Information criteria and cross validation for Bayesian inference in regular and singular cases

Sumio Watanabe¹

Received: 28 February 2021 / Revised: 30 March 2021 / Accepted: 31 March 2021 /
Published online: 11 May 2021
© The Author(s) 2021

Abstract

In data science, an unknown information source is estimated by a predictive distribution defined from a statistical model and a prior. In an older Bayesian framework, it was explained that the Bayesian predictive distribution should be the best on the assumption that a statistical model is convinced to be correct and a prior is given by a subjective belief in a small world. However, such a restricted treatment of Bayesian inference cannot be applied to highly complicated statistical models and learning machines in a large world. In 1980, a new scientific paradigm of Bayesian inference was proposed by Akaike, in which both a model and a prior are candidate systems and they had better be designed by mathematical procedures so that the predictive distribution is the better approximation of unknown information source. Nowadays, Akaike's proposal is widely accepted in statistics, data science, and machine learning. In this paper, in order to establish a mathematical foundation for developing a measure of a statistical model and a prior, we show the relation among the generalization loss, the information criteria, and the cross-validation loss, then compare them from three different points of view. First, their performances are compared in singular problems where the posterior distribution is far from any normal distribution. Second, they are studied in the case when a leverage sample point is contained in data. And last, their stochastic properties are clarified when they are used for the prior optimization problem. The mathematical and experimental comparison shows the equivalence and the difference among them, which we expect useful in practical applications.

Keywords Information criterion · Cross validation · Bayesian model choice

✉ Sumio Watanabe
swatanab@c.titech.ac.jp

¹ Department of Mathematical and Computing Science, Tokyo Institute of Technology, 2-12-1 Oookayama, Meguro-ku, Tokyo, W8-42, 152-8552, Japan

1 Introduction

In data science, we estimate an unknown information source by a predictive distribution defined from a statistical model and a prior. In an older framework of Bayesian statistics in the 20th century, it was assumed that a statistical model is convinced to be correct and a prior is given by a subjective belief, resulting that the predictive distribution is believed to be subjectively optimal solution without any check or test. However, such a restricted treatment of Bayesian inference cannot be the foundation of today's data science in the real world, because almost all statistical models and learning machines are different from the data generating distributions, and their priors are chosen not by any *a priori* knowledge or subjective belief, but by the mathematical laws for the minimum generalization loss, the maximum marginal likelihood, or other optimization methods.

In statistics, a new paradigm was already proposed by Akaike (1980a, 1980b) that both a statistical model and a prior are understood as only candidate systems and they had better be optimized so that the predictive distributions are controlled to be the better approximations of the unknown probability distribution. Also it was proposed by Good (1952); Gelman et al. (2013); Gelman and Shalizi (2013) that the marginal likelihood or the predictive distribution had better be checked and tested because Bayesian confirmation theory does not ensure the predictive distribution is suitable for the unknown distribution when a statistical model is too simple or too redundant. Nowadays, their proposal is widely accepted in statistics, data science, and machine learning, on which many statistical systems and learning machines are being applied to scientific and practical problems. For example, see Antonia Amaral Turkman et al. (2019); Congdon (2019); Hobbs et al. (2015); Korner-Nievergelt (2015); Lambert (2018); Martin (2016); McElreath (2020); Reich and Ghosh (2019); Wang et al. (2018).

In this paper, to develop the scientific evaluation measures of both a statistical model and a prior, we study the mathematical relation among the generalization loss, the information criteria AIC by Akaike (1974), DIC by Spiegelhalter et al. (2002), WAIC by Watanabe (2010), and the leave-one-out cross-validation loss by Gelfand et al. (1992); Vehtari and Lampinen (2002) in Bayesian inference from the following three points of view.

First, they are compared in the singular condition that the posterior distribution can not be approximated by any normal distribution. In many statistical models which have hierarchical structures or latent variables, the posterior distributions can not be approximated by any normal distribution, even asymptotically. In such cases, the generalization loss can not be estimated by information criteria AIC or DIC, however, it can be estimated by WAIC and the cross validation.

Second, they are studied in the case when a leverage sample point is contained in data. If a sample consists of independently and identically distributed random variables, then information criteria and the leave-one-out cross-validation loss are asymptotically equivalent, however, if otherwise, they are not. If a conditional probability is estimated using not an i.i.d. input sample, then there are cases when information criteria are better estimators of generalization loss than the cross-validation loss.

And last, their properties are clarified when they are used for the prior optimization problem. Information criteria and cross validation loss can be understood as the evaluation measures of the prior design. Minimization of WAIC and leave-one-out cross validation loss makes the average generalization loss, however, it does not minimize the generalization loss as a random variable.

This paper consists of five sections. In the second section, we define the generalization loss, the information criteria, and the cross-validation loss. In the third section, they are compared from the three different points of view. In the fourth section, their statistical difference is discussed, and in the last section, we conclude the paper.

2 Definitions of information criteria

In this section, we define the generalization loss, information criteria, and cross validation.

2.1 Definitions of statistical inference

Let $q(x)$ be a probability density function of $x \in \mathbb{R}^N$ and $X_1, X_2, \dots, X_n$ be an i.i.d. sample, in other words, a sample is a set of independent random variables whose probability density function is $q(x)$. In this paper, we mainly study i.i.d. cases. For the case when a sample is not i.i.d., see the Sect. 3.2. The cross validation procedure needs the i.i.d. condition, whereas information criteria can be used in several not i.i.d. cases as shown in Watanabe (2021).

A statistical inference is defined by a map

$$X^n = (X_1, X_2, \dots, X_n) \mapsto p^*(x),$$

where $p^*(x)$ is an estimated probability density function on $\mathbb{R}^N$, which is often called a predictive distribution. There are many procedures in statistical inference, for example, the maximum likelihood method, the maximum *a posteriori* (MAP) method, the Bayesian method, the sparse estimation method, and so on. In this paper, we study the Bayesian method.

