Quantum Statistical Inference

R. N. Silver
MS B262 Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545
e-mail: rns@loke.lanl.gov

Abstract

Can the mathematical structure of quantum statistical mechanics be adapted to statistical inference - the human problem of reasoning from incomplete and uncertain data? A unified theory of quantum statistical mechanics and Bayesian statistical inference is proposed to address ill-posed inverse problems, in which the goal is inference of a continuous density function. The relative information measure is a combination of Von Neumann entropy and Fisher information, with the balance set by a local smoothing hyperparameter. The classical limit is Kullback-Liebler cross entropy. There is no global smoothing parameter. An optimal image is estimated along with statistical and incompleteness errors. QSI yields significantly better images than the maximum entropy method, because it accounts explicitly for image continuity.

1 Introduction

The hypothesis that principles of reasoning may be related to principles of physics is a possibility so important and tantalizing that it should be thoroughly explored. The Physics of Computation Workshop examined the intersect between physics, information theory and statistics concepts. Such cross-fertilization could be very stimulating for all three disciplines. Many relations are already known, such as the fundamental role of entropy in both information theory and statistical mechanics. A more recently proposed relation [1, 2] is between the Schrödinger equation of quantum mechanics and a variational principle [3] on Fisher information. This functional was originally introduced [4] as a measure of inverse uncertainty of a position parameter determined by maximum likelihood statistical estimation. It is remarkable that Fisher information is related to entropy in the same manner as surfaces to volumes of sets [5]. This paper adapts such relations to provide a novel technique for statistical inference [6].

The inference of a density function from incomplete and uncertain data is a common statistics problem. Examples include image reconstruction, density function estimation (sampling), interpolation, etc. These are ill-posed inverse problems; they have no unique solution. Of many approaches in the statistics and applications literature [7, 8, 9], the most successful have used Bayesian statistical inference (see, e.g., [10]). Data are systematically combined with available prior knowledge to restrict the hypothesis space. Outputs of Bayesian calculations are probabilities, in the sense of degree-of-belief, assigned to possible density functions which are consistent with data. Probabilities are conditional on the prior knowledge assumed. The most probable density function is usually presented as the *image*. Both statistical and incompleteness errors on the image may also be estimated. A subset of non-parametric Bayesian inference is statistical regularization (or penalized likelihood). Prior knowledge is incorporated in a regularization functional which is maximized, subject to constraints of data.

The non-parametric approach closest to information theory is the maximum entropy method (ME) [11]. The regularization functional is the (Kullback-Liebler or Shannon-Jaynes) cross entropy of the density function defined relative to a default model. ME applies only a global smoothing constraint [7], which maximizes cross entropy. It may be justified by a requirement (see, e.g., [12]) of no prior covariance between points in an image. It often yields excellent solutions to diverse inverse problems. However, in many cases ME images exhibit spurious artifacts and overly sharp structure [13]; nor does ME work for problems in which local smoothing is critical such as interpolation and density function estimation. To correct such deficiencies, developers [13] of the leading ME code have incorporated local smoothing in the form of a user-chosen intrinsic correlation function blurring of a

0-8186-3420-0/93 $03.00

hidden ME image. This approach has been a successful way to improve image quality in many cases. But their approach is open to several criticisms: choice of the intrinsic correlation function is *ad hoc*; negative cross entropy as an information measure is insensitive to local smoothness; there is no metric in the space of observations; there is an *ad hoc* global smoothing parameter, etc.

A far more intriguing approach to local smoothing is suggested by the close mathematical analogy between ME and classical statistical mechanics. In turn, classical statistical mechanics is a well-defined limit of the more general theory of quantum statistical mechanics (see, e.g., [14]). It has been developed to provide a probabilistic description of physical phenomena on the microscopic scale of atoms and nuclei. Quantum statistical mechanics can be derived from a maximum entropy principle [11, 14], provided the concept of a density function is generalized to a *density matrix* and Shannon entropy is generalized to *Von Neumann entropy*. Diagonal elements of the density matrix correspond to a density function, which are smoother than classical density functions for the same constraints. Off-diagonal elements correspond to uncertainty in position. Quantum statistical mechanics may be Nature's choice for a probability theory, because it has been empirically validated to many orders of magnitude accuracy in diverse physical systems.

Neither quantum mechanics nor statistical mechanics have been successfully formulated as abstract mathematics, despite many notable efforts. Instead, they are usually regarded as a set of empirically derived 'cookbook rules' that, nevertheless, always work. For physical problems there is no uncertainty about principles for making practical predictions, and accuracy is limited only by calculational ability. Continuing controversy about the interpretation of quantum theory is not due to any empirical failure; it is due to the counterintuitive nature of quantum concepts and the paucity of mathematical foundations.

Is there something to learn about probability theory from quantum mechanics? Is the empirical success of quantum theory a big clue to more efficient principles of statistical inference? With these hopes, this paper adapts the mathematical structure of quantum statistical mechanics, such as it is, to statistics. The mathematical constructs are motivated by analogy to an extremely successful physical probability theory. The result, *Quantum Statistical Inference*(QSI), should be a useful addition to statistical methodology. In addition, this effort may encourage a fruitful exchange of ideas between statistics, physics and information theory.

This paper presents the theory of QSI in statistics language. Sec. 2 describes the quantum representation of a density function. Sec. 3 presents the method for solving ill-posed inverse problems in a Bayesian framework. Sec. 4 presents applications to interpolation and deconvolution. Physics language is avoided until Sec. 5, where the physical origins of QSI are identified. Sec. 6 concludes.

2 Quantum Density Functions

Consider a density function $f(x)$ defined in the domain $0 \leq x \leq 1$ which satisfies $f(x) \geq 0$ and $\int_0^1 dx f(x) = 1$. It is to be inferred from a calculus of variations problem: minimize Fisher information subject to data constraints. Such a penalized likelihood method for local smoothing has been previously explored in the statistics literature (e.g., [7], p. 150; [9], p. 110), but the implementation in QSI will be quite different.

The variational problem is to minimize

$$Q_1 \equiv \frac{\gamma^2}{4}\int_0^1 dx \frac{1}{f(x)}\left(\frac{\partial f(x)}{\partial x}\right)^2 + \int_0^1 dx U(x) f(x) - \varepsilon \int_0^1 dx f(x) \quad . \tag{1}$$

The first term is Fisher information with Lagrange multiplier $\gamma^2/4$. Use of a differential operator ensures the density function is locally smooth. The second term represents linear Lagrange constraints $U(x)$ due to data. For example, if constraints are a set of $\mathcal{E}(R_k) \equiv \int_0^1 dx R_k(x) f(x)$, then choose $U(x) = \sum_k \lambda_k R_k(x)$ where λ_k are Lagrange multipliers. The third term is a normalization constraint with Lagrange multiplier ε.

