Journal of Machine Learning Research 11 (2010) 3571-3594 Submitted 4/10; Revised 10/10; Published 12/10

Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory

Sumio Watanabe SWATANAB@PI.TITECH.AC.JP
Precision and Intelligence Laboratory
Tokyo Institute of Technology
Mailbox R2-5, 4259 Nagatsuta, Midori-ku
Yokohama 226-8503, Japan

Editor: Manfred Opper

Abstract

In regular statistical models, the leave-one-out cross-validation is asymptotically equivalent to the Akaike information criterion. However, since many learning machines are singular statistical models, the asymptotic behavior of the cross-validation remains unknown. In previous studies, we established the singular learning theory and proposed a widely applicable information criterion, the expectation value of which is asymptotically equal to the average Bayes generalization loss. In the present paper, we theoretically compare the Bayes cross-validation loss and the widely applicable information criterion and prove two theorems. First, the Bayes cross-validation loss is asymptotically equivalent to the widely applicable information criterion as a random variable. Therefore, model selection and hyperparameter optimization using these two values are asymptotically equivalent. Second, the sum of the Bayes generalization error and the Bayes cross-validation error is asymptotically equal to $2\lambda/n$, where λ is the real log canonical threshold and n is the number of training samples. Therefore the relation between the cross-validation error and the generalization error is determined by the algebraic geometrical structure of a learning machine. We also clarify that the deviance information criteria are different from the Bayes cross-validation and the widely applicable information criterion.

Keywords: cross-validation, information criterion, singular learning machine, birational invariant

1. Introduction

A statistical model or a learning machine is said to be regular if the map taking parameters to probability distributions is one-to-one and if its Fisher information matrix is positive definite. If a model is not regular, then it is said to be singular. Many learning machines, such as artificial neural networks (Watanabe, 2001b), normal mixtures (Yamazaki and Watanabe, 2003), reduced rank regressions (Aoyagi and Watanabe, 2005), Bayes networks (Rusakov and Geiger, 2005; Zwiernik, 2010), mixtures of probability distributions (Lin, 2010), Boltzmann machines (Aoyagi, 2010), and hidden Markov models (Yamazaki and Watanabe, 2005), are not regular but singular (Watanabe, 2007). If a statistical model or a learning machine contains a hierarchical structure, hidden variables, or a grammatical rule, then the model is generally singular. Therefore, singular learning theory is necessary in modern information science.

The statistical properties of singular models have remained unknown until recently, because analyzing a singular likelihood function had been difficult (Hartigan, 1985; Watanabe, 1995). In singular statistical models, the maximum likelihood estimator does not satisfy asymptotic normality.

Consequently, AIC is not equal to the average generalization error (Hagiwara, 2002), and the Bayes information criterion (BIC) is not equal to the Bayes marginal likelihood (Watanabe, 2001a), even asymptotically. In singular models, the maximum likelihood estimator often diverges, or even if it does not diverge, makes the generalization error very large. Therefore, the maximum likelihood method is not appropriate for singular models. On the other hand, Bayes estimation was proven to make the generalization error smaller if the statistical model contains singularities. Therefore, in the present paper, we investigate methods for estimating the Bayes generalization error.

Recently, new statistical learning theory, based on methods from algebraic geometry, has been established (Watanabe, 2001a; Drton et al., 2009; Watanabe, 2009, 2010a,c; Lin, 2010). In singular learning theory, a log likelihood function can be made into a common standard form, even if it contains singularities, by using the resolution theorem in algebraic geometry. As a result, the asymptotic behavior of the posterior distribution is clarified, and the concepts of BIC and AIC can be generalized onto singular statistical models. The asymptotic Bayes marginal likelihood was proven to be determined by the real log canonical threshold (Watanabe, 2001a), and the average Bayes generalization error was proven to be estimable by the widely applicable information criterion (Watanabe, 2009, 2010a,c).

Cross-validation is an alternative method for estimating the generalization error (Mosier, 1951; Stone, 1977; Geisser, 1975). By definition, the average of the cross-validation is equal to the average generalization error in both regular and singular models. In regular statistical models, the leave-one-out cross-validation is asymptotically equivalent to AIC (Akaike, 1974) in the maximum likelihood method (Stone, 1977; Linhart, 1986; Browne, 2000). However, the asymptotic behavior of the cross-validation in singular models has not been clarified.

In the present paper, in singular statistical models, we theoretically compare the Bayes cross-validation, the widely applicable information criterion, and the Bayes generalization error and prove two theorems. First, we show that the Bayes cross-validation loss is asymptotically equivalent to the widely applicable information criterion as a random variable. Second, we also show that the sum of the Bayes cross-validation error and the Bayes generalization error is asymptotically equal to $2\lambda/n$, where λ is the real log canonical threshold and n is the number of training samples. It is important that neither λ or n is a random variable. Since the real log canonical threshold is a birational invariant of the statistical model, the relationship between the Bayes cross-validation and the Bayes generalization error is determined by the algebraic geometrical structure of the statistical model.

The remainder of the present paper is organized as follows. In Section 2, we introduce the framework of Bayes learning and explain singular learning theory. In Section 3, the Bayes cross-validation is defined. In Section 4, the main theorems are proven. In Section 5, we discuss the results of the present paper, and the differences among the cross-validation, the widely applicable information criterion, and the deviance information criterion are investigated theoretically and experimentally. Finally, in Section 6, we summarize the primary conclusions of the present paper.

2. Bayes Learning Theory

In this section, we summarize Bayes learning theory for singular learning machines. The results presented in this section are well known and are the fundamental basis of the present paper. Table 1 lists variables, names, and equation numbers in the present paper.

Variable	Name	Equation number
$\mathbb{E}_w[\ ]$	posterior average	Equation (1)
$\mathbb{E}_w^{(i)}[\ ]$	posterior average without X_i	Equation (16)
$L(w)$	log loss function	Equation (8)
L_0	minimum loss	Equation (9)
L_n	empirical loss	Equation (10)
$B_gL(n)$	Bayes generalization loss	Equation (3)
$B_tL(n)$	Bayes training loss	Equation (4)
$G_tL(n)$	Gibbs training loss	Equation (7)
$C_vL(n)$	cross-validation loss	Equation (17)
$B_g(n)$	Bayes generalization error	Equation (11)
$B_t(n)$	Bayes training error	Equation (12)
$C_v(n)$	cross-validation error	Equation (28)
$V(n)$	functional variance	Equation (5)
$Y_k(n)$	kth functional cumulant	Equation (18)
WAIC(n)	WAIC	Equation (6)
λ	real log canonical threshold	Equation (29)
ν	singular fluctuation	Equation (30)

Table 1: Variables, Names, and Equation Numbers

2.1 Framework of Bayes Learning

First, we explain the framework of Bayes learning.

Let $q(x)$ be a probability density function on the N dimensional real Euclidean space $\mathbb{R}^N$. The training samples and the testing sample are denoted by random variables $X_1, X_2, ..., X_n$ and X, respectively, which are independently subject to the same probability distribution as $q(x)dx$. The probability distribution $q(x)dx$ is sometimes called the true distribution.

A statistical model or a learning machine is defined as 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 the set of all parameters. In Bayes estimation, we prepare a probability density function $\varphi(w)$ on W. Although $\varphi(w)$ is referred to as a prior distribution, in general, $\varphi(w)$ does not necessary represent an *a priori* knowledge of the parameter.

