Bayesian predictive density operators for exchangeable quantum-statistical models

Fuyuhiko Tanaka* and Fumiyasu Komaki
Department of Mathematical Informatics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
(Received 30 December 2004; published 20 May 2005)

Quantum state estimation has been widely investigated and there are mainly two approaches proposed: One is based on the point estimation of an unknown parameter and the other is based on the Bayesian method. We adopt the relative entropy from the true state to a predictive density operator as a loss function. We consider exchangeable quantum models with an arbitrary chosen measurement and show that Bayesian predictive density operators are the best predictive density operators when we evaluate them by using the average relative entropy based on a prior. This result is a quantum version of Aitchison's result in classical statistics.

DOI: 10.1103/PhysRevA.71.052323 PACS number(s): 03.67.−a, 03.65.Yz

I. INTRODUCTION

In classical statistics, the problem of predicting an unobserved variable y by using an observed variable x has been investigated. Suppose that a parametric model

$$\mathcal{P} = \{p(y|\theta): \theta \in \Theta\},$$

which is a set of probability densities, is given, where Θ is a parameter space. Random variables x and y are distributed according to the same true probability density $p(\cdot|\theta)$ in $\mathcal{P}$. We predict the unobserved variable y with a predictive density $\hat{p}(y;x)$ constructed by using the observed variable x. The closeness of the true density $p(y|\theta)$ and a predicted density $\hat{p}(y;x)$ is evaluated by using the Kullback-Leibler divergence

$$D(p\|\hat{p}) := \int p(y|\theta)\log\frac{p(y|\theta)}{\hat{p}(y;x)}\mathrm{d}y.$$

Aitchison [1] showed that a Bayesian predictive density $p_\pi(y|x) := \int_\Theta p(y|\theta)\pi(\theta|x)d\theta$, where $\pi(\theta|x)$ is a posterior distribution, is the best predictive density when we evaluate a predictive density $\hat{p}(y;x)$ by using the average Kullback-Leibler divergence $\int\pi(\theta)\int D(p\|\hat{p})p(x|\theta)\mathrm{d}x\mathrm{d}\theta$, where $\pi(\theta)$ is a probability density. Intuitively speaking, if we have some uncertainty on θ, then moderate averaged estimation from the data x is better than one based on a point estimation. We extend this result in classical statistics to the quantum setting.

In quantum statistics, problems of statistical inference and state estimation have received a lot of attention over the past several years with recent developments of experimental techniques. Historically speaking, the parameter estimation problem on quantum systems dates back 30 years, when Helstrom, Holevo, and other researchers vigorously investigated the topic and gave some extension of mathematical statistical concepts on classical probability.

The Bayesian approach for quantum statistics has also been investigated [2,3]. Jones [4] has derived a quantum Bayes rule for pure states with the uniform prior. Later, Bužek *et al.* [5] pointed out that it can be applied to mixed states with a purification ansatz. Schack *et al.* [6] extended their result to a more general framework of exchangeable states. They showed that a quantum state after a measurement can be interpreted as the state averaged over the posterior. Bužek *et al.* [5] recommended using the Bayesian technique, especially when the sample size of the experimental data is small. They proposed using a posterior state corresponding to a posterior distribution in classical counterparts.

From the viewpoints of information quantity and the Bayes rule, however, Bayesian estimation on quantum states has not been fully discussed. Performances of the Bayesian approach compared with other approaches such as the maximal likelihood method have not been discussed theoretically. In the present paper, we show that the Bayesian method has a better performance than the plug-in method when exchangeable states are considered. To our knowledge, our proof has not been given in the general framework. The main result can be regarded as the quantum version of the widely known result by Aitchison in classical statistics.

II. PRELIMINARY

We briefly summarize some notations of quantum measurements. Let $\mathcal{H}$ be a separable (possibly infinite-dimensional) Hilbert space of a quantum system. A Hermitian operator ρ on $\mathcal{H}$ is called a *state* or *density operator* if it satisfies

$$\mathrm{Tr}\rho = 1, \quad \rho \geqslant 0.$$

We denote the set of all states on $\mathcal{H}$ as $\mathcal{S}(\mathcal{H})$.

