Characterization of the Bayes Estimator and the MDL Estimator for Exponential Families

Jun-ichi Takeuchi

***Abstract*— We analyze the relationship between a Minimum Description Length (MDL) estimator (posterior mode) and a Bayes estimator for exponential families. We show the following results concerning these estimators: a) Both the Bayes estimator with Jeffreys prior and the MDL estimator with the uniform prior with respect to the expectation parameter are nearly equivalent to a bias-corrected maximum-likelihood estimator with respect to the canonical parameter. b) Both the Bayes estimator with the uniform prior with respect to the canonical parameter and the MDL estimator with Jeffreys prior are nearly equivalent to the maximum-likelihood estimator (MLE), which is unbiased with respect to the expectation parameter. These results together suggest a striking symmetry between the two estimators, since the canonical and the expectation parameters of an exponential family form a dual pair from the point of view of information geometry. Moreover, a) implies that we can approximate a Bayes estimator with Jeffreys prior simply by deriving an appropriate MDL estimator or an appropriate bias-corrected MLE. This is important because a Bayes mixture density with Jeffreys prior is known to be maximin in universal coding [7].**

***Index Terms*—Bayes estimator, exponential family, higher order asymptotic theory, information geometry, Minimum Description Length principle, universal source coding.**

I. Introduction

In this study, we examine the estimation of parameters of probability densities in the general class of regular exponential families [5]. In particular, we analyze the relationship between the following two estimators and reveal a symmetry between the two: a Minimum Description Length (MDL) estimator [12] (also called a Minimum Message Length estimator [15], [16], which has a posterior mode interpretation) and the estimator which is Bayes with respect to Kullback–Leibler divergence (KL divergence for short) between the parameterized densities (it can be obtained by projecting the Bayes posterior mixture density onto the original exponential family).

In the field of universal source coding, Bayes mixture densities have recently become a popular subject of study for two main reasons: 1) Bayes decision theory can be used to determine the code which will achieve the minimax redundancy [7]–[9], [11], and 2) codes based on a Bayes mixture will be superior to two-step codes [13]. The MDL estimator we study here is based on a two-step code (This is one form of Rissanen's MDL principle [12], which more generally also encompasses mixture-based codes).

In two-step coding, we first use observed data to estimate a source distribution, which we then encode. Next, on the basis of the estimate, we encode the observed data. In the MDL principle, an optimal estimate is defined as that which gives the shortest total code length for given data, and we refer to estimators based on the MDL principle as MDL estimators. MDL estimators have been shown to be especially effective when the complexity (i.e., number of parameters) of the actual source distribution is unknown [2], [3], [18], [19]. (Here, however, we will focus on the case of parameter estimation in a fixed family.)

By way of contrast, Bayes codes do not encode the source distribution and are a fundamentally different from the two-step type. Consider a parametric family of probability densities, $S = \{p(x|\theta)|\theta \in \Theta\}$ on a space $\mathcal{X}$, and a predictive density estimator $f(x_{t+1}|x^t)$ for S (i.e., $\sum_{x\in\mathcal{X}} f(x|x^t) = 1$). The cumulative risk with KL divergence is defined as

$$R_N(\theta, f) \equiv \sum_{t=1}^{N} \int p(x^t|\theta) \ln(p(x_t|\theta)/f(x_t|x^{t-1}))\, dx^t$$

where "ln" denotes the natural logarithm,

$$p(x^t|\theta) = \prod_{u=1}^{t} p(x_u|\theta)$$

and θ is the actual parameter. This cumulative risk is also the redundancy of the code based on the joint density

$$f(x^N) = \prod_{t=0}^{N-1} f(x_{t+1}|x^t)$$

(see [6]). Next, the Bayes risk (or Bayes redundancy) of f with respect to the prior $w(\theta)d\theta$ is defined as

$$R_N(w, f) \equiv \int R_N(\theta, f) w(\theta)\, d\theta.$$

The Bayes predictor with this prior uses

$$f_w(x_{t+1}|x^t) \equiv \int p(x|\theta) w(\theta|x^N)\, d\theta$$

which achieves $\min_f R_N(w, f)$ (where $w(\theta|x^N)$ is the posterior density).

The Bayes predictor with the *Jeffreys prior* [10] (denoted by w_J) is of special importance among the Bayes procedures, since f_{w_J} asymptotically maximin for the redundancy

Manuscript received July 27, 1995; revised October 12, 1996. The material in this paper was presented at the IEEE International Symposium on Information Theory, Whistler, BC, Canada, September 18–22, 1995.

The author is with the Theory NEC Laboratory, RWCP (Real World Computing Partnership) c/o C&C Research Laboratories, NEC Corporation. 4-1-1 Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216, Japan.

Publisher Item Identifier S 0018-9448(97)03780-2.

$R_N(\theta, f)$ and modifications of it are asymptotically minimax under certain conditions [7], [17], [4].

For Bernoulli sources ($\mathcal{X} = \{0, 1\}$, $p(1|r) = r$), we have

$$f_{w_J}(1|x^N) = (k + 0.5)/(N + 1)$$

(k is the number of occurrences of "1" in N trials), i.e., a classic Laplace estimator is maximin for the Bayes redundancy. Moreover, this estimator can be derived from the MDL principle combined with the assumption that we either use the binary digit expansion of the value of r itself as the code for the source distribution or assume that the prior is uniform over the range of r [20], [16]. Bayes procedures for coding (unlike the two-step code) necessarily involve mixtures and the corresponding predictive density estimates suffer from the problem that, in general, they do not belong to the class of the original source. When an estimate belongs to the original class (i.e., densities estimated by plugging in parameter estimators), we say that the estimator is "proper." Bayes predictors are not always proper, while MDL estimators are: An MDL estimate (based on a two-step code) always belongs to the original class.

Inspired by the circumstances mentioned above, we analyze the relationship between f_w and the MDL estimator for general exponential families. Since we do not have the means to analyze the relationship between them directly, we consider the proper estimator f minimizing $R_N(w, f)$. We let $\tilde{f}_w(\cdot|x^N) = p(\cdot|\hat{\theta}_w)$ denote such an estimator. Note that $\hat{\theta}_w$ is equal to

$$\arg\min_{\theta'} \int D(p(\cdot|\theta)||p(\cdot|\theta'))w(\theta|x^N)d\theta.$$

(We let $D(p||q)$ denote KL divergence of q with respect to p.) By simple manipulation, we can see $\tilde{f}_w(\cdot|x^N) = p(\cdot|\hat{\theta}_w)$ can be obtained by projecting $f_w(\cdot|x^N)$ by KL divergence to the original exponential family. Hence, we refer to $\tilde{f}_w$ as the projected Bayes estimator. In information geometry [1] it is known that the -1-geodesic connecting $\tilde{f}_w(\cdot|x^N)$ and $f_w(\cdot|x^N)$ is orthogonal to the family S and $\tilde{f}_w(\cdot|x^N)$ is referred to as the -1-projection of $f_w(\cdot|x^N)$ onto S.

In reference to the MDL estimator, in order to specify an estimator we have to specify a coding scheme for source distributions. Specifying a coding scheme is equivalent to specifying a discreet prior distribution, often obtained by discretizing a continuous distribution w for the parameters. As discussed in Section V, the MDL estimates take the form $(*)$

$$\hat{\theta}_{mdl} = \arg\min_{\theta}\big(-\ln p(x^N|\theta) - \ln(w(\theta)/\sqrt{I(\theta)})\big)$$

where $I(\theta)$ is the determinant of Fisher information matrix (see also [15], [16], [2]). We refer to the estimator specified in such a way as the MDL estimator with the prior w. When w is uniform over the coordinates, e.g., ξ, we say that the MDL estimator is taken with respect to the coordinate system ξ, since using that prior is equivalent to using the decimal expansion of ξ's value for coding the distribution. (A strict coding interpretation would require a proper prior that integrals (sums) to one over all θ. Here we will use $(*)$ also for improper priors such as the uniform over $\Re$.)

We have obtained the following result: Under a certain weak condition, "the MDL estimator with respect to the expectation parameter (as defined in Section II)" coincides with $\tilde{f}_{w_J}$ up to the $O(1/N)$ term. This is the generalization of the equivalence between the two estimators for Bernoulli sources, because the parameter $r = p(1|r)$ is the expectation parameter in that case. We also have shown that these estimators coincide with the bias-corrected maximum likelihood estimator (bias-corrected MLE, for short) [1] with respect to the *canonical parameter* (as defined in Section II) up to the $O(1/N)$ term. These results not only supply an easy way to approximate the projected Bayes estimator, which is hard to calculate strictly, but characterize the maximin estimator on the basis of information geometry. That is, $\tilde{f}_{w_J}$ is nearly unbiased with respect to the canonical parameters. Moreover, we show that both the projected Bayes estimator with the uniform prior over the canonical parameter and the MDL estimator with Jeffreys prior equal the MLE ignoring terms of order $o(1/N)$. Noting that the MLE is unbiased with respect to the expectation parameter, these results throw light on the symmetry between the projected Bayes estimator and the MDL estimator, because the natural and the expectation parameters of an exponential family form a dual pair from the point of view of information geometry [1].

