Journal of Machine Learning Research 14 (2013) 867-897 Submitted 8/12; Revised 2/13; Published 3/13

A Widely Applicable Bayesian Information Criterion

Sumio Watanabe SWATANAB@DIS.TITECH.AC.JP
Department of Computational Intelligence and Systems Science
Tokyo Institute of Technology
Mailbox G5-19, 4259 Nagatsuta, Midori-ku
Yokohama, Japan 226-8502

Editor: Manfred Opper

Abstract

A statistical model or a learning machine is called regular if the map taking a parameter to a probability distribution is one-to-one and if its Fisher information matrix is always positive definite. If otherwise, it is called singular. In regular statistical models, the Bayes free energy, which is defined by the minus logarithm of Bayes marginal likelihood, can be asymptotically approximated by the Schwarz Bayes information criterion (BIC), whereas in singular models such approximation does not hold.

Recently, it was proved that the Bayes free energy of a singular model is asymptotically given by a generalized formula using a birational invariant, the real log canonical threshold (RLCT), instead of half the number of parameters in BIC. Theoretical values of RLCTs in several statistical models are now being discovered based on algebraic geometrical methodology. However, it has been difficult to estimate the Bayes free energy using only training samples, because an RLCT depends on an unknown true distribution.

In the present paper, we define a widely applicable Bayesian information criterion (WBIC) by the average log likelihood function over the posterior distribution with the inverse temperature $1/\log n$, where n is the number of training samples. We mathematically prove that WBIC has the same asymptotic expansion as the Bayes free energy, even if a statistical model is singular for or unrealizable by a statistical model. Since WBIC can be numerically calculated without any information about a true distribution, it is a generalized version of BIC onto singular statistical models.

Keywords: Bayes marginal likelihood, widely applicable Bayes information criterion

1. Introduction

A statistical model or a learning machine is called regular if the map taking a parameter to a probability distribution is one-to-one and if its Fisher information matrix is always positive definite. If otherwise, it is called singular. Many statistical models and learning machines are not regular but singular, for example, artificial neural networks, normal mixtures, binomial mixtures, reduced rank regressions, Bayesian networks, and hidden Markov models. In general, if a statistical model contains hierarchical layers, hidden variables, or grammatical rules, then it is singular. In other words, if a statistical model is devised so that it extracts hidden structure from a random phenomenon, then it naturally becomes singular. If a statistical model is singular, then the likelihood function cannot be approximated by any normal distribution, resulting that neither AIC, BIC, nor MDL can be used

in statistical model evaluation. Hence constructing singular learning theory is an important issue in both statistics and learning theory.

A statistical model or a learning machine is represented by a probability density function $p(x|w)$ of $x \in \mathbb{R}^N$ for a given parameter $w \in W \subset \mathbb{R}^d$, where W is a set of all parameters. A prior probability density function is denoted by $\varphi(w)$ on W. Assume that training samples $X_1, X_2, ..., X_n$ are independently subject to a probability density function $q(x)$, which is called a true distribution. The log loss function or the minus log likelihood function is defined by

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|w). \tag{1}$$

Also the Bayes free energy $\mathcal{F}$ is defined by

$$\mathcal{F} = -\log \int \prod_{i=1}^{n} p(X_i|w)\varphi(w)dw. \tag{2}$$

This value $\mathcal{F}$ can be understood as the minus logarithm of marginal likelihood of a model and a prior, hence it plays an important role in statistical model evaluation. In fact, a model or a prior is often optimized by maximization of the Bayes marginal likelihood (Good, 1965), which is equivalent to minimization of the Bayes free energy.

If a statistical model is regular, then the posterior distribution can be asymptotically approximated by a normal distribution, resulting that

$$\mathcal{F} = nL_n(\hat{w}) + \frac{d}{2}\log n + O_p(1), \tag{3}$$

where $\hat{w}$ is the maximum likelihood estimator, d is the dimension of the parameter space, and n is the number of training samples. The right hand side of Equation (3) is the well-known Schwarz Bayesian information criterion (BIC) (Schwarz, 1978).

If a statistical model is singular, then the posterior distribution is different from any normal distribution, hence the Bayes free energy cannot be approximated by BIC in general. Recently, it was proved that, even if a statistical model is singular,

$$\mathcal{F} = nL_n(w_0) + \lambda \log n + O_p(\log\log n),$$

where w_0 is the parameter that minimizes the Kullback-Leibler distance from a true distribution to a statistical model, and $\lambda > 0$ is a rational number called the real log canonical threshold (RLCT) (Watanabe, 1999, 2001a, 2009, 2010a) .

The birational invariant RLCT, which was firstly found by a research of singular Schwartz distribution (Gelfand and Shilov, 1964), plays an important role in algebraic geometry and algebraic analysis (Bernstein, 1972; Sato and Shintani, 1974; Kashiwara, 1976; Varchenko, 1976; Kollár, 1997; Saito, 2007). In algebraic geometry, it represents a relative property of singularities of a pair of algebraic varieties. In statistical learning theory, it shows the asymptotic behaviors of the Bayes free energy and the generalization loss, which are determined by a pair of an optimal parameter set and a parameter set W.

If a triple of a true distribution, a statistical model, and a prior distribution is fixed, then there is an algebraic geometrical procedure which enables us to find an RLCT (Hironaka, 1964). In

fact, RLCTs for several statistical models and learning machines are being discovered. For example, RLCTs have been studied in artificial neural networks (Watanabe, 2001b; Aoyagi and Nagata, 2012), normal mixtures (Yamazaki and Watanabe, 2003), reduced rank regressions (Aoyagi and Watanabe, 2005), Bayes networks with hidden variables (Rusakov and Geiger, 2005; Zwiernik, 2010, 2011), binomial mixtures, Boltzmann machines (Yamazaki and Watanabe, 2005), and hidden Markov models. To study singular statistical models, new algebraic geometrical theory is constructed (Watanabe, 2009; Drton et al., 2009; Lin, 2011; Király et al., 2012).

Based on such researches, the theoretical behavior of the Bayes free energy is clarified. These results are very important because they indicate the quantitative difference of singular models from regular ones. However, in general, an RLCT depends on an unknown true distribution. In practical applications, we do not know a true distribution, hence we cannot directly apply the theoretical results to statistical model evaluation.

In the present paper, in order to estimate the Bayes free energy without any information about a true distribution, we propose a widely applicable Bayesian information criterion (WBIC) by the following definition:

$$\mathrm{WBIC} = \mathbb{E}_w^{\beta}[nL_n(w)], \quad \beta = \frac{1}{\log n}, \tag{4}$$

where $\mathbb{E}_w^{\beta}[\ ]$ shows the expectation value over the posterior distribution on W that is defined by, for an arbitrary integrable function $G(w)$,

$$\mathbb{E}_w^{\beta}[G(w)] = \frac{\int G(w) \prod_{i=1}^{n} p(X_i|w)^{\beta}\, \varphi(w) dw}{\int \prod_{i=1}^{n} p(X_i|w)^{\beta}\, \varphi(w) dw}. \tag{5}$$

In this definition, $\beta > 0$ is called the inverse temperature. Then the main purpose of this paper is to show

$$\mathcal{F} = \mathrm{WBIC} + O_p(\sqrt{\log n}).$$

To establish mathematical support of WBIC, we prove three theorems. Firstly, in Theorem 3 we show that there exists a unique inverse temperature β^* which satisfies

$$\mathcal{F} = \mathbb{E}_w^{\beta^*}[nL_n(w)].$$

The optimal inverse temperature β^* is a random variable which satisfies the convergence in probability, $\beta^* \log n \to 1$ as $n \to \infty$. Secondly, in Theorem 4 we prove that, even if a statistical model is singular,

$$\mathrm{WBIC} = nL_n(w_0) + \lambda \log n + O_p(\sqrt{\log n}).$$

In other words, WBIC has the same asymptotic behavior as the Bayes free energy even if a statistical model is singular. And lastly, in Theorem 5 we prove that, if a statistical model is regular, then

$$\mathrm{WBIC} = nL_n(\hat{w}) + \frac{d}{2}\log n + O_p(1),$$

which shows WBIC coincides with BIC in regular statistical models. Moreover, it is expected that a computational cost in numerical calculation of WBIC is far smaller than that of the Bayes free

Variable	Name	Equation Number
$\mathcal{F}$	Bayes free energy	Equation (2)
$\mathcal{G}$	Generalization loss	Equation (29)
WBIC	WBIC	Equation (4)
WAIC	WAIC	Equation (30)
$\mathbb{E}_w^{\beta}[\ ]$	posterior average	Equation (5)
β^*	optimal inverse temperature	Equation (18)
$L(w)$	log loss function	Equation (6)
$L_n(w)$	empirical loss	Equation (1)
$K(w)$	Average log likelihood ratio	Equation (7)
$K_n(w)$	empirical log likelihood ratio	Equation (8)
λ	real log canonical threshold	Equation (15)
m	multiplicity	Equation (16)
$Q(K(w), \varphi(w))$	parity of model	Equation (17)
$(\mathcal{M}, g(u), a(x,u), b(u))$	resolution quartet	Theorem 1

Table 1: Variable, Name, and Equation Number

energy. These results show that WBIC is a generalized version of BIC onto singular statistical models and that RLCTs can be estimated even if a true distribution is unknown.

This paper consists of eight sections. In Section 2, we summarize several notations. In Section 3, singular learning theory and the standard representation theorem are introduced. The main theorems and corollaries of this paper are explained in Section 4, which are mathematically proved in Section 5. As the purpose of the present paper is to prove the mathematical support of WBIC, Sections 4 and 5 are the main sections. In section 6, a method how to use WBIC in statistical model evaluation is illustrated using an experimental result. In section 7 and 8, we discuss and conclude the present paper.

2. Statistical Models and Notations

In this section, we summarize several notations. Table 1 shows variables, names, and equation numbers in this paper. The average log loss function $L(w)$ and the entropy of the true distribution S are respectively defined by

$$\begin{aligned} L(w) &= -\int q(x) \log p(x|w) dx, \\ S &= -\int q(x) \log q(x) dx. \end{aligned} \tag{6}$$

Then $L(w) = S + D(q||p_w)$, where $D(q||p_w)$ is the Kullback-Leibler distance defined by

$$D(q||p_w) = \int q(x) \log \frac{q(x)}{p(x|w)} dx.$$

Then $D(q||p_w) \geq 0$, hence $L(w) \geq S$. Moreover, $L(w) = S$ if and only if $p(x|w) = q(x)$.