Let Ω be a space of all possible outcomes of an experiment (e.g., $\Omega = \mathbf{R}^n$) and suppose that a σ algebra $\mathcal{B} := \mathcal{B}(\Omega)$ of subsets of Ω is given. An affine map μ from $\mathcal{S}(\mathcal{H})$ into a set of probability distributions on Ω, $\mathcal{P} = \{\mu(\mathrm{d}x)\}$ is called a *measurement*. There is a one-to-one correspondence between a measurement and a resolution of the identity [3]. A map from $\mathcal{B}$ into the set of positive Hermitian operators

$$E: B \mapsto E(B),$$

where E satisfies

$$E(\phi) = O, \quad E(\Omega) = I, \tag{1}$$

*Electronic address: ftanaka@stat.t.u-tokyo.ac.jp

$$E(\cup_i B_i) = \sum_i E(B_i), \quad B_i \cap B_j = \phi, \quad \forall B_i \in \mathcal{B}, \qquad (2)$$

is called a positive operator valued measure (POVM). Any physical measurement can be represented by a POVM.

The rule describing a post-measurement state is as follows (e.g., Nielsen and Chuang [7]). We consider only discrete outcome cases, where Ω is a countable set. Then, a family of linear operators $\{A_x\}$ satisfying

$$\sum_{x\in\Omega} A_x^* A_x = I$$

describes a measurement when considering $\{E_x = A_x^* A_x\}$ as POVM. Performing such a measurement for an arbitrarily fixed ρ yields an outcome x with probability $p_x := \mathrm{Tr}\rho E_x = \mathrm{Tr}\rho A_x^* A_x$ and the quantum state ρ changes to

$$\frac{A_x \rho A_x^*}{p_x}$$

after the outcome x is observed.

Now we describe our setting of state estimation. Assume that a state ρ_θ on $\mathcal{H}$ is characterized by an unknown finite-dimensional parameter $\theta \in \Theta \subset \mathbf{R}^n$. If $\dim \mathcal{H} < \infty$, θ may cover the full range (often called the full model).

A quantum state for N systems, $\rho^{(N)}$, is described on the N-fold tensor product Hilbert space $\mathcal{H}^{\otimes N}$. Suppose that a system composed of $N+M$ subsystems is given and that a measurement is performed only for selected N subsystems with the other M subsystems left. Then, the measurement is described by $\{A_x \otimes I\}$, where $\{A_x\}$ is a family of linear operators on $\mathcal{H}^{\otimes N}$ such that $\{E_x := A_x^* A_x\}$ is a POVM and I is the identity operator on $\mathcal{H}^{\otimes M}$. Note that in contrast to classical cases, the measurement could affect the remaining M subsystems.

Our aim is to estimate the true state $\sigma_\theta := \rho_\theta^{\otimes M}$ of the remaining M subsystems by using a measurement $\{E_x\}$ on the selected N subsystems $\rho_\theta^{\otimes N}$. We fix an arbitrarily chosen measurement. Note that E is given as a POVM on $\mathcal{H}^{\otimes N}$. It is not necessarily in the form of a tensor product $E_x^{\otimes N}$, which represents a repetition of the same measurement E_x for each system. Thus, all possible measurements on N subsystems, which may use entanglement, are considered.

The performance of a predictive density operator $\hat{\sigma}(x)$ is evaluated by the relative entropy $D(\sigma_\theta \| \hat{\sigma}(x))$, a quantum analogue of the Kullback-Leibler divergence in classical statistics. The quantum relative entropy from ρ to σ is defined by

$$D(\rho \| \sigma) := \mathrm{Tr}[\rho(\log \rho - \log \sigma)]. \qquad (3)$$

It satisfies the positivity condition $D(\rho\|\sigma) \geqslant 0$ and $D(\rho\|\sigma) = 0 \Leftrightarrow \rho = \sigma$. For other properties and useful inequalities, see, e.g., [7]. Thus, it can be used as a measure for the goodness of state estimation.

We can regard a state ρ as a quantum analogue of a probability distribution in classical statistics. Indeed, when $[\rho, \sigma] = 0$, both density operators are simultaneously decomposed as

$$\rho = \sum_x p_x E_x, \quad \sigma = \sum_x q_x E_x,$$

where E_x is a projection operator onto a common eigenspace to two eigenvalues p_x and q_x. (If $\dim \mathcal{H} < \infty$, the formulas above reduce to simultaneous diagonalization of two Hermitian matrices.) Then,

$$\begin{aligned} D(\rho \| \sigma) &= \mathrm{Tr}[\rho(\log \rho - \log \sigma)] \\ &= \sum_x [p_x(\log p_x - \log q_x)] \\ &= D(p \| q). \end{aligned}$$

Thus, the quantum relative entropy (3) is equal to the Kullback-Leibler divergence in classical statistics. Even if $[\rho, \sigma] \neq 0$, it is known that the quantum relative entropy asymptotically (i.e., $N \to \infty$) reduces to the Kullback-Leibler divergence [8].