Define a 'wave function' by $\psi(x) = \pm\sqrt{f(x)}$. The Euler-Lagrange equation for $\psi(x)$ is

$$\frac{\partial Q_1}{\partial \psi(x)} - \frac{d}{dx}\left(\frac{\partial Q_1}{\partial(d\psi(x)/dx)}\right) = 0 \quad . \tag{2}$$

The result is a wave equation,

$$-\gamma^2 \frac{d^2\psi(x)}{dx^2} + U(x)\psi(x) = \varepsilon\psi(x) \quad . \tag{3}$$

The differential operator on the l.h.s. of Eq. (3) is

$$\hat{\mathbf{H}} = \gamma^2 \hat{\mathbf{I}}_F + \hat{\mathbf{U}} \equiv -\gamma^2 \frac{\partial^2}{\partial x^2} + U(x) \quad . \tag{4}$$

Eq. (3) is a special case of a Sturm-Liouville equation. It is an eigenvalue problem for $\hat{\mathbf{H}}$ with an infinite number of eigenfunctions, $\psi_n(x)$, and eigenvalues ε_n. Appropriate boundary conditions are defined by $df(x)/dx = 0$ at $x = 0, 1$. These conditions are imposed by expanding the domain to $0 \leq x \leq 2$ with periodic boundary conditions. Set $U(2.0 - x) = U(x)$ and $f(2.0 - x) = f(x)$. Then $f_n(x) = \psi_n^2(x)$. The $\psi_n(x)$ are either symmetric or antisymmetric under $x \leftrightarrow 2.0 - x$. Calculus of variations arguments (e.g. [3], Sec. 12-3) show the lowest eigenfunction, $\psi_o(x)$, to be an absolute minimum of Fisher information. All other $\psi_n(x)$ are minima of Fisher information with respect to variations orthogonal to other $\psi_n(x)$ with lower ε_n. $\psi_o(x)$ is nodeless. The number of nodes in $\psi_n(x)$ increases with ε_n. The $\psi_n(x)$ satisfy orthogonality relations with inner product,

$$\int_0^2 dx \psi_n(x)\psi_m(x) = \delta_{n,m} \quad . \tag{5}$$

They also satisfy completeness relations,

$$\sum_{n=0}^{\infty} \psi_n(x)\psi_n(x') = \delta(x - x') \quad . \tag{6}$$

For example, if $U(x) = 0$, symmetric eigenfunctions $\phi_k(x)$ are $\cos(\pi m x)$ and antisymmetric ones are $\sin(\pi m x)$ with m's integer. The lowest eigenfunction is $\phi_o(x) = 1/\sqrt{2}$. Eigenvalues are $(\gamma\pi m)^2$.

Orthogonality and completeness together imply the eigenfunctions form a Hilbert space. Any eigenfunction, $\psi_n(x)$, of Eq. (3) for $U(x) \neq 0$ can be expanded as a linear combination of $U(x) = 0$ solutions which defines x-basis functions,

$$\psi_n(x) = \sum_{k=0}^{\infty} u_{n,k}\phi_k(x) \quad . \tag{7}$$

The matrix $u_{n,k}$ is the unitary transformation which diagonalizes $\hat{H}$. A density function is a linear combination of eigenfunctions,

$$f(x) = 2\sum_n w_n \psi_n^2(x) = 2\sum_{k,l} \rho_{k,l}\phi_k(x)\phi_l(x) \quad . \tag{8}$$

Normalization is $\sum_n w_n = 1$. The factor 2 ensures $\int_0^1 dx f(x) = 1$. The real symmetric matrix $(\hat{\rho})_{k,l} = \rho_{k,l}$ is termed the 'density matrix'. In the x-basis (image space)

$$\hat{\rho}(x, x') = \sum_{k,l} \rho_{k,l}\phi_k(x)\phi_l(x') \quad . \tag{9}$$

In contrast, previous penalized likelihood methods for local smoothing in the statistics literature [7] consider only lowest, $\psi_o(x)$, solutions of the Fisher information variational problem.

Differential operators, $\hat{\mathbf{O}}$, may be represented by matrices in the x-basis,

$$\hat{\mathbf{O}}(x, x') = \sum_{k,l} \phi_k(x) \left[\int_0^2 dx'' \phi_k(x'')\hat{\mathbf{O}}\phi_l(x'')\right] \phi_l(x') \quad . \tag{10}$$

For example, the matrix form of $\hat{\mathbf{U}}(x, x') = U(x)\delta(x - x')$, and the matrix form for $\hat{\mathbf{I}}_F(x, x') = \sum_m (\pi m)^2 \cos(\pi m(x - x'))$ (although in practice a finite difference approximation is preferred). Matrix representations of operators in other bases may be obtained by unitary transformations. For example, $\hat{\mathbf{I}}_F$ is diagonal in the Fourier transform of the x-basis, and the $\phi_k(x)$ are its eigenfunctions. Expectation value of an operator $\hat{\mathbf{O}}$ is obtained from a trace,

$$\mathcal{E}(\hat{\mathbf{O}}) \equiv Tr\{\hat{\mathbf{O}}\hat{\rho}\} = \int_0^2 dx \int_0^2 dx' \hat{\mathbf{O}}(x, x')\hat{\rho}(x', x) \quad . \tag{11}$$

For example, $\mathcal{E}(\hat{\mathbf{U}}) = \int_0^1 dx U(x) f(x)$.

How does QSI achieve local smoothing? Suppose one took the first $2N$ $\phi_k(x)$ as basis functions. The highest eigenfunction would be $\phi_{2N-1}(x) = \sin(\pi N x)/\sqrt{2}$. Then one could form $2N$ pixel basis functions to span the domain $0 \leq x \leq 2$ by a unitary transformation. Let pixels be centered at $x_j = (j - 0.5)/N$. Pixel basis functions are

$$\tilde{\phi}_j(x) = \frac{1}{\sqrt{N}} \sum_{k=0}^{2N-1} \phi_k(x)\phi_k(x_j) \quad . \tag{12}$$

These functions are orthonormal, peak at x_j, and are zero at $x = x_{j'}$ for any $j' \neq j$. If the sum in the completeness relation, Eq. (6), is truncated at $2N$, the Dirac delta function is replaced with a peaked function whose width is proportional to $1/N$. Density functions formed out of a finite number of pixels would have a finite x-resolution. In essence, QSI achieves local smoothing by limiting the number of Fourier components which can contribute to a density function. In practice, it is a bad idea to abruptly truncate a Fourier series in this way because of the Gibbs phenomenon. Instead, the series is gradually truncated by applying a constraint on $\mathcal{E}(\hat{\mathbf{I}}_F)$, as follows.

The more general variational principle than Eq. (1) is to maximize the Von Neumann entropy of the density matrix,

$$S_Q(\hat{\rho}) \equiv -\mathcal{E}(\ln \hat{\rho}) = -\sum_n w_n \ln(w_n) \quad , \tag{13}$$

subject to any applicable constraints. This expression is not a cross entropy, because all eigenfunctions are assigned equal prior probability in accord with the unitary invariance of Hilbert space. The variational problem is to solve $\delta Q_2(\hat{\rho})/\delta\hat{\rho} = 0$ with

$$Q_2(\hat{\rho}) \equiv S_Q - \gamma^2 \mathcal{E}(\hat{\mathbf{I}}_F) - \mathcal{E}(\hat{\mathbf{U}}) - \ln Z_Q \mathcal{E}(\hat{\mathbf{1}}) \quad . \quad (14)$$

The Fisher information term imposes local smoothing with Lagrange multiplier γ^2. The next term imposes Lagrange constraints due to data. The normalization constraint $\mathcal{E}(\hat{\mathbf{1}}) = 1$ has Lagrange multiplier $-\ln Z_Q$. A maximum, $Q_2(\hat{\rho}) = 0$, is obtained for

$$\hat{\rho} = \frac{e^{-\hat{\mathbf{H}}}}{Z_Q} \qquad Z_Q = Tr\{e^{-\hat{\mathbf{H}}}\} \equiv e^{-F} \quad . \quad (15)$$

This expression defines an exponential family of density matrices parameterized by $U(x)$ and γ. F is a cumulant generating function, from which all other quantities of interest can be derived by taking appropriate derivatives. For example, in the 'flat limit' of $U(x) = 0$ these expressions are $f(x) = 1$, $\hat{\rho}(x,x') \simeq \exp(-\delta x^2/4\gamma^2)$ for $\gamma^2 \ll 1$, and $Z_Q \simeq 1/2\gamma\sqrt{\pi}$. The density matrix becomes diagonal for $\gamma = 0$. For $U(x) \neq 0$ one has $f(x) \neq 1$.