For a given function $f(w)$ on W, the expectation value of $f(w)$ with respect to the posterior distribution is defined as

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

where $0 < \beta < \infty$ is the inverse temperature. The case in which $\beta = 1$ is most important because this case corresponds to strict Bayes estimation. The Bayes predictive distribution is defined as

$$p^*(x) \equiv \mathbb{E}_w[p(x|w)]. \tag{2}$$

In Bayes learning theory, the following random variables are important. The Bayes generalization loss $B_gL(n)$ and the Bayes training loss $B_tL(n)$ are defined, respectively, as

$$B_gL(n) = -\mathbb{E}_X[\log p^*(X)], \tag{3}$$

$$B_tL(n) = -\frac{1}{n}\sum_{i=1}^{n}\log p^*(X_i), \tag{4}$$

where $\mathbb{E}_X[\ ]$ gives the expectation value over X. The *functional variance* is defined as

$$V(n) = \sum_{i=1}^{n}\Big\{\mathbb{E}_w[(\log p(X_i|w))^2] - \mathbb{E}_w[\log p(X_i|w)]^2\Big\}, \tag{5}$$

which shows the fluctuation of the posterior distribution. In previous papers (Watanabe, 2009, 2010a,b), we defined the widely applicable information criterion

$$\mathrm{WAIC}(n) \equiv B_tL(n) + \frac{\beta}{n}V(n), \tag{6}$$

and proved that

$$\mathbb{E}[B_gL(n)] = \mathbb{E}[\mathrm{WAIC}(n)] + o(\frac{1}{n}),$$

holds for both regular and singular statistical models, where $\mathbb{E}[\ ]$ gives the expectation value over the sets of training samples.

Remark 1 *Although the case in which $\beta = 1$ is most important, general cases in which $0 < \beta < \infty$ are also important for four reasons. First, from a theoretical viewpoint, several mathematical relations can be obtained using the derivative of β. For example, using the Bayes free energy or the Bayes stochastic complexity,*

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

the Gibbs training loss

$$G_tL(n) = -\mathbb{E}_w\Big[\frac{1}{n}\sum_{i=1}^{n}\log p(X_i|w)\Big] \tag{7}$$

can be written as

$$G_tL(n) = \frac{\partial\mathcal{F}}{\partial\beta}.$$

Such relations are useful in investigating Bayes learning theory. We use $\partial^2\mathcal{F}/\partial\beta^2$ to investigate the deviance information criteria in Section 5. Second, the maximum likelihood method formally corresponds to $\beta = \infty$. The maximum likelihood method is defined as

$$p^*(x) = p(x|\hat{w}),$$

instead of Equation (2), where $\hat{w}$ is the maximum likelihood estimator. Its generalization loss is also defined in the same manner as Equation (3). In regular statistical models, the asymptotic Bayes generalization error does not depend on $0 < \beta \leq \infty$, whereas in singular models it strongly depends on β. Therefore, the general case is useful for investigating the difference between the maximum likelihood and Bayes methods. Third, from an experimental viewpoint, in order to approximate the posterior distribution, the Markov chain Monte Carlo method is often applied by controlling β. In particular, the identity

$$\mathcal{F}(1) = \int_0^1 \frac{\partial F}{\partial \beta}\, d\beta$$

is used in the calculation of the Bayes marginal likelihood. The theoretical results for general β are useful for monitoring the effect of controlling β (Nagata and Watanabe, 2008). Finally, in the regression problem, β can be understood as the variance of the unknown additional noise (Watanabe, 2010c) and so may be optimized as the hyperparameter. For these reasons, in the present paper, we theoretically investigate the cases for general β.

2.2 Notation

In the following, we explain the notation used in the present study.

The log loss function $L(w)$ and the entropy S of the true distribution are defined, respectively, as

$$\begin{aligned} L(w) &= -\mathbb{E}_X[\log p(X|w)], \\ S &= -\mathbb{E}_X[\log q(X)]. \end{aligned} \tag{8}$$

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

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

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 the present paper, we assume that there exists a parameter $w_0 \in W$ that minimizes $L(w)$,

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

Note that such w_0 is not unique in general because the map $w \mapsto p(x|w)$ is, in general, not a one-to-one map in singular learning machines. In addition, we 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 a single element but rather an analytic or algebraic set with singularities. Here, a set in $\mathbb{R}^d$ is said to be an analytic or algebraic set if and only if the set is equal to the set of all zero points of an analytic or algebraic function, respectively. For simple notations, the minimum log loss L_0 and the empirical log loss L_n are defined, respectively, as

$$L_0 = -\mathbb{E}_X[\log p_0(X)], \tag{9}$$

$$L_n = -\frac{1}{n}\sum_{i=1}^{n} \log p_0(X_i). \tag{10}$$

Then, by definition, $L_0 = \mathbb{E}[L_n]$. Using these values, Bayes generalization error $B_g(n)$ and Bayes training error $B_t(n)$ are defined, respectively, as

$$B_g(n) = B_g L(n) - L_0, \tag{11}$$
$$B_t(n) = B_t L(n) - L_n. \tag{12}$$

Let us define a log density ratio function as:

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

which is equivalent to

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

Then, it immediately follows that

$$\begin{aligned} B_g(n) &= -\mathbb{E}_X[\log \mathbb{E}_w[\exp(-f(X,w))]], \\ B_t(n) &= -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_w[\exp(-f(X_i,w))], \\ V(n) &= \sum_{i=1}^{n} \Big\{ \mathbb{E}_w[f(X_i,w)^2] - \mathbb{E}_w[f(X_i,w)]^2 \Big\}. \end{aligned}$$

Therefore, the problem of statistical learning is characterized by the function $f(x,w)$.

Definition 2 *(1) If $q(x) = p_0(x)$, then $q(x)$ is said to be realizable by $p(x|w)$. Otherwise, $q(x)$ is said to be unrealizable.*
(2) If the set W_0 consists of a single point w_0 and if the Hessian matrix $\nabla\nabla L(w_0)$ is strictly positive definite, then $q(x)$ is said to be regular for $p(x|w)$. Otherwise, $q(x)$ is said to be singular for $p(x|w)$.

Bayes learning theory was investigated for a realizable and regular case (Schwarz, 1978; Levin et al., 1990; Amari, 1993). The WAIC was found for a realizable and singular case (Watanabe, 2001a, 2009, 2010a) and for an unrealizable and regular case (Watanabe, 2010b). In addition, WAIC was generalized for an unrealizable and singular case (Watanabe, 2010d).

2.3 Singular Learning Theory

We summarize singular learning theory. In the present paper, we assume the followings.

2.3.1 ASSUMPTIONS

(1) The set of parameters W is a compact set in $\mathbb{R}^d$, the open kernel[1] of which is not the empty set. The boundary of W is defined by several analytic functions,

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

1. The open kernel of a set A is the largest open set that is contained in A.

(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 $s \geq 8$ and let

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

be a Banach space. The map $W \ni w \mapsto f(x,w)$ is an $L^s(q)$ valued analytic function.
(4) A nonnegative function $K(w)$ is defined as

$$K(w) = \mathbb{E}_X[f(X,w)].$$

The set W_ε is defined as

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

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

$$(\forall w \in W_\varepsilon) \quad \mathbb{E}_X[f(X,w)] \geq c\, \mathbb{E}_X[f(X,w)^2]. \tag{13}$$

Remark 3 *In ordinary learning problems, if the true distribution is regular for or realizable by a learning machine, then assumptions (1), (2), (3) and (4) are satisfied, and the results of the present paper hold. If the true distribution is singular for and unrealizable by a learning machine, then assumption (4) is satisfied in some cases but not in other cases. If the assumption (4) is not satisfied, then the Bayes generalization and training errors may have asymptotic behaviors other than those described in Lemma 1 (Watanabe, 2010d).*

The investigation of cross-validation in singular learning machines requires singular learning theory. In previous papers, we obtained the following lemma.

