WAIC AND WBIC ARE INFORMATION CRITERIA FOR SINGULAR STATISTICAL MODEL EVALUATION

Sumio Watanabe[1]

[1] Department of Computational Intelligence and Systems Science
Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama, 226-8502, Japan
E-mail:swatanab@dis.titech.ac.jp

ABSTRACT

Many statistical models and learning machines which have hierarchical structures, hidden variables, and grammatical rules are not regular but singular statistical models. In singular models, the log likelihood function can not be approximated by any quadratic form of a parameter, resulting that conventional information criteria such as AIC, BIC, TIC or DIC can not be used for model evaluation. Recently, new information criteria, WAIC and WBIC, were proposed based on singular learning theory. They can be applicable even if a true distribution is singular for or unrealizable by a statistical model. In this paper, we introduce definitions of WAIC and WBIC and discuss their fundamental properties.

1. INTRODUCTION

A statistical model is said to be *regular* if the map taking parameters to probability distributions is one-to-one and if Fisher information matrix is always positive definite. If otherwise it is called *singular*. Many statistical models and learning machines which have hierarchical structures, hidden variables, and grammatical rules are not regular but singular. For example, artificial neural networks [1, 2], normal mixtures [3], Boltzmann machines [4], hidden Markov models, Bayesian networks [5, 6], and reduced rank regressions [7] are singular statistical models.

In singular models, the log likelihood function can not be approximated by any quadratic form of the parameter, resulting that asymptotic normality of the maximum likelihood estimator does not hold. In Bayes estimation, the posterior distribution can not be approximated by any normal distribution even asymptotically. Hence the average of AIC [8] or DIC [9] is not equal to that of the generalization loss, or BIC [10] is not an asymptotic estimator of the log Bayes marginal likelihood.

Recently, we established singular learning theory [11, 12], by which asymptotic properties of both regular and singular models can be derived using algebraic geometry. New information criteria WAIC [13, 14] and WBIC [15] were proposed, which are applicable in singular statistical model evaluation.

In this paper, we introduce singular learning theory, and explain mathematical properties of WAIC and WBIC. We show the following facts hold for both regular or singular statistical models.
(1) The expectation value of WAIC is asymptotically equal to that of the Bayes generalization error.
(2) WBIC is asymptotically equivalent to the Bayes marginal likelihood as a random variable.
(3) If a true distribution is regular for and realizable by a statistical model, then WAIC is asymptotically equal to AIC as a random variable.
(4) If a true distribution is regular for a statistical model, then WBIC is asymptotically equal to BIC as a random variable.

2. BAYES STATISTICS

Let $\mathbb{R}^N$ be an N-dimensioanl Euclidean space. X_1, X_2, ..., X_n be a sequence of $\mathbb{R}^N$-valued random variables which are independently subject to the same probability distribution as $q(x)dx$. Here $q(x)dx$ is referred to as a true distribution. We use a notation

$$X^n = (X_1, X_2, ..., X_n).$$

The expectation values over $q(x)dx$ and X^n are respectively denoted by $\mathbb{E}_X[\ ]$ and $\mathbb{E}[\ ]$. The entropy and empirical entropy of a true distribution are respectively given by

$$\begin{aligned} S &= -\int q(x) \log q(x) dx, \\ S_n &= -\frac{1}{n} \sum_{i=1}^{n} \log q(X_i). \end{aligned}$$

A set of parameters is denoted by $W \subset \mathbb{R}^d$. A parametric model and a prior are respectively defined by $p(x|w)$ and $\varphi(w)$. Our purpose is to make an evaluation method of the pair $(p(x|w), \varphi(w))$ for a given random samples X^n without the regularity condition. The posterior distribution is defined by

$$p(w|X^n) = \frac{1}{Z_n}\ \varphi(w) \prod_{i=1}^{n} p(X_i|w),$$

where Z_n is a normalizing constant.

The expectation value over the posterior distribution is denoted by $\mathbb{E}_w[\ ]$. The *predictive distribution* is defined by

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

The Bayes *generalization and training losses* are respectively defined by

$$\begin{aligned} G_n &= -\mathbb{E}_X[\log p(X|X^n)], \\ T_n &= -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|X^n). \end{aligned}$$

Then

$$\mathbb{E}[G_n] = S + \mathbb{E}\Big[\int q(x) \log \frac{q(x)}{p(x|X^n)} dx\Big].$$

Hence the expectation value of G is equal to the sum of the entropy of the true distribution and the Kullback-Leibler distance of the true and predictive distributions.