Let $p(x|w)$ and $\varphi(w)$ be a statistical model and a prior, respectively, where $w \in \mathbb{R}^d$ is a parameter. In the maximum likelihood method, the predictive distribution is defined by $p^*(x) = p(x|w^*)$, where w^* is the maximum likelihood estimator. In the MAP method, the maximum likelihood estimator is replaced by the maximum *a posteriori* estimator. In the Bayesian method, the posterior distribution is defined by

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

where Z_n is a normalizing constant referred to as the marginal likelihood. For an arbitrary function of $f(w)$, let $\mathbb{E}_w[f(w)]$ and $\mathbb{V}_w[f(w)]$ denote the posterior average and

the posterior variance of $f(w)$, respectively. The Bayesian predictive distribution is defined by

$$p^*(x) = p(x|X^n) = \mathbb{E}_w[p(x|w)]. \quad (2)$$

Note that, in data science, we do not know whether a statistical model and a prior are appropriate or not.

For a given predictive distribution $p^*(x)$, its generalization loss G_n is defined by

$$G_n = - \int q(x) \log p^*(x) dx. \quad (3)$$

Since X^n is a random variable, G_n is also a random variable. Then minimizing G_n is equivalent to minimizing Kullback-Leibler divergence $K(q||p^*)$, where

$$K(q||p^*) = \int q(x) \log \frac{q(x)}{p^*(x)} dx.$$

Hence the generalization loss G_n gives the quantitative evaluation of a pair of a model and a prior according to the Kullback-Leibler divergence, however, since $q(x)$ is unknown in general, we need mathematical theory to estimate G_n . The training loss T_n is defined by

$$T_n = -\frac{1}{n} \sum_{i=1}^n \log p^*(X_i).$$

The generalization loss G_n can not be estimated by T_n , because the expectation values of G_n and T_n are not equal to each other. If a more complicated model is employed, then the training loss decreases, but the generalization loss may increase.

Remark It is shown by the Bayesian decision theory that, if w is generated from some true $\Phi(w)$ and if X^n is generated from some true $\prod_{i=1}^n P(X_i|w)$, then the expectation value of the generalization loss is made smallest if and only if $\phi(w) = \Phi(w)$ and $p(x|w) = P(x|w)$. However, in almost all situations of statistical inference, both the true prior and the true model are unknown, hence the decision theory can not be the base of Bayesian inference of unknown information source $q(x)$ in a real world. If $q(x)$ is located outside of a set of statistical model, then min max theory can not be used for designing the prior.

2.2 Asymptotic generalization loss

In this subsection, we explain the asymptotic behavior of the generalization loss of Bayesian inference. We need several mathematical definitions. Let $W_0 \subset \mathbb{R}^d$ be the set of all parameters that minimize

$$L(w) = - \int q(x) \log p(x|w) dx.$$

If there exists $w_0 \in W_0$ such that $q(x) = p(x|w_0)$, then an unknown distribution $q(x)$ is said to be realizable by a statistical model $p(x|w)$. If W_0 consists of a single element w_0 and the Hessian matrix

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

is positive definite, then $q(x)$ is said to be regular for $p(x|w)$. Many statistically asymptotic theories assume the regularity condition, however, if a statistical model has a hierarchical structure or latent variables, then regularity condition is not satisfied. In fact, neural networks, normal mixtures, hidden Markov models, matrix factorizations, latent Dirichlet allocations, and so on do not satisfy the regularity condition, resulting that their posterior distributions are far from any normal distribution. Even in such singular cases, the generalization performance of Bayesian inference is clarified based on algebraic geometry by Watanabe (2009).

Let $L_n(w)$ be the empirical log loss function,

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

We assume that, for an arbitrary parameter $w_0 \in W_0$, $p(x|w_0)$ represents the same probability density function. Note that, if the unknown distribution is realizable by a statistical model, then $L(w_0)$ and $L_n(w_0)$ are equal to the entropy S and the empirical entropy S_n of $q(x)$, respectively.

Even if an unknown distribution $q(x)$ is neither realizable by nor regular for a statistical model, it is proved by singular learning theory in Watanabe (2009, 2018) that there exists a random variable ξ_n such that

$$G_n = L_n(w_0) + \xi_n/n,$$

which satisfies the convergence in distribution $\xi_n \rightarrow \xi$. We need the zeta function of Bayesian statistics for $z \in \mathbb{C}$ by

$$\zeta(z) = \int (L(w) - L(w_0))^z \varphi(w) dw.$$

It is proved based on the resolution theorem in Hironaka (1964) that $\zeta(z)$ is a holomorphic function in $\Re(z) > 0$ which can be analytically continued to the unique meromorphic function onto the entire complex plane, whose poles are all real and negative values. Let $(-\lambda(w_0))$ be the largest pole of the zeta function. Then, by using algebraic geometry, it is proved that

$$\lim_{n \rightarrow \infty} \mathbb{E}[\xi_n] = \mathbb{E}[\xi] = \lambda(w_0).$$

Therefore,

$$\mathbb{E}[G_n] = L(w_0) + \lambda(w_0)/n + o(1/n). \quad (4)$$

This equation shows that the generalization loss is the sum of the bias $L(w_0)$ and the variance $\lambda(w_0)/n$. If an unknown distribution is regular for a statistical model, then $\lambda(w_0) = d/2$, where d is the dimension of the parameter space, if otherwise, $\lambda(w_0) \leq d/2$ with an assumption $\varphi(w_0) > 0$ for some $w_0 \in W_0$. The constant $\lambda(w_0)$ is referred to as the real log canonical threshold, which is a birational invariant, in other words, the generalization loss of Bayes inference is determined by algebro-geometric structure of a statistical model.

Concrete values of them were found in the matrix factorization by Aoyagi and Watanabe (2005), the Poisson mixture by Sato and Watanabe (2019), the latent Dirichlet allocation by Hayashi (2021), and many statistical models and learning machines such as Yamazaki and Watanabe (2003); Yamazaki (2016); Yamazaki and Kaji (2013); Zwiernik (2011); Watanabe (2009). They are useful in the design of Markov chain Monte Carlo by Nagata and Watanabe (2008) and the singular Bayesian information criterion by Drton and Plummer (2017).