Lemma 1 *Assume that assumptions (1), (2), (3), and (4) are satisfied. Then, the followings hold.*
(1) Three random variables $nB_g(n)$, $nB_t(n)$, and $V(n)$ converge in law, when n tends to infinity. In addition, the expectation values of these variables converge.
(2) For $k = 1,2,3,4$, we define

$$M_k(n) \equiv \sup_{|\alpha| \leq 1+\beta} \mathbb{E}\Big[\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{E}_w[|f(X_i,w)|^k \exp(\alpha f(X_i,w))]}{\mathbb{E}_w[\exp(\alpha f(X_i,w))]} \Big],$$

where $\mathbb{E}[\ ]$ gives the average over all sets of training samples. Then,

$$\mathrm{limsup}_{n\to\infty} \Big(n^{k/2}\, M_k(n) \Big) < \infty. \tag{14}$$

(3) The expectation value of the Bayes generalization loss is asymptotically equal to the widely applicable information criterion,

$$\mathbb{E}[B_g L(n)] = \mathbb{E}[WAIC(n)] + o(\frac{1}{n}). \tag{15}$$

Proof For the case in which $q(x)$ is realizable by and singular for $p(x|w)$, this lemma was proven in Watanabe (2010a) and Watanabe (2009). In fact, the proof of Lemma 1 (1) is given in Theorem 1 of Watanabe (2010a). Also Lemma 1 (2) can be proven in the same manner as Equation (32) in Watanabe (2010a) or Equation (6.59) in Watanabe (2009). The proof of Lemma 1 (3) is given in Theorem 2 and the discussion of Watanabe (2010a). For the case in which $q(x)$ is regular for and unrealizable by $p(x|w)$, this lemma was proven in Watanabe (2010b). For the case in which $q(x)$ is singular for and unrealizable by $p(x|w)$, these results can be generalized under the condition that Equation (13) is satisfied (Watanabe, 2010d). ■

3. Bayes Cross-Validation

In this section, we introduce the cross-validation in Bayes learning.

The expectation value $\mathbb{E}_w^{(i)}[\ \ ]$ using the posterior distribution leaving out X_i is defined as

$$\mathbb{E}_w^{(i)}[\ \ ] = \frac{\int (\ \ ) \prod_{j\neq i}^{n} p(X_j|w)^{\beta}\, \varphi(w) dw}{\int \prod_{j\neq i}^{n} p(X_j|w)^{\beta}\, \varphi(w) dw}, \tag{16}$$

where $\prod_{j\neq i}^{n}$ shows the product for $j = 1,2,3,..,n$, which does not include $j = i$. The predictive distribution leaving out X_i is defined as

$$p^{(i)}(x) = \mathbb{E}_w^{(i)}[p(x|w)].$$

The log loss of $p^{(i)}(x)$ when X_i is used as a testing sample is

$$-\log p^{(i)}(X_i) = -\log \mathbb{E}_w^{(i)}[p(X_i|w)].$$

Thus, the log loss of the Bayes cross-validation is defined as the empirical average of them,

$$C_v L(n) = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_w^{(i)}[p(X_i|w)]. \tag{17}$$

The random variable $C_v L(n)$ is referred to as the *cross-validation loss*. Since $X_1, X_2, ..., X_n$ are independent training samples, it immediately follows that

$$\mathbb{E}[C_v L(n)] = \mathbb{E}[B_g L(n-1)].$$

Although the two random variables $C_v L(n)$ and $B_g L(n-1)$ are different,

$$C_v L(n) \neq B_g L(n-1),$$

their expectation values coincide with each other by the definition. Using Equation (15), it follows that

$$\mathbb{E}[C_v L(n)] = \mathbb{E}[\mathrm{WAIC}(n-1)] + o(\frac{1}{n}).$$

Therefore, three expectation values $\mathbb{E}[C_v L(n)]$, $\mathbb{E}[B_g L(n-1)]$, and $\mathbb{E}[\mathrm{WAIC}(n-1)]$ are asymptotically equal to each other. The primary goal of the present paper is to clarify the asymptotic behaviors of three random variables, $C_v L(n)$, $B_g L(n)$, and $\mathrm{WAIC}(n)$, when n is sufficiently large.

Remark 4 *In practical applications, the Bayes generalization loss $B_g L(n)$ indicates the accuracy of Bayes estimation. However, in order to calculate $B_g L(n)$, we need the expectation value over the testing sample taken from the unknown true distribution, hence we cannot directly obtain $B_g L(n)$ in practical applications. On the other hand, both the cross-validation loss $C_v L(n)$ and the widely applicable information criterion $WAIC(n)$ can be calculated using only training samples. Therefore, the cross-validation loss and the widely applicable information criterion can be used for model selection and hyperparameter optimization. This is the reason why comparison of these random variables is an important problem in statistical learning theory.*

4. Main Results

In this section, the main results of the present paper are explained. First, we define functional cumulants and describe their asymptotic properties. Second, we prove that both the cross-validation loss and the widely applicable information criterion can be represented by the functional cumulants. Finally, we prove that the cross-validation loss and the widely applicable information criterion are related to the birational invariants.

4.1 Functional Cumulants

Definition 5 *The generating function $F(\alpha)$ of functional cumulants is defined as*

$$F(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_w[p(X_i|w)^{\alpha}].$$

The kth order functional cumulant $Y_k(n)$ $(k = 1,2,3,4)$ is defined as

$$Y_k(n) = \frac{d^k F}{d\alpha^k}(0). \tag{18}$$

Then, by definition,

$$\begin{aligned} F(0) &= 0, \\ F(1) &= -B_t L(n), \\ Y_1(n) &= -G_t L(n), \\ Y_2(n) &= V(n)/n. \end{aligned}$$

For simple notation, we use

$$\ell_k(X_i) = \mathbb{E}_w[(\log p(X_i|w))^k] \quad (k = 1,2,3,4).$$

Lemma 2 *Then, the following equations hold:*

$$Y_1(n) = \frac{1}{n}\sum_{i=1}^{n} \ell_1(X_i), \tag{19}$$

$$Y_2(n) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\ell_2(X_i) - \ell_1(X_i)^2\Big\}, \tag{20}$$

$$Y_3(n) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\ell_3(X_i) - 3\ell_2(X_i)\ell_1(X_i) + 2\ell_1(X_i)^3\Big\}, \tag{21}$$

$$Y_4(n) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\ell_4(X_i) - 4\ell_3(X_i)\ell_1(X_i) - 3\ell_2(X_i)^2 + 12\ell_2(X_i)\ell_1(X_i)^2 - 6\ell_1(X_i)^4\Big\}. \tag{22}$$

Moreover,

$$Y_k(n) = O_p(\frac{1}{n^{k/2}}) \quad (k = 2,3,4).$$

In other words,

$$\mathrm{limsup}_{n\to\infty}\mathbb{E}[n^{k/2}\,|Y_k(n)|] < \infty \quad (k = 2,3,4). \tag{23}$$

Proof First, we prove Equations (19) through (22). Let us define

$$g_i(\alpha) = \mathbb{E}_w[p(X_i|w)^\alpha].$$

Then, $g_i(0) = 1$,

$$g_i^{(k)}(0) \equiv \frac{d^k g_i}{d\alpha^k}(0) = \ell_k(X_i) \quad (k = 1,2,3,4),$$

and

$$F(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\log g_i(\alpha).$$

For arbitrary natural number k,

$$\Big(\frac{g_i(\alpha)^{(k)}}{g_i(\alpha)}\Big)' = \frac{g_i(\alpha)^{(k+1)}}{g_i(\alpha)} - \Big(\frac{g_i(\alpha)^{(k)}}{g_i(\alpha)}\Big)\Big(\frac{g_i(\alpha)'}{g_i(\alpha)}\Big).$$