The Bayes *free energy* or the Bayes *stochastic complexity* is defined by

$$F_n = -\log Z_n.$$

Then

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

where $q(x^n)$ and $p(x^n)$ are respectively probability density functions of x^n defined by

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

The expectation value of F_n is equal to the sum of the entropy of the true distribution and the Kullback-Leibler distance of the true and estimated distributions. By these reasons, the generalization error G_n and the Bayes free energy F_n are important random variables, which are employed in statistical model evaluation.

3. REGULAR CASES

Let us define two conditions.

Definition. (1) If there exists a parameter $w_{00} \in W$ such that $q(x) = p(x|w_{00})$, then a true distribution $q(x)$ is said to be *realizable* by a statistical model $p(x|w)$. If otherwise it is called unrealizable.
(2) The log loss function $L(w)$ is defined by

$$L(w) = -\mathbb{E}_X[\log p(X|w)].$$

The set of optimal parameters W_0 is defined by

$$W_0 = \{w \in W\ ;\ \min_{w'} L(w') = L(w)\}.$$

If W_0 consists of a single element w_0 and if the Hessian matrix

$$J_{ij} = \frac{\partial^2}{\partial w_i \partial w_j} L(w_0)$$

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

The log likelihood loss $L_n(w)$ is defined by

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

The well known information criteria AIC, DIC, BIC are respectively defined by

$$\begin{aligned} \text{AIC} &= L_n(\hat{w}) + \frac{d}{n}, \\ \text{DIC} &= T_n + \frac{d_{DIC}}{n}, \\ \text{BIC} &= nL_n(\hat{w}) + \frac{d}{2}\log n. \end{aligned}$$

where $\hat{w}$ is the maximum likelihood estimator, d is the dimension of the parameter space, and

$$d_{DIC} \equiv 2n(\mathbb{E}_w[L_n(w)] - L_n(\mathbb{E}_w[w]))$$

is the effective dimension estimated by DIC. If a true distribution $q(x)$ is realizable by and regular for $p(x|w)$, then G_n and T_n respectively satisfy

$$\begin{aligned} \mathbb{E}[G_n] &= S + \frac{d}{2n} + o(\frac{1}{n}), \\ \mathbb{E}[T_n] &= S - \frac{d}{2n} + o(\frac{1}{n}). \end{aligned}$$

In such a case, AIC and DIC respectively satisfy

$$\begin{aligned} \mathbb{E}[\text{AIC}] &= S + \frac{d}{2n} + o(\frac{1}{n}), \\ \mathbb{E}[\text{DIC}] &= S + \frac{d}{2n} + o(\frac{1}{n}). \end{aligned}$$

Also if a true distribution is realizable by and regular for a statistical model, F_n and BIC respectively satisfy

$$\begin{aligned} F_n &= nS_n + \frac{d}{2}\log n + O_p(1), \\ \text{BIC} &= nS_n + \frac{d}{2}\log n + O_p(1). \end{aligned}$$

However, if otherwise, such equations do not hold.

4. SINGULAR CASES

Many statistical models are not regular but singular. For example, a normal mixture of $x \in \mathbb{R}^1$ for a parameter (a, b)

$$p(x|a, b) = (1-a)\mathcal{N}(0, 1) + a\mathcal{N}(b, 1)$$

is not a regular model, where $\mathcal{N}(v, 1)$ is the normal distribution with average v and variance one. Figure 1 (1), (2),