Remark From the mathematical point of view, the real log canonical threshold $\lambda(w_0)$ can be understood as a volume dimension of the neighborhood of a set $\{w; L(w) - L(w_0) = 0\}$. In fact, it is proved in Watanabe (2009) that

$$\lambda(w_0) = \lim_{\varepsilon \rightarrow +0} \frac{1}{\log \varepsilon} \log \left(\int_{L(w) - L(w_0) < \varepsilon} \varphi(w) dw \right).$$

The prediction accuracy of Bayes inference is determined by the volume of the set of the almost optimal parameters.

2.3 Information criteria and cross validation

In this subsection, we introduce several information criteria to estimate the generalization loss. For the other information criteria to estimate the marginal likelihood, see Sect. 4. The asymptotic behavior of the generalization loss was clarified by Eq.(3), however, the real log canonical threshold depends on the unknown distribution, resulting that w_0 is also unknown, hence it cannot be directly applied to practical problems.

Remark In this paper, we study the scale of information criteria and cross validation as estimators of the generalization loss $\mathbb{E}[G_n]$. In many practical problems, it should be remarked that their scales are defined so as to estimate $2n \times \mathbb{E}[G_n]$, according to the Akaike's pioneer work.

The concept of the information criterion was firstly proposed by Akaike (1974). The definition of AIC is given by

$$\text{AIC} = T_n + d/n,$$

which is called Akaike information criterion (AIC). At first, AIC was made for estimating the generalization loss of the maximum likelihood method, but it can be employed in Bayes, because if the unknown distribution is realizable by and regular for a statistical model, then the order of the difference between the above AIC and AIC using the maximum likelihood estimator is smaller than $1/n$. For Bayesian estimation, the deviance information criterion (DIC) was proposed by Spiegelhalter et al. (2002),

$$\text{DIC} = \frac{1}{n} \sum_{i=1}^n \log p(X_i | \mathbb{E}_w[w]) - \frac{2}{n} \sum_{i=1}^n \mathbb{E}_w[\log p(X_i | w)],$$

by which the Bayesian generalization loss can be estimated, if an unknown distribution is realizable by and regular for a statistical model. The widely applicable information criterion (WAIC) was proposed by

$$\text{WAIC} = T_n + \frac{1}{n} \sum_{i=1}^n \mathbb{V}_w[\log p(X_i | w)].$$

The value $\mathbb{V}_w[\log p(X_i | w)]$ shows the fluctuation of $\log p(X_i | w)$ according to the posterior distribution. The mathematical relation between the generalization loss and this value is shown by using the functionally partial integration over the limit Gaussian process of the empirical process in Watanabe (2009). This relation has the same mathematical structure as the fluctuation-dissipation theorem which shows the correspondence from the system is equal to the fluctuation of the system.

The leave-one-out cross-validation loss in Gelfand et al. (1992); Vehtari and Lampinen (2002); Vehtari et al. (2017) is defined by

$$\text{LOO} = -\frac{1}{n} \sum_{i=1}^n \log p(X_i | X^n \setminus X_i),$$

where $X^n \setminus X_i$ is the sample which does not contain X_i . If the posterior distribution is precisely realized, then LOO can be numerically approximated by using the importance sampling cross-validation loss,

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

Even if an unknown distribution is unrealizable by or singular for a statistical model, the generalization loss can be estimated by WAIC, LOO, and ISCV. Note that AIC, DIC, and WAIC are calculated numerically using Markov chain Monte Carlo (MCMC) method, if the posterior variance $V(i) = \mathbb{V}_w[\log p(X_i | w)]$ is finite. The value $V(i)$ shows the leverage of a sample point X_i which will be discussed in Sect. 3.2. Note that LOO needs n different posterior distributions for $X^n \setminus X_i$, whereas ISCV can be calculated by one posterior distribution for X^n . However, in ISCV, the average $\mathbb{E}_w[1/p(X_i | w)]$ or variance $\mathbb{V}_w[1/p(X_i | w)]$ may be larger or infinite if a leverage sample point is contained, which was proved by Peruggia (1997); Epifani et al. (2008), resulting that the posterior calculation by MCMC method fails,

and a numerical improvement was developed by Gelman et al. (2014). The difference of information criteria and the leave-one-out cross validation in influential observation cases is explained in Sect. 3.2.

3 Comparison of information criteria and cross validation

In this section, we compare information criteria and cross validation loss in three different problems, that is to say, (1) in singular posterior cases, (2) in influential observation cases, and (3) in hyperparameter optimization cases.

3.1 Regular and singular cases

Firstly, we compare information criteria and leave-one-out cross-validation loss in regular and singular cases.

Regular Cases. If an unknown distribution $q(x)$ is realizable by and regular for a statistical model $p(x|w)$, then the posterior distribution can be approximated by a normal distribution when the sample size n tends to infinity. On such a regularity condition, AIC, DIC, WAIC, LOO, and ISCV are asymptotically equivalent to each other as random variables. That is to say,

$$\begin{aligned} \text{AIC} &= \text{DIC} + o_p(1/n) = \text{WAIC} + o_p(1/n) \\ &= \text{ISCV} + o_p(1/n) = \text{LOO} + o_p(1/n). \end{aligned}$$

They are all asymptotically unbiased estimators of the generalization loss,

$$\begin{aligned} \mathbb{E}[G_n] &= \mathbb{E}[\text{AIC}] + o(1/n) = \mathbb{E}[\text{DIC}] + o(1/n) = \mathbb{E}[\text{WAIC}] + o(1/n) \\ &= \mathbb{E}[\text{LOO}] + o(1/n) = \mathbb{E}[\text{ISCV}] + o(1/n). \end{aligned}$$

However, the generalization loss and AIC have asymptotically inverse correlation as is shown by Watanabe (2018),

$$(G_n - L(w_0)) + (\text{AIC} - L_n(w_0)) = \frac{d}{n} + o_p(1/n) \quad (5)$$

and DIC, WAIC, LOO, and ISCV satisfy the same equation as Eq.(5). This property strongly affects the model selection procedures by AIC, DIC, WAIC, LOO, and ISCV, which is the weak point of both the information criteria and the cross-validation loss. It should be emphasized that, since Eq.(5) is not trivial, many users of the information criteria and the cross validation loss may not be aware of such a risk.