By applying this relation recursively, Equations (19), (20), (21), and (22) are derived. Let us prove Equation (23). The random variables $Y_k(n)$ $(k = 2,3,4)$ are invariant under the transform,

$$\log p(X_i|w) \mapsto \log p(X_i|w) + c(X_i), \tag{24}$$

for arbitrary $c(X_i)$. In fact, by replacing $p(X_i|w)$ by $p(X_i|w)e^{C(X_i)}$, we define

$$\hat{F}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\log\mathbb{E}_w[p(X_i|w)^\alpha\, e^{\alpha c(X_i)}].$$

Then, the difference between $F(\alpha)$ and $\hat{F}(\alpha)$ is a linear function of α, which vanishes by higher-order differentiation. In particular, by selecting $c(X_i) = -\log p_0(X_i)$, we can show that $Y_k(n)$ $(k = 2,3,4)$ are invariant by the following replacement,

$$\log p(X_i|w) \mapsto f(X_i, w).$$

In other words, $Y_k(n)$ $(n=2,3,4)$ are invarianrt by the replacement,

$$\ell_k(X_i) \mapsto \mathbb{E}_w[f(X_i,w)^k].$$

Using the Cauchy-Schwarz inequality, for $1 \leq k' \leq k$,

$$\mathbb{E}_w[|f(X_i,w)|^{k'}]^{1/k'} \leq \mathbb{E}_w[|f(X_i,w)|^k]^{1/k}.$$

Therefore, for $k=2,3,4$,

$$\mathbb{E}[|Y_k(n)|] \leq \mathbb{E}\Big[\frac{C_k}{n}\sum_{i=1}^n \mathbb{E}_w[|f(X_i,w)|^k]\Big] \leq C_k M_k(n),$$

where $C_2=2, C_3=6, C_4=26$. Then, using Equation (14), we obtain Equation (23). ■

Remark 6 *Using Equation (24) with* $c(X_i) = -\mathbb{E}_w[\log p(X_i|w)]$ *and the normalized function defined as*

$$\ell_k^*(X_i) = \mathbb{E}_w[(\log p(X_i|w) - c(X_i))^k],$$

it follows that

$$\begin{aligned} Y_2(n) &= \frac{1}{n}\sum_{i=1}^n \ell_2^*(X_i), \\ Y_3(n) &= \frac{1}{n}\sum_{i=1}^n \ell_3^*(X_i), \\ Y_4(n) &= \frac{1}{n}\sum_{i=1}^n \Big\{\ell_4^*(X_i) - 3\ell_2^*(X_i)^2\Big\}. \end{aligned}$$

These formulas may be useful in practical applications.

4.2 Bayes Cross-validation and Widely Applicable Information Criterion

We show the asymptotic equivalence of the cross-validation loss $C_vL(n)$ and the widely applicable information criterion $\mathrm{WAIC}(n)$.

Theorem 1 *For arbitrary* $0 < \beta < \infty$, *the cross-validation loss* $C_vL(n)$ *and the widely applicable information criterion* $\mathrm{WAIC}(n)$ *are given, respectively, as*

$$\begin{aligned} C_vL(n) &= -Y_1(n) + \Big(\frac{2\beta-1}{2}\Big)Y_2(n) \\ &\quad -\Big(\frac{3\beta^2-3\beta+1}{6}\Big)Y_3(n) + O_p(\frac{1}{n^2}), \\ \mathrm{WAIC}(n) &= -Y_1(n) + \Big(\frac{2\beta-1}{2}\Big)Y_2(n) \\ &\quad -\frac{1}{6}Y_3(n) + O_p(\frac{1}{n^2}). \end{aligned}$$

Proof First, we consider $C_v L(n)$. From the definitions of $\mathbb{E}_w[\ ]$ and $\mathbb{E}_w^{(i)}[\ ]$, we have

$$\mathbb{E}_w^{(i)}[(\ \ )] = \frac{\mathbb{E}_w[(\ \ )p(X_i|w)^{-\beta}\,]}{\mathbb{E}_w[p(X_i|w)^{-\beta}\,]}. \tag{25}$$

Therefore, by the definition of the cross-validation loss, Equation (17),

$$C_v L(n) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\mathbb{E}_w[\,p(X_i|w)^{1-\beta}\,]}{\mathbb{E}_w[\,p(X_i|w)^{-\beta}\,]}.$$

Using the generating function of functional cumulants $F(\alpha)$,

$$C_v L(n) = F(-\beta) - F(1-\beta). \tag{26}$$

Then, using Lemma 1 (2) for each $k = 2,3,4$, and $|\alpha| < 1+\beta$,

$$\begin{aligned}
\mathbb{E}[|F^{(k)}(\alpha)|] &\leq \mathbb{E}\Big[\frac{C_k}{n}\sum_{i=1}^{n}\frac{\mathbb{E}_w[|f(X_i,w)|^k \exp(\alpha f(X_i,w))]}{\mathbb{E}_w[\exp(\alpha f(X_i,w))]}\Big] \\
&\leq C_k M_k(n),
\end{aligned}$$

where $C_2 = 2, C_3 = 6, C_4 = 26$. Therefore,

$$|F^{(k)}(\alpha)| = O_p(\frac{1}{n^{k/2}}). \tag{27}$$

By Taylor expansion of $F(\alpha)$ among $\alpha = 0$, there exist β^*, β^{**} ($|\beta^*|, |\beta^{**}| < 1+\beta$) such that

$$\begin{aligned}
F(-\beta) &= F(0) - \beta F'(0) + \frac{\beta^2}{2}F''(0) \\
&\quad -\frac{\beta^3}{6}F^{(3)}(0) + \frac{\beta^4}{24}F^{(4)}(\beta^*), \\
F(1-\beta) &= F(0) + (1-\beta)F'(0) + \frac{(1-\beta)^2}{2}F''(0) \\
&\quad +\frac{(1-\beta)^3}{6}F^{(3)}(0) + \frac{(1-\beta)^4}{24}F^{(4)}(\beta^{**}).
\end{aligned}$$

Using $F(0) = 0$ and Equations (26) and (27), it follows that

$$\begin{aligned}
C_v L(n) &= -F'(0) + \frac{2\beta-1}{2}F''(0) \\
&\quad -\frac{3\beta^2 - 3\beta + 1}{6}F^{(3)}(0) + O_p(\frac{1}{n^2}).
\end{aligned}$$

Thus, we have proven the first half of the theorem. For the latter half, by the definitions of $\mathrm{WAIC}(n)$, Bayes training loss, and the functional variance, we have

$$\begin{aligned}
\mathrm{WAIC}(n) &= B_t L(n) + (\beta/n)V(n), \\
B_t L(n) &= -F(1), \\
V(n) &= nF''(0).
\end{aligned}$$

Therefore,

$$\mathrm{WAIC}(n) = -F(1) + \beta F''(0).$$

By Taylor expansion of $F(1)$, we obtain

$$\mathrm{WAIC}(n) = -F'(0) + \frac{2\beta - 1}{2} F''(0) - \frac{1}{6} F^{(3)}(0) + O_p(\frac{1}{n^2}),$$

which completes the proof. ■

From the above theorem, we obtain the following corollary.

