How New Information Criteria WAIC and WBIC Worked for MLP Model Selection

Seiya Satoh[1] and Ryohei Nakano[2]
[1]*National Institute of Advanced Industrial Science and Tech, 2-4-7 Aomi, Koto-ku, Tokyo, 135-0064, Japan*
[2]*Chubu University, 1200 Matsumoto-cho, Kasugai, 487-8501, Japan*
seiya.satoh@aist.go.jp, nakano@cs.chubu.ac.jp

Keywords: Information Criteria, Model Selection, Multilayer Perceptron, Singular Model.

Abstract: The present paper evaluates newly invented information criteria for singular models. Well-known criteria such as AIC and BIC are valid for regular statistical models, but their validness for singular models is not guaranteed. Statistical models such as multilayer perceptrons (MLPs), RBFs, HMMs are singular models. Recently WAIC and WBIC have been proposed as new information criteria for singular models. They are developed on a strict mathematical basis, and need empirical evaluation. This paper experimentally evaluates how WAIC and WBIC work for MLP model selection using conventional and new learning methods.

1 INTRODUCTION

A statistical model is called regular if the mapping from a parameter vector to a probability distribution is one-to-one and its Fisher information matrix is always positive definite; otherwise, it is called singular. Many useful statistical models such as multilayer perceptrons (MLPs), RBFs, HMMs, Gaussian mixtures, are all singular.

Given data, we sometimes have to select the best statistical model that has the optimum trade-off between goodness of fit and model complexity. This task is called model selection, and many information criteria have been proposed as measures for this task.

Most information criteria such as AIC (Akaike's information criterion) (Akaike, 1974), BIC (Bayesian information criterion) (Schwarz, 1978), and BPIC (Bayesian predictive information criterion) (Ando, 2007) are for regular models. These criteria assume the asymptotic normality of maximum likelihood estimator; however, in singular models this assumption does not hold any more. Recently Watanabe established singular learning theory (Watanabe, 2009), and proposed new criteria WAIC (widely applicable information criteria) (Watanabe, 2010) and WBIC (widely applicable Bayesian information criterion) (Watanabe, 2013), applicable to singular models. WAIC and WBIC have been developed on a strict mathematical basis, and how they work for singular models needs to be investigated hereafter.

Let MLP(J) be an MLP having J hidden units; note that an MLP model is determined by the number J. When evaluating MLP model selection experimentally, we have to run learning methods for different MLP models. There can be two ways to perform this learning: independent learning and successive learning. In the former, we run a learning method repeatedly and independently for each MLP(J), whereas in the latter MLP(J) learning inherits solutions from MLP($J-1$) learning. As J gets larger, a model gets more complex having more fitting capability. This means training error should monotonically decrease as J gets larger. However, independent learning will not guarantee this monotonicity. A new learning method called SSF (Singularity Stairs Following) (Satoh and Nakano, 2013a; Satoh and Nakano, 2013b) realizes successive learning by utilizing singular regions to stably find excellent solutions, and can guarantee the monotonicity.

This paper experimentally evaluates how new criteria WAIC and WBIC work for MLP model selection, compared with conventional criteria AIC and BIC, using a conventional learning method called BPQ (Back Propagation based on Quasi-Newton) (Saito and Nakano, 1997) and the new learning method SSF for search and sampling. BPQ is a kind of quasi-Newton method with BFGS (Broyden-Fletcher-Goldfarb-Shanno) update.

Satoh, S. and Nakano, R.
How New Information Criteria WAIC and WBIC Worked for MLP Model Selection.
DOI: 10.5220/0006120301050111
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 105-111
ISBN: 978-989-758-222-6

2 INFORMATION CRITERIA FOR MODEL SELECTION

Let a statistical model be $p(\mathbf{x}|\mathbf{w})$, where $\mathbf{x}$ is an input vector and $\mathbf{w}$ is a parameter vector. Let given data be $D = \{\mathbf{x}^\mu,\ \mu = 1, \cdots, N\}$, where N indicates data size, the number of data points.

AIC and BIC:
AIC (Akaike information criterion) (Akaike, 1974) and BIC (Bayesian information criterion) (Schwarz, 1978) are famous information criteria for regular models. Both deal with the trade-off between goodness of fit and model complexity.