II. Preliminaries

The exponential family is defined as follows [5], [1].

Definition 1: Let ν be a σ-finite measure on the Borel subsets of $\Re^n$ and $\mathcal{X}$ be the support of ν. Define

$$\Theta_a \equiv \big\{\theta|\theta \in \Re^n, \int_{\mathcal{X}} \exp(\theta^i x_i)\nu(dx) < \infty\big\}.$$

Let Θ denote a subset of Θ_a. Define a function ψ and a probability density p on $\mathcal{X}$ with respect to ν by

$$\psi(\theta) \equiv \ln \int_{\mathcal{X}} \exp(\theta^i x_i)\nu(dx)$$

and

$$p(x|\theta) \equiv \exp\big(\theta^i x_i - \psi(\theta)\big).$$

We refer to the set $S(\Theta) \equiv \{p(x|\theta)|\theta \in \Theta\}$ as an exponential family.

In the above definition and hereafter, we use Einstein's convention about summation, i.e., $\theta^i x_i$ denotes $\sum_{i=1}^n \theta^i x_i$ (x_i denotes the ith component of x). Exponential families include many common statistical models such as Gauss distributions, Poisson distributions, Bernoulli sources, etc., where the role of x is played by a suitable function of the original variables in these cases [5]. It is known that Θ_a is a convex set. Let $\mathcal{W}$ denote the closure of the convex hull of $\mathcal{X}$. It is known that we can assume that $\dim \mathcal{W} = \dim \Theta_a = n$ holds without loss of generality [5]. An exponential family which satisfies this condition is said to be minimal. We assume $S(\Theta_a)$ is minimal in this paper.

It is known that $\psi(\theta)$ is of class C^∞ and strictly convex on Θ_a° (where A° denotes the interior of $A \subseteq \Re^n$). We refer to θ as the canonical parameter (or θ-coordinates). We define the expectation parameter (or η-coordinates) as $\eta_i \equiv E_\theta(x_i)$, where E_θ denotes the expectation with respect to $p(x|\theta)$. It is known that the function on Θ_a° mapping $\theta \mapsto \eta$ is an injection and of class C^∞. Let $H_a = \{\eta(\theta)|\theta \in \Theta_a^\circ\}$ be the range of this map. The parameters θ and η have the

geometrical interpretation as follows: θ-coordinates are the affine coordinates for the 1-connection and η-coordinates is the affine coordinates for the -1-connection (see [1]).

We let ∂_i and ∂^i, respectively, denote the differential operators $\partial/\partial\theta^i$ and $\partial/\partial\eta_i$. Note that $\partial_i\psi=\eta_i$, $\partial_i\partial_j\psi=g_{ij}$, and $\partial_i\partial_j\partial_k\psi=T_{ijk}$ hold on Θ_a°, where

$$g_{ij}=E_\theta((x_i-\eta_i)(x_j-\eta_j))$$

and

$$T_{ijk}=E_\theta((x_i-\eta_i)(x_j-\eta_j)(x_k-\eta_k)).$$

Further, g_{ij} is the Fisher information matrix with respect to θ. We let g^{ij} denote the Fisher information matrix with respect to η. Then, $g_{ij}g^{jk}=\delta_i^k$ holds, i.e., g^{ij} is the inverse matrix of g_{ij}. (δ_i^k denotes the Kronecker's delta.) Finally, we note that the following holds:

$$\partial_i \ln(\det|g_{ij}|)^{1/2}=T_{ijk}g^{jk}/2. \tag{1}$$

In this paper, we refer to a function f which maps

$$\bigcup_{i=0,1,\cdots}\mathcal{X}^i$$

to $\mathcal{H}$ (any set of probability distributions, referred to as a hypothesis set) as an estimator. Let f be an estimator. We let $f[x^N]$ denote the image of x^N by f and $f[x^N](x)$ (or $f(x|x^N)$) denote the density of $f[x^N]$ at x. In particular, when $\mathcal{H}\subseteq S(\Theta_a)$, f is said to be proper.

We let $\hat{f}$ denote the MLE, i.e.,

$$\hat{f}[x^N]=\arg\max_{p\in S(\Theta_a)}p(x^N).$$

We let $\hat{\eta}$ and $\hat{\theta}$, respectively, denote $\eta(\hat{f}[x^N])$ and $\theta(\hat{f}[x^N])$ for simplicity. Moreover, let $\hat{g}_{in}$, $\hat{T}_{ijk},\cdots$, respectively, denote their values at $\hat{f}[x^N]$. We let

$$\bar{x}=\sum_{t=1}^N x_t/N$$

where x_t denotes the tth observed value of x (not the tth component). This $\bar{x}$ is a sufficient statistic for Θ. It is known that if $\bar{x}\in H_a$ holds, then the MLE $\hat{\eta}$ equals $\bar{x}$. Note that if an exponential family $S(\Theta_a)$ is "steep," then $H_a=\mathcal{W}^\circ$ holds. (If $E_\theta(|x|)=\infty$ holds for any $\theta\in\Theta_a-\Theta_a^\circ$, then $S(\Theta_a)$ is said to be steep, see for example [5].) When $S(\Theta_a)$ is steep, by the definition of $\mathcal{W}$, $\bar{x}\in\mathcal{W}$, $=\bar{H}_a$ holds (where $\bar{A}$ denotes the closure of $A\subseteq\Re^n$). Strictly speaking, there does not exist $\hat{\eta}$ for $\bar{x}\in\partial H_a$, however, we define $\hat{\eta}\equiv\bar{x}$ for $\bar{x}\in\bar{H}_a$ for minimal steep exponential families (where ∂A denotes the boundary of $A\subseteq\Re^n$).

Now we define prior distributions. Let Θ ($\subseteq\Theta_a$) be a connected open set of $\Re^n$ and $w(\theta)d\theta$ a prior distribution whose support is $\bar{\Theta}$. In the sequel, when we simply say "the prior w," it is supposed that w is defined with respect to the measure $d\theta$. We assume that w is of class C^2 on Θ and that $\int_\Theta w(\theta)d\theta$ does not necessarily equal 1. Moreover, we permit the case that $\int_\Theta w(\theta)d\theta=\infty$ (improper prior). We define Jeffreys prior as $w_J(\theta)\equiv(\det|g_{ij}(\theta)|)^{1/2}$.

III. The Bayes Estimator

We define the Bayes mixture estimator (Bayes predictor) with the prior w as

$$f_w[x^N](x)\equiv\int_\Theta p(x|\theta)w(\theta|x^N)d\theta$$

where $w(\theta|x^N)$ is defined as

$$p(x^N|\theta)w(\theta)/\int p(x^N|\theta)w(\theta)d\theta$$

provided

$$\int p(x^N|\theta)w(\theta)d\theta<\infty.$$

We define the projection of f_w to $S(\Theta_a)$ (denoted by $\tilde{f}_w$) as follows, where $D(p||q)$ is defined as

$$D(p||q)=\int p(x)\ln(p(x)/q(x))\nu(dx)$$

(KL divergence):

$$\tilde{f}_w[x^N]\equiv\arg\min_{p\in S(\Theta_a)}D(f_w[x^N]||p).$$

We refer to $\tilde{f}_w$ as the projected Bayes estimator with the prior w. As we mentioned in the Introduction, $\tilde{f}_w$ equals the proper estimator f minimizing $R_N(w,f)$.

The following Proposition holds.

Proposition 1: Suppose

$$\int_\Theta p(x^N|\theta)w(\theta)d\theta<\infty.$$

Define

$$\tilde{\eta}\equiv\int_\Theta\eta(\theta)w(\theta|x^N)d\theta.$$

If $\tilde{\eta}$ is finite and belongs to H_a, then $\eta_i(\tilde{f}_w[x^N])=\tilde{\eta}_i$ holds.