Corollary 1 *For arbitrary $0 < \beta < \infty$, the cross-validation loss $C_v L(n)$ and the widely applicable information criterion* $\mathrm{WAIC}(n)$ *satisfy*

$$C_v L(n) = \mathrm{WAIC}(n) + O_p(\frac{1}{n^{3/2}}).$$

In particular, for $\beta = 1$,

$$C_v L(n) = \mathrm{WAIC}(n) + O_p(\frac{1}{n^2}).$$

More precisely, the difference between the cross-validation loss and the widely applicable information criterion is given by

$$C_v L(n) - \mathrm{WAIC}(n) \cong \Big(\frac{\beta - \beta^2}{2}\Big) Y_3(n).$$

If $\beta = 1$,

$$C_v L(n) - \mathrm{WAIC}(n) \cong \frac{1}{12} Y_4(n).$$

4.3 Generalization Error and Cross-validation Error

In the previous subsection, we have shown that the cross-validation loss is asymptotically equivalent to the widely applicable information criterion. In this section, let us compare the Bayes generalization error $B_g(n)$ given in Equation (11) and the cross-validation error $C_v(n)$, which is defined as

$$C_v(n) = C_v L(n) - L_n. \tag{28}$$

We need mathematical concepts, the real log canonical threshold, and the singular fluctuation.

Definition 7 *The zeta function $\zeta(z)$ $(Re(z) > 0)$ of statistical learning is defined as*

$$\zeta(z) = \int K(w)^z \varphi(w) dw,$$

where

$$K(w) = \mathbb{E}_X[f(X,w)]$$

is a nonnegative analytic function. Here, $\zeta(z)$ can be analytically continued to the unique meromorphic function on the entire complex plane $\mathbb{C}$. All poles of $\zeta(z)$ are real, negative, and rational numbers. The maximum pole is denoted as

$$(-\lambda) = \textit{ maximum pole of } \zeta(z). \tag{29}$$

Then, the positive rational number λ is referred to as the real log canonical threshold. The singular fluctuation is defined as

$$\nu = \nu(\beta) = \lim_{n\to\infty} \frac{\beta}{2}\mathbb{E}[V(n)]. \tag{30}$$

Note that the real log canonical threshold does not depend on β, whereas the singular fluctuation is a function of β.

Both the real log canonical threshold and the singular fluctuation are birational invariants. In other words, they are determined by the algebraic geometrical structure of the statistical model. The following lemma was proven in a previous study (Watanabe, 2010a,b,d).

Lemma 3 *The following convergences hold:*

$$\lim_{n\to\infty} n\mathbb{E}[B_g(n)] = \frac{\lambda-\nu}{\beta} + \nu, \tag{31}$$

$$\lim_{n\to\infty} n\mathbb{E}[B_t(n)] = \frac{\lambda-\nu}{\beta} - \nu, \tag{32}$$

Moreover, convergence in probability

$$n(B_g(n) + B_t(n)) + V(n) \to \frac{2\lambda}{\beta} \tag{33}$$

holds.

Proof For the case in which $q(x)$ is realizable by and singular for $p(x|w)$, Equations (31) and (32) were proven by in Corollary 3 in Watanabe (2010a). The equation (33) was given in Corollary 2 in Watanabe (2010a). For the case in which $q(x)$ is regular for $p(x|w)$, these results were proved in Watanabe (2010b). For the case in which $q(x)$ is singular for and unrealizable by $p(x|w)$ they were generalized in Watanabe (2010d). ■

4.3.1 EXAMPLES

If $q(x)$ is regular for and realizable by $p(x|w)$, then $\lambda = \nu = d/2$, where d is the dimension of the parameter space. If $q(x)$ is regular for and unrealizable by $p(x|w)$, then λ and ν are given by Watanabe (2010b). If $q(x)$ is singular for and realizable by $p(x|w)$, then λ for several models are obtained by resolution of singularities (Aoyagi and Watanabe, 2005; Rusakov and Geiger, 2005; Yamazaki and Watanabe, 2003; Lin, 2010; Zwiernik, 2010). If $q(x)$ is singular for and unrealizable by $p(x|w)$, then λ and ν remain unknown constants.

We have the following theorem.

Theorem 2 *The following equation holds:*

$$\lim_{n\to\infty} n\mathbb{E}[C_v(n)] = \frac{\lambda-\nu}{\beta}+\nu,$$

The sum of the Bayes generalization error and the cross-validation error satisfies

$$B_g(n)+C_v(n) = (\beta-1)\frac{V(n)}{n}+\frac{2\lambda}{\beta n}+o_p(\frac{1}{n}).$$

In particular, if $\beta=1$,

$$B_g(n)+C_v(n) = \frac{2\lambda}{n}+o_p(\frac{1}{n}).$$

Proof By Equation (31),

$$\mathbb{E}[B_g(n-1)] = \Big(\frac{\lambda-\nu}{\beta}+\nu\Big)\frac{1}{n}+o(\frac{1}{n}).$$

Since $\mathbb{E}[C_v(n)] = \mathbb{E}[B_g(n-1)]$,

$$\begin{aligned}\lim_{n\to\infty} n\mathbb{E}[C_v(n)] &= \lim_{n\to\infty} n\mathbb{E}[B_g(n-1)]\\ &= \frac{\lambda-\nu}{\beta}+\nu.\end{aligned}$$

From Equation (33) and Corollary 1,

$$B_t(n) = C_v(n)-\frac{\beta}{n}V(n)+O_p(\frac{1}{n^{3/2}}),$$

and it follows that

$$(B_g(n)+C_v(n)) = (\beta-1)\frac{V(n)}{n}+\frac{2\lambda}{\beta n}+o_p(\frac{1}{n}),$$

which proves the Theorem. ■

This theorem indicates that both the cross-validation error and the Bayes generalization error are determined by the algebraic geometrical structure of the statistical model, which is extracted as the real log canonical threshold. From this theorem, in the strict Bayes case $\beta=1$, we have

$$\begin{aligned}\mathbb{E}[B_g(n)] &= \frac{\lambda}{n}+o(\frac{1}{n}),\\ \mathbb{E}[C_v(n)] &= \frac{\lambda}{n}+o(\frac{1}{n}),\end{aligned}$$

and

$$B_g(n)+C_v(n) = \frac{2\lambda}{n}+o_p(\frac{1}{n}). \tag{34}$$

Therefore, the smaller cross-validation error $C_v(n)$ is equivalent to the larger Bayes generalization error $B_g(n)$. Note that a regular statistical model is a special example of singular models, hence both Theorems 1 and 2 also hold in regular statistical models. In Watanabe (2009), it was proven that the random variable $nB_g(n)$ converges to a random variable in law. Thus, $nC_v(n)$ converges to a random variable in law. The asymptotic probability distribution of $nB_g(n)$ can be represented using a Gaussian process, which is defined on the set of true parameters, but is not equal to the χ^2 distribution in general.

Remark 8 *The relation given by Equation (34) indicates that, if $\beta = 1$, the variances of $B_g(n)$ and $C_v(n)$ are equal. If the average value $2\nu = \mathbb{E}[V(n)]$ is known, then $B_t(n) + 2\nu/n$ can be used instead of $C_v(n)$, because both average values are asymptotically equal to the Bayes generalization error. The variance of $B_t(n) + 2\nu/n$ is smaller than that of $C_v(n)$ if and only if the variance of $B_t(n)$ is smaller than that of $B_g(n)$. If a true distribution is regular for and realizable by the statistical model, then the variance of $B_t(n)$ is asymptotically equal to that of $B_g(n)$. However, in other cases, the variance of $B_t(n)$ may be smaller or larger than that of $B_g(n)$.*

5. Discussion

Let us now discuss the results of the present paper.