In this paper, we assume that there exists a parameter w_0 in the interior of W which minimizes $L(w)$,

$$L(w_0) = \min_{w \in W} L(w),$$

where the interior of a set S is the defined by the largest open set that is contained in S. Note that such w_0 is not unique in general, because the map $w \mapsto p(x|w)$ is not one-to-one in general in singular statistical models. We also assume that, for an arbitrary w that satisfies $L(w) = L(w_0)$, $p(x|w)$ is the same probability density function. Let $p_0(x)$ be such a unique probability density function. In general, the set

$$W_0 = \{w \in W; p(x|w) = p_0(x)\}$$

is not a set of single element but an analytic set or an algebraic set with singularities. Let us define a log density ratio function,

$$f(x,w) = \log \frac{p_0(x)}{p(x|w)},$$

which is equivalent to

$$p(x|w) = p_0(x) \exp(-f(x,w)).$$

Two functions $K(w)$ and $K_n(w)$ are respectively defined by

$$K(w) = \int q(x) f(x,w) dx, \tag{7}$$

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, w). \tag{8}$$

Then it immediately follows that

$$\begin{aligned} L(w) &= L(w_0) + K(w), \\ L_n(w) &= L_n(w_0) + K_n(w). \end{aligned}$$

The expectation value over all sets of training samples $X_1, X_2, ..., X_n$ is denoted by $\mathbb{E}[\ ]$. For example, $\mathbb{E}[L_n(w)] = L(w)$ and $\mathbb{E}[K_n(w)] = K(w)$. The problem of statistical learning is characterized by the log density ratio function $f(x,w)$. In fact,

$$\mathbb{E}_w^{\beta}[nL_n(w)] = nL_n(w_0) + \mathbb{E}_w^{\beta}[nK_n(w)], \tag{9}$$

$$\mathbb{E}_w^{\beta}[nK_n(w)] = \frac{\int nK_n(w) \exp(-n\beta K_n(w)) \varphi(w) dw}{\int \exp(-n\beta K_n(w)) \varphi(w) dw}. \tag{10}$$

The main purpose of the present paper is to prove

$$\mathcal{F} = nL_n(w_0) + \mathbb{E}_w^{\beta}[nK_n(w)] + O_p(\sqrt{\log n})$$

for $\beta = 1/\log n$.

Definition.

(1) If $q(x) = p_0(x)$, then $q(x)$ is said to be *realizable* by $p(x|w)$. If otherwise, it is said to be *unrealizable*.

(2) If the set W_0 consists of a single element w_0 and if the Hessian matrix

$$J_{ij}(w) = \frac{\partial^2 L}{\partial w_i \partial w_j}(w) \tag{11}$$

at $w = w_0$ is strictly positive definite, then $q(x)$ is said to be *regular* for $p(x|w)$. If otherwise, it is said to be *singular* for $p(x|w)$.

Note that the matrix $J(w)$ is equal to the Hessian matrix of $K(w)$ and that $J(w_0)$ is equal to the Fisher information matrix if the true distribution is realizable by a statistical model. Also note that, if $q(x)$ is realizable by $p(x|w)$, then $K(w)$ is Kullback-Leibler divergence of $q(x)$ and $p(x|w)$. However, if $q(x)$ is not realizable by $p(x|w)$, then it is not.

3. Singular Learning Theory

In this section we summarize singular learning theory. In the present paper, we assume the following conditions.

Fundamental Conditions.

(1) The set of parameters W is a compact set in $\mathbb{R}^d$ whose interior is not the empty set. Its boundary is defined by several analytic functions $\pi_1(w)$, $\pi_2(w)$, ..., $\pi_k(w)$, in other words,

$$W = \{w \in \mathbb{R}^d; \pi_1(w) \geq 0, \pi_2(w) \geq 0, ..., \pi_k(w) \geq 0\}.$$

(2) The prior distribution satisfies $\varphi(w) = \varphi_1(w)\varphi_2(w)$, where $\varphi_1(w) \geq 0$ is an analytic function and $\varphi_2(w) > 0$ is a C^∞-class function.

(3) Let $s \geq 6$ and

$$L^s(q) = \{f(x); \|f\|_s \equiv \Big(\int |f(x)|^s q(x)dx\Big)^{1/s} < \infty\}$$

be a Banach space. There exists an open set $W' \supset W$ such that the map $W' \ni w \mapsto f(x,w)$ is an $L^s(q)$-valued analytic function.

(4) The set W_ε is defined by

$$W_\varepsilon = \{w \in W \ ; \ K(w) \leq \varepsilon\}.$$

It is assumed that there exist constants $\varepsilon, c > 0$ such that

$$(\forall w \in W_\varepsilon) \quad \mathbb{E}_X[f(X,w)] \geq c\, \mathbb{E}_X[f(X,w)^2]. \tag{12}$$

Remark. (1) These conditions allow that the set of optimal parameters

$$W_0 = \{w \in W \ ; \ p(x|w) = p(x|w_0)\} = \{w \in W \ ; \ K(w) = 0\}$$

may contain singularities, and that the Hessian matrix $J(w)$ at $w \in W_0$ is not positive definite. Therefore $K(w)$ can not be approximated by any quadratic form in general.

(2) The condition Equation (12) is satisfied if a true distribution is realizable by or regular for a statistical model (Watanabe, 2010a). If a true distribution is unrealizable by and singular for a statistical model, there is an example which does not satisfy this condition. In the present paper, we study the case when Equation (12) is satisfied.

Lemma 1 *Assume Fundamental Conditions (1)-(4). Let*

$$\beta = \frac{\beta_0}{\log n},$$

where $\beta_0 > 0$ is a constant and let $0 \leq r < 1/2$. Then, as $n \to \infty$,

$$\int_{K(w) \geq 1/n^r} \exp(-n\beta K_n(w))\varphi(w)dw = o_p(\exp(-\sqrt{n})), \tag{13}$$

$$\int_{K(w) \geq 1/n^r} nK_n(w)\exp(-n\beta K_n(w))\varphi(w)dw = o_p(\exp(-\sqrt{n})). \tag{14}$$

The proof of Lemma 1 is given in Section 5.

Let $\varepsilon > 0$ be a sufficiently small constant. Lemma 1 shows that integrals outside of the region W_ε do not affect the expectation value $\mathbb{E}_w^\beta[nK_n(w)]$ asymptotically, because in the following theorems, we prove that integral in the region W_ε have larger orders than the integral outside of W_ε. To study integrals in the region W_ε, we need algebraic geometrical method, because the set $\{w; K(w) = 0\}$ contains singularities in general. There are quite many kinds of singularities, however, the following theorem makes any singularities be a same standard form.

Theorem 1 *(Standard Representation) Assume Fundamental Conditions (1)-(4). Let $\varepsilon > 0$ be a sufficiently small constant. Then there exists an quartet $(\mathcal{M}, g(u), a(x,u), b(u))$, where*
(1) $\mathcal{M}$ is a d-dimensional real analytic manifold,
(2) g is a proper analytic function $g : \mathcal{M} \to W'_\varepsilon$, where W'_ε is the set that is defined by the largest open set contained in W_ε and $g : \{u \in \mathcal{M}; K(g(u)) \neq 0\} \to \{w \in W'_\varepsilon; K(w) \neq 0\}$ is a bijective map,
(3) $a(x,u)$ is an $L^s(q)$-valued analytic function,
(4) and $b(u)$ is an infinitely many times differentiable function which satisfies $b(u) > 0$,
such that the following equations are satisfied in each local coordinate of $\mathcal{M}$:

$$\begin{aligned} K(g(u)) &= u^{2k}, \\ f(x, g(u)) &= u^k a(x,u), \\ \varphi(w)dw &= \varphi(g(u))|g'(u)|du = b(u)|u^h|du, \end{aligned}$$

where $k = (k_1, k_2, ..., k_d)$ and $h = (h_1, h_2, ..., h_d)$ are multi-indices made of nonnegative integers. At least one of k_j is not equal to zero.

Remark. (1) In this theorem, for $u = (u_1, u_2, \cdots, u_d) \in \mathbb{R}^d$, notations u^{2k} and $|u^h|$ respectively represent

$$\begin{aligned} u^{2k} &= u_1^{2k_1} u_2^{2k_2} \cdots u_d^{2k_d}, \\ |u^h| &= |u_1^{h_1} u_2^{h_2} \cdots u_d^{h_d}|. \end{aligned}$$

The singularity $u = 0$ in $u^{2k} = 0$ is said to be normal crossing. Theorem 1 shows that any singularities can be made normal crossing by using an analytic function $w = g(u)$.
(2) A map $w = g(u)$ is said to be proper if, for an arbitrary compact set C, $g^{-1}(C)$ is also compact.
(3) The proof of Theorem 1 is given in Theorem 6.1 of a book (Watanabe, 2009, 2010a). In order to prove this theorem, we need the Hironaka resolution of singularities (Hironaka, 1964; Atiyah, 1970) that is the fundamental theorem in algebraic geometry. The function $w = g(u)$ is often referred to as a resolution map.
(4) In this theorem, a quartet $(k, h, a(x,u), b(u))$ depends on a local coordinate in general. For a given function $K(w)$, there is an algebraic recursive algorithm which enables us to find a resolution

map $w = g(u)$. However, even for a fixed $K(w)$, a resolution map is not unique, resulting that a quartet $(\mathcal{M}, g(u), a(x,u), b(u))$ is not unique.

Definition. (Real Log Canonical Threshold) Let $\{\mathcal{U}_\alpha; \alpha \in \mathcal{A}\}$ be a system of local coordinates of a manifold $\mathcal{M}$,

$$\mathcal{M} = \bigcup_{\alpha \in \mathcal{A}} \mathcal{U}_\alpha.$$

The real log canonical threshold (RLCT) is defined by

$$\lambda = \min_{\alpha \in \mathcal{A}} \min_{j=1}^{d} \Big(\frac{h_j + 1}{2k_j} \Big), \tag{15}$$

where we define $1/k_j = \infty$ if $k_j = 0$. The multiplicity m is defined by

$$m = \max_{\alpha \in \mathcal{A}} \#\Big\{ j; \frac{h_j + 1}{2k_j} = \lambda \Big\}, \tag{16}$$

where $\#S$ shows the number of elements of a set S.

The concept of RLCT is well known in algebraic geometry and statistical learning theory. In the following definition we introduce a parity of a statistical model.

Definition. (Parity of Statistical Model) The support of $\varphi(g(u))$ is defined by

$$\mathrm{supp}\ \varphi(g(u)) = \overline{\{u \in \mathcal{M}\ ;\ g(u) \in W_\varepsilon,\ \ \varphi(g(u)) > 0\}},$$

where $\overline{S}$ shows the closure of a set S. A local coordinate $\mathcal{U}_\alpha$ is said to be an essential local coordinate if both equations

$$\begin{aligned} \lambda &= \min_{j=1}^{d} \Big(\frac{h_j + 1}{2k_j} \Big), \\ m &= \#\{ j; (h_j + 1)/(2k_j) = \lambda \}, \end{aligned}$$

hold in its local coordinate. The set of all essential local coordinates is denoted by $\{\mathcal{U}_\alpha; \alpha \in \mathcal{A}^*\}$. If, for an arbitrary essential local coordinate, there exist both $\delta > 0$ and a natural number j in the set $\{ j\ ;\ (h_j + 1)/(2k_j) = \lambda \}$ such that
(1) k_j is an odd number,
(2) $\{(0,0,..,0,u_j,0,0,..,0)\ ;\ |u_j| < \delta\} \subset \mathrm{supp}\ \varphi(g(u))$,
then we define $Q(K(g(u)), \varphi(g(u))) = 1$. If otherwise, $Q(K(g(u)), \varphi(g(u))) = 0$. If there exists a resolution map $w = g(u)$ such that $Q(K(g(u)), \varphi(g(u))) = 1$, then we define

$$Q(K(w), \varphi(w)) = 1. \tag{17}$$

If otherwise $Q(K(w), \varphi(w)) = 0$. If $Q(K(w), \varphi(w)) = 1$, then the parity of a statistical model is said to be odd, otherwise even.