A relative information measure (or discrimination statistic) may be defined between two density matrices in this family. Let $Q_2^o(\hat{\rho})$ stand for Eq. (14) using default values of $U_o(x)$ and γ_o, and let $\hat{\rho}_o$ be the default model density matrix which maximizes it. The natural quantum information measure for $\hat{\rho}$ relative to $\hat{\rho}_o$ is

$$I_Q \equiv -Q_2^o(\hat{\rho}) = F_o(\hat{\rho}) - F_o(\hat{\rho}_o) \quad , \quad (16)$$

which measures the misfit of the variational principle, Eq. (14). Here,

$$F_o(\hat{\rho}) \equiv \mathcal{E}(\hat{\mathbf{H}}_o) - S_Q \quad ; \quad \hat{\mathbf{H}}_o \equiv \gamma_o^2 \hat{\mathbf{I}}_F + \hat{\mathbf{U}}_o \, . \quad (17)$$

Thus, for $U_o(x) = 0$ information is a simple sum of expectation values for Fisher information and quantum negentropy. For $U_o(x) \neq 0$, $\hat{H}_o$ may be regarded as a generalized Fisher information because of the identity,

$$\gamma^2 \int_0^1 dx \left(\frac{\partial\psi(x)}{\partial x}\right)^2 + \int_0^1 dx U_o(x)\psi^2(x) = \gamma^2 \int_0^1 dx w^2(x) \left[\frac{\partial}{\partial x}\left(\frac{\psi(x)}{w(x)}\right)\right]^2 \quad , \quad (18)$$

where $w(x)$ is the lowest eigenfunction for $U(x) = U_o(x)$. As shown in the next section, use of a relative information measure in QSI ensures global smoothing toward a default model representing prior knowledge of the density function.

An argument for assigning a prior probability can be developed as follows. Notice $e^{-F_o(\hat{\rho}_o)} = \sum_n e^{-\varepsilon_n^o}$, so that eigenfunctions for $\varepsilon_n \ll 1$ contribute 1 and those for $\varepsilon_n \gg 1$ contribute little. Thus e^{-F_o} simply counts the number of eigenfunctions of $\hat{\mathbf{H}}_o$ which can contribute within constraints. Eq. (16) then says that a prior probability of the form,

$$P[\vec{f}|\hat{\mathbf{U}}_o, \gamma_o] \propto \exp(-I_Q) \quad , \quad (19)$$

would be the ratio of the number of contributing eigenfunctions for $\hat{\rho}$ compared to $\hat{\rho}_o$. This Hilbert space argument is telling us how to count; a probability should be proportional to the number of different ways the system can be realized within constraints. Artificial insertion of a global smoothing (statistical regularization) hyperparameter multiplying I_Q in Eq. (19) would raise this number to some ad hoc power, and would therefore be incorrect. A proportional relation is used in Eq. (19) because the metric for the space of $\vec{f}$ has not yet been defined.

The cumulant generating function $F(\hat{\rho})$ and the information measure $I_Q(\hat{\rho})$ are related by a Legendre transformation,

$$I_Q = F - F_o(\hat{\rho}_o) - \mathcal{E}(\hat{\mathbf{U}} - \hat{\mathbf{U}}_o) - (\gamma^2 - \gamma_o^2)\mathcal{E}(\hat{\mathbf{I}}_F) \quad . \quad (20)$$

Application of perturbation theory (e.g., [3], chpt. 10) shows F to be a natural function of Lagrange multipliers, $U(x)$ and γ. I_Q is a natural function of expectation values of operators, $f(x') = \mathcal{E}(\delta(x - x'))$ and $\mathcal{E}(\hat{\mathbf{I}}_F)$. First order infinitesimal variations are

$$\delta F = \int_0^1 dx f(x)\delta U(x) \quad ,$$
$$\delta I_Q = \int_0^1 dx[-U(x) + U_o(x)]\delta f(x) \quad . \quad (21)$$

Second order infinitesimal variations are

$$\delta^2 F = -\frac{1}{2}\int_0^1 dx \int_0^1 dx' \hat{\mathbf{g}}(x,x')\delta U(x)\delta U(x') \, ,$$
$$\delta^2 I_Q = \frac{1}{2}\int_0^1 dx \int_0^1 dx' \hat{\mathbf{g}}^{-1}(x,x')\delta f(x)\delta f(x') \quad . \quad (22)$$

Let $\hat{\mathbf{d}}(\tau) \equiv \exp(-\hat{\mathbf{H}}\tau)$. The perturbation theory result for $\hat{\mathbf{g}}(x,x')$ is

$$\frac{1}{Z_Q}\int_0^1 d\tau \hat{\mathbf{d}}(x,x';\tau)\hat{\mathbf{d}}(x,x';1-\tau) - f(x)f(x') \quad . \quad (23)$$

The matrix $\hat{\mathbf{g}}$ is positive semi-definite and real symmetric. Then Eq. (22) implies I_Q is a positive convex

function of $f(x)$ which has a minimum of $I_Q = 0$ at the default model $f_o(x)$. Similarly, F is a concave function of $\delta U(x) = U(x) - U_o(x)$.

For example, in the 'flat limit' for $U(x) = 0$ and $\gamma^2 \ll 1$,

$$\hat{\mathbf{g}}(x,x') \simeq \frac{\sqrt{\pi}}{4\gamma} J\left(\frac{\delta x}{\gamma}\right) - 1 \quad ,$$
$$J(y) = \frac{2}{\pi}\int_0^{\frac{\pi}{2}} d\theta \exp\left[-\frac{y^2}{2\sin\theta}\right] \quad . \tag{24}$$

This function is non-Gaussian, peaked at $\delta x = 0$ and has a δx-variance equal to $\gamma^2/3$. As γ is increased, $\hat{\mathbf{g}}$ becomes broader. And as $\gamma \to \infty$, $\hat{\mathbf{g}}(x,x') \to 0$.

Eqs. (21,22) together imply

$$\delta f(x) = -\int_0^1 dx' \hat{\mathbf{g}}(x,x')\delta U(x') \quad . \tag{25}$$

Therefore, $\hat{\mathbf{g}}(x,x')$ may also be regarded as a 'linear response function' for changes in $f(x)$ induced by changes in $U(x')$. The completeness relation forces $\int_0^2 dx' \hat{\mathbf{d}}(x,x';\tau)\hat{\mathbf{d}}(x,x';1-\tau) = \hat{\mathbf{d}}(x,x;1)$. By Eq. (15), $f(x) = \hat{\mathbf{d}}(x,x;1)/Z_Q$. Hence, $\int_0^1 dx' \hat{\mathbf{g}}(x,x') = 0$. This property of $\hat{\mathbf{g}}$ ensures variation of $U(x)$ preserves the normalization of $f(x)$, and a shift of $U(x)$ by a constant does not change $f(x)$. And $\det[\hat{\mathbf{g}}]$ is the Jacobian for transformations of variables between $f(x)$ and $U(x)$.

The $\hat{\mathbf{g}}^{-1}(x,x')$ is a Fisher information matrix in the space of $\vec{f}$, which should be clearly distinguished from Fisher information as introduced [1] in Eq. (1) which is in the space of x. Eqs. (19,22) imply the following uncertainty relations:

$$C[f(x), f(x')] = \hat{\mathbf{g}}(x,x') \quad ,$$
$$C[U(x), U(x')] = \hat{\mathbf{g}}^{-1}(x,x') \quad ,$$
$$C[f(x), U(x')] = \delta(x-x') \quad . \tag{26}$$

Here $C[A,B]$ stands for covariance of A and B. The equivalence between the linear response function, Eq. (25), and the covariance for $f(x)$, Eq. (26), is an important concept. If $U(x)$ is fixed, $f(x)$ is uncertain, and vice versa. Individual eigenfunctions do not satisfy data constraints, but expectation values involving sums over eigenfunctions do. At fixed $U(x)$ the most probable value of $f(x)$ is equal to $\mathcal{E}(f(x))$, which is a sufficient statistic to determine $U(x)$. The exponential family of density functions given by Eq. (15) achieves equality limits of more general uncertainty relations termed Cramer-Rao inequalities [2, 14] in statistics.

