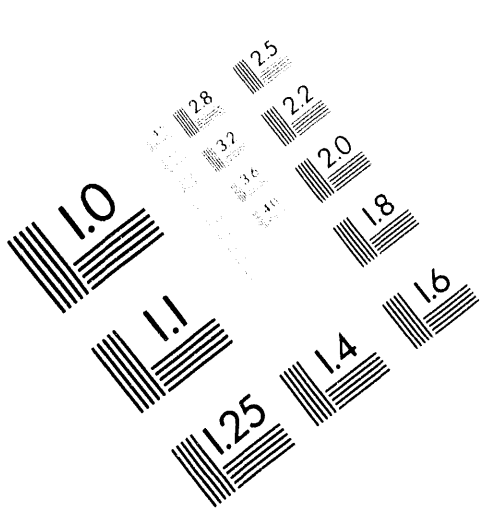
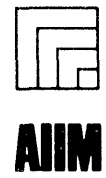
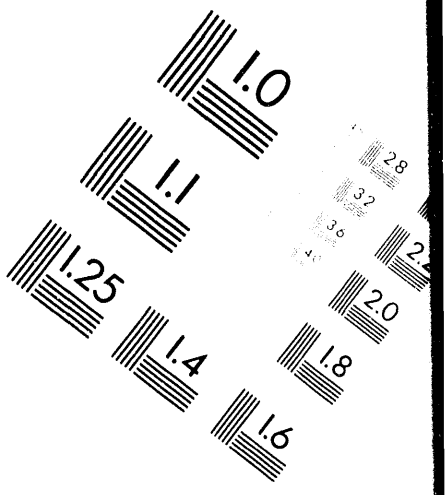


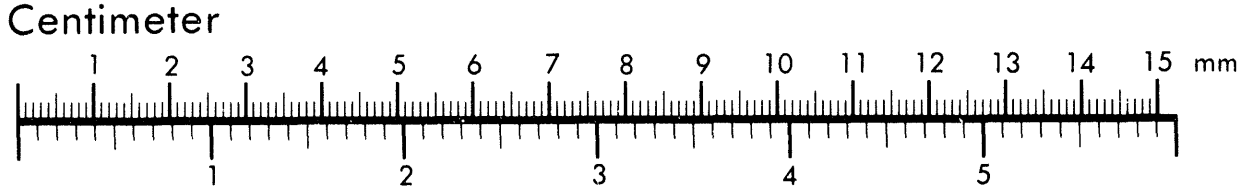


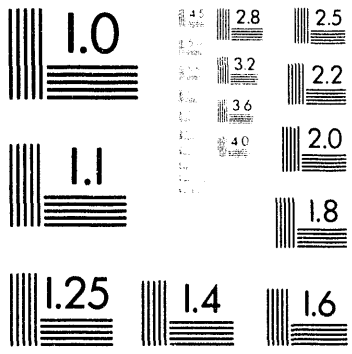
Association for Information and Image Management

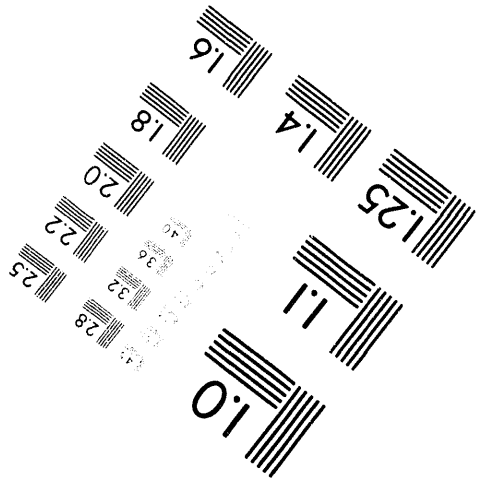
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910

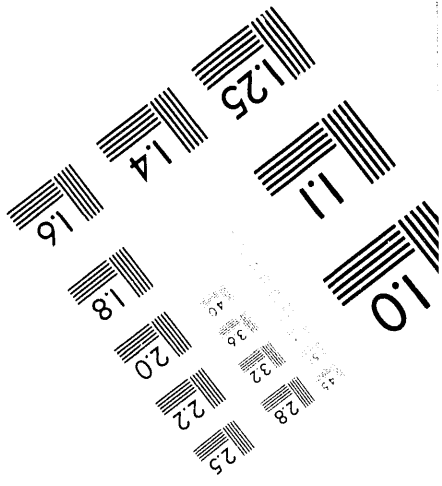
301 587 8202











MANUFACTURED TO AIIM STANDARDS
BY APPLIED IMAGE, INC.

1 of 1

LA-UR-94-2168

Conf-9408107--2

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

TITLE: APPLICATIONS OF QUANTUM ENTROPY TO STATISTICS

AUTHOR(S): R. N. Silver
H. F. Martz

SUBMITTED TO: Proceedings - invited paper at American Statistical Association, Toronto, Canada - August 12-18, 1994

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.

By acceptance of this article, the publisher recognized that the U S Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution or to allow others to do so for U S Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U S Department of Energy.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

Los Alamos

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

FORM NO. 836 R4
ST. NO. 2629 5/81

MASTER

APPLICATIONS OF QUANTUM ENTROPY TO STATISTICS

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

KEY WORDS: maximum entropy, quantum entropy, penalized likelihood, statistical regularization, Bayesian statistics, inverse problems, density estimation, image reconstruction.

Abstract

This paper develops two generalizations of the maximum entropy (ME) principle. First, Shannon classical entropy is replaced by von Neumann quantum entropy to yield a broader class of information divergences (or penalty functions) for statistics applications. Negative relative quantum entropy enforces convexity, positivity, non-local extensivity and prior correlations such as smoothness. This enables the extension of ME methods from their traditional domain of ill-posed inverse problems to new applications such as non-parametric density estimation. Second, given a choice of information divergence, a combination of ME and Bayes rule is used to assign both prior and posterior probabilities. Hyperparameters are interpreted as Lagrange multipliers enforcing constraints. Conservation principles are proposed to set statistical regularization and other hyperparameters, such as conservation of information and smoothness. ME provides an alternative to heirarchical Bayes methods.

