Proof of a Modified Jaynes's Estimation Theory

R.F. Streater
King's College London
Department of Mathematics
Strand, London WC2R 2LS

01 April 2010

Abstract

It is proved that the state of maximum entropy having observed values for the n observables, $X_1, \ldots, X_n$, is the same state that minimises the matrix of covariances of any n locally unbiased estimators for n parameters for the probability distribution of $X_1, \ldots, X_n$. We sketch how to get a similar result in quantum theory, in which $X_1, \ldots, X_n$ are (not necessarily commuting) quadratic forms that are bounded relative to a positive self-adjoint operator H such that $\exp(-\beta H)$ is of trace-class for some positive β.

Contents

This paper has been accepted for publication in "Open Systems and Information Dynamics".

1 The Estimator of Jaynes

Let X be a random variable with sample-space Ω and Borel ring $\mathcal{B}$. What is the best choice of probability measure μ on Ω such that the mean of X,

$$\mathbf{E}[X] := \int_\Omega X(\omega)\mu(d\omega) \tag{1}$$

has a given value? If this value is indeed known, then Jaynes [6] has argued that if no other information about the probability distribution of X is known, then the answer is the distribution of maximum entropy among all distributions giving this mean. The requirement that the entropy is maximum is motivated by the attempt not to impose any information about the distribution other than the given mean. Clearly, we cannot find out anything about $(\Omega, \mu, \mathcal{B})$ from $\mathbf{E}[X]$ that is not obtainable from the probability distribution of X. Let us give Jaynes's method.

Let us assume that Ω is discrete; let $\omega \mapsto \mu(\omega)$ be a probability measure on Ω, and let $\rho(x) = \mu\{\omega : X(\omega) = x$ be the probability distribution of X. The entropy of ρ is

$S(\rho) = -\sum_x \rho(x)\log(\rho(x))$ We must maximise this under the condition that the mean of X is known to be x_0, and the sum of probabilities is 1. We use Lagrange's multipliers λ_1, λ_2: maximise

$S_\lambda = S(\rho) + \lambda_1 \sum_x x\rho(x) + \lambda_2 \sum_x \rho(x)$,

subject to no conditions on the values of $\rho(x)$, to get the function $\rho_\lambda(x)$. We then find λ_1, λ_2 by the equations

$$\begin{aligned} \sum_x x\rho_\lambda(x) &= x_0 \\ \sum_x \rho_\lambda(x) &= 1. \end{aligned}$$

We find that

$$\frac{\partial S_\lambda}{\partial \rho(x)} = -\log\rho(x) - 1 + \lambda_1 x + \lambda_2 = 0 \tag{2}$$

for all x. Thus we see that the Jaynes state has an exponential form,

$$\rho(x) = e^{\lambda_1 x + \lambda_2 - 1}. \tag{3}$$

We get a similar result, an exponential form for the density function, if there are N random variables of known means; thus, let $X_1, \cdots, X_N$ be N linearly independent random variables of known means. Then the state of maximum entropy among all states having these means has the form

$$\rho(x) = \exp\{\lambda_1 X_1 + \ldots + \lambda_N X_N + \text{const.}\}. \tag{4}$$

We can also drop the assumption that the variables $X_1, \cdots, X_N$ have only discrete values; then we replace the sum, $\sum_j$ by the integral over x, and the entropy by the differential entropy, to get again the exponential form for the density. The problem is, in estimation theory we do not know the means; the best we can do is to measure the value of each of the variables in our sample, and then take these values as approximations to the means. This does not give us exactly the answer wanted by Jaynes. However, we now show that this does give us the best estimate in the sense of R. A. Fisher!