Singular Cases. In general, the posterior distribution can not be approximated by any normal distribution even if the sample size tends to infinity in Watanabe (2018).

Even if the posterior distribution is far from any normal distribution, it was proved in Watanabe (2010) that

$$\text{WAIC} = \text{LOO} + O_p(1/n^2),$$

and that

$$\mathbb{E}[G_n] = \mathbb{E}[\text{WAIC}] + O(1/n^2) = \mathbb{E}[\text{LOO}] + O(1/n^2). \quad (6)$$

In such singular cases, neither AIC nor DIC satisfies these equations. In singular cases, the posterior expectation value of the parameter $\mathbb{E}_w[w]$ in DIC is not appropriate for statistical inference. The generalization loss and WAIC have asymptotically inverse correlation as is shown by Watanabe (2010),

$$(G_n - L(w_0)) + (\text{WAIC} - L_n(w_0)) = \frac{2\lambda(w_0)}{n} + o_p(1/n), \quad (7)$$

and both LOO and ISCV satisfy the same equation as Eq.(7). In singular cases, the real log canonical threshold $\lambda(w_0)$ is not larger than the regular cases. Hence, even if a statistical model $p(x|w)$ is redundant for an information source $q(x)$, the increase of the Bayesian generalization loss is smaller than the regular cases, which is one of the good properties of Bayesian inference. In the model selection problem, we should remark the fact that the increases of WAIC, LOO, and ISCV are also smaller than the regular cases.

Example 1 (Normal mixture in regular and singular cases). Let us compare AIC, DIC, WAIC, and ISCV in regular and singular cases. A normal distribution of $x \in \mathbb{R}^N$ which has a mean b_k and a variance and $1/s_k$ is denoted by

$$\mathcal{N}(x|b_k, s_k) = \left(\frac{s_k}{2\pi}\right)^{N/2} \exp\left(-\frac{s_k\|x - b_k\|^2}{2}\right).$$

We study a statistical model and a prior given in the following equations, where $w = (a, b, s)$ is a parameter, $a = \{a_k\}$, $b = \{b_k\}$, and $s = \{s_k\}$,

$$p(x|w) = \sum_{k=1}^K a_k \mathcal{N}(x|b_k, s_k),$$

$$\varphi(a, b, s) \propto \prod_{k=1}^K \{(a_k)^{\alpha-1} (s_k)^r \exp(-(s_k/2)(\rho + \mu\|b_k\|^2))\}.$$

Here K is the number of components in the statistical model. Note that the mixture ratio $a = \{a_k\}$ satisfies $a_k \geq 0$ and $\sum_k a_k = 1$, and (α, r, ρ, μ) is a hyperparameter. A normal mixture is a typical singular model, because if a model is redundant, then W_0 is a set with singularities as pointed out by Yamazaki and Watanabe (2003). Let us show simple experimental results for regular and singular cases as is shown

in Watanabe (2021). An experiment was set as $N = 3$, $n = 100$, $r = N/2$, $\alpha = 1$, and $\rho = \mu = 1$. A distribution $q(x)$ was set as $p(x|w_0)$,

$$p(x|w_0) = (1/K_0) \sum_{k=1}^{K_0} \mathcal{N}(x|b_{k0}, 1),$$

where $b_{10} = (3, 0, 0)$, $b_{20} = (0, 3, 0)$, $b_{30} = (3, 0, 0)$, $b_{40} = (-1, -1, -1)$. For a regular case, $K_0 = 4$ was used, whereas, for a singular case, $K_0 = 2$ was used. These two distributions were estimated by a normal mixture of $K = 4$ components. The posterior distributions were approximated by using the Gibbs sampler of the simultaneous distribution of the parameter and latent variable. In Table 1, means and standard deviations of $G_n - S$, $AIC - S_n$, $DIC - S_n$, $WAIC - S_n$ and $ISCV - S_n$ are shown which were calculated by using 100 independent trials, where $S = L(w_0)$ and $S_n = L_n(w_0)$ are entropy and empirical entropy of $p(x|w_0)$. In the regular case, all information criteria and ISCV could approximate the generalization loss, whereas, in the singular case, averages of AIC and DIC were larger and smaller than the generalization loss, respectively, whereas averages of WAIC and ISCV were almost equal to the generalization loss. The sample size $n = 100$ is not sufficiently large according to the asymptotic theory, however, information criteria and cross validation loss were almost equal to each other.

3.2 Influential observation

It is often assumed that a sample is independently and identically distributed (i.i.d.) in theoretical studies of information criteria. Such an i.i.d. condition is necessary for the use of the cross validation loss, however, information criteria can be employed in several not i.i.d. cases, for example, conditional independent cases.

In this subsection, we study the statistical inference of a conditional probability density for dependent inputs. Assume that $x^n = \{x_i; i = 1, 2, \dots, n\}$ is not a set of random variables but a constant sequence and that $\{Y_i; i = 1, 2, \dots, n\}$ is a set of an

Table 1 Means and standard deviations of information criteria in regular and singular cases

Case	M/S	$G_n - S$	$AIC - S_n$	$DIC - S_n$	$WAIC - S_n$	$ISCV - S_n$
Regular	MEAN	0.0930	0.1232	0.1011	0.0993	0.0998
Regular	STD	0.0326	0.0264	0.0394	0.0262	0.0263
Singular	MEAN	0.0612	0.1491	-0.4357	0.0584	0.0586
Singular	STD	0.0262	0.0302	0.3794	0.0255	0.0256