It was proved in Theorem 2.4 of a book (Watanabe, 2009) that, for a given set (q, p, φ), λ and m are independent of a choice of a resolution map. Such a value is called a birational invariant. The RLCT is a birational invariant.

Lemma 2 *If a true distribution $q(x)$ is realizable by a statistical model $p(x|w)$, then the value $Q(K(g(u)), \varphi(g(u)))$ is independent of a choice of a resolution map $w = g(u)$.*

Proof of this lemma is shown in Section 5. Lemma 2 indicates that, if a true distribution is realizable by a statistical model, then $Q(K(g(u)),\varphi(g(u)))$ is a birational invariant. The present paper proposes a conjecture that $Q(K(g(u)),\varphi(g(u)))$ is a birational invariant in general. By Lemma 2, this conjecture is proved if we can show the proposition that, for an arbitrary nonnegative analytic function $K(w)$, there exist $q(x)$ and $p(x|w)$ such that $K(w)$ is the Kullback-Leibler distance from $q(x)$ to $p(x|w)$.

Example. Let $w=(a,b,c)\in\mathbb{R}^3$ and

$$K(w)=(ab+c)^2+a^2b^4,$$

which is the Kullback-Leibler distance of a neural network model in Example 1.6 of a book (Watanabe, 2009), where a true distribution is realizable by a statistical model. The prior $\varphi(w)$ is defined by some nonzero function on a sufficiently large compact set. In a singular statistical model, compactness of the prior is necessary in general, because, if the parameter set is not compact, then it is not easy to mathematically treat the integration on the neighborhood among the infinite point.

Let a system of local coordinates be

$$\mathcal{U}_i=\{(a_i,b_i,c_i)\in\mathbb{R}^3\}\quad (i=1,2,3,4).$$

A resolution map $g:\mathcal{U}_1\cup\mathcal{U}_2\cup\mathcal{U}_3\cup\mathcal{U}_4\to\mathbb{R}^3$ in each local coordinate is defined by

$$\begin{aligned}
a=a_1c_1,\quad & b=b_1,\quad & c=c_1,\\
a=a_2,\quad & b=b_2c_2,\quad & c=a_2(1-b_2)c_2,\\
a=a_3,\quad & b=b_3,\quad & c=a_3b_3(b_3c_3-1),\\
a=a_4,\quad & b=b_4c_4,\quad & c=a_4b_4c_4(c_4-1).
\end{aligned}$$

This map g is made of recursive blowing-ups whose centers are smooth manifolds, hence it is one-to-one as a map $g:\{u;K(g(u))>0\}\to\{w;K(w)>0\}$. Then

$$\begin{aligned}
K(a,b,c) &= c_1^2\{(a_1b_1+1)^2+a_1^2b_1^4\}=a_2^2c_2^2(1+b_2^2c_2^2)\\
&= a_3^2b_3^4(c_3^2+1)=a_4^2b_4^2c_4^4(1+b_4^2).
\end{aligned}$$

Therefore integration over W can be calculated using integration over the manifold. The Jacobian determinant $|g'(u)|$ is

$$\begin{aligned}
|g'(u)| &= |c_1|=|a_2c_2|\\
&= |a_3b_3^2|=|a_4b_4c_4|^2.
\end{aligned}$$

In other words, in each local coordinate,

$$\begin{aligned}
(k_1,k_2,k_3) &= (0,0,1),(1,0,1),(1,2,0),(1,1,2),\\
(h_1,h_2,h_3) &= (0,0,1),(1,0,1),(1,2,0),(2,2,2).
\end{aligned}$$

Therefore

$$\Big(\frac{h_1+1}{2k_1},\frac{h_2+1}{2k_2},\frac{h_2+1}{2k_2}\Big)=(\infty,\infty,1),(1,0,1),(1,\frac{3}{4},\infty),(\frac{3}{2},\frac{3}{2},\frac{3}{4}).$$

The smallest value among them is $3/4$ which is equal to λ and the multiplicty is $m=1$. The essential local coordinates are $\mathcal{U}_3$ and $\mathcal{U}_4$. In $\mathcal{U}_3$ and $\mathcal{U}_4$, the sets $\{u_j\,;\,(h_j+1)/(2k_j)=3/4\}$ are respectively $\{u_2\}$ and $\{u_3\}$, where $2k_j=4$ in both cases. Consequently, both k_j are even, thus the parity is given by $Q(K(w),\varphi(w))=0$.

Lemma 3 *Assume that the Fundamental Conditions (1)-(4) are satisfied and that a true distribution $q(x)$ is regular for a statistical model $p(x|w)$. If w_0 is contained in the interior of W and if $\varphi(w_0) > 0$, then*

$$\lambda = \frac{d}{2}, \quad m = 1,$$

and

$$Q(K(w), \varphi(w)) = 1.$$

Proof of this lemma is shown in Section 5.

Theorem 2 *Assume that the Fundamental Conditions (1)-(4) are satisfied. Then the following holds.*

$$\mathcal{F} = nL_n(w_0) + \lambda \log n - (m-1)\log\log n + R_n,$$

where λ is a real log canonical threshold, m is its multiplicity, and $\{R_n\}$ is a sequence of random variables which converges to a random variable in law, when $n \to \infty$.

Theorem 2 was proved in the previous papers. In the case when $q(x)$ is realizable by and singular for $p(x|w)$, the expectation value of $\mathcal{F}$ is given by algebraic analysis (Watanabe, 2001a). The asymptotic behavior of $\mathcal{F}$ as a random variable was shown in a book (Watanabe, 2009). These results were generalized (Watanabe, 2010a) for the case that $q(x)$ is unrealizable.

Remark. In practical applications, we do not know the true distribution, hence λ and m are unknown. Therefore, we can not directly apply Theorem 2 to such cases. The main purpose of the present paper is to make a new method how to estimate $\mathcal{F}$ even if the true distribution is unknown.

4. Main Results

In this section, we introduce the main results of the present paper.

Theorem 3 (*Unique Existence of the Optimal Parameter*) *Assume that $L_n(w)$ is not a constant function of w. Then the followings hold.*
(1) The value $\mathbb{E}_w^{\beta}[nL_n(w)]$ is a decreasing function of β.
(2) There exists a unique β^* $(0 < \beta^* < 1)$ which satisfies

$$\mathcal{F} = \mathbb{E}_w^{\beta^*}[nL_n(w)]. \tag{18}$$

Note that the function $L_n(w)$ is not a constant function in an ordinary statistical model with probability one. The Proof of Theorem 3 is given in Section 5. Based on this theorem, we define the optimal inverse temperature.

Definition. The unique parameter β^* that satisfies Equation (18) is called the optimal inverse temperature.

In general, the optimal inverse temperature β^* depends on a true distribution $q(x)$, a statistical model $p(x|w)$, a prior $\varphi(w)$, and training samples. Therefore β^* is a random variable. In the present paper, we study its probabilistic behavior. Theorem 4 is a mathematical base for such a purpose.

Theorem 4 (*Main Theorem*) *Assume Fundamental Conditions (1)-(4) and that*

$$\beta = \frac{\beta_0}{\log n},$$

where β_0 is a constant. Then there exists a random variable U_n such that

$$\mathbb{E}_w^{\beta}[nL_n(w)] = nL_n(w_0) + \frac{\lambda \log n}{\beta_0} + U_n\sqrt{\frac{\lambda \log n}{2\beta_0}} + O_p(1),$$

where λ is the real log canonical threshold and $\{U_n\}$ is a sequence of random variables, which satisfies $\mathbb{E}[U_n] = 0$, converges to a Gaussian random variable in law as $n \to \infty$. Moreover, if a true distribution $q(x)$ is realizable by a statistical model $p(x|w)$, then $\mathbb{E}[(U_n)^2] < 1$.

The proof of Theorem 4 is given in Section 5. Theorem 4 with $\beta_0 = 1$ shows that

$$\mathrm{WBIC} = nL_n(w_0) + \lambda \log n + U_n\sqrt{\frac{\lambda \log n}{2}} + O_p(1),$$

whose first two main terms are equal to those of $\mathcal{F}$ in Theorem 2. From Theorem 4 and its proof, three important corollaries are derived.

Corollary 1 *If the parity of a statistical model is odd, $Q(K(w), \varphi(w)) = 1$, then $U_n = 0$.*

Corollary 2 *Let β^* be the optimal inverse temperature. Then*

$$\beta^* = \frac{1}{\log n}\Big(1 + \frac{U_n}{\sqrt{2\lambda \log n}} + o_p\Big(\frac{1}{\sqrt{\log n}}\Big)\Big).$$

Corollary 3 *Let $\beta_1 = \beta_{01}/\log n$ and $\beta_2 = \beta_{02}/\log n$, where β_{01} and β_{02} are positive constants. Then the convergence in probability*

$$\frac{\mathbb{E}_w^{\beta_1}[nL_n(w)] - \mathbb{E}_w^{\beta_2}[nL_n(w)]}{1/\beta_1 - 1/\beta_2} \to \lambda \tag{19}$$

holds as $n \to \infty$, where λ is the real log canonical threshold.

Proofs of these corollaries are given in Section 5. Note that, if the expectation value $\mathbb{E}_w^{\beta_1}[\ ]$ is calculated by some numerical method, then $\mathbb{E}_w^{\beta_2}[\ ]$ can be estimated using $\mathbb{E}_w^{\beta_1}[\ ]$ by using

$$\mathbb{E}_w^{\beta_2}[nL_n(w)] = \frac{\mathbb{E}_w^{\beta_1}[nL_n(w)\exp(-(\beta_2 - \beta_1)nL_n(w))]}{\mathbb{E}_w^{\beta_1}[\exp(-(\beta_2 - \beta_1)nL_n(w))]}. \tag{20}$$

Therefore RLCT can be estimated by the same computational cost as WBIC. In Bayes estimation, the posterior distribution is often approximated by some numerical method. If we know theoretical values of RLCTs, then we can confirm the approximated posterior distribution by comparing theoretical values with estimated ones.

The well-known Schwarz BIC is defined by

$$\mathrm{BIC} = nL_n(\hat{w}) + \frac{d}{2}\log n,$$

where $\hat{w}$ is the maximum likelihood estimator. WBIC can be understood as the generalized BIC onto singular statistical models, because it satisfies the following theorem.

Theorem 5 *If a true distribution $q(x)$ is regular for a statistical model $p(x|w)$, then*

$$\mathrm{WBIC} = nL_n(\hat{w}) + \frac{d}{2}\log n + o_p(1).$$

Proof of Theorem 5 is given in Section 5. This theorem shows that the difference of WBIC and BIC is smaller than a constant order term, if a true distribution is regular for a statistical model. This theorem holds even if a true distribution $q(x)$ is unrealizable by $p(x|w)$.