Proof: Noting $p(x|\theta)=\exp(\theta^i x_i-\psi(\theta))$, we have

$$\begin{aligned}
&\int_{x\in\mathcal{X}}f_w[x^N](x)\ln p(x|\theta)\nu(dx)\\
&\quad=\int_{x\in\mathcal{X}}\int_{\xi\in\Theta}p(x|\xi)w(\xi|x^N)d\xi\cdot(\theta^i x_i-\psi(\theta))\nu(dx)\\
&\quad=\int_{\xi\in\Theta}\int_{x\in\mathcal{X}}p(x|\xi)(\theta^i x_i-\psi(\theta))\nu(dx)\cdot w(\xi|x^N)d\xi\\
&\quad=\int_{\xi\in\Theta}(\theta^i\eta_i(\xi)-\psi(\theta))w(\xi|x^N)d\xi\\
&\quad=\theta^i\int_{\xi\in\Theta}\eta_i(\xi)w(\xi|x^N)d\xi-\psi(\theta)\\
&\quad=\theta^i\cdot\tilde{\eta}_i-\psi(\theta).
\end{aligned}\tag{2}$$

Note that the estimate $\tilde{f}_w[x^N]$ maximizes

$$\int_{\mathcal{X}}f_w[x^N](x)\ln p(x)\nu(dx)$$

among p in S. Since $\theta(\tilde{\eta})$ belongs to Θ_a by the assumption, θ which maximizes (2) is $\theta(\tilde{\eta})$, i.e., we obtain the proposition. Q.E.D.

Remark: If $S(\Theta_a)$ is steep, then $\tilde{\eta} \in H_a$ always holds whenever $\tilde{\eta}$ is finite.

We let H_{in} denote a fixed compact set included in the interior of H_a and x^∞ an infinite sequence $x_1 x_2 \cdots$. Let x^N denote the sequence which consists of the first N elements of x^∞. We define a class of x^∞, T_N as

$$T_N = \left\{ x^\infty \mid \forall I \geq N, \sum_{t=1}^{I} x_t / I \in H_{in} \right\}.$$

We define a family of real-valued function F_α on $\mathcal{W} \times \Theta$ as

$$F_\alpha(z, \theta) = \exp(\alpha(\theta^i z_i - \psi(\theta)))$$

where $\alpha \in \Re$. We can write $p(x^N|\theta) = F_N(\bar{x}, \theta)$.

We let $\Theta/\{\theta^i\}$ denote a set

$$\{(\theta^1, \cdots, \theta^{i-1}, \theta^{i+1}, \cdots, \theta^n) | \theta \in \Theta\}$$

and for

$$\zeta(i) = (\zeta^1, \cdots, \zeta^{i-1}, \zeta^{i+1}, \cdots, \zeta^n) \in \Theta/\{\theta^i\}$$

let $l_{\zeta(i)}$ denote a line in $\Re^n$ which is parallel to the ith axis and goes through the point $(\zeta^1, \cdots, \zeta^{i-1}, 0, \zeta^{i+1}, \cdots, \zeta^n)$.

Now, we make the following assumptions:

Assumption 1: For any $i \in \{1, \cdots, n\}$ and almost all $\zeta(i) \in \Theta/\{\theta^i\}$, $l_{\zeta(i)} \cap \partial\Theta$ is a finite set.

Assumption 2: When N is not less than a certain integer N_1, for any $z \in H_{in}$, $F_N(z, \theta)w(\theta)$ and $\eta(\theta)F_N(z, \theta)w(\theta)$ are integrable on Θ.

Assumption 3: When N is not less than a certain integer N_2, for any $z \in H_{in}$, $(\partial_i w(\theta))F_N(z, \theta)$ is integrable on Θ.

For example, if Θ is a convex set of $\Re^n$, then Assumption 1 holds. It also holds, if Θ can be decomposed to finite number of n-dimensional convex sets, i.e., $\bar{\Theta} = \bigcup_{1 \leq i \leq r} \bar{\Theta}_i$ and $\Theta_i \cap \Theta_j = \emptyset$ $(i \neq j)$ hold, where r is a certain positive integer and each Θ_i is an n-dimensional open convex set.

Assumption 2 and 3 can be checked by using the following proposition.

Proposition 2: Let ξ be a real-valued function on Θ. For all sufficiently large N, $\forall z \in H_{in}$

$$\int_\Theta |\xi(\theta) F_N(z, \theta)| d\theta < \infty$$

holds, if and only if $\exists z \in H_{in}$

$$\int_\Theta |\xi(\theta) F_N(z, \theta)| d\theta < \infty$$

holds for a certain real N.

The proof is given in the Appendix.

Concerning the projected Bayes estimator, we obtain the following theorem.

Theorem 1: Suppose Assumptions 1–3 hold for an exponential family $S(\Theta)$ and a prior w. For any $N' \in \mathbb{N}$

$$\eta_i(\tilde{f}_w[x^N]) = \hat{\eta}_i + \partial_i \ln w(\hat{\theta})/N + O(\sqrt{\ln N}/N\sqrt{N})$$

holds unformly for $x^\infty \in T_{N'}$. Especially when $w(\theta)$ is uniform over Θ

$$\eta_i(\tilde{f}_w[x^N]) = \hat{\eta}_i + O(e^{-CN})$$

holds.

Remark: If x^∞ is a sample of $p^* \in S(\Theta^\circ_{in})$, then we can show

$$\lim_{N \to \infty} \Pr\{x^\infty \in T_N\} = 1.$$

We give the proof in the next session.

We have the following corollary.

Corollary 1: Suppose Assumptions 1–3 hold for an exponential family $S(\Theta)$ and Jeffreys prior. For any $N' \in \mathbb{N}$

$$\eta_i(f_{w_J}[x^N]) = \hat{\eta}_i + \hat{T}_{ijk}\hat{g}^{jk}/2N + O(\sqrt{\ln N}/N\sqrt{N})$$

holds, uniformly for $x^\infty \in T_{N'}$.

Proof: Let $w(\theta) = w_J(\theta)$ in Theorem 1. Using (1), we obtain the claim of the corollary. Q.E.D.

IV. Proof of Theorem 1

By differentiating

$$F_N(\bar{x}, \theta)w(\theta) = \exp(N(\theta^i \bar{x}_i - \psi(\theta)))w(\theta)$$

with respect to θ^k, we have

$$\begin{aligned}
&\partial_k \big(\exp(N(\theta^i \bar{x}_i - \psi(\theta)))w(\theta)\big) \\
&\quad = N(\bar{x}_k - \eta_k(\theta)) \exp(N(\theta^i \bar{x}_i - \psi(\theta)))w(\theta) \\
&\qquad + \exp(N(\theta^i \bar{x}_i - \psi(\theta)))\partial_k \ln w(\theta) \cdot w(\theta) \\
&\quad = \big(N(\bar{x}_k - \eta_k(\theta)) + \partial_k \ln w(\theta)\big) F_N(\bar{x}, \theta)w(\theta).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\partial_k(F_N(\bar{x}, \theta)w(\theta)) = {} & N\bar{x}_k F_N(\bar{x}, \theta)w(\theta) \\
& - N\eta_k(\theta)F_N(\bar{x}, \theta)w(\theta) \\
& + \partial_k \ln w(\theta) \cdot F_N(\bar{x}, \theta)w(\theta).
\end{aligned}$$

Since the fact that $\bar{x} \in H_{in}$ holds for $N \geq N'$ and Assumptions 2 and 3, we can see that the right-hand side is integrable for sufficiently large N. Therefore, the left hand-side is also integrable. Hence, integrating both sides over Θ, we have

$$\begin{aligned}
\int \partial_k(F_N(\bar{x}, \theta)w(\theta))\, d\theta = {} & N\bar{x}_k \int F_N(\bar{x}, \theta)w(\theta)\, d\theta \\
& - N \int \eta_k(\theta)F_N(\bar{x}, \theta)w(\theta)\, d\theta \\
& + \int \partial_k \ln w(\theta) \cdot F_N(\bar{x}, \theta)w(\theta)\, d\theta.
\end{aligned}$$

Dividing both sides by $\int_\Theta F_N(\bar{x}, \theta)w(\theta)d\theta$ and letting

$$w_N(\theta|z) \equiv F_N(z, \theta)w(\theta) / \int F_N(z, \theta)w(\theta)d\theta$$

($w_N(\theta|\bar{x})$ equals the posterior density), we have

$$\int_\Theta \partial_k w_N(\theta|\bar{x})d\theta = N\bar{x}_k - N\tilde{\eta}_k + \int_\Theta \partial_k \ln w(\theta) \cdot w_N(\theta|\bar{x})d\theta.$$

Namely, we obtain

$$\begin{aligned}
\tilde{\eta}_k = {} & \bar{x}_k + \int \partial_k \ln w(\theta) \cdot w_N(\theta|\bar{x})d\theta/N \\
& - \int \partial_k w_N(\theta|\bar{x})d\theta/N. \qquad (3)
\end{aligned}$$

We can prove the following two formulas:

$$\int_\Theta \partial_k \ln w(\theta)\cdot w_N(\theta|z)d\theta = \partial_k \ln w(\theta_z) + O(\sqrt{\ln N}/\sqrt{N}) \tag{4}$$

and

$$\int_\Theta \partial_k w_N(\theta|z)d\theta = O(\exp(-DN)) \tag{5}$$

where we let θ_z denote $\theta|_{\eta=z}$ and D is a certain constant. These hold uniformly for $z \in H_{in}$. By the assumption, $\bar{x}$ in (3) belongs to H_{in} for $N \geq N'$. Therefore, the right-hand side of (3) belongs to H_a for sufficiently large N under (4) and (5), i.e., we can obtain the claim of the theorem.