5.1 From Regular to Singular

First, we summarize the regular and singular learning theories.

In regular statistical models, the generalization loss of the maximum likelihood method is asymptotically equal to that of the Bayes estimation. In both the maximum likelihood and Bayes methods, the cross-validation losses have the same asymptotic behaviors. The leave-one-out cross-validation is asymptotically equivalent to the AIC, in both the maximum likelihood and Bayes methods.

On the other hand, in singular learning machines, the generalization loss of the maximum likelihood method is larger than the Bayes generalization loss. Since the generalization loss of the maximum likelihood method is determined by the maximum value of the Gaussian process, the maximum likelihood method is not appropriate in singular models (Watanabe, 2009). In Bayes estimation, we derived the asymptotic expansion of the generalization loss and proved that the average of the widely applicable information criterion is asymptotically equal to the Bayes generalization loss (Watanabe, 2010a). In the present paper, we clarified that the leave-one-out cross-validation in Bayes estimation is asymptotically equivalent to WAIC.

It was proven (Watanabe, 2001a) that the Bayes marginal likelihood of a singular model is different from BIC of a regular model. In the future, we intend to compare the cross-validation and Bayes marginal likelihood in model selection and hyperparameter optimization in singular statistical models.

5.2 Cross- validation and Importance Sampling

Second, let us investigate the cross-validation and the importance sampling cross-validation from a practical viewpoint.

In Theorem 1, we theoretically proved that the leave-one-out cross-validation is asymptotically equivalent to the widely applicable information criterion. In practical applications, we often approx-

imate the posterior distribution using the Markov Chain Monte Carlo or other numerical methods. If the posterior distribution is precisely realized, then the two theorems of the present paper hold. However, if the posterior distribution was not precisely approximated, then the cross-validation might not be equivalent to the widely applicable information criterion.

In Bayes estimation, there are two different methods by which the leave-one-out cross-validation is numerically approximated. In the former method, CV_1 is obtained by realizing all posterior distributions $\mathbb{E}_w^{(i)}[\ ]$ leaving out X_i for $i = 1, 2, 3, ..., n$, and the empirical average

$$CV_1 = -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_w^{(i)}[p(X_i|w)]$$

is then calculated. In this method, we must realize n different posterior distributions, which requires heavy computational costs.

In the latter method, the posterior distribution leaving out X_i is estimated using the posterior average $\mathbb{E}_w[\ ]$, in the same manner as Equation (25),

$$\mathbb{E}_w^{(i)}[p(X_i|w)] \cong \frac{\mathbb{E}_w[p(X_i|w)\ p(X_i|w)^{-\beta}\ ]}{\mathbb{E}_w[p(X_i|w)^{-\beta}\ ]}.$$

This method is referred to as the importance sampling leave-one-out cross-validation (Gelfand et al., 1992), in which only one posterior distribution is needed and the leave-one-out cross-validation is approximated by CV_2,

$$CV_2 \cong -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\mathbb{E}_w[p(X_i|w)\ p(X_i|w)^{-\beta}\ ]}{\mathbb{E}_w[p(X_i|w)^{-\beta}\ ]}.$$

If the posterior distribution is completely realized, then CV_1 and CV_2 coincide with each other and are asymptotically equivalent to the widely applicable information criterion. However, if the posterior distribution is not sufficiently approximated, then the values CV_1, CV_2, and WAIC(n) might be different.

The average values using the posterior distribution may sometimes have infinite variances (Peruggia, 1997) if the set of parameters is not compact. Moreover, in singular learning machines, the set of true parameters is not a single point but rather an analytic set, hence we must restrict the parameter space to be compact for well-defined average values. Therefore, we adopted the assumptions in Section 2.3 that the parameter space is compact and the log likelihood function has the appropriate properties. Under these conditions, the observables studied in the present paper have finite variances.

5.3 Comparison with the Deviance Information Criteria

Third, let us compare the deviance information criterion (DIC) (Spiegelhalter et al., 2002) to the Bayes cross-validation and WAIC, because DIC is sometimes used in Bayesian model evaluation. In order to estimate the Bayesian generalization error, DIC is written by

$$DIC_1 = B_t L(n) + \frac{2}{n}\sum_{i=1}^{n} \Big\{ -E_w[\log p(X_i|w)] + \log p(X_i|E_w[w]) \Big\},$$

where the second term of the right-hand side corresponds to the "effective number of parameters" of DIC divided by the number of parameters. Under the condition that the log likelihood ratio function

in the posterior distribution is subject to the χ^2 distribution, a modified DIC was proposed (Gelman et al., 2004) as

$$DIC_2 = B_t L(n) + \frac{2}{n}\Big[E_w[\Big\{\sum_{i=1}^{n}\log p(X_i|w)\Big\}^2] - E_w[\sum_{i=1}^{n}\log p(X_i|w)]^2\Big],$$

the variance of which was investigated previously (Raftery, 2007). Note that DIC_2 is different from WAIC. In a singular learning machine, since the set of optimal parameters is an analytic set, the correlation between different true parameters does not vanish, even asymptotically.

We first derive the theoretical properties of DIC. If the true distribution is regular for the statistical model, then the set of the optimal parameter is a single point w_0. Thus, the difference of $E_w[w]$ and the maximum *a posteriori* estimator is asymptotically smaller than $1/\sqrt{n}$. Therefore, based on the results in Watanabe (2010b), if $\beta = 1$,

$$\mathbb{E}[DIC_1] = L_0 + (3\lambda - 2\nu(1))\frac{1}{n} + o(\frac{1}{n}).$$

If the true distribution is realizable by or regular for the statistical model and if $\beta = 1$, then the asymptotic behavior of DIC_2 is given by

$$\mathbb{E}[DIC_2] = L_0 + (3\lambda - 2\nu(1) + 2\nu'(1))\frac{1}{n} + o(\frac{1}{n}), \tag{35}$$

where $\nu'(1) = (d\nu/d\beta)(1)$. Equation (35) is derived from the relations (Watanabe, 2009, 2010a,b,d),

$$\begin{aligned} DIC_2 &= B_t L(n) - 2\frac{\partial}{\partial\beta}G_t L(n), \\ \mathbb{E}[G_t L(n)] &= L_0 + \Big(\frac{\lambda}{\beta} - \nu(\beta)\Big)\frac{1}{n} + o(\frac{1}{n}), \end{aligned}$$

where $G_t L(n)$ is given by Equation (7).

Next, let us consider the DIC for each case. If the true distribution is regular for and realizable by the statistical model and if $\beta = 1$, then $\lambda = \nu = d/2$, $\nu'(1) = 0$, where d is the number of parameters. Thus, their averages are asymptotically equal to the Bayes generalization error,

$$\begin{aligned} \mathbb{E}[DIC_1] &= L_0 + \frac{d}{2n} + o(\frac{1}{n}), \\ \mathbb{E}[DIC_2] &= L_0 + \frac{d}{2n} + o(\frac{1}{n}). \end{aligned}$$

In this case, the averages of DIC_1, DIC_2, CV_1, CV_2, and WAIC have the same asymptotic behavior.