A normal mixture of four components was used as a statistical model, and two different distributions were estimated. In a singular case, AIC and DIC were larger and smaller than the generalization loss, respectively. WAIC and ISCV could estimate the generalization loss even if the posterior distribution was singular

independent random variables whose probability distribution is $\prod_{i=1}^n q(y_i|x_i)$. In this situation, the posterior average $\mathbb{E}_w[\cdot]$ and the variance $\mathbb{V}_w[\cdot]$ are defined by the posterior distribution

$$p(w|x^n, Y^n) = \frac{1}{Z_n} \varphi(w) \prod_{i=1}^n p(Y_i|x_i, w).$$

The generalization and training losses are defined by

$$G_n = -\frac{1}{n} \sum_{i=1}^n \int q(y_i|x_i) \log \mathbb{E}_w[p(y_i|x_i, w)] dy_i,$$

$$T_n = -\frac{1}{n} \sum_{i=1}^n \log \mathbb{E}_w[p(Y_i|x_i, w)].$$

The information criteria and leave-one-out cross-validation loss are defined by the same way as foregoing equations,

$$\text{AIC} = T_n + \frac{d}{n}, \quad (8)$$

$$\text{DIC} = \frac{1}{n} \sum_{i=1}^n \log p(Y_i|x_i, \mathbb{E}_w[w]) - \frac{2}{n} \sum_{i=1}^n \mathbb{E}_w[\log p(Y_i|x_i, w)], \quad (9)$$

$$\text{WAIC} = T_n + \frac{1}{n} \sum_{i=1}^n \mathbb{V}_w[\log p(Y_i|x_i, w)], \quad (10)$$

$$\text{ISCV} = \frac{1}{n} \sum_{i=1}^n \log \mathbb{E}_w[1/p(Y_i|x_i, w)]. \quad (11)$$

Note that, if x^n is independent and identically distributed (i.i.d.), then $\mathbb{E}[\text{ISCV}] = \mathbb{E}[G_{n-1}]$ by definition, whereas, if otherwise, $\mathbb{E}[\text{ISCV}] \neq \mathbb{E}[G_{n-1}]$. For example, if x^n is fixed or dependent, ISCV is not an unbiased estimator of the generalization loss in general. On the other hand, even for a fixed x^n , information criteria are asymptotically unbiased estimators of the generalization loss if the sample size is sufficiently large. If an unknown distribution is realizable by and regular for a statistical model, then AIC, DIC, and WAIC can be employed, which is shown by Watanabe (2018).

Example 2 (Influential observation in linear regression) We study a simple regression model of $x, y \in \mathbb{R}$ and a prior $\varphi(w)$,

$$p(y|x, a, s) = (s/(2\pi))^{1/2} \exp(-(s/2)(y - ax)^2), \quad (12)$$

$$\varphi(a, s|\mu) \propto s \exp(-(\mu/2)s(1 + a^2)), \quad (13)$$

where $a \in \mathbb{R}$, $s > 0$ are parameters and $\mu = 0.01$ is a hyperparameter. Let the sample size $n = 10$, and we set the x^n as $x_i = i/10$ ($i = 1, 2, \dots, 9$) and

$$x_{10} = 1 \text{ or } 4.$$

We compare information criteria and importance sampling leave-one-out cross-validation loss for two cases $x_{10} = 4$, which is a leverage sample point, and $x_{10} = 1$, which is not. A regression problem in which a leverage sample point is contained is often called an influential observation problem. The output Y^n are independently taken from $\prod_i p(y_i|x_i, a_0, s_0)$, where $a_0 = 0.1$, $1/s_0 = 0.01$. Then S and S_n are conditional entropy and empirical one, respectively. The average and standard deviations of the generalization loss, information criteria, and importance sampling leave-one-out cross-validation loss are shown in Table 2. In an influential observation case, the posterior average and variance of $\mathbb{E}_w[1/p(Y_i|X_i, w)]$ are larger or infinite by Peruggia (1997); Epifani et al. (2008), and both the average and variance of the $\mathbb{E}[\text{ISCV}]$ were larger than that of the generalization loss, where the averages of information criteria were asymptotically equal to that of generalization loss. In general, whether a sample point is leverage one or not depends on not only a sample point itself but also a statistical model. The author would like to recommend that both information criteria and cross validation had better be calculated and compared. If they are different, a leverage sample point is contained in the data, which can be found because it makes the functional variance

$$V(i) = \mathbb{V}_w[\log p(Y_i|x_i, w)].$$

larger than the other sample points, which is shown in Watanabe (2018).

Example 3 (High dimensional regression) We study a high dimensional regression model of $y \in \mathbb{R}$, $x, a \in \mathbb{R}^M$ and a prior $\varphi(a)$,

$$p(y|x, a) = (1/(2\pi))^{1/2} \exp(-(1/2)(y - a \cdot x)^2), \quad (14)$$

$$\varphi(a) \propto \exp(-(1/2)\|a\|^2). \quad (15)$$

Table 2 Means and standard deviations of information criteria for fixed inputs

Case		$G_n - S$	$\text{AIC} - S_n$	$\text{DIC} - S_n$	$\text{WAIC} - S_n$	$\text{ISCV} - S_n$
Ordinary	MEAN	0.1033	0.1095	0.0764	0.1015	0.1178
Ordinary	STD	0.1250	0.1068	0.1091	0.1199	0.1261
Leverage	MEAN	0.0959	0.1270	0.0770	0.0869	0.1274
Leverage	STD	0.1142	0.1121	0.1133	0.1217	0.1332

The case $x_{10} = 4$ is a leverage sample point, which is the influential observation, but $x_{10} = 1$ is not. Cross-validation loss can not be used if an input is a leverage sample point

In this example, we study the difference of two generalization losses,

$$G_n^{(1)} = - \int \int q(x)q(y|x) \log \mathbb{E}_a[p(y|x, a)] dx dy,$$

$$G_n^{(2)} = - \frac{1}{n} \sum_{i=1}^n \int q(y|x_i) \log \mathbb{E}_a[p(y|x_i, a)] dy.$$