2 Fisher's Method

Suppose that X is a random variable on a sample space $(\Omega, \mathcal{X})$, and that the probability density function of X is one of the family $\{\rho_\lambda; \lambda \in \mathbf{R}\}$. We take a sample of X and use it to estimate the best value for λ. Thus, an estimator for λ is chosen; it is a function, say f of the value, say x, found for X. Thus, we would find the parameter λ to be $f(x)$. This, being a function of a random variable, is a random variable. Fisher suggested [3] that the best estimators should be first, *locally unbiased*, and secondly, *of minimal variance*. We say that f is unbiased if the average value gives the correct answer:

$$\mathbf{E}_{\rho_\lambda}[f(X)] = \lambda. \tag{5}$$

It is convenient to require a weaker condition, and replace this by its derivative, to arrive at the condition called locally unbiased:

$$\int \frac{\partial \rho_\lambda(x)}{\partial \lambda} f(x)\, dx = \mathbf{E}_{\rho_\lambda}\left[\frac{\partial \log \rho_\lambda}{\partial \lambda} f(X)\right] = 1. \tag{6}$$

Thus the condition, locally unbiased, may be written as the condition

$$\langle \partial_\lambda \log \rho_\lambda, f\rangle = 1. \tag{7}$$

Here, $\langle g, h\rangle = \mathbf{E}_{\rho_\lambda}[gh]$ furnishes the real vector space of functions of X with a scalar product.

Fisher's second condition is that the variance V of f should be as small as possible. This will reduce the chance that the measurement of f, taken to be our estimate of λ, is very far from the correct value. He proved that V is bounded below by G^{-1}, where G is the variance of $L = \partial_\lambda \log \rho_\lambda$. This has zero mean:

$$\int \rho_\lambda \partial_\lambda \log \rho_\lambda\, dx \quad = \quad \int \partial_\lambda \rho_\lambda\, dx$$

$$
\begin{aligned}
&= \partial_\lambda \int \rho_\lambda \, dx \\
&= \partial_\lambda \, 1 \\
&= 0.
\end{aligned}
$$

Fisher called G the *information* of the family of probabilities $\{\rho_\lambda\}$; he said that an estimate is 100 % efficient if $VG = 1$.

We now prove Fisher's inequality: consider the pair of random variables f, L. Since L has zero mean, the off-diagonal elements of the joint covariance matrix are $\mathbf{E}[Lf]$, which is 1 by the locally unbiased condition, eq.(7). Thus the covariance matrix $\mathcal{C}$ of f, L is

$$
\mathcal{C} = \begin{pmatrix} V & 1 \\ 1 & G. \end{pmatrix}. \tag{8}
$$

Since $\mathcal{C}$ is positive semidefinite, we have that $\det \mathcal{C} \geq 0$, which gives us Fisher's inequality $VG \geq 1$. In general, for a choice of the family $\{\rho_\lambda\}$, no choice of λ will give equality, that is, $VG > \kappa$ for some $\kappa > 1$ will hold for all λ. An exception is the exponential family, $\rho_\lambda(x) = Z^{-1} e^{-x/\lambda}$, $x \geq 0$, provided we choose $f(x) = x$. This is an unbiased estimate, and also gives $VG = 1$.

For the proof, note that with this choice,

$$
Z = \int_0^\infty e^{-x/\lambda} dx = \lambda, \tag{9}
$$

and

$$
\mathbf{E}_\lambda[f] = Z^{-1} \int_0^\infty x e^{-x/\lambda} dx = \lambda. \tag{10}
$$

Thus, f is unbiased. Also, we note that

$$
\mathbf{E}[x^2] = \lambda^{-1} \int_0^\infty x^2 e^{-x/\lambda} dx = 2\lambda^2, \tag{11}
$$

from which we find that $V = \mathbf{E}[x^2] - (\mathbf{E}[x])^2 = \lambda^2$. For the Fisher information $L = \partial_\lambda \log \rho_\lambda$, we find that $L = -1/\lambda + x/\lambda^2$. This gives that

$$
G = \mathcal{V}[L] = \lambda^{-4} \mathcal{V}[x] = \lambda^{-2}; \tag{12}
$$

it follows that $VG = 1$, and the estimate is 100 % efficient. Thus, if we follow Jaynes in maximising the entropy, subject to requiring that the mean value is fixed as equal to the measured value, rather than to the exact mean, we obtain the same answer as by the most efficient Fisher method.