There are mainly two approaches on inference of state σ_θ for the parametric model above. One approach is to use $\sigma_{\hat{\theta}(x)}$, where $\hat{\theta}(x)$ is an estimator of θ, depending on the observation x. The other approach corresponds to the Bayesian predictive density approach in classical statistics [4,5]. We shall briefly review the idea. First, we assume a probability density $\pi(\theta)$ on the parameter space. In mathematical statistics, $\pi(\theta)$ is usually called a *prior density*. When there is no knowledge about parameter θ, which is often called *noninformative*, several people have discussed what kind of prior should be used [9,10]. From the data x obtained from a measurement, a posterior distribution $\pi(\theta|x)$ is constructed as

$$\pi(\theta|x) := \frac{p(x|\theta)\pi(\theta)}{\int \mathrm{d}\theta p(x|\theta)\pi(\theta)},$$

where $p(x|\theta) = \mathrm{Tr}\sigma_\theta E_x$. Next, taking an average of σ_θ with $\pi(\theta|x)$, one can obtain the Bayesian estimator

$$\sigma|_x = \int \mathrm{d}\theta \sigma_\theta \pi(\theta|x).$$

We call this state estimator, as in classical statistics, a *Bayesian predictive density operator*. In order to distinguish two estimators, we call $\sigma_{\hat{\theta}}$, an estimator based on $\hat{\theta}$, a *plug-in predictive density operator*. In the next section, we show that Bayesian predictive density operators are better than plug-in predictive density operators.

If we assume a prior probability density $\pi(\theta)$ on the parameter space Θ, the mixture state is given by

$$\rho^{(N)} := \int \mathrm{d}\theta \pi(\theta) \rho_\theta^{\otimes N}. \qquad (4)$$

A state of the form (4) is called an *exchangeable state* [6], and arises, e.g., if each subsystem is prepared in the same unknown way, as in quantum state tomography.

In a quantum exchangeable model (4), as Schack *et al.* [6] showed, a posterior distribution $\pi(\theta|x)$ naturally arises. As described above, a post-measurement state with outcome x obtained is given by

$$\rho_x^{(N+M)} = \frac{1}{p_x}\left[(A_x \otimes I)\left(\int \mathrm{d}\theta \pi(\theta) \rho_\theta^{\otimes(N+M)} \right)(A_x^* \otimes I) \right].$$

After the measurement of the selected N subsystems, we restrict our attention only to the remaining M subsystems. Taking a partial trace, we obtain the resulting state ρ_x^M on $\mathcal{H}^{\otimes M}$ (for partial trace, see, e.g., [7]).

The final state ρ_x^M can be rewritten using a posterior $\pi(\theta|x)$ in the form of an exchangeable model [6],

$$\begin{aligned}
\rho_x^M &= \mathrm{Tr}_N[\rho_x^{N+M}] \\
&= \frac{1}{p_x}\mathrm{Tr}_N\left[(A_x \otimes I) \int \mathrm{d}\theta \pi(\theta) \rho_\theta^{\otimes(N+M)} (A_x^* \otimes I) \right] \\
&= \frac{1}{p_x}\mathrm{Tr}_N\left[\int \mathrm{d}\theta \pi(\theta) \rho_\theta^{\otimes(N+M)} E_x \otimes I \right] \\
&= \frac{1}{p_x} \int \mathrm{d}\theta \pi(\theta) \{ \mathrm{Tr}_N[\rho_\theta^{\otimes(N+M)} E_x \otimes I] \}.
\end{aligned}$$

Since $\mathrm{Tr}_N[(A \otimes B)(C \otimes D)] = \mathrm{Tr}[AC]BD$ holds for A, C on $\mathcal{H}^{\otimes N}$ and B, D on $\mathcal{H}^{\otimes M}$, the partial trace above is rewritten as

$$\mathrm{Tr}_N[\rho_\theta^{\otimes(N+M)}(E_x \otimes I)] = \mathrm{Tr}_N[\rho_\theta^{\otimes N} E_x] \rho_\theta^{\otimes M} = p(x|\theta) \rho_\theta^{\otimes M}.$$