The first one $G_n^{(1)}$ is the generalization loss for the case $\{(X_i, Y_i)\}$ is the set of independent random variables subject to $q(x)q(y|x)$, whereas the second one $G_n^{(2)}$ is that for the case $\{(Y_i|x_i)\}$ is the set of conditionally independent random variables subject to $q(y_i|x_i)$ for given $\{x_i\}$. If $M/n \rightarrow 0$, then $G_n^{(1)} - G_n^{(2)} \rightarrow 0$, however, if otherwise, then $G_n^{(1)} \neq G_n^{(2)}$. Two entropies are defined by

$$S^{(1)} = - \int \int q(x)q(y|x) \log q(y|x) dx dy,$$

$$S^{(2)} = - \frac{1}{n} \sum_{i=1}^n \int q(y_i|x_i) \log q(y_i|x_i) dy_i.$$

Information criteria AIC, DIC, WAIC, and ISCV are defined by the same way as Eqs.(8)–(11). LOO is defined by

$$\text{LOO} = - \frac{1}{n} \sum_{i=1}^n \log p(Y_i|X_i, X^n \setminus X_i, Y^n \setminus Y_i).$$

In the experiment, $\{x_i\}$ was independently taken from

$$q(x) = \frac{1}{(2\pi)^{M/2}} \exp\left(-\frac{\|x\|^2}{2}\right)$$

and fixed. $\{Y_i\}$ was independently taken from $p(y|x_i, a_0)$, where $a_0 = (1, 1, \dots, 1) \in \mathbb{R}^M$. The sample size $n = 50$ was fixed, and the dimension M was set as $M = 10, 20, 50, 100, 200$. In high dimensional cases, we can not assume that n is sufficiently large and there are many leverage sample points. Two different generalization losses were calculated. In Table 3, GE1 and GE2 denote $G_n^{(1)} - S^{(1)}$ and $G_n^{(2)} - S^{(2)}$, respectively. LOO, ISCV, WAIC, DIC, and AIC show the values of $\text{LOO}-S_n$, $\text{ISCV}-S_n$, $\text{WAIC}-S_n$, $\text{DIC}-S_n$, and $\text{AIC}-S_n$, respectively. Note that the empirical entropy does not depend on the assumption of inputs. In the experiment, LOO and WAIC estimated averages of GE1 and GE2, respectively. The standard deviations of LOO were larger than other criteria. In a high dimensional case, ISCV was not equal to LOO, because the posterior average in ISCV by MCMC was not calculated precisely.

Table 3 Means and standard deviations in a high dimensional regression problem. M is the dimension of x and n is a sample size

	M	n	GE1	GE2	LOO	ISCV	WAIC	DIC	AIC
Mean	10	50	0.112	0.090	0.108	0.108	0.099	0.096	0.127
Mean	20	50	0.226	0.156	0.246	0.242	0.193	0.200	0.298
Mean	50	50	1.014	0.312	1.044	0.682	0.340	0.432	0.847
Mean	100	50	1.965	0.344	1.974	0.753	0.344	0.488	1.844
Mean	200	50	2.500	0.350	2.531	0.751	0.336	0.486	3.836
Std	10	50	0.050	0.037	0.055	0.055	0.054	0.046	0.046
Std	20	50	0.069	0.046	0.076	0.076	0.070	0.061	0.063
Std	50	50	0.160	0.050	0.223	0.151	0.114	0.103	0.102
Std	100	50	0.067	0.054	0.169	0.116	0.101	0.101	0.100
Std	200	50	0.038	0.056	0.158	0.122	0.112	0.111	0.111

Two kinds of generalization losses GE1 and GE2 are defined, which are estimated by LOO, ISCV, WAIC, DIC, and AIC. the experimental result shows that LOO and WAIC estimated GE1 and GE2, respectively. Standard deviation of LOO was larger than other criteria

3.3 Prior optimization problem

In this subsection, we study a prior optimization problem in regular cases. That is to say, for a fixed statistical model, the generalization loss, information criteria, and cross-validation loss are understood as functionals of priors, and their minimization problem is studied.

In this section, a candidate prior $\varphi(w)$ is assumed to be a general nonnegative function of a parameter $w \in \mathbb{R}^d$ which may be improper, that is to say $\int \varphi(w)dw$ may be infinite. Even in such cases, the posterior and predictive distributions can be defined by the same equations as Eqs. (1), (2). Let $G(\varphi)$, WAIC(φ), and LOO(φ) be the generalization loss, WAIC, and the leave-one-out cross-validation loss which are functionals of a candidate prior $\varphi(w)$.

Remark In this subsection, we assume the regularity condition that an unknown distribution is regular for a statistical model, however, it may be unrealizable by a statistical model. The general theory for singular cases is not yet constructed. If a posterior distribution is singular, then it has the phase transition according to hyperparameter control as is explained in Watanabe (2018), in other words, the asymptotic support of the posterior distribution drastically changes at some critical value of the hyperparameter.

Let $\varphi_0(w)$ be an arbitrary fixed prior. For example, one can choose $\varphi_0(w) = 1$ for an arbitrary w . For a given candidate prior $\varphi(w)$, we define

$$\phi(w) = \varphi(w)/\varphi_0(w).$$

If $\varphi_0(w) \equiv 1$ is chosen, then $\phi(w) = \varphi(w)$. Assume that the posterior distribution can be approximated by a normal distribution. Then it was proved in Watanabe (2018) that there exists a function $\mathcal{M}(\phi, w)$ of $\phi(w)$ and w which satisfies

$$\mathbb{E}[G(\varphi)] = \mathbb{E}[G(\varphi_0)] + \frac{\mathcal{M}(\phi, w_0)}{n^2} + o\left(\frac{1}{n^2}\right), \quad (16)$$

$$\mathbb{E}[\text{LOO}(\varphi)] = \mathbb{E}[G(\varphi_0)] + \frac{d/2 + \mathcal{M}(\phi, w_0)}{n^2} + o\left(\frac{1}{n^2}\right), \quad (17)$$

$$\mathbb{E}[\text{WAIC}(\varphi)] = \mathbb{E}[G(\varphi_0)] + \frac{d/2 + \mathcal{M}(\phi, w_0)}{n^2} + o\left(\frac{1}{n^2}\right). \quad (18)$$