Now, we show (4). Hereafter in this section, we let $\hat{\theta}$ denote $\theta_z = \theta|_{\eta=z}$. Let G_N denote the left-hand side of (4) and let $h(\theta) \equiv \partial_k \ln w(\theta)$, then we have

$$G_N = \int h(\theta)F_N(z,\theta)w(\theta)d\theta / \int F_N(z,\theta)w(\theta)d\theta.$$

Let $\delta = ((n+1)\ln N/(2N))^{1/2}$ and define

$$M_\delta \equiv \{\theta | \hat{g}_{ij}(\theta^i - \hat{\theta}^i)(\theta^j - \hat{\theta}^j) \leq \delta^2\}.$$

($\hat{g}_{ij}$ denotes its value at $\eta = z$.) By Taylor's theorem, we have

$$\begin{aligned}\ln F_N(z,\theta) = \ln F_N(z,\hat{\theta}) - (N/2)\hat{g}_{ij}(\theta^i - \hat{\theta}^i)(\theta^j - \hat{\theta}^j) \\ + NO(|\theta - \hat{\theta}|^3).\end{aligned}$$

Since Θ_{in} is compact and g_{ij} is of class C^∞ on Θ_a°, $|\partial_t g_{ij}(\theta)| < \infty$ holds uniformly on Θ_{in}. Then, the following holds for an arbitrary θ in M_δ:

$$\begin{aligned}F_N(z,\theta) &= F_N(z,\hat{\theta})e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)+O(N\delta^3)} \\ &= (1+O(N\delta^3))F_N(z,\hat{\theta})e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}.\end{aligned} \tag{6}$$

Here, though the term $O(\delta^3)$ depends also on θ, the order of δ is uniform with respect to θ. Hence, hereafter, we use the order notation for δ in the same sense.

We also have $\forall\theta \in \partial M_\delta$

$$\begin{aligned}\ln F_N(z,\hat{\theta} - \ln F_N(z,\theta) &\geq (N/2)\hat{g}_{ij}(\theta^i - \hat{\theta}^i)(\theta^j - \hat{\theta}^j) \\ &\quad + N\cdot O(|\theta - \hat{\theta}|^3) \\ &= N\delta^2/2 + O(N\delta^3).\end{aligned}$$

Hence, by the fact that $\psi(\theta) - \theta^i\bar{x}_i$ is strictly convex on a convex set Θ_a°, the following holds for an arbitrary θ in $\Theta - M_\delta$:

$$F_N(z,\theta) \leq (1+O(N\delta^3))F_N(z,\hat{\theta})\exp(-N\delta^2/2). \tag{7}$$

Now, we evaluate the numerator of G_N. We have

$$\begin{aligned}\int_\Theta h(\theta)F_N(z,\theta)w(\theta)d\theta &= \int_{M_\delta} h(\theta)F_N(z,\theta)w(\theta)d\theta \\ &\quad + \int_{\Theta - M_\delta} h(\theta)F_N(z,\theta)w(\theta)d\theta.\end{aligned} \tag{8}$$

By Taylor's theorem, $\forall\theta \in M_\delta$

$$\ln w(\theta) = \ln w(\hat{\theta}) + (\theta^i - \hat{\theta}^i)\partial_i w(\theta')/w(\theta')$$

holds, where $\theta' = \alpha\theta + (1-\alpha)\hat{\theta}$ for a certain $\alpha \in [0,1]$. Since w is of class C^2 and $w > 0$ on Θ and Θ_{in} is compact, we have $|\partial_i w(\theta)/w(\theta)| < \infty$ on Θ_{in}. Therefore, $\forall\theta \in M_\delta$

$$\ln w(\theta) = \ln w(\hat{\theta}) + O(\delta)$$

i.e., $\forall\theta \in M_\delta$

$$w(\theta) = (1+O(\delta))w(\hat{\theta})$$

holds. Similarly, we also obtain $\forall\theta \in M_\delta$

$$h(\theta) = h(\hat{\theta}) + O(\delta).$$

Then, using (6), we can transform the first term of the right-hand side of (8) as follows:

$$\begin{aligned}\int_{M_\delta} h(\theta)F_N(z,\theta)w(\theta)d\theta &= ABC\cdot F_N(z,\hat{\theta})w(\hat{\theta}) \\ &\quad \cdot \int_{M_\delta} e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}d\theta\end{aligned}$$

where we let

$$\begin{aligned}A &= h(\hat{\theta}) + O(\delta) \\ B &= 1 + O(\delta)\end{aligned}$$

and

$$C = 1 + O(N\delta^3).$$

Moreover, we have

$$\begin{aligned}&\int_{M_\delta} e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}d\theta \\ &\quad = \int_{\Re^n} e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}d\theta \\ &\qquad - \int_{M_\delta^c} e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}d\theta \\ &\quad = (2\pi/\hat{g}N)^{n/2} - \int_{M_\delta^c} e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}d\theta\end{aligned}$$

where we let $\hat{g}$ denote $\det|\hat{g}_{ij}|$ and M_δ^c denote $\Re^n - M_\delta$. Here, we have

$$\begin{aligned}&\int_{M_\delta^c} e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}d\theta \\ &\leq e^{-(N-1)\delta^2/2}\int_{M_\delta^c} e^{-(1/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}d\theta = O(e^{-N\delta^2}).\end{aligned}$$

Therefore, we have

$$\int_{M_\delta} e^{-(N/2)\hat{g}_{ij}(\theta^i-\hat{\theta}^i)(\theta^j-\hat{\theta}^j)}d\theta = (2\pi/\hat{g}N)^{n/2} - O(e^{-N\delta^2}).$$

Hence, noting $ABC = O(1)$, we obtain

$$\begin{aligned}&\int_{M_\delta} h(\theta)F_N(z,\theta)w(\theta)d\theta \\ &\quad = F_N(z,\hat{\theta})w(\hat{\theta})(ABC(2\pi/\hat{g}N)^{n/2} - O(e^{-N\delta^2})).\end{aligned} \tag{9}$$

Noting (7) and

$$\int_\Theta |h(\theta)F_N(z,\theta)w(\theta)|d\theta < \infty \quad \text{for } N \geq \max\{N_2, N_3\}$$

we have

$$\begin{aligned}\left|\int_{\Theta - M_\delta} h(\theta)F_N(z,\theta)w(\theta)d\theta\right| &\leq \int_{\Theta - M_\delta} |h(\theta)F_N(z,\theta)w(\theta)|d\theta \\ &= F_N(z,\hat{\theta})O\left(e^{-N\delta^2}\right) \\ &= F_N(z,\hat{\theta})w(\hat{\theta})O\left(e^{-N\delta^2}\right).\end{aligned}$$

(The last equality follows from the fact that $1/w(\theta) < \infty$ on Θ_{in}.) Now, we obtain

$$\begin{aligned}&\int_\Theta h(\theta)F_N(z,\theta)w(\theta)d\theta \\ &\quad = F_N(z,\hat{\theta})w(\hat{\theta})(ABC(2\pi/\hat{g}N)^{n/2} + O(e^{-N\delta^2})). \quad (10)\end{aligned}$$

Next, we consider the denominator of G_N. We can similarly evaluate that by plugging in 1 to A except for the evaluation of

$$\int_{\Theta - M_\delta} F_N(z,\theta)w(\theta)d\theta.$$

Now, noting (7) and

$$\int_\Theta |F_N(z,\theta)w(\theta)|d\theta < \infty, \qquad \text{for } N \geq N_1$$

we have

$$\begin{aligned}\int_{\Theta - M_\delta} F_N(z,\theta)w(\theta)d\theta &= F_N(z,\hat{\theta})O\left(e^{-N\delta^2}\right) \\ &= F_N(z,\hat{\theta})w(\hat{\theta})O\left(e^{-N\delta^2}\right).\end{aligned}$$

Namely, we obtain

$$\begin{aligned}&\int_\Theta F_N(z,\theta)w(\theta)d\theta \\ &\quad = F_N(z,\hat{\theta})w(\hat{\theta})(B'C'(2\pi/\hat{g}N)^{n/2} + O(e^{-N\delta^2})) \quad (11)\end{aligned}$$

where $B' = 1 + O(\delta)$ and $C' = 1 + O(N\delta^3)$.