Remark. Since the set of parameters W was assumed to be compact, it is proved in Main Theorem 6.4 of a book (Watanabe, 2009) that $nL_n(w_0) - nL_n(\hat{w})$ is a constant order random variable in general. If a true distribution is regular for and realizable by a statistical model, its average is asymptotically equal to $d/2$, where d is the dimension of parameter. If a true distribution is singular for a statistical model, then it is sometimes much larger than $d/2$, because it is asymptotically equal to the maximum value of the Gaussian process. Whether replacement of $nL_n(w_0)$ by $nL(\hat{w})$ is appropriate or not depends on the statistical model and its singularities (Drton, 2009).

5. Proofs of Main Results

In this section, we prove the main theorems and corollaries.

5.1 Proof of Lemma 1

Let us define an empirical process,

$$\eta_n(w) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(K(w) - f(X_i, w)).$$

It was proved in Theorem 5.9 and 5.10 of a book (Watanabe, 2009) that $\eta_n(w)$ converges to a random process in law and

$$\|\eta_n\| \equiv \sup_{w\in W}|\eta_n(w)|$$

also converges to a random variable in law. If $K(w) \geq 1/n^r$, then

$$\begin{aligned} nK_n(w) &= nK(w) - \sqrt{n}\,\eta_n(w) \\ &\geq n^{1-r} - \sqrt{n}\,\|\eta_n\|. \end{aligned}$$

By the condition $1-r > 1/2$ and $\beta = \beta_0/\log n$,

$$\begin{aligned} \exp(\sqrt{n})\int_{K(w)\geq 1/n^r}\exp(-n\beta K_n(w))\varphi(w)dw \\ \leq \exp(-n^{1-r}\beta + \sqrt{n} + \sqrt{n}\beta\|\eta_n\|), \end{aligned}$$

which converges to zero in probability, which shows Equation (13). Then, let us prove Equation (14). Since the set of parameter W is compact, $\|K\| \equiv \sup_w K(w) < \infty$. Therefore,

$$\begin{aligned} |nK_n(w)| &\leq n\|K\| + \sqrt{n}\|\eta_n\| \\ &= n\,(\|K\| + \|\eta_n\|/\sqrt{n}). \end{aligned}$$

Hence

$$\begin{aligned}
&\exp(\sqrt{n})\int_{K(w)\geq 1/n^r}|nK_n(w)|\exp(-n\beta K_n(w))\varphi(w)dw\\
&\leq (\|K\|+\|\eta_n\|/\sqrt{n})\\
&\times\exp(-n^{1-r}\beta+\sqrt{n}+\sqrt{n}\beta\|\eta_n\|+\log n),
\end{aligned}$$

which converges to zero in probability. (Q.E.D.)

5.2 Proof of Lemma 3

Without loss of generality, we can assume $w_0=0$. Since $q(x)$ is regular for $p(x|w)$, there exists w^* such that

$$K(w)=\frac{1}{2}w\cdot J(w^*)w,$$

where $J(w)$ is given in Equation (11). Since $J(w_0)$ is a strictly positive definite matrix, there exists $\varepsilon>0$ such that, if $K(w)\leq\varepsilon$, then $J(w^*)$ is positive definite. Let ℓ_1 and ℓ_2 be respectively the minimum and maximum eigen values of $\{J(w^*);K(w)\leq\varepsilon\}$. Then

$$\frac{1}{4}\ell_1\sum_{j=1}^d w_j^2\leq\frac{1}{2}w\cdot J(w^*)w\leq\ell_2\sum_{j=1}^d w_j^2.$$

By using a blow-up $g:\mathcal{U}_1\cup\cdots\cup\mathcal{U}_d\to W$ which is represented on each local coordinate $\mathcal{U}_i=(u_{i1},u_{i2},...,u_{id})$,

$$\begin{aligned}
w_i &= u_{ii},\\
w_j &= u_{ii}u_{ij}\quad(j\neq i),
\end{aligned}$$

it follows that

$$\frac{\ell_1\,u_{ii}^2}{4}(1+\sum_{j\neq i}u_{ij}^2)\leq\frac{u_{ii}^2}{2}(\hat{u},J(w^*)\hat{u})\leq\ell_2\,u_{ii}^2(1+\sum_{j\neq i}u_{ij}^2),$$

where $\hat{u}_{ij}=u_{ij}$ $(j\neq i)$ and $\hat{u}_{ii}=1$. These inequalities show that $k_i=1$ in $\mathcal{U}_i$, therefore $Q(K(w),\varphi(w))=1$. The Jacobian determinant of the blow-up is

$$|g'(u)|=|u_{ii}|^{d-1},$$

hence $\lambda=d/2$ and $m=1$. (Q.E.D.)

5.3 Proof of Theorem 3

Let us define a function $F_n(\beta)$ of $\beta>0$ by

$$F_n(\beta)=-\log\int\prod_{i=1}^n p(X_i|w)^\beta\varphi(w)dw.$$

Then, by the definition, $\mathcal{F}=F_n(1)$ and

$$\begin{aligned}
F_n'(\beta) &= \mathbb{E}_w^\beta[nL_n(w)],\\
F_n''(\beta) &= -\mathbb{E}_w^\beta[(nL_n(w))^2]+\mathbb{E}_w^\beta[nL_n(w)]^2.
\end{aligned}$$

By the Cauchy-Schwarz inequality and the assumption that $L_n(w)$ is not a constant function,

$$F_n''(\beta) < 0,$$

which shows (1). Since $F_n(0) = 0$,

$$\mathcal{F} = F_n(1) = \int_0^1 F_n'(\beta) d\beta.$$

By using the mean value theorem, there exists β^* $(0 < \beta^* < 1)$ such that

$$\mathcal{F} = F_n'(\beta^*) = \mathbb{E}_w^{\beta^*}[nL_n(w)].$$

Here $F_n'(\beta)$ is a decreasing function, β^* is unique, which completes Theorem 3. (Q.E.D.)

5.4 First Preparation for Proof of Theorem 4

In this subsection, we prepare the proof of Theorem 4. By using Equation (9) and Equation (10), the proof of Theorem 4 results in evaluating $E_w^{\beta}[nK_n(w)]$. By Lemma 1,

$$E_w^{\beta}[nK_n(w)] = \frac{B_n + o_p(\exp(-\sqrt{n}))}{A_n + o_p(\exp(-\sqrt{n}))}, \tag{21}$$

where A_n and B_n are respectively defined by

$$A_n = \int_{K(w)<\varepsilon} \exp(-n\beta K_n(w))\varphi(w) dw, \tag{22}$$

$$B_n = \int_{K(w)<\varepsilon} nK_n(w) \exp(-n\beta K_n(w))\varphi(w) dw. \tag{23}$$

By Theorem 1, an integral over $\{w \in W; K(w) < \varepsilon\}$ is equal to that over $\mathcal{M}$. For a given set of local coordinates $\{\mathcal{U}_\alpha\}$ of $\mathcal{M}$, there exists a set of C^∞ class functions $\{\varphi_\alpha(g(u))\}$ such that, for an arbitrary $u \in \mathcal{M}$,

$$\sum_{\alpha \in \mathcal{A}} \varphi_\alpha(g(u)) = \varphi(g(u)).$$

By using this fact, for an arbitrary integrable function $G(w)$,

$$\int_{K(w)<\varepsilon} G(w)\varphi(w) dw = \sum_{\alpha \in \mathcal{A}} \int_{\mathcal{U}_\alpha} G(g(u))\varphi_\alpha(g(u))|g'(u)| du.$$

Without loss of generality, we can assume that $\overline{\mathcal{U}_\alpha \cap \mathrm{supp}\, \varphi(g(u))}$ is isomorphic to $[-1,1]^d$. Moreover, by Theorem 1, there exists a function $b_\alpha(u) > 0$ such that

$$\varphi_\alpha(g(u))|g'(u)| = |u^h| b_\alpha(u),$$

in each local coordinate. Consequently,

$$\int_{K(w)<\varepsilon} G(w)\varphi(w) dw = \sum_{\alpha \in \mathcal{A}} \int_{[-1,1]^d} du\, G(g(u))\, |u^h|\, b_\alpha(u).$$

In each local coordinate,

$$K(g(u)) = u^{2k}.$$

We define a function $\xi_n(u)$ by

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{u^k - a(X_i, u)\}.$$

Then

$$K_n(g(u)) = u^{2k} - \frac{1}{\sqrt{n}} u^k \xi_n(u).$$

Note that

$$u^k = \int a(x,u) q(x) dx$$

holds, because

$$u^{2k} = K(g(u)) = \int f(x, g(u)) q(x) dx = u^k \int a(x,u) q(x) dx.$$

Therefore, for an arbitrary u,

$$\mathbb{E}[\xi_n(u)] = 0.$$

The function $\xi_n(u)$ can be understood as a random process on $\mathcal{M}$. On Fundamental Conditions (1)-(4), it is proved in Theorem 6.1, Theorem 6.2, and Theorem 6.3 of a book (Watanabe, 2009) that (1) $\xi_n(u)$ converges to a Gaussian random process $\xi(u)$ in law and

$$\mathbb{E}[\sup_u \xi_n(u)^2] \to \mathbb{E}[\sup_u \xi(u)^2].$$

(2) If $q(x)$ is realizable by $p(x|w)$, and if $u^{2k} = 0$, then

$$\mathbb{E}[\xi_n(u)^2] = \mathbb{E}_X[a(X,u)^2] = 2. \tag{24}$$

By using the random process $\xi_n(u)$, the two random variables A_n and B_n can be represented by integrals over $\mathcal{M}$,

$$\begin{aligned} A_n &= \sum_{\alpha \in \mathcal{A}} \int_{[-1,1]^d} du \, \exp(-n\beta u^{2k} + \sqrt{n}\beta u^k \xi_n(u)) |u^h| b_\alpha(u), \\ B_n &= \sum_{\alpha \in \mathcal{A}} \int_{[-1,1]^d} du \, (nu^{2k} - \sqrt{n} u^k \xi_n(u)) \\ &\quad \times \exp(-n\beta u^{2k} + \sqrt{n}\beta u^k \xi_n(u)) |u^h| b_\alpha(u) \end{aligned}$$

To prove Theorem 4, we study asymptotics of these two random variables.

5.5 Second Preparation for Proof of Theorem 4

To evaluate two integrals A_n and B_n as $n \to \infty$, we have to study the asymptotic behavior of the following Schwartz distribution,

$$\delta(t - u^{2k}) \, |u|^h$$

for $t \to 0$. Without loss of generality, we can assume that, in each essential local coordinate,

$$\lambda = \frac{h_1+1}{2k_1} = \frac{h_2+1}{2k_2} = \cdots = \frac{h_m+1}{2k_m} < \frac{h_j+1}{2k_j},$$

for an arbitrary j such that $m < j \leq d$. A variable $u \in \mathbb{R}^d$ is denoted by

$$u = (u_a, u_b) \in \mathbb{R}^m \times \mathbb{R}^{d-m}.$$

We define a measure du^* by

$$du^* = \frac{(\prod_{j=1}^{m} \delta(u_j))\ (\prod_{j=m+1}^{d} (u_j)^{\mu_j})\ du}{2^m\ (m-1)!\ (\prod_{j=1}^{m} k_j)}, \tag{25}$$

where $\delta(\ )$ is the Dirac delta function, and $\mu = (\mu_{m+1}, \mu_2, ..., \mu_d)$ is a multi-index defined by

$$\mu_j = -2\lambda k_j + h_j \quad (m+1 \leq j \leq d).$$

Then $\mu_j > -1$, hence Equation (25) defines a measure on $\mathcal{M}$. The support of du^* is $\{u = (u_a, u_b)\ ;\ u_a = 0\}$.

