 221-260.

DATE

FILMED

12/27/93

END