Now, using (11) and (10), we obtain

$$\begin{aligned}G_N &= \frac{ABC(2\pi/\hat{g}N)^{n/2} + O(e^{-N\delta^2})}{B'C'(2\pi/\hat{g}N)^{n/2} + O(e^{-N\delta^2})} \\ &= \frac{ABC + O(N^{n/2}e^{-N\delta^2})}{B'C' + O(N^{n/2}e^{-N\delta^2})}.\end{aligned}$$

Hence, noting $\delta^2 = (n+1)\ln N/(2N)$, we have

$$\begin{aligned}N^{n/2}e^{-N\delta^2} &= O(1/\sqrt{N}) \\ \delta &= O((\ln N/N)^{1/2})\end{aligned}$$

and

$$N\delta^3 = O((\ln N/N)^{1/2}).$$

Then, we have

$$B'C' = 1 + O((\ln N/N)^{1/2})$$

and

$$ABC = h(\hat{\theta}) + O((\ln N/N)^{1/2}).$$

Therefore, we have

$$\begin{aligned}G_N &= (h(\hat{\theta}) + O((\ln N/N)^{1/2}))/(1 + O((\ln N/N)^{1/2})) \\ &= h(\hat{\theta}) + O((\ln N/N)^{1/2}).\end{aligned}$$

Namely, we have (4).

Next, we show (5). Suppose $k = 1$ for simplicity. Let

$$I \equiv \int_\Theta (\partial_1 w_N(\theta|z))d\theta.$$

By Assumption 1, for almost all $\zeta(1) \in \Theta/\{\theta^1\}$, $l_{\zeta(1)} \cap \partial\Theta$ is a finite set. Let $Q(\zeta(1))$ denote $l_{\zeta(1)} \cap \partial\Theta$. We let Z denote the largest subset of $\Theta/\{\theta^1\}$, such that $Q(\zeta(1))$ is a finite set for all $\zeta(1) \in Z$. Then, integrating $\partial_1 w_N(\theta|z)$ with respect to θ^1 (Fubini's theorem), we obtain

$$I = \int_{\zeta \in Z} \sum_{\mu \in Q(\zeta)} \epsilon(\mu)(w_N(\mu|z))d\zeta$$

where $\epsilon(\mu)$ is $+1$ or -1. Hence, we have

$$\begin{aligned}|I| &\leq \int_{\zeta \in Z} \sum_{\mu \in Q(\zeta)} w_N(\mu|z)d\zeta \\ &= \frac{\displaystyle\int_{\zeta \in Z} \sum_{\mu \in Q(\zeta)} F_N(z,\mu)w(\mu)d\zeta}{\displaystyle\int_{\theta \in \Theta} F_N(z,\theta)w(\theta)d\theta}. \quad (12)\end{aligned}$$

Since $x^\infty \in T_{N'}$, $\hat{\theta} = \theta|_{\eta=z} \in \Theta_{in} \subset \Theta^\circ$ holds for all $N \geq N'$. Hence, $\partial\Theta \cap M_\delta = \emptyset$ holds for large N. Hence from (7), we have $\forall\theta \in \partial\Theta$

$$F_N(z,\theta) \leq (1 + O(N\delta^3))F_N(z,\hat{\theta})\exp(-N\delta^2/2).$$

Therefore, noting $\mu \in Q(\zeta) \subset \partial\Theta$ and

$$\int_{\zeta \in Z} \sum_{\mu \in Q(\zeta)} F_N(z,\mu)w(\mu)d\zeta < \infty$$

for $N > \max\{N_1, N_2\}$, we have

$$\int_{\zeta \in Z} \sum_{\mu \in Q(\zeta)} F_N(z,\mu)w(\mu)d\zeta = F_N(z,\hat{\theta})O(\exp(-D'N)).$$

(D' is a certain constant.) Together with (12) and (11), we can write

$$|I| = O(N^{n/2}\exp(-D'N)) = O(\exp(-DN))$$

where D is a certain constant. This concludes (5). Now we have obtained the claim of the theorem. Q.E.D.

V. The MDL Estimator

In this section, we construct an estimator for the n-dimensional model $S = \{p(x|\theta)|\theta \in \Theta\}$ (it is not necessarily an exponential family) based on the MDL principle, i.e., we determine the code (quantization) for parameters based on the MDL principle. The argument in this section is essentially parallel to those of Rissanen [12], Barron [2], and Wallace and Freeman [16]. The point that is unique to our derivation is that we determine the code so as to minimize the average total code length. In [12] and [2], the *typical* value of total code length was minimized. In [16], the *expectation* (by the prior distribution) of the average total code length was minimized.

Suppose the prior $d\phi = w(\theta)d\theta$ over Θ. The coding of the parameter θ consists of two parts: 1) quantizing Θ (to obtain a countable set) and 2) describing quantized points. Since the optimal coding depends on the data size N, we let the quantization depend on N and let Θ_N denote the set of quantized points. (We suppose that N is known prior to encoding.) Let $r(\bar{\theta})$ denote the region represented by the quantized point $\bar{\theta}$ and let $v(\bar{\theta}) = \int_{\theta \in r(\bar{\theta})} w(\theta)d\theta$. Moreover, we give the code length $l_N(\bar{\theta}) = -\ln v(\bar{\theta})$ to the quantized point $\bar{\theta}$. We let $\mathcal{L}_N$ denote the set of such coding schema, obtained by varying the quantizations Θ_N. Now, we give the definition of MDL estimator.

Definition 2: Define θ_{mdl} as follows:

$$\begin{aligned}\hat{\theta}_{l_N} &\equiv \arg\min_{\bar{\theta}\in\Theta_N}(-\ln p(x^N|\bar{\theta}) + l_N(\bar{\theta})),\\ l_N^* &\equiv \arg\min_{l_N\in\mathcal{L}_N} E_\theta(-\ln p(x^N|\hat{\theta}_{l_N}) + l_N(\hat{\theta}_{l_N})),\\ \theta_{mdl}(x^N) &\equiv \arg\min_{\bar{\theta}}(-\ln p(x^N|\bar{\theta}) + l_N^*(\bar{\theta})).\end{aligned}$$

We refer to the function which maps $x^N \in \mathcal{X}^N$ to $f^w_{mdl}[x^N] \equiv p(\cdot|\theta_{mdl}(x^N))$ as the MDL estimator with the prior w. Especially, when w is uniform over Θ, we refer to $f^{wd\theta}_{mdl}$ as the MDL estimator with respect to the coordinate system θ and let f^θ_{mdl} denote it.

Remark: In the following approximation, the code length l^* does not depend on θ.

Let us determine on approximation to l_N^*. To this end we shall obtain the conditional expected value of total description length DL $(\bar{\theta})$ under the condition $\hat{\theta}_{l_N} \in r(\bar{\theta})$. The code length for the parameter is given by $-\ln v(\bar{\theta})$ and the code length for the data x^N is given by $-\ln p(x^N|\hat{\theta}_{l_N})$. Let g_{ij} denote the Fisher information matrix with respect to θ. We suppose that $r(\bar{\theta})$'s can be approximated by rectangles each axis of which lies in the direction of principal axis of $g_{ij}(\bar{\theta})$. (If not the case, the description length becomes longer. See [16].) Let X_α denote the unit tangent vector along the αth principal axis of $g_{ij}(\theta)$, and λ_α denote the eigenvalue associated with the αth principal axis. Let d_α be a smooth function defined on Θ and suppose that $d_\alpha(\bar{\theta})$ equals the length of the X_α direction's axis of $r(\bar{\theta})$. Then

$$v(\bar{\theta}) \sim \prod_{\alpha=1}^n d_\alpha(\bar{\theta})w(\bar{\theta})$$

holds. (Hereafter, we define

$$v(\hat{\theta}) \equiv \prod_{\alpha=1}^n d_\alpha(\hat{\theta})w(\hat{\theta})$$

for $\hat{\theta} \notin \Theta_N$.) By Taylor expansion, we have

$$\ln p(x^N|\hat{\theta}_{l_N}) \sim \ln p(x^N|\hat{\theta}) - (N/2)g_{ij}(\hat{\theta}^i_{l_N} - \hat{\theta}^i)(\hat{\theta}^j_{l_N} - \hat{\theta}^j).$$

We suppose that the conditional distribution of $\hat{\theta}$ given $\hat{\theta} \in r(\bar{\theta})$ is approximately uniform so that

$$E_\theta[g_{ij}(\hat{\theta}^i_{l_N} - \hat{\theta}^i)(\hat{\theta}^j_{l_N} - \hat{\theta}^j)|\hat{\theta}_{l_N} = \bar{\theta}]$$

almost equals

$$\sum_\alpha \lambda_\alpha(\bar{\theta})d_\alpha(\bar{\theta})^2/12.$$

Then, we obtain

$$\begin{aligned}\mathrm{DL}\,(\bar{\theta}) &\sim E_\theta[-\ln p(x^N|\hat{\theta})|\hat{\theta}_{l_N} = \bar{\theta}]\\ &\quad + (N/24)\sum_\alpha \lambda_\alpha(\bar{\theta})d_\alpha(\bar{\theta})^2 - \ln v(\bar{\theta}).\end{aligned}$$