If the true distribution is regular for and unrealizable by the statistical model and if $\beta = 1$, then $\lambda = d/2$, $\nu = (1/2)\mathrm{tr}(IJ^{-1})$, and $\nu'(1) = 0$ (Watanabe, 2010b), where I is the Fisher information matrix at w_0, and J is the Hessian matrix of $L(w)$ at $w = w_0$. Thus, we have

$$\begin{aligned} \mathbb{E}[DIC_1] &= L_0 + \Big(\frac{3d}{2} - \mathrm{tr}(IJ^{-1})\Big)\frac{1}{n} + o(\frac{1}{n}), \\ \mathbb{E}[DIC_2] &= L_0 + \Big(\frac{3d}{2} - \mathrm{tr}(IJ^{-1})\Big)\frac{1}{n} + o(\frac{1}{n}). \end{aligned}$$

In this case, as shown in Lemma 3, the Bayes generalization error is given by $L_0 + d/(2n)$ asymptotically, and so the averages of the deviance information criteria are not equal to the average of the Bayes generalization error.

If the true distribution is singular for and realizable by the statistical model and if $\beta = 1$, then

$$\begin{aligned} \mathbb{E}[DIC_1] &= C + o(1), \\ \mathbb{E}[DIC_2] &= L_0 + (3\lambda - 2\nu(1) + 2\nu'(1))\frac{1}{n} + o(\frac{1}{n}), \end{aligned} \tag{36}$$

where C $(C \neq L_0)$ is, in general, a constant. Equation (36) is obtained because the set of true parameters in a singular model is not a single point, but rather an analytic set, so that, in general, the average $E_w[w]$ is not contained in the neighborhood of the set of the true parameters. Hence the averages of the deviance information criteria are not equal to those of the Bayes generalization error.

The averages of the cross-validation loss and WAIC have the same asymptotic behavior as that of the Bayes generalization error, even if the true distribution is unrealizable by or singular for the statistical model. Therefore, the deviance information criteria are different from the cross-validation and WAIC, if the true distribution is singular for or unrealizable by the statistical model.

5.4 Experiment

In this section, we describe an experiment. The purpose of the present paper is to clarify the theoretical properties of the cross-validation and the widely applicable information criterion. An experiment was conducted in order to illustrate the main theorems.

Let $x, y \in \mathbb{R}^3$. We considered a statistical model defined as

$$p(x,y|w) = \frac{s(x)}{(2\pi\sigma^2)^{3/2}} \exp(-\frac{\|y - R_H(x,w)\|^2}{2\sigma^2}),$$

where $\sigma = 0.1$ and $s(x)$ is $\mathcal{N}(0, 2^2 I)$. Here, $\mathcal{N}(m,A)$ exhibits a normal distribution with the average vector m and the covariance matrix A, and I is the identity matrix. Note that the distribution $s(x)$ was not estimated. We used a three-layered neural network,

$$R_H(x,w) = \sum_{h=1}^{H} a_h \tanh(b_h \cdot x),$$

where the parameter was

$$w = \{(a_h \in \mathbb{R}^3, b_h \in \mathbb{R}^3)\ ;\ h = 1,2,...,H\} \in \mathbb{R}^{6H}.$$

In the experiment, a learning machine with $H = 3$ was used and the true distribution was set with $H = 1$. The parameter that gives the distribution is denoted as w_0, which denotes the parameters of both models $H = 1, 3$. Then, $R_H(x,w_0) = R_{H_0}(x,w_0)$. Under this condition, the set of true parameters

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

is not a single point but an analytic set with singularities, resulting that the regularity condition is not satisfied. In this case, the log density ratio function is equivalent to

$$f(x,y,w) = \frac{1}{2\sigma^2}\Big\{\|y - R_H(x,w)\|^2 - \|y - R_H(x,w_0)\|^2\Big\}.$$

In this model, although the Bayes generalization error is not equal to the average square error

$$\mathrm{SE}(n) = \frac{1}{2\sigma^2}\mathbb{E}\mathbb{E}_X\Big[\,\| R_H(X,w_0) - \mathbb{E}_w[R_H(X,w)]\,\|^2\Big],$$

asymptotically $\mathrm{SE}(n)$ and $B_g(n)$ are equal to each other (Watanabe, 2009).

The prior distribution $\varphi(w)$ was set as $\mathcal{N}(0,10^2 I)$. Although this prior does not have compact support mathematically, it can be understood in the experiment that the support of $\varphi(w)$ is essentially contained in a sufficiently large compact set.

In the experiment, the number of training samples was fixed as $n = 200$. One hundred sets of 200 training samples each were obtained independently. For each training set, the strict Bayes posterior distribution $\beta = 1$ was approximated by the Markov chain Monte Carlo (MCMC) method. The Metropolis method, in which each random trial was taken from $\mathcal{N}(0,(0.005)^2 I)$, was applied, and the average exchanging ratio was obtained as approximately 0.35. After 100,000 iterations of Metropolis random sampling, 200 parameters were obtained in every 100 sampling steps. For a fixed training set, by changing the initial values and the random seeds of the software, the same MCMC sampling procedures were performed 10 times independently, which was done for the purpose of minimizing the effect of the local minima. Finally, for each training set, we obtained $200 \times 10 = 2,000$ parameters, which were used to approximate the posterior distribution.

Table 2 shows the experimental results. We observed the Bayes generalization error $BG = B_g(n)$, the Bayes training error $BT = B_t(n)$, importance sampling leave-one-out cross-validation $CV = CV_2 - L_n$, the widely applicable information criterion $\mathrm{WAIC} = \mathrm{WAIC}(n) - L_n$, two deviance information criteria, namely, $DIC1 = DIC_1 - L_n$ and $DIC2 = DIC_2 - L_n$, and the sum $BG + CV = B_g(n) + C_v(n)$. The values AVR and STD in Table 2 show the average and standard deviation of one hundred sets of training data, respectively. The original cross-validation CV_1 was not observed because the associated computational cost was too high.

The experimental results reveal that the average and standard deviation of BG were approximately the same as those of CV and WAIC, which indicates that Theorem 1 holds. The real log canonical threshold, the singular fluctuation, and its derivative of this case were estimated as

$$\begin{aligned} \lambda &\approx 5.6, \\ \nu(1) &\approx 7.9, \\ \nu'(1) &\approx 3.6. \end{aligned}$$

Note that, if the true distribution is regular for and realizable by the statistical model, $\lambda = \nu(1) = d/2 = 9$ and $\nu'(1) = 0$. The averages of the two deviance information criteria were not equal to that of the Bayes generalization error. The standard deviation of $BG + CV$ was smaller than the standard deviations of BG and CV, which is in agreement with Theorem 2.

Note that the standard deviation of BT was larger than those of CV and WAIC, which indicates that, even if the average value $\mathbb{E}[C_v(n) - B_t(n)] = 2\nu/n$ is known and an alternative cross-validation, such as the AIC,

$$CV_3 = B_t L(n) + 2\nu/n,$$

is used, then the variance of $CV_3 - L_n$ was larger than the variances of $C_v L(n) - L_n$ and $\mathrm{WAIC}(n) - L_n$.

	BG	*BT*	*CV*	WAIC	*DIC*1	*DIC*2	$BG+CV$
AVR	0.0264	-0.0511	0.0298	0.0278	-35.1077	0.0415	0.0562
STD	0.0120	0.0165	0.0137	0.0134	19.1350	0.0235	0.0071

Table 2: Average and standard deviation

	BG	*BT*	*CV*	WAIC	*DIC*1	*DIC*2	$BG+CV$
BG	1.000	-0.854	-0.854	-0.873	0.031	-0.327	0.043
BT		1.000	0.717	0.736	0.066	0.203	-0.060
CV			1.000	0.996	-0.087	0.340	0.481
WA				1.000	-0.085	0.341	0.443
*DIC*1					1.000	-0.069	-0.115
*DIC*2						1.000	0.102

Table 3: Correlation matrix

Table 3 shows the correlation matrix for several values. The correlation between *CV* and WAIC was 0.996, which indicates that Theorem 1 holds. The correlation between *BG* and *CV* was -0.854, and that between *BG* and WAIC was -0.873, which corresponds to Theorem 2.