Fisher stated, and Cramér [2] and Rao [12] later proved, that an inequality similar to $VG \geq 1$ holds if the number of parameters is greater than one. Thus, let $\lambda \in \mathbf{R}^N$ be N parameters, and let ρ_λ be a probability density on $\mathbf{R}^N$, and we measure N random variables $X_1, \cdots, X_N$ (on some space) and intend to estimate λ from their measurements x_j. We need N functions $f_1, \cdots, f_N$ of the $X_1, \cdots, X_N$, which do not depend on λ, which are locally unbiased estimators. That is, we require that

$$\mathbf{E}_{\rho_\lambda}[\partial_j \log \rho_\lambda f_k] = \delta_{jk} \tag{13}$$

Again, Fisher defines the information of the family of distributions ρ_λ to be the N random variables, which are scores: thus

$$L_j = \frac{\partial \log \rho_\lambda}{\partial \lambda_j} \tag{14}$$

and again eq.(13) shows us that $\mathbf{E}[L_j f_k] = \delta_{jk}$; more, that this is the covariance of L_j and f_k, since L_j has zero mean. The covariance matrix is

$$\begin{pmatrix} V & 1 \\ 1 & G \end{pmatrix} \tag{15}$$

where $G_{ij} = \mathcal{C}(L_i L_j)$. One can then show [16] that $V \geq G^{-1}$ with the meaning that for $N \times N$ matrices A, B we write $A \geq B$ iff $A - B$ is positive-semi-definite. We say, following Fisher, that a method of estimation for which $VG = 1$ is 100% efficient.

By the same mathematics as for one variable, the exponential family is 100% efficient. Thus, Jaynes's method, modified by using the actual values of the random variables $X_1, \cdots, X_N$ as their means, gives the same answer as the Fisher method. More, we can treat the cases in which some of the variables X_j are equal. For example, we might measure X_1 $r > 1$ times from independent samples, and take the average of the measurements as the best approximation for our estimate of the mean. Instead, we introduce r Lagrange parameters for these variables, and find them by the same equations as if the copies of X_1 were different. We then put all copies of X_1 to be the same after the parameters are found. If we measure some X_j more times than others, then more weight should be given to those random variables with more accurate information about them. The method just described does this.

There is no reason why we must choose the variables $X_1, \cdots, X_N$ to be independent. For example, we might include the cross-moment $\mathbf{E}[X_1 X_2]$ by choosing $X_3 = X_1 X_2$. This can be evaluated if we measure X_1 and X_2 from the same sample. This is not always possible in quantum probability.

3 Quantum estimation theory

Chentsov argued that a measure of distinguishability of two states of a classical system should decrease (or stay the same) if both systems are subject to the same stochastic map. Here, a stochastic map on the Hilbert space of random variables is linear, positive and maps 1 to 1. For, a stochastic map increases our uncertainty of the system, making it more difficult to distinguish the states. He and Morozova proved that the variance is the unique metric (up to a factor) on the space of random variables with this property. For quantum estimation, Chentsov required a good measure of distinguishability to decrease under all completely positive unity-preserving linear maps, which is the quantum version of stochastic maps. However, Petz [8] showed that this condition failed to determine the wanted quantum information uniquely up to a factor. Soon afterwards, Petz [9] found all metrics with the "Chentsov property", but had by then committed himself to choosing one, the BKM-metric [10, 8]. Grasselli and the author [4] showed that this metric is the unique (up to a factor) Euclidean metric on matrix-space, among all metrics with the Chentsov property, such that the dual vector-bundle of states is torsion-free. We adopt this metric here.