Similar uncertainty relations exist between γ^2 and $\hat{\mathbf{I}}_F$. In particular, $\partial\mathcal{E}(\hat{\mathbf{I}}_F)/\partial\gamma^2$ is strictly negative. Therefore, $\mathcal{E}(\hat{\mathbf{I}}_F)$ increases as γ^2 decreases, is maximal in the 'classical' limit of $\gamma^2 = 0$, and it is zero in the 'quantum' limit of $\gamma^2 \to \infty$.

The symmetry exhibited in Eqs. (21,22,26) between Legendre transform conjugate variables is known as 'duality' [8], and it is satisfied only in the absence of a global smoothing hyperparameter. Here, duality is more than just a calculational method for convex optimization problems, but it is intrinsic to the theory. More generally, QSI satisfies the framework of information geometry in the statistics literature [15, 16, 2]. I_Q is an 'information divergence'. The Fisher information matrix $\hat{\mathbf{g}}^{-1}$ is equivalent to a 'Riemann metric'. Use of these differential geometry concepts helps to ensure invariance of QSI to the choice of coordinate system.

The Jeffrey's prior, or 'Riemann volume' $\sqrt{\det[\hat{\mathbf{g}}^{-1}]}$, defines a coordinate-invariant metric for integration over the space of $\vec{f}$. Thus, the prior probability at fixed $\gamma = \gamma_o$ should be:

$$P[\vec{f}|\hat{\mathbf{U}}_o, \gamma] \approx \sqrt{\det[\hat{\mathbf{g}}^{-1}]}\exp(-I_Q) \quad . \tag{27}$$

Use of the $\approx$ symbol is a mathematical warning: this expression is ill-defined unless the x-basis is discretized. Even then $\hat{\mathbf{g}}$ has a zero-eigenvalue in order to preserve the normalization of $f(x)$. However, these potential difficulties cancel out of all practical applications of this expression, as shall be shown in the next section.

To better understand this formulation, it is helpful to consider the 'classical limit' defined by $\gamma = \gamma_o \to 0$ and $U(x) \neq 0$. $\hat{\mathbf{H}}$ becomes diagonal in the x-basis. The density matrix $\hat{\rho}$ reduces to a diagonal matrix whose elements are a ME image given by $f(x) = \exp(-U(x))/Z$. I_Q reduces to negative (Kullback-Liebler or Shannon-Jaynes) cross entropy,

$$S_C = -\int_0^1 dx f(x)\ln\left(\frac{f(x)}{f_o(x)}\right) \quad , \tag{28}$$

with $f_o(x)$ the default model. The $\hat{\mathbf{g}}(x,x')$ reduces to $f(x)\delta(x-x') - f(x)f(x')$. There is no prior covariance between points in an image, apart from normalization. Thus, the classical limit is ME. If the x-basis is discretized into N equal steps with $f_i = f(x_i)/N$, then

$$P[\vec{f}|U_o] \to \frac{1}{\prod_{i=1}^N \sqrt{2\pi f_i}}\exp(S_C)\,\delta\left(1 - \sum_i f_i\right) \quad . \tag{29}$$

The zero eigenvalue of $\hat{\mathbf{g}}$ has been removed by the introduction of the δ-function. This limit differs somewhat from 'Classic MaxEnt' [13], because $f(x)$ is normalized to unity and a statistical regularization (or

global smoothing) hyperparameter is absent. In Cambridge parlance, $\alpha = 1$ is known *a priori*, which eliminates controversy about how to determine it. As discussed in Sec. 4, this ME limit is recommended only for a restricted set of data analysis problems involving very sharp, isolated structure in $f(x)$.

What happens in the general case of $\gamma = \gamma_o \neq 0$ and fixed $U(x) \neq 0$? The $\gamma = 0$ limit is the ME image discussed above which has maximum structure in $f(x)$, maximum I_Q, and $\hat{\mathbf{g}}(x, x')$ a δ-function. As γ is increased the image becomes blurred on a length scale characterized by $\gamma/\sqrt{3}$, I_Q falls, and $\hat{\mathbf{g}}$ becomes broader. In the 'quantum limit' of $\gamma \to \infty$, $f(x) \to 1$, $I_Q \to 0$, and $\hat{\mathbf{g}} \to 0$.

3 Bayesian Statistical Inference

Of the many possible applications for QSI, this paper considers only the class of data analysis problems of form

$$D_k = \int_0^1 dx R_k(x) f(x) + N_k \quad . \tag{30}$$

This is the k'th component of a vector equation, where $\vec{D}$ are N_d data which are typically incomplete, $\vec{R}(x)$ is an integral transform, and $\vec{N}$ represents noise.

Consider application of QSI to this ill-posed inverse problem, with $\gamma = \gamma_o$. Introduce Lagrange multipliers λ_k for each datum,

$$U(x) = \vec{\lambda} \cdot \vec{R}(x) + U_o(x) \quad . \tag{31}$$

Use of N_d Lagrange multipliers exhausts the number of free parameters in $U(x)$ which can be determined uniquely by N_d data. Inserting Eq. (31) into Eq. (15) defines a sub-family of density matrices parameterized by $\vec{\lambda}$ and γ. From Eq. (21), first order variations are

$$\frac{\partial F}{\partial \lambda_k} = \mathcal{E}(R_k) \qquad \frac{\partial I_Q}{\partial \mathcal{E}(R_k)} = -\lambda_k \quad . \tag{32}$$

From Eq. (22), second order variations are

$$\frac{\partial^2 F}{\partial \lambda_k \partial \lambda_l} = \mathbf{K}_{k,l} \qquad \frac{\partial^2 I_Q}{\partial \mathcal{E}(R_k) \partial \mathcal{E}(R_l)} = \mathbf{K}^{-1}_{k,l} \quad . \tag{33}$$

From Eq. (25), the Fisher information matrix in the space of observations, $\mathcal{E}(R_k)$, is

$$\mathbf{K}_{k,l} = -\frac{\partial \mathcal{E}(R_k)}{\partial \lambda_l} \equiv \int_0^1 dx \int_0^1 dx' R_k(x) \mathbf{g}(x, x') R_l(x') \quad . \tag{34}$$

Alternatively $\mathbf{K}$ is the Riemann inner product of two observables R_k and R_l. It is a real symmetric positive semi-definite matrix. Analogues of uncertainty relations, Eq. (26), are

$$\begin{aligned} C[R_k, R_l] &= \mathbf{K}_{k,l} \quad , \\ C[R_k, \lambda_l] &= \delta_{k,l} \quad , \\ C[\lambda_k, \lambda_l] &= \mathbf{K}^{-1}_{k,l} \quad , \end{aligned} \tag{35}$$

where $\delta R_k(x) \equiv R_k(x) - \mathcal{E}(R_k)$. The appropriate prior probability corresponding to Eq. (27) is

$$P[\mathcal{E}(\vec{R})|\hat{\mathbf{U}}_o, \gamma] = \frac{1}{\sqrt{(2\pi)^{N_d} \det[\mathbf{K}]}} \exp(-I_Q) \quad , \tag{36}$$

which is normalized to unit integral over the space of $\mathcal{E}(\vec{R})$ within Gaussian approximations for the integral. The Jacobian for transformations between integrals over $\mathcal{E}(\vec{R})$ and $\vec{\lambda}$ is $\sqrt{\det[\mathbf{K}]}$.