The accuracy of numerical approximation of the posterior distribution depends on the statistical model, the true distribution, the prior distribution, the Markov chain Monte Carlo method, and the experimental fluctuation. In the future, we intend to develop a method by which to design experiments. The theorems proven in the present paper may be useful in such research.

5.5 Birational Invariant

Finally, we investigate the statistical problem from an algebraic geometrical viewpoint.

In Bayes estimation, we can introduce an analytic function of the parameter space $g : U \to W$,

$$w = g(u).$$

Let $|g'(u)|$ be its Jacobian determinant. Note that the inverse function g^{-1} is not needed if g satisfies the condition that $\{u \in U; |g'(u)| = 0\}$ is a measure zero set in U. Such a function g is referred to as a birational transform. It is important that, by the transform,

$$\begin{aligned} p(x|w) &\mapsto p(x|g(u)), \\ \varphi(w) &\mapsto \varphi(g(u))|g'(u)|, \end{aligned}$$

the Bayes estimation on W is equivalent to that on U. A constant defined for a set of statistical models and a prior is said to be a birational invariant if it is invariant under such a transform $w = g(u)$.

The real log canonical threshold λ is a birational invariant (Atiyah, 1970; Hiroanaka, 1964; Kashiwara, 1976; Kollór et al., 1998; Mustata, 2002; Watanabe, 2009) that represents the algebraic geometrical relation between the set of parameters W and the set of the optimal parameters W_0. Although the singular fluctuation is also a birational invariant, its properties remain unknown. In the present paper, we proved in Theorem 1 that

$$\mathbb{E}[B_g L(n)] = \mathbb{E}[C_v L(n)] + o(1/n). \tag{37}$$

On the other hand, in Theorem 2, we proved that

$$B_g(n) + C_v(n) = \frac{2\lambda}{n} + o_p(1/n). \tag{38}$$

In model selection or hyperparameter optimization, Equation (37) shows that minimization of the cross-validation makes the generalization loss smaller on average. However, Equation (38) shows that minimization of the cross-validation does not ensure minimum generalization loss. The widely applicable information criterion has the same property as the cross-validation. The constant λ appears to exhibit a bound, which can be attained by statistical estimation for a given pair of a statistical model and a prior distribution. Hence, clarification of the algebraic geometrical structure in statistical estimation is an important problem in statistical learning theory.

6. Conclusion

In the present paper, we have shown theoretically that the leave-one-out cross-validation in Bayes estimation is asymptotically equal to the widely applicable information criterion and that the sum of the cross-validation error and the generalization error is equal to twice the real log canonical threshold divided by the number of training samples. In addition, we clarified that cross-validation and the widely applicable information criterion are different from the deviance information criteria. This result indicates that, even in singular statistical models, the cross-validation is asymptotically equivalent to the information criterion, and that the asymptotic properties of these models are determined by the algebraic geometrical structure of a statistical model.

References

H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 19:716-723, 1974.

S. Amari. A universal theorem on learning curves, *Neural Networks*, 6(2):161-166, 1993.

M. Aoyagi. Stochastic complexity and generalization error of a restricted Boltzmann machine in Bayesian estimation. *Journal of Machine Learning Research*, 11:1243-1272, 2010.

M. Aoyagi, 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.

M.W. Browne. Cross-Validation Methods. *Journal of Mathematical Psychology*, 44:108-132, 2000.

H. Cramer. *Mathematical Methods of Statistics.* Princeton University Press, 1949.

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

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320-328, 1975.

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

A.E. Gelfand, D.K. Dey, H. Chang. Model determination using predictive distributions with implementation via sampling-based method. *Bayesian Statistics*, 4:147-167, Oxford University Press, Oxford, 1992.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin. *Bayesian Data Analysis.* Chapman & Hall CRC, Boca Raton, 2004.

K. Hagiwara. On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation*, 14:1979-2002, 2002.

J. A. Hartigan. A failure of likelihood asymptotics for normal mixtures. In Proceedings of the *Barkeley Conference in Honor of J. Neyman and J. Kiefer*, Vol. 2, pages 807–810, 1985.

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.

J. Kollór, S.Mori, C.H.Clemens, A.Corti. *Birational Geometry of Algebraic Varieties*. Cambridge Tract in Mathematics Cambridge University Press, Cambridge, 1998.

C.I. Mosier. Problems and designs of cross-validation. *Educational and Psychological Measurement*, 11:5-11, 1951.

M. Mustata. Singularities of pairs via jet schemes. *Journal of the American Mathematical Society*, 15:599-615. 2002.

E. Levin, N. Tishby, S.A. Solla. A statistical approaches to learning and generalization in layered neural networks. *Proceedings of IEEE,* 78(10):1568-1574. 1990.

S. Lin. Asymptotic approximation of marginal likelihood integrals. *arXiv:1003.5338*, 2010.

H. Linhart, W. Zucchini. *Model Selection*. John Wiley and Sons, New York, 1986.

K. Nagata and S. Watanabe, Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *International Journal of Neural Networks*, 21(7):980-988, 2008.

T. Oaku. Algorithms for the b-function and D-modules associated with a polynomial. *Journal of Pure Applied Algebra,* 117:495-518, 1997.

M. Peruggia. On the variability of case-detection importance sampling weights in the Bayesian linear model. *Journal of American Statistical Association*, 92:199-207, 1997.

A.E. Raftery, M.A. Newton, J.M. Satagopan, P.N. Krivitsly. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, 8:1-45, Oxford University Press, Oxford, 2007.

D. Rusakov, 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.

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

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, A. Linde. Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, Series B*, 64(4):583-639, 2002.

M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society*. 39(B):44-47, 1977.

A. Takemura, T.Kuriki. On the equivalence of the tube and Euler characteristic methods for the distribution of the maximum of the gaussian fields over piecewise smooth domains. *Annals of Applied Probability,* 12(2):768-796, 2002.

A. W. van der Vaart, J. A. Wellner. *Weak Convergence and Empirical Processes.* Springer, 1996.

A. Vehtari, J. Lampinen. Bayesian Model Assessment and Comparison Using Cross-Validation Predictive Densities. *Neural Computation*, 14(10):2439-2468, 2002.

S. Watanabe. Generalized Bayesian framework for neural networks with singular Fisher information matrices. In the *Proceedings of the International Symposium on Nonlinear Theory and Its applications,* pages 207–210, 1995.

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. Almost all learning machines are singular. In the *Proceedings of the IEEE Int. Conf. FOCI*, pages 383–388, 2007.

S. Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, Cambridge, UK, 2009.

S. Watanabe. Equations of states in singular statistical estimation. *Neural Networks*. 23(1):20-34, 2010a.

S. Watanabe. Equations of states in statistical learning for an unrealizable and regular case. *IEICE Transactions*. E93A(3):617-626, 2010b.

S. Watanabe. A limit theorem in singular regression problem. *Advanced Studies of Pure Mathematics*. 57:473-492, 2010c.

S. Watanabe. Asymptotic learning curve and renormalizable condition in statistical learning theory. *Journal of Physics Conference Series*, 233, No.012014, 2010d.

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

K. Yamazaki, S. Watanabe. Algebraic geometry and stochastic complexity of hidden Markov models. *Neurocomputing*, 69:62-84, 2005.

P. Zwiernik. An asymptotic approximation of the marginal likelihood for general Markov models. *arXiv:1012.0753v1*, 2010.