We start with a separable Hilbert space $\mathcal{H}$, and a faithful state given by the density operator ρ on $\mathcal{B}(\mathcal{H})$. Since ρ is positive, we can write $\rho = e^{-H}$ for some self-adjoint H, which is bounded below. We require that ρ^β should be of trace-class for some $\beta < 1$. We consider the neighbourhood of ρ of the form of density matrices of the form

$$\rho_\lambda = Z_\lambda^{-1} e^{-H+\lambda X}. \tag{16}$$

Here, X is a closed positive quadratic form, obeying

1. $\operatorname{Dom} H^{\frac{1}{2}} \subseteq \operatorname{Dom} X$
2. $X \leq aH + b$ for some small enough positive a and some real b.

One can then prove that ρ_λ is of trace-class for all $\lambda \in \mathbf{R}$ sufficiently close to zero. We say that as X runs over this set, we get a neighbourhood of ρ_0. We measure operators of the form $H + X_1$ in the state ρ_λ, and seek the best estimate for λX.

This theory is termed *non-parametric estimation*, as we do not prescribe the choice of X beyond being Kato-small relative to H.

Note that the choice of $\rho^{-1}\partial_\lambda \rho$ as a version of Fisher information would not be the best; the fact is, this expression is not hermitian unless the operators ρ and $\partial_\lambda \rho$ commute, so cannot be suitable for the information in

general. Helstrom [5] and Petz and Toth [10] get round this by using the idea of a *logarithmic derivative*. Let g be a real or complex scalar product on the space of $N \times N$ matrices; we say that a matrix L is the g-logarithmic derivative of the family ρ_λ of density matrices if for any matrix F, we have

$$\mathrm{Tr}\{\partial_\lambda \rho_\lambda \, F\} = g(L, F). \tag{17}$$

In these terms, we say that an operator F is an unbiased estimator for λ given the family of density matrices $\{\rho_\lambda\}$, if

$$\mathrm{Tr}(\rho_\lambda \, F) = \lambda. \tag{18}$$

We say that F is locally unbiased if

$$\mathrm{Tr}(\partial_\lambda \rho_\lambda \, F) = 1. \tag{19}$$

We use eq. (17) to get the condition that F is locally unbiased:

$$g(L, F) = 1. \tag{20}$$

This leads to the quantum version of the Fisher inequality:

$$1 = g(L, F) \leq g(L, L)^{1/2} g(F, F)^{1/2}, \tag{21}$$

by the Schwarz inequality. We define the information to be L and use $g(F, F)$ as the uncertainty of the estimator F. The Schwarz inequality becomes an equality only if the two vectors F and L are proportional.

There are many choices of scalar product g that obey the Chentsov property. Helstom used $g(A, B) = \rho_\lambda \left((AB + BA)/2\right)$, but this violates duality [4]. This means that the dual connection involved has torsion. However, it provides the largest value for the quantum Fisher information of a family of estimators involving a single parameter, and has recently been advocated by Petz. Originally, Petz and Toth chose the following, the BKM metric, for scores A and B relative to the density operator ρ:

$$g_\rho(A, B) = \int_0^1 \mathrm{Tr}\left[\rho^\alpha A \rho^{1-\alpha} B\right] d\alpha. \tag{22}$$

This appeared in the theory of Kubo [7] and Mori; it was proved to be positive-definite by Bogoliubov [1], and is, up to a factor, the only one satisfying the Chentsov property which is torsion-free [4]. Thus, with this choice, we see that L_B is defined such that

$$\mathrm{Tr}\left(\partial_\lambda \rho_\lambda X\right) = \int_0^1 \mathrm{Tr}\left(\rho_\lambda^\alpha L_B \rho_\lambda^{1-\alpha}\right) d\alpha. \tag{23}$$

Petz and Toth [10] gives the form for L_B:

$$L_B = \int_0^\infty (\alpha + \rho_\lambda)^{-1} \partial_\lambda \rho_\lambda (\alpha + \rho_\lambda)^{-1} d\alpha. \tag{24}$$

We see that if ρ and $\partial\lambda\rho$ commute, we can do the α-integration to get the Fisher information $\partial_\lambda \log \rho_\lambda$.