The log-likelihood is defined as follow:

$$L_N(\mathbf{w}) = \sum_{\mu=1}^{N} \log p(\mathbf{x}^\mu|\mathbf{w}). \quad (1)$$

Let $\widehat{\mathbf{w}}$ be a maximum likelihood estimator. AIC is given below as an estimator of a compensated log-likelihood using the asymptotic normality of $\widehat{\mathbf{w}}$. Here M is the number of parameters.

$$\begin{aligned} \mathrm{AIC} &= -2L_N(\widehat{\mathbf{w}}) + 2M \\ &= -2\sum_{\mu=1}^{N} \log p(\mathbf{x}^\mu|\widehat{\mathbf{w}}) + 2M \quad (2) \end{aligned}$$

BIC is obtained as an estimator of free energy $F(D)$ shown below. Here $p(D)$ is called evidence and $p(\mathbf{w})$ is a prior distribution of $\mathbf{w}$.

$$F(D) = -\log p(D), \quad (3)$$

$$p(D) = \int p(\mathbf{w}) \prod_{\mu=1}^{N} p(\mathbf{x}^\mu|\mathbf{w})\, d\mathbf{w} \quad (4)$$

BIC is derived using the asymptotic normality and Laplace approximation.

$$\begin{aligned} \mathrm{BIC} &= -2L_N(\widehat{\mathbf{w}}) + M \log N \\ &= -2\sum_{\mu=1}^{N} \log p(\mathbf{x}^\mu|\widehat{\mathbf{w}}) + M \log N \quad (5) \end{aligned}$$

AIC and BIC can be calculated using only one point estimator $\widehat{\mathbf{w}}$.

WAIC and WBIC:
WAIC and WBIC are derived from Watanabe's singular learning theory (Watanabe, 2009) as new information criteria for singular models. Watanabe introduced the following four quantities: Bayes generalization loss BL_g, Bayes training loss BL_t, Gibbs generalization loss GL_g, and Gibbs training loss GL_t.

$$BL_g = -\int p^*(\mathbf{x})\ \log p(\mathbf{x}|D) d\mathbf{x} \quad (6)$$

$$BL_t = -\frac{1}{N}\sum_{\mu=1}^{N} \log p(\mathbf{x}^\mu|D) \quad (7)$$

$$GL_g = -\int \left(\int p^*(\mathbf{x}) \log p(\mathbf{x}|\mathbf{w}) d\mathbf{x} \right) p(\mathbf{w}|D) d\mathbf{w} \quad (8)$$

$$GL_t = -\int \left(\frac{1}{N}\sum_{\mu=1}^{N} \log p(\mathbf{x}^\mu|\mathbf{w}) \right) p(\mathbf{w}|D) d\mathbf{w} \quad (9)$$

Here $p^*(\mathbf{x})$ is the true distribution, $p(\mathbf{w}|D)$ is a posterior distribution, and $p(\mathbf{x}|D)$ is a predictive distribution.

$$p(\mathbf{w}|D) = \frac{1}{p(D)}\, p(\mathbf{w}) \prod_{\mu=1}^{N} p(\mathbf{x}^\mu|\mathbf{w}) \quad (10)$$

$$p(\mathbf{x}|D) = \int p(\mathbf{x}|\mathbf{w})\ p(\mathbf{w}|D)\ d\mathbf{w} \quad (11)$$

WAIC1 and WAIC2 are given as estimators of BL_g and GL_g respectively (Watanabe, 2010). WAIC1 reduces to AIC for regular models.

$$\mathrm{WAIC1} = BL_t + 2(GL_t - BL_t) \quad (12)$$

$$\mathrm{WAIC2} = GL_t + 2(GL_t - BL_t) \quad (13)$$

WBIC is given as an estimator of free energy $F(D)$ for singular models (Watanabe, 2013), where $p_\beta(\mathbf{w}|D)$ is a posterior distribution under the inverse temperature β. In WBIC context, β is set to be $1/\log(N)$. WBIC reduces to BIC for regular models.

$$\mathrm{WBIC} = -\int \left(\sum_{\mu=1}^{N} \log p(\mathbf{x}^\mu|\mathbf{w}) \right) p_\beta(\mathbf{w}|D) d\mathbf{w} \quad (14)$$

$$p_\beta(\mathbf{w}|D) = \frac{1}{p_\beta(D)}\, p(\mathbf{w}) \prod_{\mu=1}^{N} p(\mathbf{x}^\mu|\mathbf{w})^\beta \quad (15)$$

There are two ways to calculate WAIC and WBIC: analytic approach and empirical one. We employ the latter, which requires a set of weights $\{\mathbf{w}_t\}$ which approximates a posterior distribution (Watanabe, 2009).

3 SSF: NEW LEARNING METHOD

SSF (Singularity Stairs Following) is briefly explained; for details, refer to (Satoh and Nakano, 2013a; Satoh and Nakano, 2013b). SSF finds solutions of MLP(J) successively from J=1 until J_{max} making good use of singular regions of each MLP(J). Singular regions of MLP(J) are created by utilizing the optimum of MLP($J-1$). Gradient is zero all over the region.