1 Introduction

Entropy originated in statistical physics, where Boltzmann/Gibbs entropy is a fundamental measure of uncertainty and disorder in physical systems. Later entropy was adapted to information theory, where Shannon entropy is the fundamental measure of the efficiency of encoding a communication channel. The role of entropy in statistics has been comparatively limited. Cross entropy (Kullback-Liebler) is a measure of the information divergence between two positive extensive distributions. *Maximize entropy* (ME) is a principle for assigning probabilities when the information available is incomplete. Entropy is an important regularization functional for solving ill-posed linear inverse problems, and it is the foundation of an widely used Bayesian image reconstruction method. ME data analysis methods have been especially popular among physical scientists, who recognize a mathematical analogy to statistical physics.

ME is usually applied to the generic data analysis problem of inferring a positive distribution $f(x)$ defined over a compact continuous domain $x \in X$ on the basis of some data D and our prior knowledge. ME assumes that each comparable element of a positive distribution should be regarded *a priori* as independent and equally likely, and only those correlations required to satisfy the constraints of data should be imposed. However, suppose we have prior knowledge about the local smoothness of f, i.e. f has continuous derivatives with respect to x of some order. Common examples include density estimation, image reconstruction and interpolation. Continuous derivatives require prior correlations among neighboring elements of f, rather than independence. This violates the conditions for applying classical ME. A proposed modification to ME to incorporate smoothness is to equate f to the convolution of a smoothing kernel with the distribution of a 'latent' (or 'hidden') variable [Skilling, 1989]. However, such approaches abandon use of information divergences (also termed a *penalty* or *regularization* functionals) such as entropy defined on the f manifold.

In view of the many successes of ME for practical statistics applications and the significance of entropy in both physics and information theory, statisticians should consider generalizations of entropy which can incorporate prior correlations such as smoothness. Fortunately, the required mathematics has already been developed in quantum statistical physics. The generalization was invented by J. von Neumann in 1927 to be applied to the newly emerging quantum mechanics. It is known as *quantum entropy*. Although quantum entropy has been used so far only in quantum physics, it is a concave functional which can be defined on any Hilbert space. Therefore, it can be adapted to statistical inference. We term the applications of quantum entropy to statistics as *quantum statistical inference* (QSI) methods [Silver, 1993]. As information di-

vergences, both negative classical relative entropy (or cross entropy) and negative relative quantum entropy enforce desirable properties such as global smoothing toward a default model, positivity, normalization, extensivity, and convex optimization. But, in addition, quantum entropy enforces correlations, such as local smoothing, by constraining the expected values of operators. The maximum local smoothing limit of QSI is traditional penalized likelihood [Good and Gaskins, 1980] which does not enforce extensivity. The zero local smoothing limit of QSI is classical ME.

This paper also suggests a ME approach to incorporating such information divergences into statistical inference. Set prior probabilities by maximizing their entropy subject to constraints on their expected information. The statistical regularization hyperparameter is then equivalent to a Lagrange multiplier to enforce this contraint. Set posterior probabilities using Bayes rule and the likelihood principle. Set hyperparameters by demanding the conservation of information under Bayes rule. The resulting criteria for hyperparameters are similar to maximum marginal likelihood (ML-II) in empirical Bayes, but the derivation is not an approximation to an heirarchical Bayes procedure.

Many statisticians may prefer to substitute the language of *exponential families* for ME. Indeed, many physicists would do the same, using the equivalent terminology of *canonical ensembles*. The quantum information divergences, prior probabilities and posterior probabilities we propose are all members of exponential families. Many remarkable properties of such families are already well established in statistics. As far as we know, ME, exponential family and canonical ensemble descriptions are equivalent for practical applications, although there are differences in philosophy. That said, we will continue with ME.

Section II introduces quantum entropy to statisticians. Section III discusses the maximum entropy formulation of the inference process. Sec. IV presents the example of non-parametric density estimation. Sec. V briefly discusses algorithms, focusing on how to make QSI calculations efficient. And Sec VI concludes. This paper summarizes the primary QSI results critical to statistics applications, with details and proofs to be published elsewhere. The mathematical level will be heuristic rather than rigorous, as is typical of the physics literature which provided the inspiration for much of this approach. But no knowledge of physics is assumed or required.

2 Quantum Entropy

Maximum classical entropy is commonly used to infer a density function $f(x)$ defined in a continuous compact domain $x \in X$ based on linear data constraints of the form $\int_{x\in X} U(x)f(x)dx$. In this paper, we restrict discussion to normalized f required for density estimation characterized by $\int_{x\in X} f(x)dx$, although the generalization to non-normalized f required for image reconstruction is straightforward. Then, classical relative entropy (or negative cross entropy) is

$$S_c = -\int_{x\in X}\left[f(x)\ln\left(\frac{f(x)}{m(x)}\right)\right]dx \quad , \qquad (1)$$

where $m(x)$ is a default model. $I_c \equiv -S_c$ is the information divergence for maximum classical entropy methods. The maximum is

$$f_c(x) = m(x)\frac{e^{-U(x)}}{Z_c} \qquad Z_c = \int_{x\in X} m(x)e^{-U(x)}dx \quad . \qquad (2)$$

However, maximum classical entropy is not easily extended to incorporate smoothness constraints.

To generalize the maximum entropy principle, we introduce a concept that is new to statistics, but not to physics, the ***density matrix***, $\rho(x, x')$. It is defined to be a real symmetric, positive semidefinite and $\infty \times \infty$ matrix. A density function is defined to be the diagonal elements of a density matrix

$$f(x) = \rho(x, x) \quad . \qquad (3)$$

The density matrix will be determined uniquely by constraints on the density function and a maximum *quantum* entropy principle.