We saw above that an estimator in the quantum case is 100% efficient only when F is proportional to L_B. The exponential family of states is the family of the form

$$\rho_\lambda(A) = Z_\lambda^{-1} \exp\{-H_0 + \xi(\lambda)F\}. \tag{25}$$

For this family we see that for any score A we have

$$\mathrm{Tr}\,(\partial_\lambda \rho_\lambda A) = \mathrm{Tr}\left(\int_0^1 \rho_\lambda^\alpha \frac{\partial \xi}{\partial \lambda} F \rho^{1-\alpha} A \, d\alpha \right), \tag{26}$$

and this is equal to $g(L_B, A)$, which is by definition

$$\int_0^1 \mathrm{Tr}\left(\rho^\alpha L_B \rho^{1-\alpha} A \right) d\alpha. \tag{27}$$

Comparing, we see that the exponential family has $L_B = \partial_\lambda \xi F$ on the space of scores; it is the only family with this property, by the definiteness of the scalar product. The value of ξ is found by applying the condition that F be unbiased. This leads to the condition of Jaynes, but with the experimentally found value of F replacing the true mean.

The above theory can be generalised to the estimation of n parameters $\lambda = (\lambda_1, \cdots, \lambda_n)$ which are used to define a family ρ_λ of quantum states; we choose n estimators $F = (F_1, \cdots, F_n)$ which are locally unbiased:

$$\mathrm{Tr}[\partial_{\lambda_i} \rho_\lambda F_j] = \delta_{ij}. \tag{28}$$

We use the idea of logarithmic derivative for several variables:

$$\mathrm{Tr}[\partial_{\lambda_i} \rho_\lambda F_j] = g(L_i, F_j). \tag{29}$$

The family F is thus locally unbiased if

$$g(L_i, F_j) = \delta_{ij}. \tag{30}$$

We adopt the family L_i as the Fisher information for the estimator F_i. Choosing the BKM form for g we get the canonical $2n \times 2n$ covariance matrix, which is positive-definite:

$$\begin{pmatrix} g(L_i, L_j) & g(L_i, F_k) \\ g(F_\ell, L_j) & g(F_\ell, F_k) \end{pmatrix} = \begin{pmatrix} \mathcal{G}_{ij} & I \\ I & \mathcal{V} \end{pmatrix} \geq 0, \tag{31}$$

where $\mathcal{G}$ is the covariance matrix for $L_1, \cdots, L_n$, and $\mathcal{V}$ that of $F_1, \cdots, F_n$. Petz shows that this is positive-definite only if, as matrices, we have the matrix inequality

$$\mathcal{V} \geq \mathcal{G}^{-1}. \tag{32}$$

Petz then shows that equality can only hold when L_i and F_i are proportional, for all i; this is the best choice to make for the estimators. The exponential family, $\rho_\beta = Z_\beta^{-1} \exp\{\sum \beta_j F_j\}$ has this property, and the values of β_j, $j = 1, \cdots, n$ are determined by requiring that the means of F_j should be the values found for them by one measurement. Thus, this is the unique distribution with 100% efficiency in quantum estimation theory, almost as claimed by Jaynes. This result is essentially argued for in [8].

It is not necessary for the variable F_i to commute with the variable F_j, $i \neq j$. We assume that each measurement is made on a sample that is independent of all the other samples. We can choose to observe the same variable, say F_1, a number of times, say n_1 times; we will generally get a slightly different result each time. The best estimate will treat all the variables F_j, $j = 1, \cdots, n_1$ as different, until the end, when we know all the parameters; then we put them all equal to F_1.

4 Non-parametric quantum estimation

In a quantum theory based on a C^*-algebra of infinite dimension there are many pairs of mixed states whose relative entropy is infinite. The trace norm, however, allows two states of small difference in norm to have infinite relative entropy; even so, two such states could not be close to each other in a physical sense. We follow the idea of Pistone and Sempi, and introduce an Orlitz norm, for quantum mechanics in an infinite-dimensional Hilbert space.