Also it was proved in Watanabe (2018) that there exists a function $M(\phi, w)$ of $\phi(w)$ and w which satisfies

$$\text{LOO}(\varphi) = \text{LOO}(\varphi_0) + \frac{M(\phi, \hat{w})}{n^2} + O_p\left(\frac{1}{n^3}\right), \quad (19)$$

$$\text{WAIC}(\varphi) = \text{WAIC}(\varphi_0) + \frac{M(\phi, \hat{w})}{n^2} + O_p\left(\frac{1}{n^3}\right). \quad (20)$$

The concrete forms of $\mathcal{M}(\phi, w)$ and $M(\phi, w)$ are defined by using higher order differentials of the log density function $\log p(x|w)$, which are given in Watanabe (2018). They satisfy

$$M(\phi, \hat{w}) = \mathcal{M}(\phi, w_0) + O_p\left(\frac{1}{n^{1/2}}\right), \quad (21)$$

$$M(\phi, \mathbb{E}_w[w]) = M(\phi, \hat{w}) + O_p\left(\frac{1}{n}\right), \quad (22)$$

$$\mathbb{E}[M(\phi, \hat{w})] = \mathcal{M}(\phi, w_0) + O\left(\frac{1}{n}\right). \quad (23)$$

It should be emphasized that the generalization loss as a random variable has different behavior from the information criteria and leave-one-out cross-validation loss.

$$G(\varphi) = G(\varphi_0) + O_p\left(\frac{1}{n^{3/2}}\right). \quad (24)$$

Hence minimization of WAIC or LOO makes the average generalization loss asymptotically, however, it does not minimize the generalization loss as a random variable. Note that minimization of AIC or DIC does not minimize the generalization loss as either the average or the random variable.

Table 4 Means and standard deviations of hyperparameters which made the generalization loss, AIC, DIC, WAIC, and the leave-one-out cross validation loss smallest

μ	argmin G	argmin AIC	argmin DIC	armin WAIC	argmin ISCV
Mean	0.0368	0.0086	0.0276	0.0489	0.0524
STD	0.0287	0.0059	0.0160	0.0276	0.0303

The optimal hyperparameter that minimizes the average generalization loss was estimated by WAIC or the leave-one-out cross-validation loss

Example 4 (Prior Optimization in linear regression) A simple regression model of $x, y \in \mathbb{R}$ same as Eqs.(12) and (13) in Example.2 is studied. In this example, x^n is independently taken from the standard normal distribution, and Y^n is also independently taken from $p(y_i|x_i, a_0, s_0)$, where $n = 10$, $a_0 = 0.1$, and $1/s_0 = 0.01$. The generalization loss, AIC, DIC, WAIC, and importance sampling leave-one-out cross-validation loss of 50 candidate hyperparameters $0 < \mu \leq 0.2$ were compared, and the hyperparameters which made them smallest were obtained for 100 independent samples. They are shown in Table 4. Note that

$$\operatorname{argmin} \mathbb{E}[G(\varphi)] = 0.048,$$

$$\mathbb{E}[\operatorname{argmin} G(\varphi)] = 0.036,$$

which were different. By using WAIC or ISCV, the former was estimated, whereas it was not estimated by AIC or DIC. Based on the higher order equivalence, WAIC and ISCV can be used for prior optimization for the average generalization loss. In the experiment, the variance by WAIC was smaller than that of ISCV.

4 Discussion: WBIC as an extension of BIC

In the previous sections, we studied the information criteria and leave-one-out cross-validation loss as estimators of the generalization loss. In this discussion, we compare the AIC type criteria with BIC ones.

Before discussion, we summarize the statistical property of the marginal likelihood. The marginal likelihood Z_n is defined by Eq.(1), which can be understood as an estimated probability density of X^n by a statistical model and a prior, because it satisfies $\int Z_n dx^n = 1$. The free energy F_n , which is equal to the minus log marginal likelihood, is also defined by

$$F_n = -\log Z_n.$$

Then its expectation value is equal to

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

where S is the entropy of $q(x)$, $q(x^n) = \prod_{i=1}^n q(x_i)$ and $p(x^n) = Z_n$. Hence the expectation value of the free energy is smallest if and only if the Kullback-Leibler divergence between $q(x^n)$ and $p(x^n)$ is smallest. Moreover, it is immediately derived that

$$\mathbb{E}[G_n] = \mathbb{E}[F_{n+1}] - \mathbb{E}[F_n].$$

Its asymptotic value is given by

$$F_n = nL_n(w_0) + \lambda(w_0) \log n + o_p(\log n),$$

where $\lambda(w_0)$ is the real log canonical threshold. If the posterior distribution can be approximated by a normal distribution, then $\lambda(w_0) = d/2$ and F_n is asymptotically approximated by BIC in Schwarz (1978). Even in singular cases, F_n can be estimated by sBIC in Drton and Plummer (2017) and WBIC in Watanabe (2013).

Let us compare the information criteria and the marginal likelihood from two different points of view. For short description, the information criteria AIC, DIC, WAIC, and LOO are referred to as AICs, whereas BIC, sBIC, WBIC, and the free energy are BICs.

First, we compare AICs and BICs from the viewpoint of model selection consistency. Let $\mathcal{P} = \{p_k(x|w_k); k = 1, 2, \dots, K\}$ be a set of candidate statistical models. Assume that there exists a statistical model in $\mathcal{P}$ by which $q(x)$ is realizable. The set of models $\mathcal{P}_0$ is defined by the set of all statistical models in $\mathcal{P}$ by which $q(x)$ is realizable. For a given $q(x)$, the statistical model $p \in \mathcal{P}_0$ that has the smallest real log canonical threshold is called the smallest true model. A model selection criterion is called to have the model selection consistency if and only if the probability that the smallest true model is chosen converges to one as the sample size tends to infinity. The asymptotic properties of AICs and BICs are given by

$$\text{AICs} = T_n + O_p(1/n),$$

$$\text{BICs}/n = T_n + \lambda(w_0) \log n/n + o_p(\log n/n).$$

If $q(x)$ is unrealizable by $p(x|w)$, then the convergence in probability holds,

$$T_n - S_n \rightarrow \min_w KL(q(x)||p(x|w)) > 0.$$

If $q(x)$ is realizable by $p(x|w)$, then in both regular and singular cases, $n(T_n - S_n)$ converges to a random variable in distribution in Watanabe (2009, 2018). In AICs, the penalty term has the same order as the standard deviation of $T_n - S_n$, whereas in BICs, it is larger than that of $T_n - S_n$. It follows that AICs do not have model selection consistency but BICs have, in both regular and singular cases.