Hence, we can evaluate the (unconditional) expectation of the total description length (denoted by L) as follows:

$$\begin{aligned}L &= \sum_{\bar{\theta}} \mathrm{DL}\,(\bar{\theta})\Pr\{\hat{\theta}_{l_N} = \bar{\theta}\}\\ &\sim E_\theta(-\ln p(x^N|\hat{\theta})) + \sum_{\bar{\theta}}\Big(-\ln v(\bar{\theta})\\ &\quad + \sum_\alpha N\lambda_\alpha(\bar{\theta})d_\alpha(\bar{\theta})^2/24\Big)\Pr\{\hat{\theta}_{l_N} = \bar{\theta}\}.\end{aligned}$$

Since

$$\begin{aligned}&-\ln v(\bar{\theta}) + \sum_\alpha N\lambda_\alpha(\bar{\theta})d_\alpha(\bar{\theta})^2/24\\ &\qquad \sim E_\theta[-\ln v(\hat{\theta}) + \sum_\alpha N\lambda_\alpha(\hat{\theta})d_\alpha(\hat{\theta})^2/24|\hat{\theta} \in r(\bar{\theta})]\end{aligned}$$

and

$$\Pr\{\hat{\theta}_{l_N} = \bar{\theta}\} \sim \Pr\{\hat{\theta} \in r(\bar{\theta})\}$$

we obtain

$$\begin{aligned}&\sum_{\bar{\theta}}\left(\sum_\alpha N\lambda_\alpha(\bar{\theta})d_\alpha(\bar{\theta})^2/24 - \ln v(\bar{\theta})\right)\Pr\{\hat{\theta}_{l_N} = \bar{\theta}\}\\ &\qquad \sim E_\theta\left[\sum_\alpha (N\lambda_\alpha(\hat{\theta})d_\alpha(\hat{\theta})^2/24 - \ln d_\alpha(\hat{\theta}))\right]\end{aligned}$$

i.e., we have

$$L \sim E_\theta\left[\sum_\alpha (N\lambda_\alpha(\hat{\theta})d_\alpha(\hat{\theta})^2/24 - \ln d_\alpha(\hat{\theta}))\right] + C_\theta$$

where we let $C_\theta = E_\theta(-\ln p(x^N|\hat{\theta}) - \ln w(\hat{\theta}))$. The choice of $d_\alpha(\hat{\theta})$ minimizes this approximated code length is $d_\alpha = (12/(N\lambda_\alpha))^{1/2}$. Hence, we have

$$v(\theta) = \prod_{\alpha=1}^n d_\alpha(\theta)w(\theta) = 12^{n/2}N^{-n/2}(\det|g_{ij}|)^{-1/2}.$$

Finally, we obtain

$$l^*(\bar{\theta}) \sim (1/2)\cdot \ln\det|g_{ij}(\bar{\theta})| + (n/2)\ln N - \ln w(\bar{\theta}) - (n\ln 12)/2.$$

Then, we have

$$\theta_{mdl} = \arg\min_{\bar{\theta}\in\Theta_N}\big(-\ln p(x^N|\bar{\theta}) + (1/2)\cdot\ln\det|g_{ij}(\bar{\theta})| - \ln w(\bar{\theta})\big).$$

The above length $l^*(\bar{\theta})$ is essentially equivalent to the ones given in [15], [16], [2] (where a term of order $O(\ln\ln N)$ is included, because of the lack of the assumption that N is known).

In the sequel, we neglect the quantization error. Namely, we employ the following as the definition of the MDL estimator:

$$\hat{\theta}_{mdl} \equiv \arg\min_{\theta\in\Theta}\Big(-\ln p(x^N|\theta) + \frac{\ln\det|g_{ij}(\theta)|}{2} - \ln w(\theta)\Big).$$

We can rewrite the above definition as

$$\hat{\theta}_{mdl} = \arg\max_{\theta\in\Theta} p(x^N|\theta)w(\theta)/(\det|g_{ij}(\theta)|)^{1/2}.$$

Hence, the estimator $\hat{\theta}_{mdl}$ is equivalent to the posterior mode estimator provided that posterior density is defined with respect to the measure $(\det|g_{ij}(\theta)|)^{1/2}d\theta$, which is the natural volume element of S. Therefore, f^w_{mdl} is invariant under the transformation of coordinate system.

Let $\rho(\theta)$ denote $w(\theta)/(\det|g_{ij}|)^{1/2}$ and define a family of real-valued function $L_\alpha(z,\theta)$ on $\mathcal{W}\times\Theta$ as

$$L_\alpha(z,\theta) = -\alpha(\theta^i z_i - \psi(\theta)) - \ln\rho(\theta)$$

where $\alpha\in\Re$. Then, $L_N(\bar{x},\theta)$ equals the total description length for the MDL estimator with the prior w. We assume the following.

Assumption 4: When N is not less than a certain integer N_3, for any $z\in H_{in}$

$$\arg\min_{\theta\in\Theta} L_N(z,\theta)\in\Theta^\circ$$

holds.

This assumption can be checked by using the following proposition.

Proposition 3: Assumption 4 holds if and only if the following holds: For a certain $N_4\in\mathcal{N}$, for a certain $z\in H_{in}$

$$\arg\min_{\theta\in\Theta} L_N(z,\theta)\in\Theta^\circ$$

holds.

The proof is similar to that of Proposition 2, we omit it.

We can prove the following lemma.

Lemma 1: Suppose Assumptions 1 and 4 hold for $S(\Theta)$ and w. For any $N'\in\mathcal{N}$

$$\eta_i(f^w_{mdl}[x^N]) = \hat{\eta}_i + (\partial_i\ln w(\hat{\theta}) - \hat{T}_{ijk}\hat{g}^{jk}/2)/N + O(1/N^2)$$

holds, uniformly for $x^\infty\in T_{N'}$.

Proof: Let us prove that $\arg\min_{\eta\in H} L_N(z,\theta(\eta))$ equals $z_i + \partial_i\ln\rho(\theta)|_{\eta=z}/N + O(1/N^2)$ uniformly on H_{in}. That implies the claim of the Lemma.

Let θ_z denote $\theta|_{\eta=z}$. Define $K_\epsilon(\theta_z) = \{\theta|\ |\theta-\theta_z|<\epsilon\}$. First, we show that for any small ϵ, there exists N_a such that $\forall N>N_a$, $\arg\min_{\theta\in\Theta} L_N(z,\theta)\in K_\epsilon(\theta_z)$. Since $w(\theta)$ is of class C^2, $B\equiv\sup_{\theta\in\Theta_{in}}(-\ln\rho(\theta))<\infty$ holds. Let $l_z = \theta^i_z z_i - \psi(\theta_z)$, then we have $L_N(z,\theta_z)\le Nl_z + B$. Since Assumption 4, we have $\forall\theta\in\Theta_a$, $L_N(z,\theta)\ge A$ for $N=N_3$, where A is a certain real. Then, we have

$$-\ln\rho(\theta)\ge A + N_3(\theta^i\bar{x}_i - \psi(\theta)).$$

Hence, $L_N(z,\theta)\ge -(\theta^i z_i - \psi(\theta))(N-N_3) + A$ holds. Note that $-(\theta^i z_i - \psi(\theta))$ is strictly convex on Θ°_a and that Θ°_a is a convex set. For any small ϵ, there exists $\delta>0$ such that $\forall\theta\in(\Theta_a - M_\epsilon)$, $-(\theta^i\bar{x}_i - \psi(\theta))\ge -\hat{l}+\delta$ holds. Then, we have $\forall N>N_3$, $\forall\theta\in\Theta_a - M_\epsilon$

$$\begin{aligned} L_N(z,\theta) &\ge (-l_z+\delta)(N-N_3) + A \\ &= -Nl_z + A - N_3(-\hat{l}+\delta) + N\delta. \end{aligned}$$

Hence, for sufficiently large N, $\forall\theta\in(\Theta_a - M_\epsilon)$, $L_N(z,\theta)\ge L_N(z,\theta_z)$ holds. Therefore, $\arg\min_{\theta\in\Theta} L_N(z,\theta)\in K_\epsilon(\theta_z)$ holds. Since H_{in} is included in H°, the above argument uniformly holds for $z\in H_{in}$.