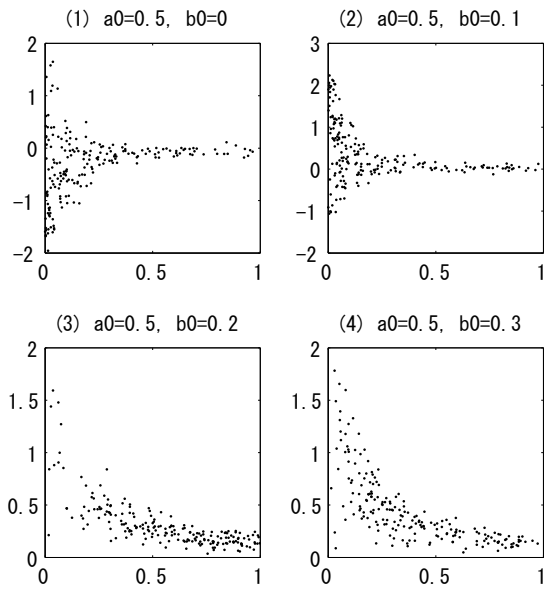

Figure 1. Posterior Distributions

(3), and (4) respectively show posterior distributions for the cases when the true distributions are given by $q(x) = p(x|a_0, b_0)$, where
(1) $(a_0, b_0) = (0.5, 0)$,
(2) $(a_0, b_0) = (0.5, 0.1)$,
(3) $(a_0, b_0) = (0.5, 0.2)$, and
(4) $(a_0, b_0) = (0.5, 0.3)$.
The number of empirical samples is $n = 500$ whereas the number of parameters is two. From the mathematical point of view, (1) is singular whereas (2),(3), and (4) are regular. However, all posterior distributions can not be approximated by any normal distribution. Sometimes one might think that, since the Lebesgue measure of the set

$$\{(a_0, b_0)\ ;\ p(x|a_0, b_0) \text{ is singular for } p(x|a, b)\}.$$

is equal to zero, statistical theory for such parameters is not necessary in practical model evaluations. However, such consideration is wrong. The set of singular parameters can not be negligible. In statistical model selection or hypothesis test, we have to determine whether a statistical model is redundant compared with a true distribution using finite samples. We have to evaluate several models under the condition that they are in almost singular states, hence statistical theory for singular cases is necessary in real-world problems.

5. WAIC AND WBIC

Recently, new statistical theory without the regularity condition was established based on algebraic geometry [12]. Let w_0 be a parameter which is contained in W_0. We define a function $K(w)$ by

$$K(w) = L(w) - L(w_0).$$

Then by definition $K(w) \geq 0$. We assume that $p(x|w_0)$ does not depend on the choice of $w_0 \in W_0$. If a true distribution is realizable by a statistical model, then $L(w_0) = S$. If otherwise $L(w_0) > S$. Assume that $K(w)$ is an analytic function of w but the Hessian matrix of $K(w)$ at w_0 is not positive definite in general. By using the fundamental theorem in algebraic geometry, there exists an analytic map $w = g(u)$ ($w \in W, u \in \mathbb{R}^d$) such that

$$\begin{aligned} K(g(u)) &= \prod_{j=1}^{d} (u_j)^{2k_j}, \\ \varphi(g(u))|g'(u)| &= b(u) \prod_{j=1}^{d} |u_j|^{h_j}, \end{aligned}$$

where $\{k_j\}$ and $\{h_j\}$ are nonnegative integer and $b(u) > 0$. This representation of a parameter is called *standard representation*, because any $K(w)$ in both regular and singular cases can be made to be this form. The *real log canonical threshold* (RLCT) λ is defined by

$$\lambda = \min_{j=1}^{d} \Big(\frac{h_j + 1}{2k_j} \Big).$$

Although such a function $w = g(u)$ is not unique, we can prove that λ does not depend on the choice of a function g, in other words, RLCT is a birational invariant. If a true distribution is regular for a statistical model, then $\lambda = d/2$, whereas, if otherwise, it is not equal to $d/2$. In order to find RLCTs, we need to find the resolution map $w = g(u)$. It is not so easy, however, RLCTs for several statistical models were found [3, 4, 7, 2, 5, 6]. Note that exchange probability in exchange Monte Carlo method is determined by RLCT [16].

By using RLCT, we can derive the asymptotic form of the posterior distribution [11, 12],

$$\begin{aligned} &\frac{n^\lambda}{(\log n)^{m-1}}\ p(w|X^n)dw \\ &\propto \int_0^\infty t^{\lambda-1} \exp(-t + \sqrt{t}\ \xi_n(u))\ b(u)du, \end{aligned}$$

where

$$\xi_n(u) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{K(g(u)) - L_n(g(u)) + L_n(w_0)}{\sqrt{K(g(u))}}.$$

is a well-defined random process which converges a gaussian process in law, and $b(u) > 0$ is a function of u. Then, even if a true distribution is singular for or unrealizable by a statistical model [13, 17],

$$\begin{aligned} \mathbb{E}[G_n] &= L(w_0) + \frac{\lambda}{n} + o(\frac{1}{n}), \\ \mathbb{E}[T_n] &= L(w_0) + \frac{\lambda - 2\nu}{n} + o(\frac{1}{n}), \\ \mathbb{E}[V_n] &= 2\nu + o(1), \end{aligned}$$

where m is a natural number and $\nu > 0$ is a birational invariant called a *singular fluctuation*. The random variable