By definition, the density matrix can be diagonalized by an orthogonal transformation,

$$\rho(x, x') = \sum_{n=0}^{\infty} \psi_n(x) w_n \psi_n(x') \quad , \qquad (4)$$

where $\int_{x\in X} \psi_n(x)\psi_m(x)dx = \delta_{nm}$. The ψ_n are complete forming a Hilbert space. ***Positive semidefinite*** means that the ***weights***, w_n, are not negative. Therefore,

$$f(x) = \sum_{n=0}^{\infty} w_n \psi_n^2(x) \geq 0 \qquad \sum_{n=0}^{\infty} w_n = 1 \quad . \qquad (5)$$

Linear constraints on a density function may be rewritten in terms of traces of matrix operators, $Tr\{\mathrm{M}\} \equiv \int_{x\in X} M(x, x)dx$, times the density matrix. For example, data constraints are

$$\int_{x\in X} U(x)f(x)dx = Tr\{\mathrm{U}\rho\} \quad , \qquad (6)$$

where $(\mathbf{U})_{x,x'} = U(x)\delta(x-x')$. For linear inverse problems the data consist of a set of $\int_{x\in X} O_k(x)f(x)dx$; then $U(x) = \sum_k \lambda_k O_k(x)$ for Lagrange multipliers λ_i. The normalization constraint is $Tr\{\rho\} = 1$.

The new constraint we introduce to enforce local smoothing is defined in terms of an Hermitian differential operator, $\mathbf{K}$. In this paper, we specifically consider quadratic, $\mathbf{K}_2 \equiv -\partial^2/\partial x^2$, and quartic, $\mathbf{K}_4 \equiv \partial^4/\partial x^4$, forms. A constraint on

$$Tr\{\mathbf{K}\rho\} = \sum_{n=0}^{\infty} w_n \int_{x\in X} \psi_n(x)\mathbf{K}\psi_n(x)dx \quad . \quad (7)$$

is an implicit local smoothing constraint on the density function, to be discussed below.

Such constraints are still not sufficient to uniquely specify the density matrix, so now we invoke a maximum *quantum entropy*,

$$S_Q \equiv -Tr\{\rho\ln(\rho)\} = -\sum_{n=0}^{\infty}[w_n\ln(w_n)] \quad , \quad (8)$$

principle. This is invariant to orthonormal transformations of the Hilbert space, and it is a concave function of ρ [Wehrl, 1978]. Using the method of Lagrange multipliers, maximize

$$Q_1 \equiv S_Q - \beta Tr\{\mathbf{K}\rho\} - Tr\{\mathbf{U}\rho\} + (\mu+1)Tr\{\rho\} \quad . \quad (9)$$

The local smoothing constraint has Lagrange multiplier β, the data constraint has Lagrange multiplier U, and the normalization constraint has Lagrange multiplier $\mu+1$. The result is

$$\rho = \exp(-\mathbf{H} + \mu\mathbf{1}) \quad , \quad (10)$$

where

$$\mathbf{H} \equiv \beta\mathbf{K} + \mathbf{U} \quad . \quad (11)$$

This constitutes an exponential family of density matrices parameterized by Lagrange multipliers. Within this family, the concavity guarantees a one-to-one correspondence between a choice of density function and a corresponding density matrix. Therefore, we may write ρ_f as the unique density matrix corresponding to a density function f. Eq. (10) is the quantum generalization of the classical result, (2), to local smoothing constraints.

This development may be related to an eigenvalue problem,

$$\mathbf{H}\psi_n(x) = \epsilon_n\psi_n(x) \quad , \quad (12)$$

by diagonalizing the density matrix, (4), provided weights are

$$w_n = \exp(-\epsilon_n + \mu) \quad . \quad (13)$$

Normalization may maintained by choosing

$$\mu = -\ln\left(\sum_{n=0}^{\infty} e^{-\epsilon_n}\right) \quad . \quad (14)$$

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

$$-\beta\frac{\partial^2\psi_n(x)}{\partial x^2} + U(x)\psi_n(x) = \epsilon_n\psi_n(x) \quad . \quad (15)$$

Such eigenvalue equations may alternatively be derived from variational principles, along with boundary conditions, as developed in the Stürm-Liouville theory of differential equations. The mathematical properties of such differential equations are well established. The lowest ϵ_n corresponds to a nodeless $\psi_n(x)$, and the number of nodes in $\psi_n(x)$ increase monotonically with ϵ_n. In the limit of $U(x) = 0$, the ψ_n are simply sines and cosines. The weights w_n filter the contributions of ψ_n with large numbers of nodes to (5), resulting in limited spatial structure in f.

The information divergence (penalty or statistical regularization functional) for QSI may now be identified as negative *relative quantum entropy*,

$$I_Q(f;m,\beta) = Tr\{\rho_f\ln(\rho_f) - \rho_f\ln(\rho_m)\} \quad . \quad (16)$$

This is the quantum generalization of the classical expression, (1). It can be used as a penalty function (or regularization functional) in penalized likelihood methods. An $\hat{f}$ is sought which maximizes

$$Q_2 = L(f|D) - \alpha I_Q(f;m,\beta) \quad . \quad (17)$$

Here, $L(f|D)$ is the *log-Likelihood*, D is data, and α is a *statistical regularization* parameter. In a Bayesian interpretation, maximizing (17) gives the mode of a posterior probability.

To study the properties of I_Q, it is also convenient to define its Legendre transform

$$Z_Q(U;U_m,\beta) \equiv I_Q(f;m,\beta) + \int f(x)(U(x) - U_m(x))dx \quad . \quad (18)$$

Here, U_m corresponds to the default model m. Note that $I_Q(m;m,\beta) = Z_Q(U_m;U_m,\beta) = 0$, and that

$$Z_Q = -\ln\left(\frac{Tr\{e^{-\mathbf{H}}\}}{Tr\{e^{-\mathbf{H}_m}\}}\right) \quad . \quad (19)$$

First order infinitesimal variations of I_Q and Z_Q are

$$\delta Z_Q = \int f(x)\delta U(x)dx \quad ,$$

$$\delta I_Q = \int [-U(x) + U_m(x)]\,\delta f(x)dx \quad . \quad (20)$$

The concavity property of S_Q means that G defined by

$$\delta^2 S_Q = -\frac{1}{2}\int G(x,x')\delta f(x)\delta f(x')dxdx' \quad (21)$$

is also positive semidefinite. Second order infinitesimal variations are

$$\delta^2 I_Q = \frac{1}{2}\int G(x,x')\delta f(x)\delta f(x')dxdx' \ ,$$

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

I_Q is a convex function of f, and Z_Q is a concave function of U. Notice the duality between f and U in these relations, which is analogous to the duality between observables and Lagrange multipliers in traditional ME methods.