Let ϵ_0 be small such that $K_{\epsilon_0}(\theta)\subset\Theta$ holds for any $\theta\in\Theta_{in}$. Then, for sufficiently large N, $L_N(z,\theta)$ takes the minimum in $K_{\epsilon_0}(z)$. Next, note that $L_N(z,\theta)$ is strictly convex on Θ° (with respect to θ) for sufficiently large N. Therefore, the equations $\partial_i L_N(z,\theta)=0$ have a unique solution in $K_{\epsilon_0}(\theta_z)$, which is the minimum point. We have

$$\partial_i L_N(z,\theta)/N = -z_i + \eta_i - \partial_i\ln\rho(\theta)/N.$$

Noting $|\partial_i\ln\rho(\theta)|<\infty$ on $\bigcup_{z\in H_{in}} K_{\epsilon_0}(\theta_z)$, we can see that the solution of the equations $\partial_i L_N(z,\theta)=0$ equals $z_i + \partial_i\ln\rho(\theta_z)/N + O(1/N^2)$. This completes the proof. Q.E.D.

Remark: In particular, $f^{w_J}_{mdl} = \hat{f}$ holds exactly.

We let w_η denote the uniform prior with respect to η. Then $w_\eta(\theta)d\theta\propto\det|g_{ij}|d\theta$ holds. Hence, noting (1), we have $\partial_i w_\eta(\theta) = T_{ijk}g^{jk}$. Therefore, we can see the following lemma holds as the special case of the Lemma 1.

Lemma 2: Suppose Assumptions 1 and 4 hold for $S(\Theta)$ and w_η. For any $N'\in\mathcal{N}$

$$\eta_i(f^\eta_{mdl}[x^N]) = \hat{\eta}_i + \hat{T}_{ijk}\hat{g}^{jk}/2N + O(1/N^2)$$

holds uniformly for $x^\infty\in T_{N'}$.

VI. Bias-Corrected Maximum-Likelihood Estimator

Hereafter, we suppose that $S(\Theta_a)$ is not only minimal but steep. In this case, the expectation of $\hat{\eta}=\bar{x}$ equals η itself. Hence, we can say that the MLE is unbiased with respect to the η-coordinates. Now, we think about the other coordinates u. Let u_0 and η_0 denote the values of u and η of the actual distribution, respectively. Thinking of u as the function of η, taking Taylor expansion of u in the neighborhood of η_0, we

have

$$\hat{u}_i - u_{0i} = (\partial^k u_i)_0(\hat{\eta}_k - \eta_{0k}) + (\partial^l\partial^k u_i)_0(\hat{\eta}_k - \eta_{0k})(\hat{\eta}_l - \eta_{0l})/2 + O(|\hat{\eta} - \eta_0|^3).$$

Taking the expectation of both sides, we have

$$E_\eta(\hat{u}_i - u_{0i}) = (\partial^l\partial^k u_i)_0(g_{kl})_0/2N + O(1/N\sqrt{N}).$$

Namely, $\hat{u}$ has bias of order $1/N$. However, if we let

$$\tilde{u}_i \equiv \hat{u}_i - \partial^l\partial^k u_i(\hat{\eta})\hat{g}_{kl}/2N \tag{13}$$

it is known that $\tilde{u}$'s bias is of order $O(N^{-3/2})$. Moreover, the mean-square error of $\tilde{u}$ is least (with respect to the $1/N^2$ term) among the efficient estimators (see [1]). The function $f^u_{bc} : x^N \mapsto p(x|\tilde{u}) \in S$ is called the bias-corrected MLE with respect to the coordinate system u.

We can prove the following lemma.

Lemma 3: For any $N' \in \mathcal{N}$,

$$\eta_i(f^\theta_{bc}[x^N]) = \bar{x}_i + \hat{T}_{ijk}\hat{g}^{jk}/2N + O(1/N^2)$$

holds, uniformly for $x^\infty \in T_{N'}$.

Proof: Plugging in θ to u of (13), we have

$$\theta^i(f^\theta_{bc}[x^N]) - \theta^i(\hat{f}[x^N]) = -\partial^l\hat{g}^{ik}\hat{g}_{kl}/2N.$$

Hence, we have

$$\eta_j(f^\theta_{bc}[x^N]) - \hat{\eta}_j = -\frac{\partial\eta_j}{\partial\theta^i}(\partial^l\hat{g}^{ik})\hat{g}_{kl}/2N + O(1/N^2) = -\hat{g}_{ji}(\partial^l\hat{g}^{ik})\hat{g}_{kl}/2N + O(1/N^2). \tag{14}$$

Now, differentiating $g_{ji}g^{ik} = \delta^k_j$ with respect to η_l, we have

$$-g_{ji}\partial^l g^{ik} = (\partial^l g_{ji})g^{ik} = \frac{\partial g_{ji}}{\partial\theta^m}\frac{\partial\theta^m}{\partial\eta_l}g^{ik} = T_{jmi}g^{ml}g^{ik}.$$

Hence, we have

$$-g_{ji}(\partial^l g^{ik})g_{kl} = T_{jmi}g^{ml}g^{ik}g_{kl} = T_{jmi}g^{ml}\delta^i_l = T_{jmi}g^{mi}.$$

Together with (14), we obtain the claim of the lemma. Q.E.D.

VII. Discussion

Corollary 1 and Lemmas 2 and 3 give asymptotic forms of $\tilde{f}_{w_J}$, f^η_{mdl} and f^θ_{bc}, and yield the following theorem.

Theorem 2: Let $S(\Theta)$ be a minimal steep exponential family and Θ satisfy Assumption 1. Suppose that Assumptions 2 and 3 hold for $S(\Theta)$ and w_J, and that Assumption 4 holds for $S(\Theta)$ and w_η. For any $N' \in \mathcal{N}$, the differences between $\eta(f^\eta_{mdl}[x^N])$, $\eta(\tilde{f}_{w_J}[x^N])$, and $\eta(f^\theta_{bc}[x^N])$ are of order $O((\ln N)^{1/2}N^{-3/2})$, uniformly for $x^\infty \in T_{N'}$.

Since the geometrical fact that $\tilde{f}_{w_J}$ is a -1-projection of f_{w_J}, we can expect that Theorem 2 represents a certain property of the maximin estimator f_{w_J}. In particular, we can think that the maximin property has strong relation with unbiasness with respect to the canonical parameter. The canonical parameter has a geometrical interpretation that it is the affine parameter associated with respect to the 1-connection [1]. It would be interesting to analyze the relation between the above optimality and the 1-connection.

Using Theorem 1 and Lemma 1, we can also obtain the following theorem, which is dual to Theorem 2.

TABLE I
Dependency of Estimators on Prior

Prior $w\,d\theta$	$d\theta$	$\sqrt{\det\|g_{ij}\|}d\theta$	$d\eta$
$\tilde{f}_w$	η-unbiased	θ-unbiased	
f^w_{mdl}		η-unbiased	θ-unbiased

Theorem 3: Let $S(\Theta)$ be a minimal steep exponential family and Θ satisfy Assumption 1. Suppose that Assumptions 2 and 3 hold for $S(\Theta)$ and $d\theta$, and that Assumption 4 holds for $S(\Theta)$ and w_J. For any $N' \in \mathcal{N}$, the differences between $\eta(f^{w_J}_{mdl}[x^N])$, $\eta(\tilde{f}_{d\theta}[x^N])$, and $\hat{\eta}$ are of order $O(1/N^2)$, uniformly for $x^\infty \in T_{N'}$.

We can illustrate the above two theorems by Table I, where we ignore the terms of order $o(1/N)$. Note that η is the affine parameter with respect to the -1-connection and is dual to θ. From this table, we assume that $\tilde{f}_w$ and f^w_{mdl} form a dual pair, because if we exchanged η and θ, and the Bayes and MDL, then we would have the same table again.

Now, we give some examples.

Example 1 (Bernoulli Sources): Define $\mathcal{X} = \{0, 1\}$

$$S = \{p(x|\eta) = \eta^x(1-\eta)^{1-x} | 0 < \eta < 1\}$$

and $\theta = \ln(\eta/(1-\eta))$. We have

$$g(\eta) \equiv g^{11} = \partial\theta/\partial\eta = 1/\eta(1-\eta).$$

We let $k \equiv \sum_{t=1}^N x_t$. As it is well known, the Bayes predictor with Jeffreys prior equals $(k+0.5)/(N+1)$, denoted by η_L. Next, we consider the MDL estimator. The total description length for the MDL with respect to the η-coordinates is

$$-k\ln\eta - (N-k)\ln(1-\eta) - (\ln\eta + \ln(1-\eta))/2 = -(k+0.5)\ln\eta - (N-k+0.5)\ln(1-\eta).$$

The value of η minimizing the total description length is $\eta_{mdl} = (k+0.5)/(N+1)$, which strictly equals η_L. Finally, we consider the bias-corrected MLE. Since

$$T_{111} = E_\theta((x-\eta)^3) = \eta(1-\eta)^3 - (1-\eta)\eta^3 = \eta(1-\eta)(1-2\eta)$$

holds, we have

$$\hat{T}_{111}\hat{g}^{11} = 1 - 2\hat{\eta} = 1 - 2k/N.$$

Hence, we have

$$\eta_{bc} = k/N + (1-(2k/N))/(2N) + O(1/N^2) = \eta_L + O(1/N^2).$$

Example 2 (Normal Distributions): The family of normal distributions is defined as

$$S = \{p(z|\mu,\sigma) = (1/\sqrt{2\pi}\sigma) \cdot \exp(-(z-\mu)^2/2\sigma^2) | \mu \in \Re, \sigma^2 > 0\}.$$