Second, we study AICs and BICs from the viewpoint of asymptotic efficiency. Assume that $q(x)$ is not realizable by any model in $\mathcal{P}$. Then the optimal model for efficiency is defined by the model that minimizes $\mathbb{E}[G_n]$, which depends on the sample size n . Remark that consistency is defined for the case $n \rightarrow \infty$ whereas efficiency is for finite n . A model selection criterion is said to have more efficiency if and only if $\mathbb{E}[G_n]$ is made smaller. If BICs are employed, then the penalty term is too large for the smaller generalization loss, hence AICs have more efficiency than BICs, in both regular and singular cases. In practical applications of data science, appropriate information criteria had better be employed according to the purpose of the data analysis.

5 Conclusion

Information criteria and leave-one-out cross-validation loss in Bayesian modeling are studied, and three points are clarified. First, if the posterior distribution is far from any normal distribution, then AIC and DIC cannot be used for model evaluation, whereas WAIC and leave-one-out cross-validation loss can be employed. Second, if inputs in a sample are not independent, then leave-one-out cross-validation loss is not an asymptotic unbiased estimator of the generalization loss. And last, in the hyperparameter optimization problem, WAIC and leave-one-out cross-validation loss can be used for minimum average generalization loss.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 60.
- Akaike, H. (1980). Likelihood and Bayes procedure. *Bayesian Statistics*, 143–166.
- Akaike, H. (1980). On the transition of the paradigm of statistical inference. *The proceedings of the Institute of Statistical Mathematics*, 27, 5–12.
- Antonia Amaral Turkman, M., Carlos Daniel, P., & Peter, M. (2019). *Computational Bayesian statistics*, Cambridge University Press
- Aoyagi, M., & Watanabe, S. (2005). Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Networks*, 18, 924–933.
- Congdon, P. D. (2019). *Bayesian hierarchical models*. Boca Raton: CRC Press.
- Drton, M., & Plummer, M. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society, Series B*, 56, 1–38.
- Epifani, I., MacEchern, S. N., & Peruggia, M. (2008). Case-Deletion importance sampling estimators: Central limit theorems and related results. *Electric Journal of Statistics*, 2, 774–806.
- Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based method. Technical Report, Department of statistics, Stanford University, Vol.462, pp. 147–167
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis III*. Boca Raton: CRC Press.
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016.
- Gelman, A., & Shalizi, C. S. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66, 8–38.
- Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B*, 14, 107–114.
- Hayashi, N. (2021). The exact asymptotic form of Bayesian generalization error in latent Dirichlet allocation. Vol.137, pp.127–137

- Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero. I. *II Annals of Mathematics*, 79, 109–326.
- Hobbs, N. T., Mevin, B., & Hooten, M. B. (2015). *Bayesian models: A statistical primer for ecologists*. Princeton: Princeton University Press.
- Korner-Nievergelt, F., et al. (2015). *Bayesian data analysis in ecology using linear models with R, BUGS, and Stan*. Cambridge: Academic Press.
- Lambert, B. (2018). A student's guide to Bayesian statistics, SAGE.
- Martin, O. (2016). *Bayesian analysis with Python*. Birmingham: Packt Publishing Ltd.
- McElreath, S. (2020). *Statistical Rethinking: A Bayesian course with examples in R and STAN* (2nd ed.). Boca Raton: CRC Press.
- Nagata, K., & Watanabe, S. (2008). Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *Neural Networks*, 21(7), 980–988.
- Peruggia, M. (1997). On the variability of case-detection importance sampling weights in the Bayesian linear model. *Journal of American Statistical Association*, 92, 199–207.
- Reich, B. J., & Ghosh, S. K. (2019). *Bayesian statistical methods*. Boca Raton: CRC Press.
- Sato, K., & Watanabe, S. (2019). Bayesian generalization error of Poisson mixture and simplex Vandermonde matrix type singularity. [arXiv:1912.13289](https://arxiv.org/abs/1912.13289)
- Schwarz, G. (1978). Estimating the dimension of a model. Vol. 6, Np.2 *Annals of Statistics*, pp.461–464.
- Spiegelhalter, D. J., Best, N. G., & CarlinBP, Linde A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society Series B*, 64(4), 583–639.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5), 1413–1432.
- Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10), 2439–2468.
- Wang, X., Ryan, Y. Y., & Faraway, J. J. (2018). *Bayesian regression modeling with INLA*. Boca Raton: CRC Press.
- Watanabe, S. (2018). Higher order equivalence of Bayes cross validation and WAIC. *Springer Proceedings in Mathematics and Statistics, Information Geometry and Its Applications*, pp.47–73
- Watanabe, S. (2021). WAIC and WBIC for mixture models. *Behaviormetrika*. <https://doi.org/10.1007/s41237-021-00133-z>.
- Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*. Cambridge: Cambridge University Press.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.
- Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867–897.
- Watanabe, S. (2018). *Mathematical theory of Bayesian statistics*. Boca Raton: CRC Press.
- Yamazaki, K. (2016). Asymptotic accuracy of Bayes estimation for latent variables with redundancy. *Machine Learning*, 102, 1–28.
- Yamazaki, K., & Kaji, D. (2013). Comparing two Bayes methods based on the free energy functions in Bernoulli mixtures. *Neural Networks*, 44, 36–43.
- Yamazaki, K., & Watanabe, S. (2003). Singularities in mixture models and upper bounds of stochastic complexity. *International Journal of Neural Networks*, 16(7), 1029–1038.
- Zwiernik, P. (2011). An asymptotic behaviour of the marginal likelihood for general Markov models. *The Journal of Machine Learning Research*, 12, 3283–3310.