Noise on the data is assumed to be normally distributed with $E(N_k) = 0$ and $Cov(N_k N_l) = \delta_{k,l}\sigma_k^2$. The likelihood function is then

$$P[\vec{D}|\vec{f}] = \frac{1}{\sqrt{\prod_{k=1}^{N_d} 2\pi\sigma_k^2}} \exp\left(-\frac{\chi^2}{2}\right) \quad , \tag{37}$$

where

$$\chi^2 = \sum_{k=1}^{N_d} \frac{(D_k - \mathcal{E}(R_k))^2}{\sigma_k^2} \quad . \tag{38}$$

Invoking Bayes theorem, the posterior probability is

$$\begin{aligned} &P[\mathcal{E}(\vec{R})|\vec{D}, \hat{\mathbf{U}}_o, \gamma] \times P[\vec{D}, \hat{\mathbf{U}}_o, \gamma] \\ &\quad = P[D|\vec{f}] \times P[\mathcal{E}(\vec{R})|\hat{\mathbf{U}}_o, \gamma] \quad , \end{aligned} \tag{39}$$

which equals

$$\frac{1}{(2\pi)^{N_d}\sqrt{\left(\prod_{k=1}^{N_d}\sigma_k^2\right)\det[\mathbf{K}]}} \exp\left(-I_Q - \frac{\chi^2}{2}\right) \quad . \tag{40}$$

Regarding the first factor as a metric, the maximum a posteriori image is obtained from the minimum of

$$Q_3 \equiv I_Q + \frac{\chi^2}{2} \quad . \tag{41}$$

The minimum is found for

$$\frac{\partial Q_3}{\partial \mathcal{E}(R_k)} = -\frac{D_k}{\sigma_k^2} + \frac{\mathcal{E}(R_k)}{\sigma_k^2} - \lambda_k = 0 \quad . \tag{42}$$

The second derivative,

$$\frac{\partial^2 Q_3}{\partial \mathcal{E}(R_k) \mathcal{E}(R_l)} = \mathbf{K}^{-1}_{k,l} + \frac{\delta_{k,l}}{\sigma_k^2} \quad , \tag{43}$$

is strictly positive implying that a unique solution to Eq. (42) exists. Solutions may be found by a variety of numerical methods for convex optimization problems.

Denote by sub(super)script 'f' quantities evaluated at the solution to Eq. (42). And define a matrix $\mathcal{K}^f_{k,l} = \mathbf{K}^f_{k,l}/\sigma_k\sigma_l$. For typical data analysis problems, the eigenvalue spectrum of $\mathcal{K}^f$ is a very steep function varying over many orders of magnitude. Independent measurements can be defined as eigenvectors of $\mathcal{K}^f$, and their quality can be ranked according to the size of their eigenvalues. Following [13], *good measurements* may be defined as eigenvectors whose eigenvalues are much greater than 1. Data for good measurements are dominated by signal, so that typical values of $|\mathcal{E}(R_k)/\sigma_k| \gg 1$. *Bad measurements* may be defined as eigenvectors whose eigenvalues are much less than 1. Data for bad measurements are dominated by noise, so typical values of $|\mathcal{E}(R_k)/\sigma_k| \ll 1$. In other words, QSI fits good data and ignores bad data as should be expected.

The *number of good measurements* may be defined by

$$N_g^f \equiv Tr[\mathcal{K}^f \cdot (1+\mathcal{K}^f)^{-1}] \quad . \tag{44}$$

An error scaling consistency argument suggests

$$\chi^2_f \approx N_d - N_g^f \quad , \tag{45}$$

so χ^2 for the most probable image is less than the number of data. An important distinction between QSI and most other regularization methods [7, 13] is that no global smoothing (statistical regularization) parameter enters Eqs. (44). Hence, putting a bound on N_g would place an upper limit on the number of good measurements in an experiment, at least when analyzed according to QSI. Numerical experiments suggest the inequality,

$$N_g(\vec{\lambda} \neq \vec{0}, \gamma \neq 0) \leq N_g(\vec{\lambda} = \vec{0}, \gamma = 0) \quad . \tag{46}$$

If this conjectured relation could be proved, it would allow initial good measurements to be identified from eigenvectors of the $\mathcal{K}(\vec{\lambda} = \vec{0}, \gamma = 0)$, which can often be calculated *a priori* from the experimental setup before any data are acquired. Throwing out initial bad measurements can dramatically reduce the computational and data acquisition tasks without affecting results.

An automatic Bayesian method for determining the optimal γ^2 local smoothing hyperparameter is to maximize the marginal posterior probability $P[\vec{D}, \hat{\mathbf{U}}_o, \gamma]$ in Eq. (39) (or *evidence*). Integrate Eq. (40) over $d\mathcal{E}(\vec{R})$ and use a Gaussian approximation for the integrand around its peak value. Then evidence is

$$\frac{1}{(2\pi)^{N_d}\sqrt{\left(\prod_{k=1}^{N_d}\sigma_k^2\right)\det[\mathcal{K}^f+1]}} \exp\left(-I_Q^f - \frac{\chi^2_f}{2}\right) \quad . \tag{47}$$

Behavior of this expression depends on variation of eigenvalues of $\mathcal{K}$ with γ. At fixed $\mathcal{E}(\vec{R})$ eigenvalues are maximal in the 'classical limit' $\gamma = 0$, decrease monotonically with increasing γ, and go to zero in the 'quantum limit' $\gamma = \infty$. Eq. (47) is a product of two factors. The Jeffrey's prior, or Riemann volume factor, $\propto 1/\sqrt{\det[\mathcal{K}^f+1]}$, favors the simpler model of larger γ, which means fewer eigenfunctions (Fourier components) contributing to the image. This factor is, therefore, equivalent to an *Occam factor* [19], since it implements the principle of Occam's razor. Eq. (43) implies the *data factor*, $\propto \exp(-I_Q - \chi^2/2)$, favors the more complex model of smaller γ. It is easier to fit data if $f(x)$ can be sharper. Balance between these two opposing factors determines the optimal γ, depending on data.

Estimates of reliability should be as important an output of a Bayesian calculation as the maximum a posteriori image. Covariance of the image may be calculated from

$$\delta^2 Q_3 = \frac{1}{2}\int_0^1 dx \int_0^1 dx' \hat{\mathbf{c}}^{-1}(x,x')\delta f(x)\delta f(x') \quad , \tag{48}$$

where

$$\hat{\mathbf{c}}^{-1}(x,x') = \hat{\mathbf{g}}^{-1}(x,x') + \sum_{k=1}^{N_d} \frac{R_k(x)R_k(x')}{\sigma_k^2} \quad . \tag{49}$$

Define

$$\Gamma_k(x) \equiv \int_0^1 dx' \hat{\mathbf{g}}(x,x')\frac{R_k(x')}{\sigma_k} \quad . \tag{50}$$

Covariance of $f(x)$ is given by

$$C[f(x), f(x')] = \hat{\mathbf{c}}(x,x') = \hat{\mathbf{g}}(x,x') - \vec{\Gamma}(x)^\dagger \cdot \left(1+\mathcal{K}^f\right)^{-1} \cdot \vec{\Gamma}(x') \quad . \tag{51}$$

This should be compared to uncertainty relations, Eq. (26), obtained in absence of data. Good measurements reduce uncertainty in $f(x)$ as expected. This covariance can be proved to be strictly positive by projection superoperator methods. For $\gamma \neq 0$, Eq. (51) yields a finite covariance on points of an image, which is an important distinction between QSI and ME.