We are concerned with measuring some observables, and finding, from the values found, an estimate for the state. We assume that one of the observables to be measured, call it H, is bounded below, and that e^{-H} is of trace-class. By rescaling H with a positive multiple, we see that there is no loss in generality in assuming as well that $e^{-\beta H}$ is also of trace-class for some $\beta < 1$. We define the *Cramér class* of the state $\rho_0 = e^{-H}/\mathrm{Tr}\, e^{-H}$ as follows. Let X be a closed symmetric form that is small relative to H. Then for small enough λ the operator $\exp\{-H - \lambda X\}$ is positive and of trace class. The states, call them $\{\rho_{\lambda X}\}$, obtained from all such X and λ make up the quantum analogue of a *neighbourhood* of ρ_0. We shall call this the *first neighbourhood*. As in the classical case, in which we have the

Cramér class, one can prove that $\rho_{\lambda X}$ is holomorphic in λ for fixed X, in the neighbourhood of $\lambda = 0$. We now define an Orlicz norm on the set of H-small enough quadratic forms which makes it into a convex subset of a real Banach space. We assume that $\mathrm{Tr}\, e^{-H} = 1$.

The quantum version of the Young function $\cosh x - 1$ used by Pistone and Sempi is, we claimed in [13, 15], the function

$$Y(X) = \frac{1}{2}\mathrm{Tr}[\exp(-H + X) + \exp(-H - X)] - 1. \tag{33}$$

This is a function on a neighbourhood of zero in the real vector space of H-bounded forms, and is convex, even, vanishes at $X = 0$, and is positive elsewhere, including possibly being infinite. We define the corresponding Orlicz norm to be

$$\|X\|_Y := \inf\left\{r > 0 : Y\left(\frac{X}{r}\right) < 1\right\}. \tag{34}$$

This is finite for all X which define an element of the first neighbourhood, which is therefore flat, because it has the -1 affine structure: we can define the convex sum of two forms X_1 and X_2 as $\lambda X_1 + (1-\lambda)X_2$ to get another H-small form. Its real linear extension is a Banach space.

We choose any point ρ_1 in the first neighbourhood of ρ_0, and define the first neighbourhood of ρ_1, and the corresponding Orlicz norm for all X in the neighbourhood of ρ_1. On the overlap of the two 'hoods we have defined two norms. We prove in [16] that these topologies are the same: the two Orlicz norms are equivalent. We thus can extend yet again, to define the topology of the set of all states that are linked to ρ_H by a finite number of steps. This set is called the *Cramér class* of ρ_0. We show that the Cramér class is convex: any element in it, say σ, can be formally regarded as equal to $Z_X^{-1}\exp\{-H - X'\}$ for some X'. Indeed, X' is equal to the sum of the steps $X_1, \cdots$ taken to reach σ from ρ_H. This X' turns out to be the same whatever was the path from ρ_H to σ. Thus, the Cramér class is a convex subset of a flat Banach manifold; it is not trivial, however. The coordinates of a neighbourhood are limited to sufficiently small values. For example, in the first neighbourhood of ρ_H, X must be small enough so that $\exp\{-H - X\}$ is of trace class, as well as being such that $H + X$ is self-adjoint. Thus, $X = -H$ is too large.

In quantum nonparametric estimation, we start at one state, say ρ, and measure the values of n bilinear forms $H + X_1, \cdots, X_n$. We seek the best estimate of ρ in the quantum Cramér class of the known state ρ_H. We write

$$\rho_\lambda = Z_{\lambda X}^{-1}\exp\{-H - \lambda X\} \tag{35}$$

for our estimate of the state, where X is H-bounded. We now give details of the case when we observe one form, $X_1 = F$. More, we assume that λ is sufficiently small, so that $\rho_{\lambda X}$ lies in the first neighbourhood of ρ_H. The Fisher information L of the family ρ_λ, using the BKM metric, can be defined as

$$L_{\lambda X} = \int_0^\infty (\alpha + \rho_{\lambda X})^{-1} \partial_\lambda \rho_{\lambda X} (\alpha + \rho_{\lambda X})^{-1} d\alpha. \tag{36}$$