V_n is defined by

$$V_n = \sum_{i=1}^{n} \{\mathbb{E}_w[\ell_i(w)^2] - \mathbb{E}_w[\ell_i(w)]^2\},$$

where $\ell_i(w) = \log p(X_i|w)$. If a true distribution is regular for a statistical model [17],

$$\nu = \frac{1}{2}\mathrm{tr}(IJ^{-1}),$$

where

$$I_{ij} = \mathbb{E}_X[\partial_i \log p(X|w_0) \partial_j \log p(X|w_0)].$$

If a true distribution is regular for and realizable by a statistical model, then $I = J$ and $\nu = d/2$. If a true distribution is not regular for a statistical model, ν is still an unknown birational invariant.

Based on this theorem, the widely applicable information criterion WAIC was introduced

$$\mathrm{WAIC} = T_n + V_n/n.$$

Then it follows that

$$\mathbb{E}[G_n] = \mathbb{E}[\mathrm{WAIC}] + O(\frac{1}{n^2}).$$

If a true distribution is regular for and realizable by a statistical model, then WAIC is asymptotically equivalent to AIC and DIC, whereas, if otherwise, they are not equivalent. It was also proved that, even if a true distribution is singular for or unrealizable by a statistical model, WAIC is equivalent to the Bayes leave-one-out cross validation as a random variable [14].

For the Bayes free energy, it was proved [11, 12] that

$$\begin{aligned} F_n &= nL_n(w_0) + \lambda \log n \\ &\quad -(m-1)\log\log n + O_p(1), \end{aligned} \tag{1}$$

where m is the order of the pole $(-\lambda)$ of the zeta function

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

which can be analytically continued to a meromorphioc function on the entire complex plain and its poles are all real, negative and rational numbers. The constant $(-\lambda)$ is equal to its maximum pole. Hence m is a natural number, which is equal to one if a true distribution is regular for a statistical model. In general, RLCT depends on the true distribution, hence we can not directly apply eq.(1) to the practical problems. To overcome such difficulty we defined the widely applicable Bayesian information criterion (WBIC) by

$$\mathrm{WBIC} = \frac{\int nL_n(w) \prod_{i=1}^{n} p(X_i|w)^{\beta} \varphi(w) dw}{\int \prod_{i=1}^{n} p(X_i|w)^{\beta} \varphi(w) dw},$$

where

$$\beta = \frac{1}{\log n}.$$

It was also proved [15] that, even if a true distribution is singular for or unrealizable by a statistical model,

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

If a parity of a statistical model [15] is odd, then

$$\mathrm{WBIC} = nL_n(w_0) + \lambda \log n + O_p(1).$$

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

$$\mathrm{WBIC} = \mathrm{BIC} + o_p(1).$$

From the numerical point of view, the conventional method to calculate Bayes free energy is given by

$$F_n = -\sum_{j=1}^{J} \log \frac{\int \prod_{i=1}^{n} p(X_i|w)^{\beta_j} \varphi(w) dw}{\int \prod_{i=1}^{n} p(X_i|w)^{\beta_{j-1}} \varphi(w) dw},$$

where $\{\beta_j\}$ is a sequence of many inverse temperatures,

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

In the conventional method, J times Markov chain Monte Carlo trials are necessary, whereas one trial is used in WBIC.

Experimental results of WAIC and WBIC are respectively reported in [13, 14, 18] and [15]. In [14], comparison of WAIC with DIC is also introduced. Evaluation of WAIC from the statistical point of view is given in [19]. If a true distribution is contained in the set of candidate models, then WBIC is better than WAIC for consistency of statistical model selection. If a true distribution is not contained in the set of candidate models, then WAIC is useful to estimate the generalization error. Theoretical comparison of information criteria in both regular and singular cases is the problem for the future study.