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

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

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

We are now positioned to discuss the smoothness properties of QSI density functions, which are determined by the choice of β, the degree of the differential smoothing operator and the nature of the data constraints U. The differentiability of f may be related to that of U by the linear response relation, (23). As discussed earlier, for linear inverse problems U consists of a sum of Lagrange multipliers times point spread functions (PSF), and the differentiability of U is the same as the PSF's. For density estimation U consists of a sum of Dirac δ-functions. Let a function be of class C^M if it has M continuous derivatives; for example, C^0 corresponds to the class of continuous functions, C^1 to the class of functions having continuous first derivatives, etc. A δ-function corresponds to C^{-2}. Also let D be the dimension of the space, N the degree of the local smoothing differential operator, and C^{M_U} the differentiability class of U. Then one can show that f belongs to the class $C^{M_U+N-D+1}$. For example, for one dimensional density estimation U is C^{-2}, ME f is C^{-2}, quadratic smoothed f is C^0, and quartic smoothed f is C^2. For two dimensional density estimation, higher than quadratic smoothing is required to obtain continuous QSI density estimates. See [Wallstrom, 1993] for a more comprehensive discussion.

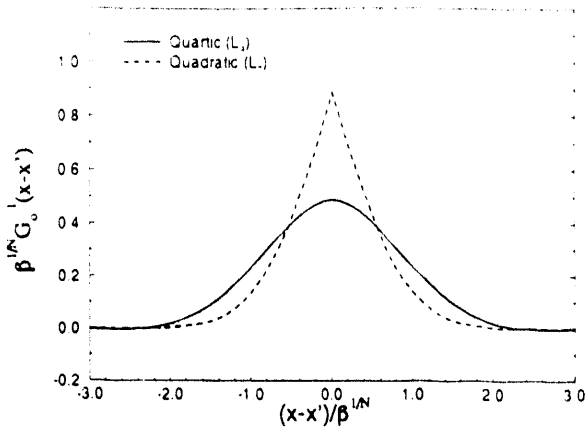


Figure 1: **Linear Response Functions** - $\mathbf{G}_o^{-1}$ for local smoothing constraints of the form $\mathbf{K}_N = \partial^N/\partial x^N$, no data constraints, and a flat default model, i.e. $U = 0$. β is the Lagrange multiplier for the local smoothing constraint on the density matrix. Results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

Readers familiar with density estimation may be tempted to identify $\mathbf{G}^{-1}$ with the kernel in a kernel density estimation procedure. However, QSI kernels are not required to be positive, and the positivity of f is enforced only by the non-linearity.

Readers familiar with ME may be tempted to identify $\mathbf{G}^{-1}$ with the *correlation function* used in the [Skilling and Gull, 1989] proposal to correct ME for local smoothing using *hidden ME images*, or the Sibisi and Skilling proposal using Derichlet priors for 'latent variables.' However, QSI is more properly termed a 'dual variable' theory because the relation between U and the observable, f, is one-to-one. In 'latent' (or 'hidden') variable theories, the relation of latent variables to observables is ill-

posed, i.e. ∞-to-one.

Classical ME ($\beta = 0$) satisfies *local extensivity*, which means that the information divergence is an additive function of the $f(x)$ at each point. QSI 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 an infinite γ.

These *convexity* and *non-local extensivity* properties of I_Q satisfy important desiderata for inverse problems, image reconstruction and density estimation. In the latter case non-local extensivity is compromised only by the added constraint on the normalization of f.

The information divergence is a concave function of β, obeying

$$\frac{\partial I_Q}{\partial \beta} = Tr\{\mathbf{K}\rho_f\} - Tr\{\mathbf{K}\rho_m\} \quad . \tag{24}$$

And $\lim_{\beta=\infty} I_Q = 0$. Although β has been introduced here as a Lagrange multiplier to constrain smoothness in a maximum entropy framework, in the next section we will interpret it as a hyperparameter in an inference procedure.

3 Entropic Inference

We propose to combine the maximum entropy principle and Bayes rule to assign probabilities and to choose optimal hyperparameters $\widehat{\alpha}$ and $\widehat{\beta}$. The development in this section applies to any information divergence, and does not require understanding quantum entropy. Our approach differs significantly from heirarchical Bayes; we do not marginalize over hyperparameters, and we do not use the full ∞-dimensional f as the fundamental variable. ([Skilling, 1994] shows that heirarchical maximum entropy methods - termed *quantified maximum entropy* - which treat f as fundamental lead to nonsense results.)

We begin with the variables and integration measure. Consider first the case of a finite number N_D of exact data; for example, $\xi_k \equiv \int_{x\in X} O_k(x) f(x) dx$. To each such observable one can associate a corresponding Lagrange multiplier λ_k, defined by $\frac{\partial I_Q}{\partial \xi_k} = -\lambda_k$. Then, (20,22) imply that information I_Q is minimized when the Lagrange multipliers for unconstrained variables remain at their default model values, i.e. $U'(x) \equiv U - \sum_{k=1}^{N_D} \lambda_k O_k(x) = U_m(x)$. Maximum entropy methods are equivalent to this minimum information condition; only measured constraints are operative in predicting unmeasured observables.

This idea must be generalized to penalized likelihood methods. The condition for maximizing (17) with respect to f is

$$\frac{\delta L}{\delta f(x)} + \alpha(U(x) - U_m(x)) = 0 \quad . \tag{25}$$

For typical data analysis problems like image reconstruction or density estimation, L is a functional of a finite number of functions of f, equal to the number of data, N_D. For example, for linear inverse problems subject to additive Gaussian noise the variables are $\xi_k \equiv \int_{x\in X} \frac{O_k(x)}{\sigma_k} f(x) dx$, where $O_k(x)$ are point spread (or resolution) functions and σ_k is the error. For density estimation, the variables are $\xi_k \equiv f(x_k)$ where the x_k are the i.i.d. samples of f. More generally, if L is concave, appropriate data space variables may be defined by principle components analysis (or singular value decomposition) of the curvature of L. The invariant integration measure may be chosen to be

$$d\Omega \equiv \sqrt{\det[\mathbf{J}]} d\xi \qquad J_{kl} \equiv -\frac{\partial^2 L}{\partial \xi_k \partial \xi_l}, \tag{26}$$

where $\mathbf{J}$ should have at most N_D nonzero eigenvalues. It is most convenient to choose the ξ_k so that $\mathbf{J}$ is diagonal. Denote its eigenvalues by J_k. Logical consistency requires independent unmeasured variables to be irrelevant, as we shall demonstrate below.