Definition. Let σ be a d-dimensional variable made of ± 1. We use the notation,

$$\sigma = (\sigma_1, \sigma_2, ..., \sigma_d) \in \mathbb{R}^d$$

where $\sigma_j = \pm 1$. The set of all such variables is denoted by $S(d)$.

$$S(d) = \{\sigma\ ;\ \sigma_j = \pm 1\ (1 \leq j \leq d)\}.$$

Also we use the notation

$$\sigma u = (\sigma_1 u_1, \sigma_2 u_2, ..., \sigma_d u_d) \in \mathbb{R}^d.$$

Then $(\sigma u)^k = \sigma^k u^k$ and $(\sigma u)^{2k} = u^{2k}$. By using this notation, we can derive the asymptotic behavior of $\delta(t - u^{2k})|u^h|$ for $t \to 0$.

Lemma 4 *Let $G(u^{2k}, u^k, u)$ be a real-valued C_1-class function of (u^{2k}, u^k, u) $(u \in \mathbb{R}^d)$. The following asymptotic expansion holds as $t \to +0$,*

$$\begin{aligned}
&\int_{[-1,1]^d} du\ \delta(t-u^{2k})|u|^h G(u^{2k}, u^k, u) \\
&= t^{\lambda-1}(-\log t)^{m-1} \sum_{\sigma \in S(d)} \int_{[0,1]^d} du^*\ G(t, \sigma^k \sqrt{t}, u) \\
&+ O\Big(t^{\lambda-1}(-\log t)^{m-2}\Big),
\end{aligned} \tag{26}$$

where du^* is a measure defined by Equation (25).

(Proof of Lemma 4) Let $Y(t)$ be the left hand side of Equation (26). Then

$$\begin{aligned} Y(t) &= \sum_{\sigma\in S(d)} \int_{[0,1]^d} \delta(t-(\sigma u)^{2k})|\sigma u|^h G((\sigma u)^{2k},(\sigma u)^k,\sigma u)d(\sigma u) \\ &= \sum_{\sigma\in S(d)} \int_{[0,1]^d} \delta(t-u^{2k})|u|^h G(t,\sigma^k\sqrt{t},u)du. \end{aligned}$$

By using Theorem 4.6 of a book (Watanabe, 2009), if $u\in[0,1]^d$, then

$$\begin{aligned} \delta(t-u^{2k})|u|^h du &= t^{\lambda-1}(-\log t)^{m-1}du^* \\ &\quad +O(t^{\lambda-1}(-\log t)^{m-2}). \end{aligned}$$

By applying this relation to $Y(t)$, we obtain Lemma 4. (Q.E.D.)

5.6 Proof of Lemma 2

Let $\Phi(w)>0$ be an arbitrary C^∞ class function on W_ε. Let $Y(t,\Phi)$ $(t>0)$ be a function defined by

$$Y(t,\Phi)\equiv\int_{K(w)<\varepsilon}\delta(t-K(w))f(x,w)\Phi(w)\varphi(w)dw,$$

whose value is independent of a choice of a resolution map. By using a resolution map $w=g(u)$,

$$Y(t,\Phi)=\sum_{\alpha\in\mathcal{A}}\sum_{\sigma\in S(d)}\int_{[-1,1]^d}du\,\delta(t-u^{2k})\,u^k\,|u|^h a(x,u)\Phi(g(u))b_\alpha(u)du.$$

By Lemma 4, and $\sigma=(\sigma_a,\sigma_b)$,

$$\begin{aligned} Y(t,\Phi) &= t^{\lambda-1/2}(\log t)^{m-1}\sum_{\alpha\in\mathcal{A}^*}\sum_{\sigma_a\in S(m)}(\sigma_a)^k\sum_{\sigma_b\in S(d-m)}(\sigma_b)^k \\ &\quad\times\int_{[0,1]^d}du^*\,a(x,\sigma u)\,\Phi(g(\sigma u))\,b_\alpha(\sigma u) \\ &\quad +O(t^{\lambda-1/2}(\log t)^{m-2}). \end{aligned}$$

By the assumption that a true distribution is realizable by a statistical model, Equation (24) shows that there exists x such that $a(x,u)\neq 0$ for $u^{2k}=0$. On the support of du^*,

$$\sigma u=(\sigma_a u_a,\sigma_b u_b)=(0,\sigma_b u_b),$$

consequently the main order term of $Y(t,\Phi)$ is determined by $\Phi(0,u_b)$.

In order to prove Lemma, it is sufficient to prove that $Q(K(g(u)),\varphi(g(u)))=1$ is equivalent to the proposition that, for an arbitrary Φ, the main term of $Y(t,\Phi)$ is equal to zero.

First, assume that $Q(K(g(u)),\varphi(g(u)))=1$. Then at least one k_j $(1\leq j\leq m)$ is odd, σ_a^k takes both values ±1, hence

$$\sum_{\sigma_a\in S(m)}\sigma_a^k=0,$$

which shows that the coefficient of the main order term in $Y(t,\Phi)$ $(t\to+0)$ is zero for an arbitrary $\Phi(w)$. Second, assume that $Q(K(g(u)),\varphi(g(u)))=0$. Then all k_j $(1\leq j\leq m)$ are even, hence

$$\sum_{\sigma_a\in S(m)}\sigma_a^k=\sum_{\sigma_a\in S(m)}1\neq 0.$$

Then there exists a function $\Phi(w)$ such that the main order term is not equal to zero. Therefore $Q(K(g(u)),\varphi(g(u)))$ does not depend on the resolution map. (Q.E.D.)

5.7 Proof of Theorem 4

In this subsection, we prove Theorem 4 using the foregoing preparations. We need to study A_n and B_n in Equation (22) and Equation (23). Firstly, we study A_n.

$$\begin{aligned} A_n &= \sum_{\alpha\in\mathcal{A}}\int_{[-1,1]^d} du\ \exp(-n\beta u^{2k}+\beta\sqrt{n}u^k\xi_n(u))|u|^h b_\alpha(u) \\ &= \sum_{\alpha\in\mathcal{A}}\int_{[-1,1]^d} du \int_0^\infty dt\ \delta(t-u^{2k})|u|^h b_\alpha(u) \\ &\quad \times\exp(-n\beta u^{2k}+\beta\sqrt{n}u^k\xi_n(u)). \end{aligned}$$

By substitution $t := t/(n\beta)$ and $dt := dt/(n\beta)$,

$$\begin{aligned} A_n &= \sum_{\alpha\in\mathcal{A}}\int_{[-1,1]^d} b_\alpha(u)du \int_0^\infty \frac{dt}{n\beta}\delta\Big(\frac{t}{n\beta}-u^{2k}\Big)|u|^h \\ &\quad \times\exp(-n\beta u^{2k}+\beta\sqrt{n}u^k\xi_n(u)). \end{aligned}$$

For simple notations, we use

$$\begin{aligned} \int_{\mathcal{M}} du^* &\equiv \sum_{\alpha\in\mathcal{A}^*}\sum_{\sigma\in S(d)}\int_{[0,1]^d} b_\alpha(u)\,du^*, \\ \xi_n^*(u) &\equiv \sigma^k\xi_n(u), \end{aligned}$$

where $\{\mathcal{U}_\alpha\ ;\ \alpha\in\mathcal{A}^*\}$ is the set of all essential local coordinates. Then by using Lemma 4, $\delta(t/n\beta - u^{2k})$ can be asymptotically expanded for $n\beta\to 0$, hence

$$\begin{aligned} A_n &= \int_{\mathcal{M}} du^* \int_0^\infty \frac{dt}{n\beta}\Big(\frac{t}{n\beta}\Big)^{\lambda-1}\Big(-\log(\frac{t}{n\beta})\Big)^{m-1} \\ &\quad \times\exp(-t+\sqrt{\beta t}\,\xi_n^*(u)) + O_p(\frac{(\log(n\beta))^{m-2}}{(n\beta)^\lambda}) \\ &= \frac{(\log(n\beta))^{m-1}}{(n\beta)^\lambda}\int_{\mathcal{M}} du^* \int_0^\infty dt\ t^{\lambda-1}\exp(-t)\ \exp(\sqrt{\beta t}\,\xi_n^*(u)) \\ &\quad +O_p(\frac{(\log(n\beta))^{m-2}}{(n\beta)^\lambda}). \end{aligned}$$

Since $\beta=\beta_0/\log n\to 0$,

$$\exp(\sqrt{\beta t}\,\xi_n^*(u)) = 1+\sqrt{\beta t}\,\xi_n^*(u)+O_p(\beta).$$

By using the gamma function,

$$\Gamma(\lambda)=\int_0^\infty t^{\lambda-1}\ \exp(-t)\,dt,$$

it follows that

$$\begin{aligned} A_n &= \frac{(\log(n\beta))^{m-1}}{(n\beta)^\lambda}\Big\{\Gamma(\lambda)\Big(\int_{\mathcal{M}} du^*\Big)+\sqrt{\beta}\Gamma(\lambda+\frac{1}{2})\Big(\int_{\mathcal{M}} du^*\xi_n^*(u)\Big)\Big\} \\ &\quad +O_p(\frac{(\log(n\beta))^{m-2}}{(n\beta)^\lambda}). \end{aligned}$$

Secondly, B_n can be calculated by the same way,

$$\begin{aligned} B_n &= \sum_{\alpha\in\mathcal{A}} \int_{[-1,1]^d} du \int_0^\infty dt\ \delta(t-u^{2k})|u|^h b_\alpha(u) \\ &\quad \times (nu^{2k} - \sqrt{n}u^k\xi_n(u)) \exp(-n\beta u^{2k} + \beta\sqrt{n}u^k\xi_n(u)). \end{aligned}$$

By substitution $t := t/(n\beta)$ and $dt := dt/(n\beta)$ and Lemma 4,

$$\begin{aligned} B_n &= \int_{\mathcal{M}} du^* \int_0^\infty \frac{dt}{n\beta} \left(\frac{t}{n\beta}\right)^{\lambda-1} \left(-\log(\frac{t}{n\beta})\right)^{m-1} \\ &\quad \times \frac{1}{\beta}(t - \sqrt{\beta t}\ \xi_n^*(u)) \exp(-t + \sqrt{\beta t}\ \xi_n^*(u)) + O_p(\frac{(\log(n\beta))^{m-2}}{\beta(n\beta)^\lambda}) \\ &= \frac{(\log(n\beta))^{m-1}}{\beta(n\beta)^\lambda} \int_{\mathcal{M}} du^* \int_0^\infty t^{\lambda-1}(t - \sqrt{\beta t}\ \xi_n^*(u)) \exp(-t) \\ &\quad \times \exp(\sqrt{\beta t}\ \xi_n^*(u)) + O_p(\frac{(\log(n\beta))^{m-2}}{\beta(n\beta)^\lambda}). \end{aligned}$$