The two probability densities p_x and $p(x|\theta)$ are related by

$$\begin{aligned}
p_x &= \mathrm{Tr}[\rho^{(N+M)}(E_x \otimes I)] \\
&= \mathrm{Tr}\left[\int \mathrm{d}\theta \pi(\theta) \rho_\theta^{\otimes(N+M)} E_x \otimes I \right] \\
&= \int \mathrm{d}\theta \pi(\theta) \mathrm{Tr}_N[\rho_\theta^{\otimes N} E_x] \mathrm{Tr}_M[\rho_\theta^{\otimes M} I] \\
&= \int \mathrm{d}\theta \pi(\theta) p(x|\theta).
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
\rho_x^M &= \frac{1}{p_x} \int \mathrm{d}\theta \pi(\theta) \{ p(x|\theta) \rho_\theta^{\otimes M} \} \\
&= \int \mathrm{d}\theta \frac{p(x|\theta)\pi(\theta)}{p_x} \rho_\theta^{\otimes M} \\
&= \int \mathrm{d}\theta \pi(\theta|x) \rho_\theta^{\otimes M}.
\end{aligned}$$

Thus, one can interpret $\pi(\theta|x)$ as a quantum analogue of the posterior distribution in classical statistics.

Now we consider comparing two methods for estimating the true state $\sigma_\theta \in \mathcal{S}(\mathcal{H}^{\otimes M})$. Let $\hat{\sigma}(x)$ and $\tilde{\sigma}(x)$ be two predictive density operators. When the difference between two estimates $\hat{\sigma}(x)$ and $\tilde{\sigma}(x) \in \mathcal{S}(\mathcal{H}^{\otimes M})$,

$$D(\sigma_\theta \| \hat{\sigma}(x)) - D(\sigma_\theta \| \tilde{\sigma}(x)) = \mathrm{Tr}\{\sigma_\theta [\log \tilde{\sigma}(x) - \log \hat{\sigma}(x)]\}, \tag{5}$$

is positive, $\hat{\sigma}(x)$ is better than $\tilde{\sigma}(x)$ as an estimate of the true state σ_θ. Since $\hat{\sigma}(x)$ and $\tilde{\sigma}(x)$ depend on observed data x for an arbitrarily chosen measurement $\{E_x\}$ on $\mathcal{H}^{\otimes N}$, the difference (5) depends on the true parameter value θ characterizing the true state and on the data x obtained from the measurement. Thus, we take an average of Eq. (5) over $p(x|\theta) := \mathrm{Tr}\sigma_\theta E_x$ and $\pi(\theta)$, and evaluate

$$E_\theta E_x \{ D(\sigma_\theta \| \hat{\sigma}(x)) - D(\sigma_\theta \| \tilde{\sigma}(x)) \}$$

in the following in order to compare plug-in predictive density operators with Bayesian predictive density operators.

III. MAIN THEOREM

In classical statistics, Aitchison [1] showed that the Bayesian predictive density $p_\pi(y|x)$ has better performance under the Kullback-Leibler divergence than any plug-in predictive density $p(y|\hat{\theta})$ when a proper prior $\pi(\theta)$ is given. We derive the corresponding result for quantum predictive density operators.

Theorem. Suppose that we perform a measurement for selected N subsystems $\rho_\theta^{\otimes N}$ of a system $\rho_\theta^{\otimes(N+M)}$ composed of $N+M$ subsystems in order to estimate the remaining M subsystems $\sigma_\theta = \rho_\theta^{\otimes M}$. The true parameter value θ is unknown and a prior probability density $\pi(\theta)$ is assumed. Let $\hat{\sigma}(x)$ be any predictive density operator, where x is an outcome of a measurement $\{E_x\}$ for the N subsystems. Performance of a predictive density operator $\hat{\sigma}(x)$ is measured with the average relative entropy

$$E_\theta E_x \{ D(\sigma_\theta \| \hat{\sigma}(x)) \} = \int \mathrm{d}\theta \pi(\theta) \int \mathrm{d}x p(x|\theta) D(\sigma_\theta \| \hat{\sigma}(x))$$

from the true state σ_θ. Then, the Bayesian predictive density operator $\sigma|_x$ based on the observation x and the prior $\pi(\theta)$ is the best predictive density operator.