In these data space variables, I_Q is a convex function of observables ξ_k, and Z_Q is a concave function of Lagrange multipliers λ_k. These are dual variables, i.e.

$$\frac{\partial I_Q}{\partial \xi_k} = -\lambda_k \qquad \frac{\partial Z_Q}{\partial \lambda_k} = \xi_k$$
$$\frac{\partial^2 I_Q}{\partial \xi_k \partial \xi_l} = -\frac{\partial \lambda_k}{\partial \xi_l} \equiv g_{kl} \quad . \tag{27}$$

Convexity means that the $N_D \times N_D$ matrix $\mathbf{g}$ is positive semidefinite. In such data space variables, the optimization condition (25) reads

$$\frac{\partial L}{\partial \xi_k} + \alpha \lambda_k = 0 \qquad U(x) \equiv \sum_{k=1}^{N_D} \lambda_k \frac{\delta \xi_k}{\delta f(x)} + U_m(x) \quad . \tag{28}$$

All calculations are performed using data space variables. A density function is an implicit function of the λ_k. Provided L is a concave function of the ξ_k, the Hessian matrix will be negative semidefinite; there will be a unique mode (or MAP solution) which may be found by non-linear convex optimization algorithms. Denote the mode by $\widehat{f}$, the corresponding mode information as $\widehat{I}_Q$, etc. The prior mode is, of course, $\widehat{f} = m$ and $\widehat{I}_Q = 0$. The posterior mode will satisfy $\widehat{f} \neq m$ and $\widehat{I}_Q > 0$, if the data disagree with the default model predictions.

Prediction typically involves adding one or two unmeasured observables of the form $\nu = \int_{x \in X} O_\nu(x) f(x) dx$ to the problem statement. The choice of $O_\nu(x)$ depends on the question being asked about f, e.g. for a point estimate at x_o choose $O_\nu(x) = \delta(x - x_o)$. The $O_\nu(x)$ are usually not related to principle components of the likelihood, and so they are irrelevant variables for determining hyperparameters as we shall show. The equivalence of the optimization conditions (25) and (28) implies that the mode with respect to ν is the same as the mode with respect to the ξ.

A prior probability $P[\xi]$ can be assigned by maximizing its classical entropy

$$S(P[\xi]) \equiv -\int P[\xi] \ln(P[\xi]) d\Omega \quad , \tag{29}$$

subject to Lagrange contraints on the expected information,

$$E(I_Q) \equiv \int I_Q(\xi; m, \beta) P[\xi] d\Omega \quad , \tag{30}$$

and the normalization, $E(1)$. Therefore, maximize

$$Q_3 \equiv S(P[\xi]) - \alpha E(I_Q) - (F-1)E(1) \quad , \tag{31}$$

where α and $F - 1$ are Lagrange multipliers. The result is

$$P[\xi; \alpha] = \exp(F(\alpha) - \alpha I_Q)$$
$$F(\alpha) \equiv -\ln\left(\int \exp(-\alpha I_Q) d\Omega\right) \quad . \tag{32}$$

Note that

$$\frac{\partial F(\alpha)}{\partial \alpha} = E(I_Q; \alpha) \quad . \tag{33}$$

$F(\alpha)$ is a concave function of α,

$$\frac{\partial^2 F}{\partial \alpha^2} = -E(I_Q^2) + (E(I_Q))^2 \leq 0 \quad , \tag{34}$$

so the relation between $E(I_Q; \alpha)$ and α is one-to-one. They are Legendre transform *dual hypervariables*; a choice of one hypervariable corresponds to a unique choice for its dual.

Next we assign a posterior probability using Bayes rule,

$$P[D, \xi; \alpha] = P[D|\xi] \times P[\xi; \alpha]$$
$$= \exp(L + F(\alpha) - \alpha I_Q) \quad . \tag{35}$$

The posterior probability of the data is

$$P[D; \alpha] \equiv \exp(-F(D; \alpha)) = \int P[D, \xi; \alpha] d\Omega \quad . \tag{36}$$

A useful identity is

$$F(D; \alpha) = \alpha E(I_Q|D; \alpha) - E(L|D; \alpha) - S(D; \alpha) \quad . \tag{37}$$

Note that

$$\frac{\partial F(D; \alpha)}{\partial \alpha} = E(I_Q|D; \alpha) \quad , \tag{38}$$

which is the posterior expected information. $F(D; \alpha)$ is also concave, and $E(I_Q|D; \alpha)$ and α are also dual hypervariables.

We propose a *conservation principle* for choosing hyperparameters: *the prior and posterior dual hypervariables should be equal.* In the case of α, it requires

$$E(I_Q|D; \widehat{\alpha}) = E(I_Q; \widehat{\alpha}) \quad . \tag{39}$$

In simpler language, *information should be conserved under Bayes rule.* In general $E(I_Q|D; \alpha)$ and $E(I_Q; \alpha)$ are different functions of α, and information will be conserved for one or a few $\widehat{\alpha}$. Using (37) and a similar relation for $F(\alpha)$,

$$P[D; \widehat{\alpha}] = \exp\left(S(D; \widehat{\alpha}) - S(\widehat{\alpha}) + E(L|D; \widehat{\alpha})\right) \quad . \tag{40}$$

This depends only on the likelihood and the entropy difference between prior and posterior probabilities. It can be used to rank hypotheses and models used in the data analysis. If more than one $\widehat{\alpha}$ can satisfy the conservation principle, solutions should be weighted according to their $P[D; \widehat{\alpha}]$. (Note: This offers the intriguing possibility of discontinuous changes in inferences.) It is also possible that no $\widehat{\alpha}$ satisfies information conservation, in which case the inference procedure is inconsistent.