Therefore,

$$\begin{aligned} B_n &= \frac{(\log(n\beta))^{m-1}}{\beta(n\beta)^\lambda} \Big\{ \Gamma(\lambda+1)\Big(\int_{\mathcal{M}} du^*\Big) + \sqrt{\beta}\,\Gamma(\lambda+\frac{3}{2})\Big(\int_{\mathcal{M}} du^*\xi_n^*(u)\Big) \\ &\quad -\sqrt{\beta}\,\Gamma(\lambda+\frac{1}{2})\Big(\int_{\mathcal{M}} du^*\xi_n^*(u)\Big)\Big\} + O_p(\frac{(\log(n\beta))^{m-2}}{\beta(n\beta)^\lambda}). \end{aligned}$$

Let us define a random variable Θ by

$$\Theta = \frac{\int_{\mathcal{M}} du^*\xi_n^*(u)}{\int_{\mathcal{M}} du^*}. \tag{27}$$

By applying results of A_n and B_n to Equation (21),

$$\mathbb{E}_w^\beta[nK_n(w)] = \frac{1}{\beta} \times \frac{\Gamma(\lambda+1) + \sqrt{\beta}\,\Theta\,\{\Gamma(\lambda+3/2) - \Gamma(\lambda+1/2)\}}{\Gamma(\lambda) + \sqrt{\beta}\,\Theta\Gamma(\lambda+1/2)} + O_p(1).$$

Note that, if a, b, c, d are constants and $\beta \to 0$,

$$\frac{c+\sqrt{\beta}\,d}{a+\sqrt{\beta}\,b} = \frac{c}{a} + \sqrt{\beta}\left(\frac{ad-bc}{a^2}\right) + O(\beta).$$

Then by using an identity,

$$\frac{\Gamma(\lambda)(\Gamma(\lambda+3/2) - \Gamma(\lambda+1/2)) - \Gamma(\lambda+1)\Gamma(\lambda+1/2)}{\Gamma(\lambda)^2} = -\frac{\Gamma(\lambda+1/2)}{2\Gamma(\lambda)},$$

we obtain

$$\mathbb{E}_w^\beta[nK_n(w)] = \frac{1}{\beta}\frac{\Gamma(\lambda+1)}{\Gamma(\lambda)} - \frac{\Theta}{\sqrt{\beta}}\frac{\Gamma(\lambda+1/2)}{2\Gamma(\lambda)} + O_p(1).$$

A random variable U_n is defined by

$$U_n = -\frac{\Theta\Gamma(\lambda+1/2)}{\sqrt{2\lambda}\Gamma(\lambda)}. \tag{28}$$

Then it follows that

$$\mathbb{E}_w^{\beta}[nK_n(w)] = \frac{\lambda}{\beta} + U_n\sqrt{\frac{\lambda}{2\beta}} + O_p(1).$$

By the definition of $\xi_n(u)$, $\mathbb{E}[\Theta] = 0$, hence $\mathbb{E}[U_n] = 0$. By using Cauchy-Schwarz inequality,

$$\Theta^2 \leq \frac{\int_{\mathcal{M}} du^*\ \xi_n^*(u)^2}{\int_{\mathcal{M}} du^*}.$$

Lastly let us study the case that $q(x)$ is realizable by $p(x|w)$. The support of du^* is contained in $u^{2k} = 0$, hence we can apply Equation (24) to Θ,

$$\mathbb{E}[\Theta^2] \leq \frac{\int_{\mathcal{M}} du^* \mathbb{E}[\xi_n^*(u)^2]}{\int_{\mathcal{M}} du^*} = 2.$$

The gamma function satisfies

$$\frac{\Gamma(\lambda+1/2)}{\Gamma(\lambda)} < \sqrt{\lambda} \quad (\lambda > 0).$$

Hence we obtain

$$\mathbb{E}[(U_n)^2] \leq \frac{\mathbb{E}[\Theta^2]}{2\lambda}\Big(\frac{\Gamma(\lambda+1/2)}{\Gamma(\lambda)}\Big)^2 < 1,$$

which completes Theorem 4. (Q.E.D.)

5.8 Proof of Corollary 1

By definition Equation (27) and Equation (28), it is sufficient to prove $\Theta = 0$, where

$$\Theta = \frac{\sum_{\alpha\in\mathcal{A}^*}\sum_{\sigma\in S(d)}\int_{[0,1]^d} b_\alpha(u)\ du^*\ \sigma^k\ \xi_n(u)}{\sum_{\alpha\in\mathcal{A}^*}\sum_{\sigma\in S(d)}\int_{[0,1]^d} b_\alpha(u)\ du^*}.$$

The support of the measure du^* is contained in the set $\{u = (0, u_b)\}$. We use a notation $\sigma = (\sigma_a, \sigma_b) \in \mathbb{R}^m \times \mathbb{R}^{d-m}$. If $Q(q,p,\varphi) = 1$ then there exists a resolution map $w = g(u)$ such that σ_a^k takes values both $+1$ and -1 in arbitrary local coordinate, hence

$$\sum_{\sigma_a\in S(m)} \sigma_a^k = 0.$$

It follows that

$$\sum_{\sigma\in S(d)} \sigma^k\xi_n(0,u_b) = \sum_{\sigma_b\in S(d-m)} \sigma_b^k\xi_n(0,u_b) \sum_{\sigma_a\in S(m)} \sigma_a^k = 0,$$

therefore, $\Theta = 0$, which completes Corollary 1. (Q.E.D.)

5.9 Proof Corollary 2

By using the optimal inverse temperature β^*, we define $T = 1/(\beta^* \log n)$. By the definition, $\mathcal{F} = \mathbb{E}_w^{\beta^*}[nL_n(w)]$. By using Theorem 2 and Theorem 4,

$$\lambda \log n = T\lambda \log n + U_n \sqrt{T\lambda \log n/2} + V_n = 0,$$

where

$$V_n = O_p(\log\log n).$$

It follows that

$$T + \frac{U_n\sqrt{T}}{\sqrt{2\lambda \log n}} - 1 + \frac{V_n}{\log n} = 0.$$

Therefore,

$$\sqrt{T} = -\frac{U_n}{\sqrt{8\lambda \log n}} + \sqrt{1 + \frac{(U_n)^2}{8\lambda \log n} - \frac{V_n}{\log n}}.$$

Since $U_n = O_p(1)$,

$$\sqrt{T} = 1 - \frac{U_n}{\sqrt{8\lambda \log n}} + o_p(\frac{1}{\sqrt{\lambda \log n}}),$$

resulting that

$$\beta^* \log n = 1 + \frac{U_n}{\sqrt{2\lambda \log n}} + o_p(\frac{1}{\sqrt{\lambda \log n}}),$$

which completes Corollary 2. (Q.E.D.)

5.10 Proof of Corollary 3

By using Theorem 4,

$$\begin{aligned}
\mathbb{E}_w^{\beta_1}[nL_n(w)] &= nL_n(w_0) + \frac{\lambda}{\beta_1} + O_p(\sqrt{\log n}), \\
\mathbb{E}_w^{\beta_2}[nL_n(w)] &= nL_n(w_0) + \frac{\lambda}{\beta_2} + O_p(\sqrt{\log n}).
\end{aligned}$$

Since $(1/\beta_1 - 1/\beta_2) = O_p(\log n)$,

$$\lambda = \frac{\mathbb{E}_w^{\beta_1}[nL_n(w)] - \mathbb{E}_w^{\beta_2}[nL_n(w)]}{1/\beta_1 - 1/\beta_2} + O_p(1/\sqrt{\log n}),$$

which shows Corollary 3. (Q.E.D.)

5.11 Proof Theorem 5

By using Equation (9) and Equation (10),

$$\mathbb{E}_w^{\beta}[nL_n(w)] = nL_n(w_0) + \mathbb{E}_w^{\beta}[nK_n(w)],$$

the proof of Theorem 5 results in evaluating $\mathbb{E}_w^{\beta}[nK_n(w)]$. By Lemma 1 for the case $r = 1/4$,

$$\mathbb{E}_w^{\beta}[nK_n(w)] = \frac{D_n + o_p(\exp(-\sqrt{n}))}{C_n + o_p(\exp(-\sqrt{n}))},$$

where C_n and D_n are respectively defined by

$$\begin{aligned} C_n &= \int_{K<1/n^{1/4}} \exp(-n\beta K_n(w))\varphi(w)dw, \\ D_n &= \int_{K<1/n^{1/4}} nK_n(w)\exp(-n\beta K_n(w))\varphi(w)dw. \end{aligned}$$

If a statistical model is regular, the maximum likelihood estimator $\hat{w}$ converges to w_0 in probability. Let $J_n(w)$ be $d \times d$ matrices defined by

$$(J_n)_{ij}(w) = \frac{\partial^2 K_n}{\partial w_i \partial w_j}(w).$$

There exists a parameter w^* such that

$$K_n(w) = K_n(\hat{w}) + \frac{1}{2}(w-\hat{w}) \cdot J_n(w^*)(w-\hat{w}).$$

Since $\hat{w} \to w_0$ in probability and $K(w) < 1/n^{1/4}$, $w^* \to w_0$ in probability. Then

$$\begin{aligned} &\|J_n(w^*) - J(w_0)\| \leq \|J_n(w^*) - J_n(w_0)\| + \|J_n(w_0) - J(w_0)\| \\ &\leq \|w^* - w_0\| \sup_{K(w)<1/n^{1/4}} \left\| \frac{\partial J_n(w)}{\partial w} \right\| + \|J_n(w_0) - J(w_0)\|, \end{aligned}$$

which converges to zero in probability as $n \to \infty$. Therefore,

$$J_n(w^*) = J(w_0) + o_p(1).$$

Since the model is regular, $J(w_0)$ is a positive definite matrix. Now we define

$$\begin{aligned} C_n &= \exp(-n\beta K_n(\hat{w})) \\ &\quad \times \int_{K(w)<n^{1/4}} \exp(-\frac{n\beta}{2}(w-\hat{w}) \cdot (J(w_0) + o_p(1))(w-\hat{w}))\varphi(w)dw. \end{aligned}$$

By substituting

$$u = \sqrt{n\beta}(w-\hat{w}),$$

it follows that

$$\begin{aligned} C_n &= \exp(-n\beta K_n(\hat{w}))(n\beta)^{-d/2} \\ &\quad \times \int \exp(-\frac{1}{2}u \cdot (J(w_0) + o_p(1))u)\varphi(\hat{w} + \frac{u}{\sqrt{n\beta}})du \\ &= \frac{(2\pi)^{d/2}\exp(-n\beta K_n(\hat{w}))(\varphi(\hat{w}) + o_p(1))}{(n\beta)^{d/2}\det(J(w_0) + o_p(1))^{1/2}}. \end{aligned}$$

H	1	2	3	4	5	6
WBIC_1 Ave.	17899.82	3088.90	71.11	78.21	83.23	87.58
WBIC_1 Std.	1081.30	226.94	3.67	3.78	3.97	4.09
WBIC_2 Ave.	17899.77	3089.03	71.18	75.43	82.54	86.83
WBIC_2 Std.	1081.30	226.97	3.54	3.89	4.03	4.08
BIC Ave.	17899.77	3089.03	71.18	83.47	91.86	94.87
BIC Std.	1081.30	226.97	3.54	3.89	4.03	4.08

Table 2: WBIC and BIC in Model Selection

In the same way,

$$\begin{aligned} D_n &= \exp(-n\beta K_n(\hat{w})) \\ &\quad \times \int_{K(w)<1/n^{1/4}} \Big(nK_n(\hat{w}) + \frac{n}{2}(w-\hat{w}) \cdot (J(w_0)+o_p(1))(w-\hat{w}) \Big) \\ &\quad \times \exp\Big(-\frac{n\beta}{2}(w-\hat{w}) \cdot (J(w_0)+o_p(1))(w-\hat{w}) \Big) \varphi(w) dw \\ &= \frac{(2\pi)^{d/2} \exp(-n\beta K_n(\hat{w}))(\varphi(\hat{w})+o_p(1))}{(n\beta)^{d/2} \det(J(w_0)+o_p(1))^{1/2}} \Big(nK_n(\hat{w}) + \frac{d}{2\beta} + o_p(1) \Big). \end{aligned}$$

Here $nK_n(\hat{w}) = O_p(1)$, because the true distribution is regular for a statistical model. Therefore,

$$\mathbb{E}_w^\beta[nL_n(w)] = nL_n(w_0) + nK_n(\hat{w}) + \frac{d}{2\beta} + o_p(1),$$

which completes Theorem 5. (Q.E.D.)