Covariances of two observables $A(x)$ and $B(x)$ follow from Eq. (51),

$$C[A,B] = K[A,B] - K[A,\vec{\Gamma}]^\dagger \cdot \left(1 + \mathcal{K}^f\right)^{-1} \cdot K[\vec{\Gamma},B] \quad , \tag{52}$$

where the Riemann inner product of observables is

$$K[A,B] \equiv \int_0^1 dx \int_0^1 dx' A(x) \hat{\mathbf{g}}(x,x') B(x') \quad . \tag{53}$$

Insight can be gained by separating covariance into a sum $C[A,B] = C_S[A,B] + C_I[A,B]$, with *statistical covariance*

$$C_S[A,B] \equiv K[A,\vec{\Gamma}]^\dagger \cdot \left(\mathcal{K}^f + \mathcal{K}^f \cdot \mathcal{K}^f\right)^{-1} \cdot K[\vec{\Gamma},B] \quad , \tag{54}$$

and an *incompleteness covariance*

$$C_I[A,B] = K[A,B] - K[A,\vec{\Gamma}]^\dagger \cdot \mathcal{K}_f^{-1} \cdot K[\vec{\Gamma},B] \quad . \tag{55}$$

This distinction makes sense in several respects. If statistical errors on the data are zero, C_I remains finite while C_S goes to zero. If A is one of the measurements, C_I goes to zero while C_S remains finite. Bad measurements are the dominant contribution to C_S, and they cancel the reduction in C_I. An example is the covariance matrix of the measurements, $C[R_k, R_l]$, for which $C_I = 0$. With the definition $\mathcal{R}_k \equiv R_k/\sigma_k$, statistical covariance is simply related to the number of good measurements,

$$Tr\{C_S[\vec{\mathcal{R}},\vec{\mathcal{R}}]\} = N_g^f \quad . \tag{56}$$

Finally, from Eq. (42) one can prove for infinitesimal variation,

$$\delta f(x) = -\int_0^1 dx' \hat{\mathbf{c}}(x,x') \delta U(x') \quad , \tag{57}$$

which should be compared to Eq. (25) in absence of data. Thus, the covariance matrix does more than describe uncertainties in the image due to incomplete or uncertain data; it also characterizes linear response of the image to changes in data or default model. Large error bars correspond to high sensitivity to input information. For example, if the image depends strongly on choice of default model, it should not be reliable. The covariance matrix is an extremely important output of the data analysis procedure.

4 Applications

This section illustrates the performance of QSI for a few inverse problems, and compares results to ME.

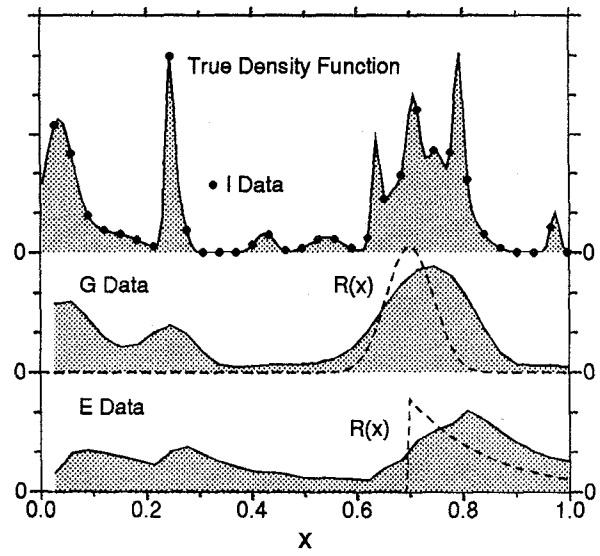

Figure 1: **Exact Data Example** - Top curve is the test density function, and dots are interpolation (I) data. Middle curves are Gaussian (G) data (solid) and resolution function $R(x)$ (dashed $\times 0.35$). Bottom curves are exponential (E) data and resolution function. Data are exact and measured at 32 points. The Gaussian has a standard deviation of 0.044 and the exponential has a decay constant of 0.15.

All examples use the test density function shown at the top of Fig. 1. The default model is taken to be flat, $U_o(x) = 0$, corresponding to $f_o(x) = 1$.

Fig. 1 also shows three data analysis problems corresponding to interpolation (I), Gaussian deconvolution (G), and exponential deconvolution (E). Data are exact and measured at 32 equally spaced x-points. ME images are obtained by the method of Lagrange multipliers and displayed using 128 pixels in Fig. 2. The ME image for the Gaussian deconvolution data is credible, but it exhibits overly sharp and occasionally spurious structure. The ME image for the E data exhibits sharp edges reflecting the sharp feature in the resolution function. The ME image for the I data equals the data for measured pixels and equals the default model value of 1.0 for unmeasured pixels. The I and E ME images are not credible because they clearly reflect how the data were measured.

These examples illustrate general characteristics of ME. It works best for problems, such as deconvolution, where $\vec{R}(x)$ is a broad function and $f(x)$ contains comparatively sharp features. However, for problems where $\vec{R}(x)$ is sharp and $f(x)$ broad, ME tends to produce spurious structure. The interpolation problem is an extreme example, since $\vec{R}(x)$ consists of a set of

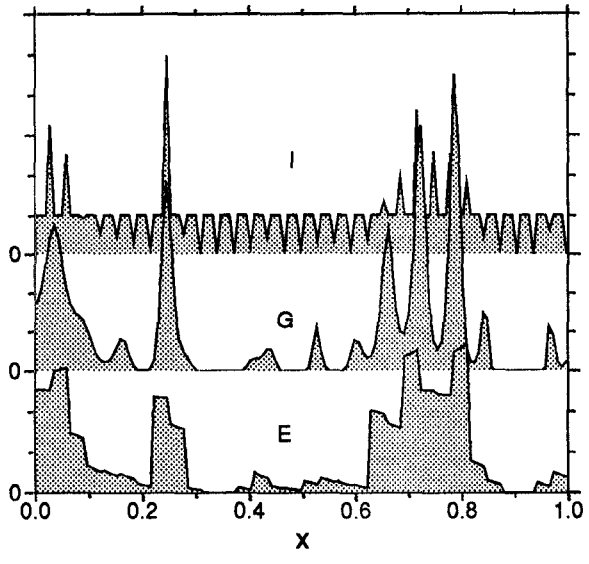

Figure 2: **Maximum Entropy Images** - For exact data example in Fig. 1, displayed in 128 pixels.

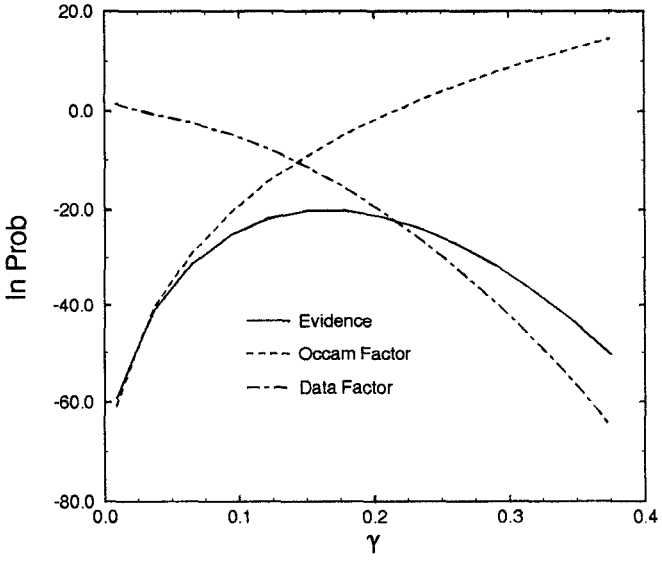

Figure 3: **Optimization of Local Smoothing Hyperparameter** - The evidence, occam factor, and data factor for the interpolation problem in Fig. 1 as a function of γ.

Dirac delta functions and the ME solution is nonsense. More generally, ME will fail for any problem where local smoothing is critical, such as density function estimation (sampling). This conclusion is true regardless of whether a global smoothing hyperparameter is used.

Application of QSI to the exact data example corresponds to setting $\mathcal{E}(\vec{R}) \rightarrow \vec{D}$ in Eq. (36). An optimal local smoothing hyperparameter γ is determined by maximizing this expression, which separates into two factors whose behavior is illustrated in Fig. 3. The Jeffrey's prior (or Riemann volume factor), $1/\sqrt{\det(\mathbf{K})}$, favors the simpler model of larger γ corresponding to fewer Fourier components in the image. This factor is therefore equivalent to an *Occam factor* in the Bayesian literature (Bretthorst, 1988), since it implements the principle of Occam's razor. The *data factor*, $\exp(-I_Q)$, favors the more complex model of smaller γ, corresponding to more Fourier components in the image. The solid curve in Fig. 3 shows the *evidence* for γ on a log-linear scale. Comparable values of $\gamma_{opt} = 0.17$ are found for all three data analysis problems in Figure 1. Figure 4 shows the optimal QSI images for the data in Figure 1. They are clearly more credible than corresponding ME ($\gamma = 0$) images shown in Figure 2. Credibility is quantitatively expressed by the many orders of magnitude larger Bayesian evidences for QSI images relative to ME images.