We are now positioned to discuss the requirement that independent unmeasured variables be irrelevant. Two variables ξ_1, ξ_2 are independent (extensive) if $I(\xi_1, \xi_2) = I(\xi_1) + I(\xi_2)$. A variable is unmeasured if $\frac{\partial L}{\partial \xi} = 0$. A variable is irrelevant if inference does not depend on its value. Suppose ξ_1 and ξ_2 are independent, and only ξ_1 is measured. Then

$$e^{-F(\alpha)} = \int d\xi_1 d\xi_2 e^{-\alpha I(\xi_1, \xi_2)} =$$
$$\int e^{-\alpha I(\xi_1)} d\xi_1 \int e^{-\alpha I(\xi_2)} d\xi_2 = e^{-F_1(\alpha) - F_2(\alpha)}$$

Using $\frac{\partial L}{\delta \xi_2} = 0$, one can show $F(D;\alpha) = F_1(D;\alpha) + F_2(\alpha)$. Hence, $F(D;\alpha) - F(\alpha) = F_1(D;\alpha) - F_1(\alpha)$, and ξ_2 is irrelevant. Similarly one can show that independent unmeasured variables are irrelevant to conservation of information, to the likelihood, to the entropy difference in (40), etc. Such arguments justify our initial choice of data space variables, since all other variables are irrelevant in this maximum entropy framework for inference.

A similar conservation principle applies to determining the local smoothing Lagrange multiplier in QSI,

$$E(Tr\{\mathbf{K}\rho\}|D;\widehat{\alpha},\widehat{\beta}) = E(Tr\{\mathbf{K}\rho\};\widehat{\alpha},\widehat{\beta}) \quad . \quad (41)$$

That is, smoothness should be conserved under Bayes rule.

These conservation principles yields the same hyperparameters as the maximum marginal likelihood (ML-II) (or *evidence*) procedure used in empirical Bayes [Good, 1983; Berger, 1985]. To prove this, maximize $P[D;\alpha,\beta]$ (36) with respect to α to yield (39), and maximize it with respect to β using (24) to yield (41). This equivalence is only valid for priors and information divergences derived by ME. Or, equivalently, it applies to priors and divergences which are members of exponential families in which hyperparameters may be viewed as Lagrange multipliers. The conservation principle for choosing hyperparameters is a postulate; it is not derived as an approximation to heirarchical Bayes. (In heirarchical procedures joint probabilities for data, f, and hyperparameters are calculated using Bayes rule, and then hyperparameters are marginalized using hyperpriors.)

Provided α is sufficiently large, we may use Gaussian approximations to estimate hyperparameters. Thus, for posterior expectation values use

$$Q_2 \approx \widehat{Q}_2 - \frac{1}{2}(\sqrt{J}\xi)^\dagger \mathbf{C}^{-1}(\xi\sqrt{J}) \quad . \quad (42)$$

And use $I_Q \approx \widehat{I_Q} + \frac{1}{2}\xi^\dagger \mathbf{g}\xi$. The inverse covariance matrix

$$\mathbf{C}_\xi^{-1} = 1 + \alpha\mathbf{M}^{-1} \qquad \mathbf{M}_{kl}^{-1} = \frac{g_{kl}}{\sqrt{J_k J_l}} \quad . \quad (43)$$

The matrix M is $N_D \times N_D$ and positive semidefinite. The *number of good measurements* may be defined by

$$N_g \equiv Tr\{\mathbf{C}_\xi\} = Tr\{\mathbf{M}(\alpha + \mathbf{M})^{-1}\} \quad . \quad (44)$$

Expectation values in this Gaussian approximation are

$$E(I_Q;\alpha) \approx \frac{N_D}{2\alpha}$$

$$E(I_Q|D;\alpha) \approx \widehat{I}_Q + \frac{N_D - N_g}{2\alpha}$$

$$E(L|D;\alpha) \approx \widehat{L} - \frac{N_g}{2} \quad . \quad (45)$$

The conservation principle requires $\widehat{\alpha}$ to be chosen to satisfy $N_g \approx 2\alpha\widehat{I}_Q$, a condition first derived by [Skilling, 1989] using the evidence procedure. This has a simple interpretation: $\widehat{I}_Q$ has increased from zero by reducing the number of degrees of freedom in the data to $N_D - N_g$.

A Gaussian approximation to the conservation principle for smoothing is not as easy to calculate, but it can be found by maximizing

$$P[D;\alpha,\beta] \approx \frac{1}{\sqrt{\det[1 + \frac{\mathbf{M}}{\alpha}]}} \exp(\widehat{Q}_2) \quad . \quad (46)$$

The relation $\widehat{Q}_2 \approx E(L - \alpha I_Q|D;\alpha) + E(\alpha I_Q;\alpha)$, which follows from (45), has been invoked. The fraction in (46) is often termed an *Occam factor* because it favors the simpler models of large α and β, but it may also be viewed as a Gaussian approximation to $\exp(S(D;\alpha) - S(\alpha))$ where S is entropy (29). The $\exp(\widehat{Q}_2)$ is the usual *data factor* which favors the more complex models of small α and β. The balance between these two terms determines the optimal hyperparameters.

To make predictions about an unmeasured observable, ν, the basic assumption is that the joint probability is

$$P[\xi,\nu,D] \propto \exp(L - \widehat{\alpha}I_Q) \quad . \quad (47)$$

For point estimates the covariance is

$$\mathbf{C}_f^{-1}(x,x') = \widehat{\alpha}G(x,x') + \sum_{k=1}^{N_D} J_k \frac{\delta\xi_k}{\delta f(x)}\frac{\delta\xi_k}{\delta f(x')} \quad (48)$$

Define

$$\Gamma_k(x) \equiv -\frac{\partial f(x)}{\partial\lambda_k} = \int_{x\in X} G^{-1}(x,x')\sqrt{J_k}\frac{\delta\xi_k}{\delta f(x')}dx' \quad . \quad (49)$$

Then, the covariance $C_f(x,x')$ is

$$\frac{G^{-1}(x,x')}{\widehat{\alpha}} - \frac{\Gamma^\dagger(x)}{\widehat{\alpha}}\left(1 + \frac{\mathbf{M}}{\widehat{\alpha}}\right)^{-1}\frac{\Gamma(x')}{\widehat{\alpha}} \quad . \quad (50)$$

$\mathbf{C}_f$ is a positive definite matrix. The variance on point estimates is given by $C_f(x,x)$. The second term in (50) gives the reduction in variance due to the data.