We fix X for now, and consider an estimator, F, for the value of λ. We measure F, say once, and take the value as the experimental mean of F. We consider the family of states ρ_λ. F is, as usual, said to be unbiased if $\mathbf{E}_{\rho_\lambda}[F] = \lambda$, and locally unbiased if $\partial_\lambda \mathbf{E}_{\rho_\lambda}[F] = 1$. The idea of the logarithmic derivative also works for this family if we use the BKM metric. Thus, we put

$$\partial_\lambda \mathrm{Tr}\,[\rho_\lambda F] = 1 = g(L, F), \tag{37}$$

where $g(L, F) = \int_0^1 \rho_\lambda^\alpha L \rho_\lambda^{1-\alpha} F \, d\alpha$ is the BKM scalar product. One can prove that this is well-defined for any X which is H-bounded, provided that λ is small enough. This would not hold if X is unbounded, if we had chosen for example the metric advocated by Helstrom [5]. Again, we get the analogue of the Fisher inequality, $g(F, F)g(L, L) \geq 1$. We get equality if and only if L and F are proportional. Thus, the distribution ρ_λ must be exponential. We then get the value of λ, by requiring that the estimate must be unbiased.

In a similar way, we may deal with n observables. Again, the exponential family gives the best estimate, verifying Jaynes's answer, provided that we use the actual values of the estimators instead of their true means to find the values of the parameters.

This method needs adjusting if the state is outside the first neighbournood of the state ρ_0.

References

[1] Bogoliubov, N, N., Phys. Abh. Soviet Union, **1**, 229-,1962

[2] Cramér, H., **Mathematical Methods of Statistics**, nineteenth printing, Princeton University Press, 1999

[3] Fisher, R. A., *Theory of Statistical Estimation*, Proceedings of the Cambridge Philosophical Society, **22**,700-725, 1925

[4] Grasselli, M. and Streater, R. F., *Uniqueness of the Chentsov metric in quantum information theory*, Infinite Dimensional Analysis, Quantum Probability and Related Fields, **4**, 173-182, 2001

[5] Helstrom, C. W., **Quantum Detection and Estimation Theory**, Academic Press, 1976.

[6] Jaynes, E. T., *Information Theory and Statistical Mechanics*, I, II, Physical Review, **106**, 620-; **108**, 171-190, 1957

[7] Kubo, R., *Statistical mechanical theory of irreversible processes*, Journal of the Physical Society of Japan, **12**, 570-586, 1957

[8] Petz, D., *Geometry of canonical correlation on the state space of a quantum system*, J. Mathematical Phys., **35**, 780-795, 1994

[9] Petz, D., *Monotone metrics on matrix spaces*, Linear Algebra and Applications, **244**, 81-96, 1996

[10] Petz, D. and Toth, G., *The Bogoliubov inner product in quantum statistics*, Letters in Mathematical Physics, **27**, 205-216, 1993

[11] Pistone, G., and Sempi, C., *An infinite-dimensional geometric structure on the space of all probability measures equivalent to a given one, Annals of Statistics*, **33**, 1543-1561, 1995

[12] Rao, C. R., *Information and accuracy attainable in the estimation of statistical parameters*, Bulletin Calcutter Mathematics Society, **37**, 81-91, 1945

[13] Streater, R. F., *The information manifold for relatively bounded potentials*, Proc. Steklov Institute of Mathematics, Moscow, **228**, 603-, 2000

[14] Streater, R. F., *Duality in information geometry*, Open Systems and Information Dynamics, **11**, 71-77, 2004

[15] Streater, R. F., *Quantum Orlicz spaces in information geometry*, Open Systems and Information Dynamics, **11**, 357-375, 2004

[16] Streater, R. F., Statistical Dynamics, second ed., Imperial College Press, 2009, Sect. 15.5