If we define a new vector valued random variable x as $x_1 = z$, $x_2 = z^2$ and let $\theta^1 = \mu/\sigma^2$, $\theta^2 = -1/(2\sigma^2)$, and

$$\psi(\theta) = -(\theta^1)^2/(4\theta^2) + (1/2)\cdot\ln(-\pi/\theta^2)$$

we can see S is an exponential family. Now, we have

$$\det|g_{ij}| = -2/(\theta^2)^3 = 16\sigma^6.$$

In reference to estimation of μ, both the projected Bayes estimator and the MDL estimator give the same result as the MLE. Therefore, it is sufficient to consider estimation of a coordinate σ^2 alone. We describe just the result for simplicity. Let $\tilde{\sigma^2}$ and $\hat{\sigma^2}_{mdl}$ denote the projected Bayes estimate (with w_J) and the MDL estimate (with w_η) of σ^2, respectively, then we have $\tilde{\sigma^2} = (N+1)v^2/(N-2)$ and $\hat{\sigma^2}_{mdl} = Nv^2/(N-3)$, where $v^2 = \bar{z^2} - \bar{z}^2$. The difference between the two is of order $O(1/N^2)$. Compare this with the so-called unbiased variance $Nv^2/(N-1)$.

In the above examples, we were able to analytically obtain the projected Bayes estimator and the MDL estimator. In general, however, it is difficult to do so. In such cases, Theorem 1 and Lemma 1 provide us with a way to approximate the projected Bayes estimator and the MDL estimator. Actually in [14], a similar method as the proof of Theorem 1 is used to approximate the Bayes codes for the Markov sources.

Appendix
Proof of Proposition 2

Suppose first that $\xi(\theta)F_{N_0}(z_0,\theta)$ is integrable on Θ, where $z_0 \in H_{in}$ holds, and N_0 and C are positive reals. In the sequel, we assume that $z \neq z_0$, since $\xi(\theta)F_N(z_0,\theta)$ is integrable when $N \geq N_0$.

Let H' denote a compact set of $\Re^n$ such that $H_{in} \subset H'^{\circ}$ and $H' \subset H_a^{\circ}$ hold. (There exists such a set, since $H_{in} \subset H_a^{\circ}$.) Let θ_z denote the value of θ which corresponds to $\eta = z$. Then, we have $F_1(z,\theta) \leq F_1(z,\theta_z)$. Moreover, the function $z \mapsto F_1(z,\theta_z)$ is continuous on H_a°. Hence, $F_1(z,\theta_z)$ is bounded on H', i.e., we have $F_1(z,\theta) \leq C_1$ for $z \in H'$ and $\theta \in \Theta_a$. Define a function $\lambda : (H_{in} - \{z_0\}) \to \Re^+$ as

$$\lambda(z) \equiv \max\{\beta | z + \beta \cdot (z - z_0) \in H'\}.$$

Let

$$d_1 = \max_{x,y \in H_{in}} |x - y|$$

and

$$d_2 = \inf_{x \in \partial H_{in}, y \in \partial H'} |x - y|.$$

($d_2 \neq 0$ by the definition of H', and $d_1 \neq 0$ holds.) Then, $\lambda(z)|z - z_0| \geq d_2$ holds, i.e., we have $\lambda(z) \geq d_2/d_1$. Let z' denote $z + \lambda(z)(z - z_0)$, then $z' \in H'$ and

$$z = (\lambda(z)z_0 + z')/(1 + \lambda(z))$$

hold. Since $z' \in H'$, we have $F_1(z',\theta) \leq C_1$ for $\theta \in \Theta_a$. Therefore, we have

$$F_1(z,\theta) = F_1(z_0,\theta)^{\lambda(z)/(1+\lambda(z))} F_1(z',\theta)^{1/(1+\lambda(z))}$$
$$\leq F_1(z_0,\theta)^{\lambda(z)/(1+\lambda(z))} C_1^{1/(1+\lambda(z))}.$$

Hence, we have

$$F_1(z,\theta)^{N_0(1+\lambda(z))/\lambda(z)} \leq F_{N_0}(z_0,\theta) C_1^{N_0/\lambda(z)}.$$

Let

$$N_1 \equiv \sup_z N_0(1 + \lambda(z))/\lambda(z) \ (< \infty)$$

and $u(z) \equiv N_1 - N_0(1 + \lambda(z))/\lambda(z)$. Since $u(z) \geq 0$ and $F_1(z,\theta) \leq C_1$ hold, we have $F_1(z,\theta)^{u(z)} \leq C_1^{u(z)}$. Hence, we obtain

$$F_{N_1}(z,\theta) = F_1(z,\theta)^{N_1}$$
$$= F_1(z,\theta)^{N_0(1+\lambda(z))/\lambda(z)} \cdot F_1(z,\theta)^{u(z)}$$
$$\leq F_{N_0}(z_0,\theta) C_1^{N_0/\lambda(z)} \cdot C_1^{u(z)}.$$

Since $F_v(z,\theta) \leq C_1^v$ hold for $v \geq 0$, we have

$$F_{N_1+v}(z,\theta) \leq F_{N_0}(z_0,\theta) C_1^{N_0/\lambda(z)} C_1^{u(z)+v}.$$

Noting that $0 < N_0/\lambda(z) < \infty$ and $0 \leq u(z) \leq N_1$ holds for $z \in Z_{in} - \{z_0\}$, we have

$$F_{N_1+v}(z,\theta) \leq F_{N_0}(z_0,\theta) C_v$$

where C_v is a real determined by v. Since $\xi(\theta)F_{N_0}(z_0,\theta)$ is integrable on Θ, $\xi(\theta)F_{N_1+v}(z,\theta)$ is integrable on Θ for any $v > 0$ and any $z \in H_{in}$. Since the converse is trivial, this implies the proposition. Q.E.D.

Acknowledgment

The author wishes to express his sincere gratitude to the following people: H. Nagaoka, T. Kawabata, N. Abe, T. Okamura, Y. Kubo, A. R. Barron, K. Nakamura, and T. Fujita.

References

[1] S.-I. Amari, *Differential-Geometrical Methods in Statistics* 2nd ed. (Lecture Notes in Statistics, vol. 28). Berlin, Germany: Springer-Verlag, 1990.
[2] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Dept. Elec. Eng., Stanford Univ., Stanford, CA, 1985.
[3] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, 1991.
[4] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. B.*, vol. 41. pp. 113–147, 1979.
[5] L. Brown, *Fundamentals of Statistical Exponential Families*. Inst. Math. Statist., 1986.
[6] B. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.
[7] ——, "Jeffreys prior is asymptotically least favorable under entropy risk," *J. Statist. Planning and Inference*, vol. 41, pp. 37–60, 1994.
[8] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783–795, Nov. 1973.
[9] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, Mar. 1980.
[10] H. Jeffreys, *Theory of Probability*, 3rd ed. Berkeley, CA: Univ. of California Press, 1961.
[11] T. Matsushima, H. Inazumi, and S. Hirasawa, "A class of distortionless codes designed by Bayes decision theory," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1288–1293, 1991.
[12] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.
[13] ——, "Stochastic complexity," *J. Roy. Statist. Soc. B*, vol. 49, pp. 223–239 and 252–265, 1987.
[14] J. Takeuchi and T. Kawabata, "Approximation of Bayes code for Markov sources," in *Proc. 1995 IEEE Int. Symp. on Information Theory*, 1995, p. 391.
[15] C. S. Wallace and D. M. Boulton, "An invariant Bayes method for point estimation," *Classification Soc. Bull.*, vol. 3, pp. 11–34, 1975.
[16] C. S. Wallace and P. R. Freeman, "Estimating and inference by compact coding," *J. Roy. Statist. Soc. B*, vol. 49. no. 3 pp. 240–265, 1987.
[17] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources,"*IEEE Trans. Inform. Theory*, to be published.
[18] K. Yamanishi, "A learning criterion for stochastic rules," *Mach. Learn.* (Special Issue on COLT '90), vol. 9, no. 2/3, pp. 165–203, 1992.
[19] J. Takeuchi, "On the convergence rate of the MDL estimator with respect to the KL-divergence" (in Japanese), in *Proc. 16th Symp. on Information Theory and its Application*, 1993, pp. 161–164.
[20] J. Takeuchi and N. Abe, "Evaluation of Laplace-like estimators in the probabilistic PAC learning model" (in Japanese), IEICE Tech. Rep. IT92-128, 1993-03. pp. 1–6, 1993.