Figure 5 shows noisy data sets for the test density function in Fig. 1 and the same Gaussian (G) and exponential (E) resolution functions. The data are measured at 64 equally spaced points with noise levels indicated. Figure 6 shows corresponding ME images obtained by setting $\gamma = 0$ in QSI. The ME images clearly show excessive overfitting and noise artifacts. Only slightly smoother images would be obtained using the Cambridge ME algorithm (MEMSYS) which includes a global smoothing parameter (Skilling, 1989). Much improved results would be obtained by correcting ME for local smoothing using intrinsic correlation functions, but it would be an *ad hoc* fix. Figure 7 shows corresponding QSI images for the same data, which are clearly superior to ME images. Although χ^2 is larger for QSI images than for ME images, the evidence for them is greater because they correspond to the simpler model of larger γ. The data sets are placed in order of increasing information content, as measured by the value of I_Q^f. This ranking also corresponds to the visual quality of the images in Figure 7 and to values of N_g^f. Note that lineshape can be more important than statistical errors in determining the information content of data, with sharper lineshapes generally yielding more information (everything else being equal).

More generally, QSI may provide a unified approach to almost all data analysis problems involving a density function discussed in the statistics literature [7, 8, 9]. These include density function estimation (sampling), multinomial smoothing, image processing, approximate solution of integral equations, etc. Statis-

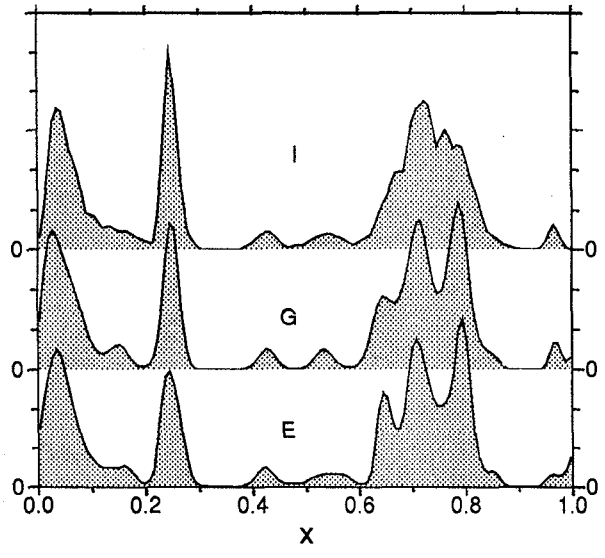

Figure 4: **Quantum Statistical Inference Images** - For the exact data example in Fig. 1, displayed in 128 pixels.

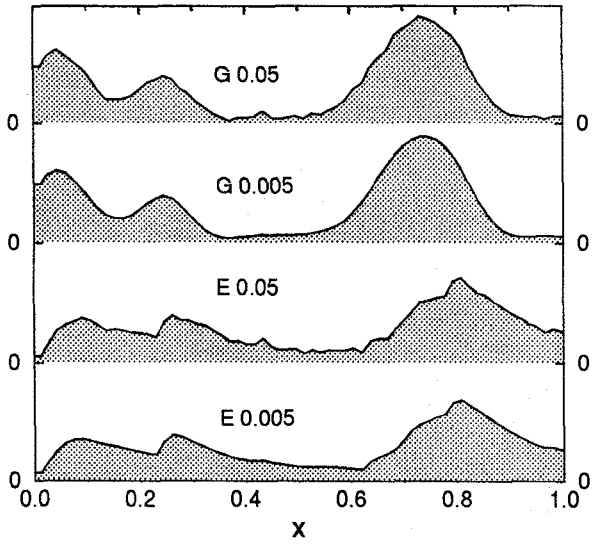

Figure 5: **Noisy Data Example** - The test density function is convolved with the same resolution functions in Fig. 1 at 64 equally spaced data points, and then Gaussian independent noise is added. The notation "E 0.05" stands for the exponential resolution function with a noise standard deviation of 0.05.

tical and incompleteness errors on the image can also be estimated. Sensitivity to changes in data and prior knowledge can be assessed. The fully Bayesian formulation permits optimization of any unknown parameters of the data analysis procedure. Prior knowledge in the form of default models and other constraints can be incorporated. A quadratic approximation to QSI termed the *quantum filter* provides a numerically efficient approximation in exchange for a possible loss of resolution and positivity. QSI may also provide a useful approach to experimental design, because it enables *a priori* identification of good and bad measurements.

5 Relation of QSI to Physics

A brief discussion of frequentist and Bayesian concepts of statistics in physics is followed by a long list which identifies physical concepts analogous to QSI concepts.

The development of concepts of probability by physicists has been largely independent of developments in statistics. Terms such as 'Bayesian' or 'frequentist' are not used, and Bayes theorem is almost unknown. But analogies may be drawn. To the extent that the scientific method is Bayesian, one may argue that use of Bayesian concepts by physicists is 'implicit'. But most probability concepts in physics are frequentist. Ordinary quantum mechanics is concerned with solving a forward problem, not an inverse problem. The goal is to calculate probabilities given the *potential energy* $V(x) = U(x)/\gamma^2$. Experiments are usually carried out in the limit of large numbers of samples, so frequentist concepts of statistics are adequate. Indeed, the quantum concept of a density function, $f(x,t) = |\psi(x,t)|^2$, is frequency of particles at point x and time t.

Quantum probabilities do not obey Bayes theorem. Feynman's path integral formulation of quantum theory is a law of conditional $\psi(x,t)$, not conditional probabilities as used in Bayes theorem. Bayes theorem is recovered from path integrals only by dropping interference terms between different histories.

The derivation of quantum statistical mechanics from a maximum entropy principle used in this paper originated with Jaynes [11]. It is discussed admirably in Balian's book [14], among others. But most derivations of statistical mechanics are frequentist; they usually invoke an unproven ergodic hypothesis and/or contact with an imagined reservoir. Frequentist arguments can work because statistical mechanics is usually concerned with a few constraints (pressure, average energy, magnetic field, etc.) applied to a system having a large number of particles (typically Avogadro's number, 6×10^{23}). Fluctuations can usually be ignored. Most, if not all, of the statis-

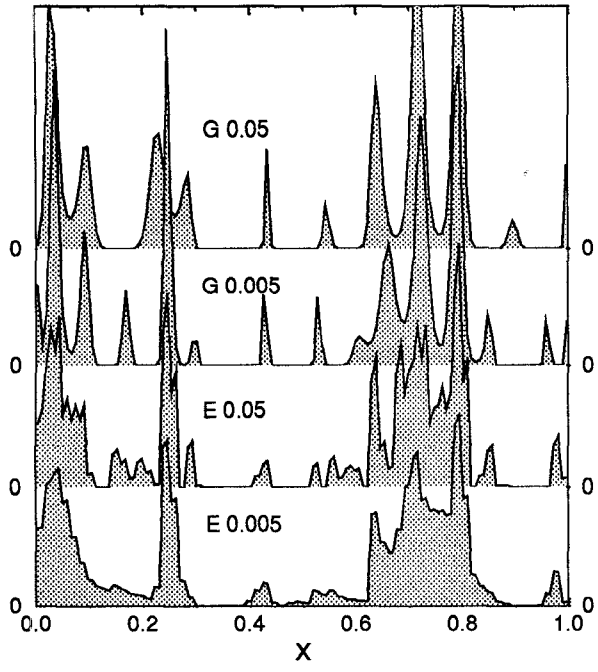

Figure 6: **Maximum Entropy Images** - For the noisy data example in Fig. 5.