6. A Method How to Use WBIC

In this section we show a method how to use WBIC in statistical model evaluation. The main theorems have already been mathematically proved, hence WBIC has a theoretical support. The following experiment was conducted not for proving theorems but for illustrating a method how to use it.

6.1 Statistical Model Selection

Firstly, we study model selection by using WBIC.

Let $x \in \mathbb{R}^M$, $y \in \mathbb{R}^N$, and $w = (A, B)$, where A is an $H \times M$ matrix and B is an $N \times H$ matrix. A reduced rank regression model is defined by

$$p(x,y|w) = \frac{r(x)}{(2\pi\sigma^2)^{N/2}} \exp\Big(-\frac{1}{2\sigma^2} \|y - BAx\|^2 \Big),$$

where $r(x)$ is a probability density function of x and σ^2 is the variance of an output. Let $\mathcal{N}_M(0,\Sigma)$ denote the M-dimensional normal distribution with the average zero and the covariance matrix Σ.

In an experiment, we set $\sigma = 0.1$, $r(x) = \mathcal{N}_M(0, 3^2 I)$, where I is the identity matrix, and $\varphi(w) = \mathcal{N}_d(0, 10^2 I)$. The true distribution was fixed as $p(x,y|w_0)$, where $w_0 = (A_0, B_0)$ was determined so

that A_0 and B_0 were respectively an $H_0 \times M$ matrix and an $M \times H_0$ matrix. Note that, in reduced rank regression models, RLCTs and multiplicities were clarified by a research (Aoyagi and Watanabe, 2005) and $Q(K(w), \varphi(w)) = 1$ for arbitrary $q(x)$, $p(x|w)$, and $\varphi(w)$. In the experiment, $M = N = 6$ and the true rank was set as $H_0 = 3$. Each element of A_0 and B_0 was taken from $\mathcal{N}_{\mathrm{J}}(0, 0.2^2)$ and fixed. From the true distribution $p(x,y|w_0)$, 100 sets of $n = 500$ training samples were generated.

The Metropolis method was employed for sampling from the posterior distribution,

$$p(w|X_1, X_2, ..., X_n) \propto \exp(-\beta n L_n(w) + \log \varphi(w)),$$

where $\beta = 1/\log n$. Every Metropolis trial was generated from a normal distribution $\mathcal{N}_d(0, (0.0012)^2 I)$, by which the acceptance probability was 0.1-0.9. First 50000 Metropolis trails were not used. After 50000 trails, $R = 2000$ parameters $\{w_r; r = 1, 2, ..., R\}$ were obtained in every 100 Metropolis steps. The expectation value of a function $G(w)$ over the posterior distribution was approximated by

$$\mathbb{E}_w^{\beta}[G(w)] = \frac{1}{R} \sum_{r=1}^{R} G(w_r).$$

The six statistical models $H = 1, 2, 3, 4, 5, 6$ were compared by the criterion,

$$\mathrm{WBIC} = \mathbb{E}_w^{\beta}[nL_n(w)], \quad (\beta = 1/\log n).$$

To compare these values among several models, we show both WBIC_1, WBIC_2, and BIC in Table 2. In the table, the average and the standard deviation of WBIC_1 defined by

$$\mathrm{WBIC}_1 = \mathrm{WBIC} - nS_n,$$

for 100 independent sets of training samples are shown, where the empirical entropy of the true distribution

$$S_n = -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i)$$

does not depend on a statistical model. Although nS_n does not affect the model selection, its standard deviation is in proportion to $\sqrt{n}$. In order to estimate the standard deviation of the essential part of WBIC, the effect of nS_n was removed. In 100 independent sets of training samples, the true model $H = 3$ was chosen 100 times in this experiment, which demonstrates a typical application method of WBIC.

Also WBIC_2 in Table 2 shows the average and the standard deviation of

$$\mathrm{WBIC}_2 = nL_n(\hat{w}) + \lambda \log n - (m-1) \log \log n - nS_n,$$

where $\hat{w}$ is the maximum likelihood estimator, λ is the real log canonical threshold, and m is the multiplicity. The maximum likelihood estimator $\hat{w} = (\hat{A}, \hat{B})$ is given in reduced rank regression (Anderson, 1951),

$$\hat{B}\hat{A} = VV^t YX^t (XX^t)^{-1},$$

where t shows the transposed matrix and X, Y, and V are matrices defined as follows.
(1) X_{ij} is the i-th coefficient of x_j
(2) Y_{ij} is the i-th coefficient of y_j.

H	1	2	3	4	5	6
Theory λ	5.5	10	13.5	15	16	17
Theory m	1	1	1	2	1	2
Average λ	5.51	9.95	13.49	14.80	15.72	16.55
Std. Dev. λ	0.17	0.31	0.52	0.65	0.66	0.72

Table 3: RLCTs for the case $H_0 = 3$

(3) The matrix V is made of eigen vectors with respect to the maximum H eigen values of the matrix,

$$YX^t(XX^t)^{-1}XY^t.$$

If we know the model that is equal to the true distribution, and if we have theoretical real log canonical threshold and multiplicity, then we can calculate WBIC_2.

The value BIC in Table 2 shows the average and the standard deviation of Shwarz BIC,

$$\mathrm{BIC} = nL_n(\hat{w}) + \frac{d}{2}\log n - nS_n,$$

where d is the essential dimension of the parameter space of the reduced rank regression,

$$d = H(M+N-H).$$

We used this dimension because the number of parameters in reduced rank regression is $H(M+N)$ and it has free dimension H^2. These results show that WBIC is a better approximator of the Bayes free energy than BIC.

6.2 Estimating RLCT

Secondly, we study a method how to estimate an RLCT. By using the same experiment as the foregoing subsection, we estimated RLCTs of reduced rank regression models by using Corollary 3. Based on Equation (19), the estimated RLCT is given by

$$\hat{\lambda} = \frac{\mathbb{E}_w^{\beta_1}[nL_n(w)] - \mathbb{E}_w^{\beta_2}[nL_n(w)]}{1/\beta_1 - 1/\beta_2},$$

where $\beta_1 = 1/\log n$ and $\beta_2 = 1.5/\log n$ and we used Equation (20) in the calculation of $\mathbb{E}_w^{\beta_2}[\ ]$. Theory λ in Table 3 shows the theoretical values of RLCTs of reduced rank regression. For the cases when true distributions are unrealizable by statistical models, RLCTs are given by half the dimension of the parameter space, $\lambda = H(M+N-H)/2$. In Table 3, averages and standard deviations of λ shows estimated RLCTs. The theoretical RLCTs were well estimated. The difference between theory and experimental results was caused by the effect of the smaller order terms than $\log n$. In the case the multiplicity $m = 2$, the term $\log\log n$ also affected the results.

7. Discussion

In this section, we discuss the widely applicable information criterion from three different points of view.

7.1 WAIC and WBIC

Firstly, let us study the difference between the free energy and the generalization error. In the present paper, we study the Bayes free energy $\mathcal{F}$ as the statistical model selection criterion. Its expectation value is given by

$$\mathbb{E}[\mathcal{F}] = nS + \int q(x^n) \log \frac{q(x^n)}{p(x^n)} dx^n,$$

where S is the entropy of the true distribution,

$$\begin{aligned} q(x^n) &= \prod_{i=1}^{n} q(x_i), \\ p(x^n) &= \int \prod_{i=1}^{n} p(x_i|w)\varphi(w) dw, \end{aligned}$$

and $dx^n = dx_1 dx_2 \cdots dx_n$. Hence minimization of $\mathbb{E}[\mathcal{F}]$ is equivalent to minimization of the Kullback-Leibler distance from the $q(x^n)$ to $p(x^n)$.

There is a different model evaluation criterion, which is the generalization loss defined by

$$\mathcal{G} = -\int q(x) \log p^*(x) dx, \tag{29}$$

where $p^*(x)$ is the Bayes predictive distribution defined by $p^*(x) = \mathbb{E}_w^{\beta}[p(x|w)]$, with $\beta = 1$. The expectation value of $\mathcal{G}$ satisfies

$$\mathbb{E}[\mathcal{G}] = S + \mathbb{E}\Big[\int q(x) \log \frac{q(x)}{p^*(x)} dx\Big].$$

Hence minimization of $\mathbb{E}[\mathcal{G}]$ is equivalent to minimization of the Kullback-Leibler distance from $q(x)$ to $p^*(x)$. Both of $\mathcal{F}$ and $\mathcal{G}$ are important in statistics and learning theory, however, they are different criteria.