There is a fundamental relation between the linear response of the mode to perturbations and the covariance matrix,

$$\delta\hat{f}(x) = -\alpha\int C_f(x,x')\delta U_p(x')dx' \quad . \quad (51)$$

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, from (23) an infinitesimal change in the default model would correspond to $\delta U_p(x) = -\int G_o(x,x')\delta m(x')dx'$. Putting (51) in words, the covariance matrix describes the sensitivity of the mode to changes in prior knowledge or data. Large errors correspond to high sensitivity to input information, and small errors correspond to low sensitivity.

We interpret

$$N_f \equiv \alpha \int \frac{\widehat{C}_f(x,x)}{\hat{f}(x)} dx \qquad (52)$$

as the *number of degrees of freedom* in $\hat{f}$. One can prove $N_f \geq 0$. In the absence of data, the prior $N_f^o = Tr\{\mathbf{G}_o^{-1}\}$ 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_f. Classical ME ($\beta = 0$) corresponds to an infinite N_f, which is why ME has infinite error bars on individual points of the MAP estimate. QSI ($\beta \neq 0$) has a finite N_f and finite error bars on individual points. The effect of the data is to reduce N_f.

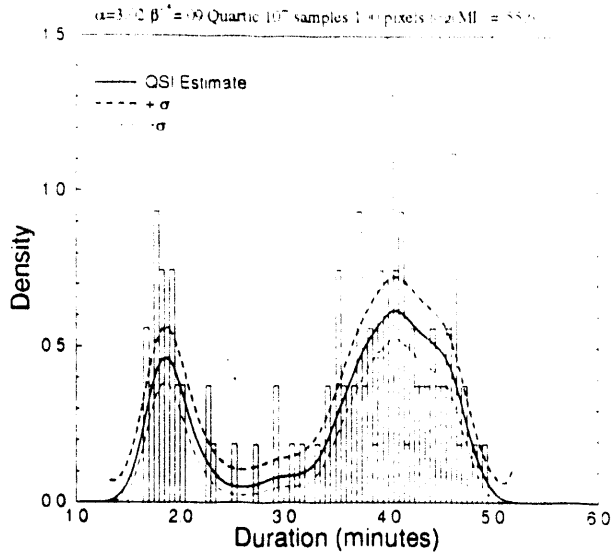


Figure 2: **Old Faithful Eruptions** - 107 measurements of the duration of geyser eruptions are displayed as a histogram with 100 bins. The solid line is the optimal QSI estimate obtained with quartic local smoothing, $\mathbf{L}_4$. The dashed lines indicate $\pm$ one standard deviation errors on the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

4 Application to Density Estimation

Non-parametric density estimation has been studied extensively by statisticians [Silverman, 1986; Izenman, 1991; Scott, 1993]. If a set of N_D observations, $\{x_k\}$, is identically and independently drawn from a probability density function $f(x)$, the problem is to estimate f when no parametric form is known. The log-likelihood function for density estimation is

$$L = \sum_{k=1}^{N_D} \ln(f(x_k)) \quad . \qquad (53)$$

We illustrate the comparative performance of maxent and QSI using the textbook example of the eruptions of the Old Faithful Geyser.

Figures 2 and 3 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. Our QSI calculations used Newton Raphson non-linear optimization and Eispack matrix diagonalization to calculate QSI images. The density estimates were also discretized into 100 bins. In Fig. 2 the solid line is the optimal QSI estimate obtained for $\alpha = 3.02$ and $\beta^{1/4} = 0.09$ with quartic local smoothing. The dashed curve shows $\pm$ one standard deviation point estimates of errors on the QSI estimate, which are calculated from (50) according to $\sigma(x) = \sqrt{C_f(x,x)}$. These provide only a partial representation of the full covariance matrix for the QSI estimate. The reader can be the judge of whether the optimal QSI estimate and errors are credible.

Figure 3 compares the optimal QSI 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. The marginal likelihood of the optimal QSI 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 choice of local smoothing differential operator. The U's for inverse problems consist of a sum of Lagrange multipliers multiplying point spread functions, whereas the U's for density estimation are

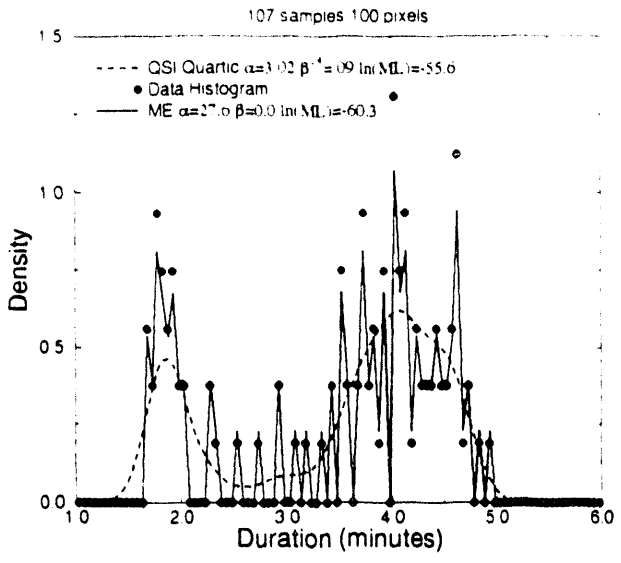


Figure 3: **Old Faithful Eruptions** - Comparison of optimal QSI (dashed) with quartic smoothing and maximum entropy (solid) which has no local smoothing. Dots are the data histogram. The ratio of marginal likelihoods (ML) favoring QSI over ME is 110.

sums of δ-functions. Typical point spread functions are already locally smooth, so that additional local smoothing is much less important. However, QSI would still be preferred over ME for most inverse problems according to the marginal likelihood, and for the practical reason that it provides point estimates of errors on f.