How to create singular regions is explained below. Consider MLP(J) with just one output unit which outputs $f_J(\mathbf{x};\theta_J) = w_0 + \sum_{j=1}^{J} w_j z_j$, where θ_J

$= \{w_0, w_j, \mathbf{w}_j, j = 1, \cdots, J\}$, $z_j \equiv g(\mathbf{w}_j^{\mathrm{T}}\mathbf{x})$, and $g(h)$ is an activation function. Given data $\{(\mathbf{x}^{\mu}, y^{\mu}), \mu = 1, \cdots, N\}$, we try to find MLP($J$) which minimizes an error function. We also consider MLP($J-1$) with $\theta_{J-1} = \{u_0, u_j, \mathbf{u}_j, j = 2, \cdots, J\}$. The output is $f_{J-1}(\mathbf{x}; \theta_{J-1}) = u_0 + \sum_{j=2}^{J} u_j v_j$, where $v_j \equiv g(\mathbf{u}_j^{\mathrm{T}}\mathbf{x})$

Now consider the following reducibility mappings α, β, and γ. Then apply α, β, and γ to the optimum $\widehat{\theta}_{J-1}$ to get regions $\widehat{\vartheta}_J^{\alpha}$, $\widehat{\vartheta}_J^{\beta}$, and $\widehat{\vartheta}_J^{\gamma}$ respectively.

$$\widehat{\theta}_{J-1} \xrightarrow{\alpha} \widehat{\vartheta}_J^{\alpha}, \quad \widehat{\theta}_{J-1} \xrightarrow{\beta} \widehat{\vartheta}_J^{\beta}, \quad \widehat{\theta}_{J-1} \xrightarrow{\gamma} \widehat{\vartheta}_J^{\gamma}$$

$$\begin{aligned}
\widehat{\vartheta}_J^{\alpha} \equiv \{\theta_J | & w_0 = \widehat{u}_0,\ w_1 = 0, \\
& w_j = \widehat{u}_j, \mathbf{w}_j = \widehat{\mathbf{u}}_j,\ j = 2, \cdots, J\} \\
\widehat{\vartheta}_J^{\beta} \equiv \{\theta_J | & w_0 + w_1 g(w_{10}) = \widehat{u}_0, \\
& \mathbf{w}_1 = [w_{10}, 0, ..., 0]^{\mathrm{T}}, \\
& w_j = \widehat{u}_j, \mathbf{w}_j = \widehat{\mathbf{u}}_j, j = 2, ..., J\} \\
\widehat{\vartheta}_J^{\gamma} \equiv \{\theta_J | & w_0 = \widehat{u}_0, w_1 + w_m = \widehat{u}_m, \\
& \mathbf{w}_1 = \mathbf{w}_m = \widehat{\mathbf{u}}_m, \\
& w_j = \widehat{u}_j, \mathbf{w}_j = \widehat{\mathbf{u}}_j, j \in \{2, ..., J\} \backslash m\}
\end{aligned}$$

Now two singular regions can be formed. One is $\widehat{\vartheta}_J^{\alpha\beta}$, the intersection of $\widehat{\vartheta}_J^{\alpha}$ and $\widehat{\vartheta}_J^{\beta}$. The parameters are as follows, where only w_{10} is free: $w_0 = \widehat{u}_0$, $w_1 = 0$, $\mathbf{w}_1 = [w_{10}, 0, \cdots, 0]^{\mathrm{T}}, w_j = \widehat{u}_j$, $\mathbf{w}_j = \widehat{\mathbf{u}}_j$, $j = 2, \cdots, J$. The other is $\widehat{\vartheta}_J^{\gamma}$, which has the restriction: $w_1 + w_m = \widehat{u}_m$.

SSF starts search from MLP(J=1) and then gradually increases J one by one until J_{max}. When starting from the singular region, the method employs eigenvector descent (Satoh and Nakano, 2012), which finds descending directions, and from then on employs BPQ (Saito and Nakano, 1997), a quasi-Newton method. SSF finds excellent solution of MLP(J) one after another for J=1,$\cdots$,J_{max}. Thus, SSF guarantees that training error decreases monotonically as J gets larger, which will be quite preferable for model selection.

4 EXPERIMENTS

Experimental Conditions:

We used artificial data since they are easy to control and their true nature is obvious. The structure of an MLP is defined as follows: the numbers of input, hidden, and output units are K, J, and I respectively. Both input and hidden layers have a bias. Values of input data were randomly selected from the range [0, 1]. Artificial data 1 and data 2 were generated using MLP(K = 5, J = 20, I = 1) and MLP(K = 10, J = 20, I = 1) respectively. Weights between input and hidden layers were set to be integers randomly selected from the range $[-10, +10]$, whereas weights between hidden and output layers were integers randomly selected from $[-20, +20]$. A small Gaussian noise with mean zero and standard deviation 0.02 was added to each MLP output. Size of training data was set to be $N = 800$, whereas test data size was set to be 1,000.