The well-known model selection criteria AIC and BIC are respectively defined by

$$\begin{aligned} \mathrm{AIC} &= L_n(\hat{w}) + \frac{d}{n}, \\ \mathrm{BIC} &= nL_n(\hat{w}) + \frac{d}{2} \log n. \end{aligned} \tag{30}$$

If a true distribution is realizable by and regular for a statistical model, then

$$\begin{aligned} \mathbb{E}[\mathrm{AIC}] &= \mathbb{E}[\mathcal{G}] + o(\frac{1}{n}), \\ \mathbb{E}[\mathrm{BIC}] &= \mathbb{E}[\mathcal{F}] + O(1). \end{aligned}$$

These relations can be generalized onto singular statistical models. We define WAIC and WBIC by

$$\begin{aligned} \mathrm{WAIC} &= T_n + V_n/n, \\ \mathrm{WBIC} &= \mathbb{E}_w^{\beta}[nL_n(w)], \quad \beta = 1/\log n, \end{aligned}$$

where

$$\begin{aligned} T_n &= -\frac{1}{n}\sum_{i=1}^{n}\log p^*(X_i|w), \\ V_n &= \sum_{i=1}^{n}\Big\{\mathbb{E}_w[(\log p(X_i|w))^2]-\mathbb{E}_w[\log p(X_i|w)]^2\Big\}. \end{aligned}$$

Then, even if a statistical model is unrealizable by and singular for a statistical model,

$$\mathbb{E}[\mathrm{WAIC}] = \mathbb{E}[\mathcal{G}]+O(\frac{1}{n^2}), \tag{31}$$

$$\mathbb{E}[\mathrm{WBIC}] = \mathbb{E}[\mathcal{F}]+O(\log\log n), \tag{32}$$

where Equation (31) was proved in a book (Watanabe, 2009, 2010b), whereas Equation (32) has been proved in the present paper. In fact, the difference between the average leave-one-out cross validation and the average generalization error is in proportion to $1/n^2$ and the difference between the leave-one-out cross validation and WAIC is also in proportion to $1/n^2$. The difference between $\mathbb{E}[\mathrm{WBIC}]$ and $\mathbb{E}[\mathcal{F}]$ is caused by the multiplicity m. If a statistical model is realizable by and regular for a statistical model, WAIC and WBIC respectively coincide with AIC and BIC,

$$\begin{aligned} \mathrm{WAIC} &= \mathrm{AIC}+o_p(\frac{1}{n}), \\ \mathrm{WBIC} &= \mathrm{BIC}+o_p(1). \end{aligned}$$

Theoretical comparison of WAIC and WBIC in singular model selection is an important problem for future study.

7.2 Other Methods How to Evaluate Free Energy

Secondly, we discuss several methods how to numerically evaluate the Bayes free energy. There are three methods other than WBIC.

Firstly, let $\{\beta_j; j=0,1,2,...,J\}$ be a sequence which satisfies

$$0=\beta_0<\beta_1<\cdots<\beta_J=1.$$

Then the Bayes free energy satisfies

$$\mathcal{F}=-\sum_{j=1}^{J}\log\mathbb{E}_w^{\beta_{j-1}}[\exp(-n(\beta_j-\beta_{j-1})L_n(w))].$$

This method can be used without asymptotic theory. We can estimate $\mathcal{F}$, if the number J is sufficiently large and if all expectation values over the posterior distributions $\{\mathbb{E}_w^{\beta_{j-1}}[\ ]\}$ are precisely calculated. The disadvantage of this method is its huge computational costs for accurate calculation. In the present paper, this method is referred to as 'all temperatures method'.

Secondly, the importance sampling method is often used. Let $H(w)$ be a function which approximates $nL_n(w)$. Then, for an arbitrary function $G(w)$, we define an expectation value $\hat{\mathbb{E}}_w[\ ]$ by

$$\hat{\mathbb{E}}_w[G(w)]=\frac{\int G(w)\exp(-H(w))\varphi(w)dw}{\int\exp(-H(w))\varphi(w)dw}.$$

Method	Asymptotics	RLCT	Comput. Cost
All Temperatures	Not used	Not Used	Huge
Importance Sampling	Not used	Not Used	Small
Two-Step	Used	Used	Small
WBIC	Used	Not Used	Small

Table 4: Comparison of Several Methods

Then

$$\begin{aligned}\mathcal{F} &= -\log \hat{\mathbb{E}}_w[\exp(-nL_n(w)+H(w))] \\ &\quad -\log \int \exp(-H(w))\varphi(w)dw,\end{aligned}$$

where the last term is the free energy of $H(w)$. Hence if we find $H(w)$ whose free energy is analytically calculated and if it is easy to generate random samples from $\hat{\mathbb{E}}_w[\ ]$, then $\mathcal{F}$ can be numerically evaluated. The accuracy of this method strongly depends on the choice of $H(w)$.

Thirdly, a two-step method was proposed (Drton, 2010). Assume that we have theoretical values about RLCTs for all cases about true distribution and statistical models. Then, in the first step, a null hypothesis model is chosen by using BIC. In the second step, the optimal model is chosen by using RLCTs with the assumption that the null hypothesis model is a true distribution. If the selected model is different from the null hypothesis model, then the same procedure is recursively applied until the null hypothesis model becomes the optimal model. In this method, asymptotic theory is necessary but RLCTs do not contain fluctuations because they are theoretical values.

Compared with these methods, WBIC needs asymptotic theory but it does not theoretical results about RLCT. The theoretical comparison of these methods is summarized in Table 4.

The effectiveness of a model selection method strongly depends on a statistical condition which is determined by a true distribution, a statistical model, a prior distribution, and a set of training samples. Under some condition, one method may be more effective, however, under the other condition, another may be. The proposed method WBIC gives a new approach in numerical calculation of the Bayes free energy which is more useful with cooperation with the conventional method. It is a future study to clarify which method is recommended in what statistical conditions.

Remark. It is one of the most important problems in Bayes statistics how to make accurate Markov Chain Monte Carlo (MCMC) process. There are several MCMC methods, for example, the Metropolis method, the Gibbs sampler method, the Hybrid Monte Carlo method, and the exchange Monte Carlo method. Numerical calculation of WBIC depends on the accuracy of MCMC process.

7.3 Algebraic Geometry and Statistics

Lastly, let us discuss a relation between algebraic geometry and statistics. In the present paper, we define the parity of a statistical $Q(K(w),\varphi(w))$ and proved that it affects the asymptotic behavior of WBIC. In this subsection we show three mathematical properties of the parity of a statistical model.

Firstly, the parity has a relation to the analytic continuation of $K(w)^{1/2}$. For example, by using blow-up, $(a,b)=(a_1,a_1b_1)=(a_2b_2,b_2)$, it follows that analytic continuation of $(a^2+b^2)^{1/2}$ is given

by

$$(a^2+b^2)^{1/2} = a_1\sqrt{1+b_1^2} = b_2\sqrt{a_2^2+1},$$

which takes both positive and negative values. On the other hand, $(a^4+b^4)^{1/2}$ takes only nonnegative value. The parity indicates such difference.

Secondly, the parity has a relation to statistical model with a restricted parameter set. For example, a statistical model

$$p(x|a) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{(x-a)^2}{2})$$

whose parameter set is given by $\{a \geq 0\}$ is equivalent to a statistical model $p(x|b^2)$ and $\{b \in \mathbb{R}\}$. In other words, a statistical model which has restricted parameter set is statistically equivalent to another even model which has unrestricted parameter set. We have a conjecture that an even statistical model has some relation to a model with a restricted parameter model.

And lastly, the parity has a relation to the difference of $K(w)$ and $K_n(w)$. As is proven (Watanabe, 2001a), the relation

$$-\log\int \exp(-nK_n(w))\varphi(w)dw = -\log\int \exp(-nK(w))\varphi(w)dw + O_p(1)$$

holds independent of the parity of a statistical model. On the other hand, if $\beta = 1/\log n$, then

$$\begin{aligned}\mathbb{E}_w^\beta[nK_n(w)] &= \frac{\int nK(w)\exp(-n\beta K(w))\varphi(w)dw}{\int \exp(-n\beta K(w))\varphi(w)} \\ &\quad + U_n\sqrt{\log n} + O_p(1).\end{aligned}$$

If the parity is odd, then $U_n = 0$, otherwise U_n is not equal to zero in general. This fact shows that the parity shows difference in a fluctuation of the likelihood function.

8. Conclusion

We proposed a widely applicable Bayesian information criterion (WBIC) which can be used even if a true distribution is unrealizable by and singular for a statistical model and proved that WBIC has the same asymptotic expansion as the Bayes free energy. Also we developed a method how to estimate real log canonical thresholds even if a true distribution is unknown.

Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture in Japan, Grant-in-Aid for Scientific Research 23500172.

References

T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22:327–351, 1951.

M. Aoyagi and K. Nagata. Learning coefficient of generalization error in Bayesian estimation and Vandermonde matrix-type singularity. *Neural Computation*, 24(6):1569–1610, 2012.

M. Aoyagi and S. Watanabe. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18(7):924–933, 2005.

M. F. Atiyah. Resolution of singularities and division of distributions. *Communications of Pure and Applied Mathematics*, 13:145–150, 1970.

I. N. J. Bernstein. Analytic continuation of distributions with respect to a parameter. *Functional Analysis and its Applications*, 6(4):26–40, 1972.

M. Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37:979–1012, 2009.

M. Drton. Reduced rank regression. In *Workshop on Singular Learning Theory*. American Institute of Mathematics, 2010.

M. Drton, B. Sturmfels, and S. Sullivant. *Lecures on Algebraic Statistics*. Birkhäuser, Berlin, 2009.

I. M. Gelfand and G. E. Shilov. *Generalized Functions. Volume I: Properties and Operations*. Academic Press, San Diego, 1964.

I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge, 1965.

H. Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, 79:109–326, 1964.

M. Kashiwara. B-functions and holonomic systems. *Inventiones Mathematicae*, 38:33–53, 1976.

F. J. Király, P. Büuau, F. C. Meinecke, D. A. J. Blythe, and K-R Müller. Algebraic geometric comparison of probability distributions. *Journal of Machine Learning Research*, 13:855–903, 2012.

J. Kollár. Singularities of pairs. In *Algebraic Geometry, Santa Cruz 1995, Proceedings of Symposia in Pure Mathematics*, volume 62, pages 221–286. American Mathematical Society, 1997.

S. Lin. *Algebraic Methods for Evaluating Integrals in Bayesian Statistics*. PhD thesis, Ph.D. dissertation, University of California, Berkeley, 2011.

D. Rusakov and D. Geiger. Asymptotic model selection for naive Bayesian network. *Journal of Machine Learning Research*, 6:1–35, 2005.

M. Saito. On real log canonical thresholds. *arXiv:0707.2308v1*, 2007.

M. Sato and T. Shintani. On zeta functions associated with prehomogeneous vector space. *Annals of Mathematics*, 100:131–170, 1974.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

A. Varchenko. Newton polyhedrons and estimates of oscillatory integrals. *Functional Analysis and its Applications*, 10(3):13–38, 1976.

S. Watanabe. Algebraic analysis for singular statistical estimation. *Lecture Notes in Computer Sciences*, 1720:39–50, 1999.

S. Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13 (4):899–933, 2001a.

S. Watanabe. Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, 14(8):1049–1060, 2001b.

S. Watanabe. *Algebraic geometry and statistical learning theory*. Cambridge University Press, Cambridge, UK, 2009.

S. Watanabe. Asymptotic learning curve and renormalizable condition in statistical learning theory. *Journal of Physics Coneference Series*, 233(1), 2010a. 012014. doi: 10.1088/1742-6596/233/1/012014.

S. Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3591, 2010b.

K. Yamazaki and S. Watanabe. Singularities in mixture models and upper bounds of stochastic complexity. *Neural Networks*, 16(7):1029–1038, 2003.

K. Yamazaki and S. Watanabe. Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities. *IEEE Transactions on Neural Networks*, 16(2): 312–324, 2005.

P. Zwiernik. Asymptotic model selection and identifiability of directed tree models with hidden variables. *CRiSM report*, 2010.

P. Zwiernik. An asymptotic behaviour of the marginal likelihood for general markov models. *Journal of Machine Learning Research*, 12:3283–3310, 2011.