5 Algorithms

The duality relations and optimization conditions for Lagrange multipliers are identical in QSI and ME. Therefore, finding the posterior mode in QSI is a non-linear convex optimization problem which may be solved using the same methods which have proven successful for classical ME [Skilling, 1993]. We find that it is most efficient to work with the dual optimization problem, and to use Newton-Raphson in conjuction with conjugate gradient inversion of the Hessian matrix. Evaluation of marginal likelihoods (36) going beyond Gaussian approximations can also use standard methods such as the Gibbs sampler.

The only new computational feature of QSI is a more complex relation than in classical ME (2) between the Lagrange multipliers and the density function. Let the QSI equations be discretized into N_p pixels using finite difference representations of differential operators. Then, naive calculation of QSI images by Eispack diagonalization of $\mathbf{H}$ (10) requires cpu time scaling as $O(N_p^3)$ and memory scaling as $O(N_p^2)$, which would be prohibitively expensive for large N_p. Fortunately, quantum physicists have developed several more efficient methods for the direct calculation of density matrices which do not require matrix diagonalization. One of the most popular is Feynman path integrals, which requires cpu time scaling and memory scaling as $O(N_p^2)$. We have recently developed a polynomial moment method which scales linearly in N_p [Silver, 1994].

Choose a and b in $\mathbf{H} = a\mathbf{X} + b$ so that all the eigenvalues of the $N_p \times N_p$ matrix $\mathbf{X}$ satisfy $-1 \leq x_n \leq +1$. Then define a new variable θ by $x = \cos(\theta)$. To calculate a density matrix such as (10), use the operator identity

$$\rho = e^{-\mathbf{H}} = e^{-b}\left[I_o(a) + 2\sum_{m=1}^{\infty}(-1)^m I_m(a) T_m(\mathbf{X})\right] \quad (54)$$

Here, $I_m(a)$ are modified Bessel functions, and $T_m(x) = \cos(m\theta)$ are chebyshev polynomials of the first kind. This is a rapidly converging expansion which may be truncated without significant error at a finite number M of terms depending on the value of a. Calculation of the $N_p \times N_p$ matrix $T_m(\mathbf{X})$ uses the polynomial recurrence relation

$$T_{m+1}(\mathbf{X}) = 2\mathbf{X}T_m(\mathbf{X}) - T_{m-1}(\mathbf{X}) \quad . \quad (55)$$

Use of the polynomial recurrence means that only two previous T_m need be stored. In a finite difference approximation to the differential smoothing operator, $\mathbf{X}$ is tridiagonal for quadratic smoothing and pentadiagonal for quartic smoothing. So the cpu time for matrix multiplications indicated in (55) scale as $O(N_p \times M)$. Because of the finite correlation length γ, the density matrix ρ is essentially band diagonal with a width proportional to $N_p \times \gamma$. Only band diagonal components of $T_m(\mathbf{X})$ of the same width need be kept in calculating the recurrence relation (55). The fact that QSI functions are smoothed over a width γ means that one can choose $N_p \propto 1/\gamma$. Hence the required width of the band is independent of γ, and the memory requirements are also linear in N_p,

This algorithm provides a controlled procedure for calculating observables from Lagrange multipliers. We find that practical QSI calculations take about a factor of 3 more cpu and memory than comparable ME calculations.

6 Discussion

QSI is a new statistical method whose applications may include ill-posed inverse problems, image reconstruction, density estimation, spectrum estimation, density function interpolation, etc. Apart from the algorithmic issues discussed in Sec. V, QSI is no more difficult to apply than other ME methods, and its domain of applicability is far greater. Quantum entropy provides a systematic way to build prior correlations into a manifold of density functions, and it may be extended to many other kinds of prior correlations. Entropic inference adds basic principles to assign prior probabilities and impose conservation in Bayes rule calculations, with hyperparameters reinterpreted as Lagrange multipliers.

We tried to introduce QSI using only statistical terms and language. This required hiding the physics intuitions which, in fact, helped to motivate our approach and provided most of the mathematics. Statisticians should be aware that strong analogies exist between QSI and the physical theory of *quantum statistical mechanics*. And they should know that the mathematics used in QSI has been empirically validated to extraordinary accuracy for numerous diverse physical systems. In Sec. II, $\mathbf{K}$ is analogous to a 'kinetic energy' operator, U to 'potential energy', β 'inverse temperature', $\mathbf{H}$ a 'Hamiltonian', exponential families are 'canonical ensembles', (15) is the time-independent Schrödinger equation, etc. In Sec. III, the ξ_k may be thought of as 'extensive variables', λ_k 'intensive variables', F 'free energies', ML-II the 'principle of minimum free energy', (51) the 'fluctuation-dissipation theorem', the conservation of information is analogous to 'conservation of energy', discontinuous changes in inference are analogous to 'phase transitions', etc. Statisticians should also be aware that the analogies with physics are incomplete. There is no mention of Bayes rule, information divergences, covariance, default models, prior and posterior probabilities, etc. in mainstream statistical mechanics textbooks. And for all its intuitive appeal, the maximum entropy principle has produced no results in statistical mechanics that were not previously obtained using other starting postulates, such as 'maximize phase space volume'.

Nevertheless, QSI demonstrates how a significant fraction of the mathematics of statistical physics may be adapted to statistical inference. Indeed, we expect QSI will become a practical tool in statistics. However, it is much more difficult to anticipate the implications of this demonstration for the philosophical foundations of both statistical inference and statistical physics, or for the additional cross-fertilization between statistics and physics which should ensue.

Acknowledgements

Research supported by the U. S. Dept. of Energy.

References

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (Second Edition), Springer-Verlag, Berlin, see p. 99.

Good, I. J. (1983), *Good Thinking: The Foundations of Probability and Its Applicatons*, 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.

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*, eds. A. Djafari, G. Demoment, 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.

Silver, R. N. and Röder, H. (1994), "Densities of States of Megadimensional Hamiltonian Matrices", to be published in Int. J. of Mod. Phys. C.

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*, ed. J. Skilling, Kluwer, Dordrecht, 45-71.

Skilling, J. and Sibisi, S., these proceedings.

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
8/1/94
END