If the variational Bayes learning or the mean field approximation is applied, the variational Bayes free energy [20] and variational generalization error [21] are respectively different from the Bayes free energy and Bayes generalization error. However, by using the importance sampling method, WAIC can be applied using variational Bayes learning [22].

6. CONCLUSION

Two statistical information criteria WAIC and WBIC were introduced. They are respectively generalized concepts of AIC and BIC, which can be used even if a true distribution is singular for or unrealizable by a statistical model.

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

7. REFERENCES

[1] Sumio Watanabe, "Algebraic geometrical methods for hierarchical learning machines," *Neural Networks*, vol. 14, no. 8, pp. 1049–1060, 2001.

[2] Miki Aoyagi and Kenji Nagata, "Learning coefficient of generalization error in Bayesian estimation and Vandermonde matrix-type singularity," *Neural Computation*, vol. 24, no. 6, pp. 1569–1610, 2012.

[3] Keisuke Yamazaki and Sumio Watanabe, "Singularities in mixture models and upper bounds of stochastic complexity," *Neural Networks*, vol. 16, no. 7, pp. 1029–1038, 2003.

[4] Keisuke Yamazaki and Sumio Watanabe, "Singularities in complete bipartite graph-type boltzmann machines and upper bounds of stochastic complexities," *IEEE Transactions on Neural Networks*, vol. 16, no. 2, pp. 312–324, 2005.

[5] Dmitry Rusakov and Dan Geiger, "Asymptotic model selection for naive Bayesian network," *Journal of Machine Learning Research*, vol. 6, pp. 1–35, 2005.

[6] Piotr Zwiernik, "Asymptotic model selection and identifiability of directed tree models with hidden variables," *CRiSM report*, 2010.

[7] Miki Aoyagi and Sumio Watanabe, "Stochastic complexities of reduced rank regression in Bayesian estimation," *Neural Networks*, vol. 18, no. 7, pp. 924–933, 2005.

[8] Hirotugu Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[9] D.J.Spiegelhalter, N.G. Best, B.P.Carlin, and A.Linde, "Bayesian measures of model complexity and fit," *Journal of Royal Statistical Society, Series B*, vol. 64, no. 4, pp. 583–639, 2002.

[10] Gideon Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[11] Sumio Watanabe, "Algebraic analysis for nonidentifiable learning machines," *Neural Computation*, vol. 13, no. 4, pp. 899–933, 2001.

[12] Sumio Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge University Press, Cambridge, UK, 2009.

[13] Sumio Watanabe, "Equations of states in singular statistical estimation," *Neural Networks*, vol. 23, no. 1, pp. 20–34, 2010.

[14] Sumio Watanabe, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, vol. 11, pp. 3571–3591, 2010.

[15] Sumio Watanabe, "A widely applicable Bayesian information criterion," *Journal of Machine Learning Research*, vol. 14, pp. 867–897, 2013.

[16] Kenji Nagata and Sumio Watanabe, "Asymptotic behavior of exchange ratio in exchange monte carlo method," *Neural Networks*, vol. 21, pp. 980–988, 2008.

[17] Sumio Watanabe, "Equations of states in statistical learning for an unrealizable and regular case," *IEICE Transactions*, vol. E93-A, pp. 617–626, 2010.

[18] Kenji Nagata, Seiji Sugita, and Masato Okada, "Bayesian spectral deconvolution with the exchange monte carlo method," *Neural Networks*, vol. 28, pp. 82–89, 2012.

[19] Aki Vehtari and Janne Ojanen, "A survey of bayesian predictive methods for model assessment, selection and comparison," *Statistics Surveys*, vol. 6, pp. 142–228, 2012.

[20] Kazuho Watanabe and Sumio Watanabe, "Stochastic complexities of gaussian mixtures in variational bayesian approximation," *Journal of Machine Learning Research*, vol. 7, pp. 625–644, 2006.

[21] Shinichi Nakajima and Sumio Watanabe, "Generalization performance of subspace bayes approach in linear neural networks," *IEICE Transactions*, vol. Vol.E89-D, no. 3, pp. 1128–1138, 2006.

[22] Koshi Yamada and Sumio Watanabe, "Information criterion for variational bayes learning in regular and singular cases," in *Proc. of The 6th International Conference on Soft Computing and Intelligent Systems, 1551, F2-55-3, 2012, Kobe*, 2012.