Proof. First, for arbitrary $\hat{\sigma}(x)$ and $\tilde{\sigma}(x)$, we rewrite the difference of two averaged Kullback-Leibler divergences as

$$\begin{aligned}
&E_\theta E_x \{ D(\sigma_\theta \| \hat{\sigma}(x)) - D(\sigma_\theta \| \tilde{\sigma}(x)) \} \\
&= \int \mathrm{d}\theta \pi(\theta) \int \mathrm{d}x p(x|\theta) \mathrm{Tr}\{\sigma_\theta [\log \tilde{\sigma}(x) - \log \hat{\sigma}(x)]\} \\
&= \int \mathrm{d}\theta \int \mathrm{d}x p_x \frac{p(x|\theta)\pi(\theta)}{p_x} \mathrm{Tr}\{\sigma_\theta [\log \tilde{\sigma}(x) - \log \hat{\sigma}(x)]\} \\
&= \int \mathrm{d}x p_x \int \mathrm{d}\theta \pi(\theta|x) \mathrm{Tr}\{\sigma_\theta [\log \tilde{\sigma}(x) - \log \hat{\sigma}(x)]\} \\
&= \int \mathrm{d}x p_x \mathrm{Tr}\left[\left\{ \int \mathrm{d}\theta \pi(\theta|x) \sigma_\theta \right\} [\log \tilde{\sigma}(x) - \log \hat{\sigma}(x)] \right] \\
&= \int \mathrm{d}x p_x \mathrm{Tr}\{ \sigma|_x [\log \tilde{\sigma}(x) - \log \hat{\sigma}(x)] \}.
\end{aligned}$$

The positivity of the above form indicates that $\tilde{\sigma}(x)$ is better than $\hat{\sigma}(x)$. We set

$$\tilde{\sigma}(x) = \sigma|_x,$$

and then we obtain

$$E_\theta E_x\{D(\sigma_\theta \| \hat{\sigma}(x)) - D(\sigma_\theta \| \sigma|_x)\}$$
$$= \int \mathrm{d}x p_x \mathrm{Tr}\{\sigma|_x[\log \sigma|_x - \log \hat{\sigma}(x)]\}$$
$$= \int \mathrm{d}x p_x D(\sigma|_x \| \hat{\sigma}(x)) \geqslant 0.$$

The last inequality holds due to the positivity of the relative entropy $D(\sigma\|\sigma') \geqslant 0$ and $p_x \geqslant 0$. Since $\hat{\sigma}(x)$ is arbitrarily chosen, it is shown that $\sigma|_x$ is better than any other $\hat{\sigma}(x)$.

IV. REMARKS

Our argument is valid even when $\dim \mathcal{H} = \infty$ if we impose some regularity conditions. For example, there are some conditions such as the exchangeability of the order of Tr and $\int \mathrm{d}\theta\, \pi(\theta)$, "measurability" of ρ_θ, and integrability of $\rho|_x = \int d\theta\, \pi(\theta|x)\rho_\theta$. Such rigorous arguments may require some mathematics, say, the theory of the Bochner integral.

Note that our argument also holds when a prepared state is described as $\rho_\theta \otimes \sigma_\theta$, where $\rho_\theta \in \mathcal{S}(\mathcal{H}^{(1)})$ and $\sigma_\theta \in \mathcal{S}(\mathcal{H}^{(2)})$, and $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ are distinct Hilbert spaces. This setting is a generalization of that introduced in Sec. II.

ACKNOWLEDGMENTS

F.T. was supported by JSPS.

[1] J. Aitchison, Biometrika **62**, 547 (1975).
[2] C. W. Helstrom, *Quantum Detection Theory* (Academic Press, New York, 1976).
[3] S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory* (North-Holland, Amsterdam, 1982).
[4] K. R. W. Jones, Ann. Phys. (N.Y.) **207**, 140 (1991).
[5] V. Bužek, R. Derka, G. Adam, and P. L. Knight, Ann. Phys. (N.Y.) **266**, 454 (1998).
[6] R. Schack, T. A. Brun, and C. M. Caves, Phys. Rev. A **64**, 014305 (2001).
[7] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2000).
[8] F. Hiai and D. Petz, Commun. Math. Phys. **143**, 99 (1991).
[9] P. Slater, J. Math. Phys. **38**, 2274 (1997).
[10] S. L. Braunstein and C. M. Caves, Phys. Rev. Lett. **72**, 3439 (1994).