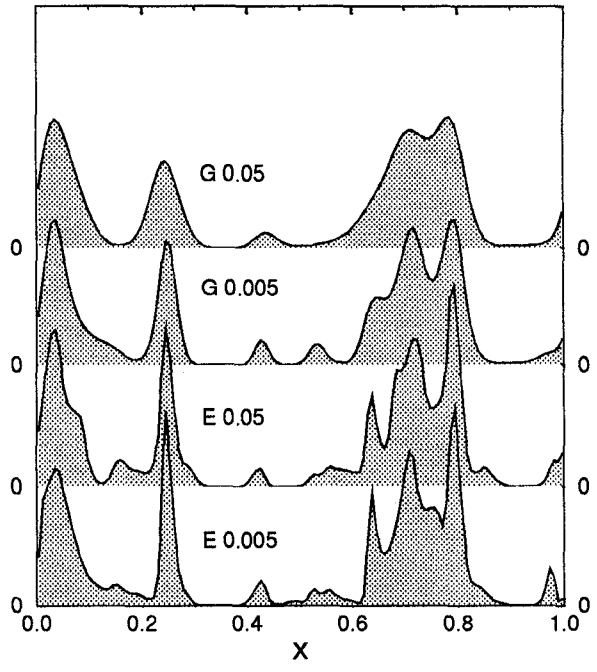

Figure 7: **Quantum Statistical Inference Images** - For the noisy data example in Fig. 5.

tical mechanics concepts which provide analogies for QSI were first derived by frequentist arguments. And it would be no surprise to see a frequentist expression of QSI.

And now for the list of analogies. Eq. (1) is a *Lagrangian.* Eq. (2) is analogous to the *Schrödinger equation.* Units correspond to setting *temperature* to 1. In Eq. (3), Fisher information is analogous to *kinetic energy*, data constraints $U(x)$ are analogous to *potential energy*, and ε is analogous to *total energy.* Differential operator $\hat{\mathbf{H}}$ in Eq. (4) is analogous to a *Hamiltonian.* The $\delta(x - x')$ in Eq. (6) is a *Dirac delta function.* Matrix representations of differential operators are primarily due to Heisenberg and Dirac.

Lagrange multipliers are analogous to *intensive variables* in statistical mechanics, and the data (or observations) are analogous to *extensive variables.* The goal of statistical mechanics is to infer the most likely macrostate of a quantum many-body system from data consisting of extensive variables and from prior knowledge of a Hilbert space of possible microstates. Concept of a *density matrix* in Eqs. (8,9) is due to Von Neumann, as is definition of its entropy, Eq. (13). The exponential family of density matrices in Eq. (15) is analogous to a *canonical ensemble* in statistical mechanics. Normalization, $Z_Q = \exp(-F)$, is analogous to a *partition function.* Cumulant generating functional F is analogous to *free energy.* The argument after Eq. (19) that probabilities should be proportional to the number of eigenfunctions that can contribute within constraints is very important in statistical mechanics. Thus, most statistical mechanics calculations begin with the partition function, from which all other quantities can be derived. To maximize system probability in a Bayesian sense is the same as the *principle of minimum free energy.*

The linear response function in Eqs. (25,57) is analogous to a *susceptibility* [17, 18]. Equivalence between the linear response function, Eq. (25,57), and the covariance for $f(x)$, Eq. (26,51), is analogous to the *fluctuation-dissipation theorem.* This is a quite fundamental statistical property of physical systems; for example, it relates electronic noise in a circuit to its resistivity, or density fluctuations in a liquid to its compressibility. Uncertainty relations, Eq. (26), are analogous to *thermodynamic uncertainty relations.* Uncertainty relations between γ^2 and $\hat{\mathbf{I}}_F$ are related to the *Heisenberg uncertainty principle* [2]. Local smoothing parameter γ is analogous to a *de Broglie wavelength.* Units chosen are equivalent to $\gamma^2 = \hbar^2/2mk_BT$, where $\hbar$ is Planck's constant, m is mass, k_B is Boltzmann's constant, T is temperature.

QSI has used only commuting observables $R_k(x)$, real wave functions, the time-independent Schrödinger equation, etc. Statistical mechanics is mathematically analogous to quantum dynamics in imaginary time $\tau \propto \gamma^2$. This is a small fraction of the rich structure of quantum theory. There is enormous potential for

the application of additional quantum concepts such as gauge fields, path integrals, semiclassical approximations, etc. to statistical inference.

6 Conclusion

Did the founders of quantum mechanics, in solving the riddles of the atom, serendipitously discover how to reason from incomplete and uncertain information? QSI is the first explicit application of quantum theory to human reasoning. The compelling motivations for QSI include mathematical analogies between statistical data analysis and statistical mechanics, the successes and limitations of the maximum entropy method, and the unquestioned success of quantum statistical mechanics in physics. The mathematical structure of quantum mechanics has been adapted to the statistical inference of continuous density functions. Applications have been demonstrated for data interpolation and deconvolution, and results are superior to maximum entropy. A vast reservoir of other physics concepts and methodology may also be applicable to statistics and information theory. They can be smoothly combined with Bayesian methods for reasoning from incomplete data and prior knowledge.

References

[1] B. R. Frieden, "Fisher Information as the Basis for the Schrödinger Wave Equation", *Am. J. Phys.* **57**, 1004 (1989).

[2] E. R. Cainiello, "Quantum and Other Physics as Systems Theory", *La Revista de Nuovo Cimento*, (1992).

[3] J. Mathews, R. L. Walker, *Mathematical Methods of Physics*, W. A. Benjamin, New York, (1964).

[4] R. A. Fisher, "Theory of Statistical Estimation", *Proc. Cambr. Phil. Soc.* **22**, 700 (1925).

[5] A. Dembo, T. M. Cover, J. A. Thomas, "Information Theoretic Inequalities", *IEEE Trans. Info. Theory* **37**, 1501 (1991).

[6] Preliminary accounts of this work were reported in: *E. T. Jaynes: Physics and Probability*, W. T. Grandy, P. W. Milonni, eds, Cambridge University Press, to be published; *Maximum Entropy and Bayesian Methods, 1992*, M. Djafari, G. Demoment, eds., Kluwer Academic Press, to be published.

[7] D. M. Titterington, "Common Structure of Smoothing Techniques in Statistics", *Int. Statist. Rev.*, **53**, 141 (1985).

[8] A. Tarantola, *Inverse Problem Theory*, Elsevier, Amsterdam (1987).

[9] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London (1986).

[10] T. Loredo, "From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics", in P. Fougere, ed., *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, 81 (1990).

[11] E. T. Jaynes, in R. D. Rosenkrantz, ed., *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, D. Reidel Publishing Co., Dordrecht (1983).

[12] Y. Tikochinsky, N. Z. Tishby, R. D. Levine, "Consistent Inference of Probabilities for Reproducible Experiments", *Phys. Rev. Letts.*, **52**, 1357 (1984).

[13] J. Skilling, S. Gull, "Classic MaxEnt", in J. Skilling, ed., *Maximum Entropy and Bayesian Methods*, Kluwer, Dordrecht, 45-71 (1989).

[14] R. Balian, *From Microphysics to Macrophysics*, Springer-Verlag, Berlin (1991).

[15] S. Amari, *Differential-geometrical Methods in Statistics*, Springer-Verlag, Berlin (1985).

[16] R. Balian, Y. Alhassid, H. Reinhardt, "Dissipation in Many-Body Systems: A Geometric Approach Based on Information Theory", *Physics Reports* **131**, 1 (1986).

[17] R. Kubo, "Some Aspects of the Statistical-Mechanical Theory of Irreversible Processes", in Britten & Dunham, eds., *Lectures in Theoretical Physics*, 1 (1959).

[18] E. Fick, G. Sauermann, *The Quantum Statistics of Dynamic Processes*, Springer-Verlag, Berlin (1990).

[19] L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Springer-Verlag, Berlin (1988).