WAIC and WBIC were compared with AIC and BIC. The empirical approach needs a sampling method; however, usual MCMC (Markov chain Monte Carlo) methods such as Metropolis algorithm will not work at all (Neal, 1996) since MLP search space is quite hard to search. Thus, we employ powerful learning methods BPQ and SSF as sampling methods. For AIC and BIC a learning method runs without any regularizer, whereas WAIC and WBIC need a weight decay regularizer whose regularization coefficient λ depends on temperature T. The temperature T was set as suggested in (Watanabe, 2010; Watanabe, 2013): $T = 1$ for WAIC and $T = \log(N)$ for WBIC. The regularization coefficient λ of WAIC is smaller than that of WBIC. WAIC and WBIC were calculated using a set of weights $\{\mathbf{w}_t\}$ approximating a posterior distribution. Test error was calculated using test data.

Our various previous experiments have shown that BPQ (Saito and Nakano, 1997) finds much better solutions than BP (Back Propagation) does, mainly because BPQ is a quasi-Newton, a 2nd-order method. Thus, we employ BPQ as a conventional learning method. We performed BPQ independently 100 times changing initial weights for each J. Moreover, we employ a newly invented learning method called SSF as well. For SSF, the maximum number of search routes was set to be 100 for each J; J was changed from 1 until 24. Each run of a learning method was terminated when the number of sweeps exceeded 10,000 or the step length got smaller than 10^{-16}.

Experimental Results:

Figures 1 to 6 show a set of results for artificial data 1. Figure 1 shows minimum training error obtained by each learning method for each J. Although SSF guarantees the monotonic decrease of minimum training error, BPQ does not in general. However, BPQ showed the monotonic decrease for this data. Figure 2 shows test error for $\widehat{\mathbf{w}}$ of the best model obtained by each learning method for each J. BPQ with λ = 0, BPQ with λ for WAIC, and BPQ with λ for WBIC got the minimum test error at J = 20, 24, and 24 respectively. SSF with λ = 0, SSF with λ for WAIC, and SSF with λ for WBIC found the minimum test error at J = 18, 19, and 20 respectively.

Figure 3 shows AIC values obtained by each

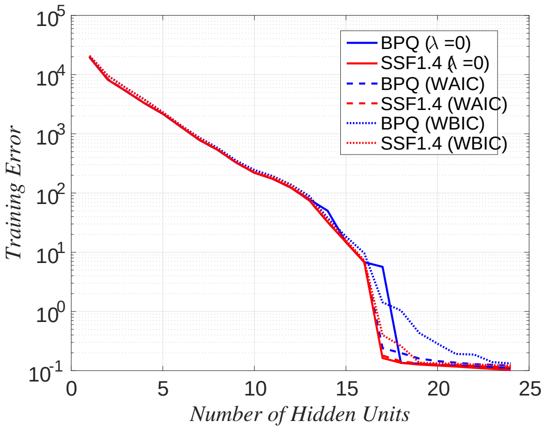

Figure 1: Training Error for Data 1.

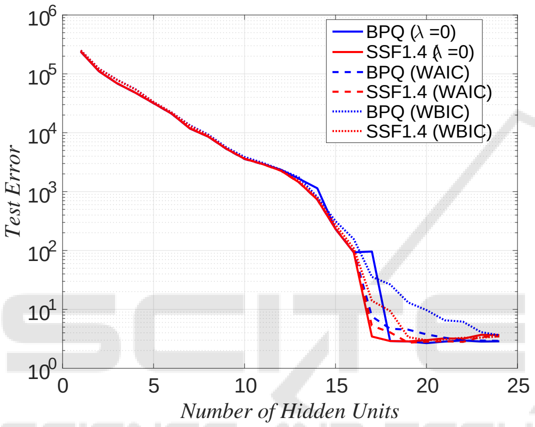

Figure 2: Test Error for Data 1.

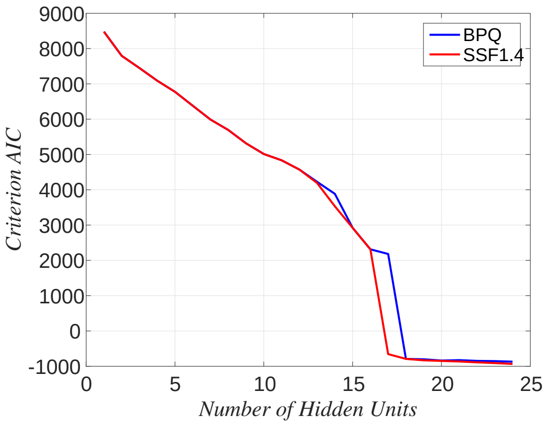

Figure 3: AIC for Data 1.

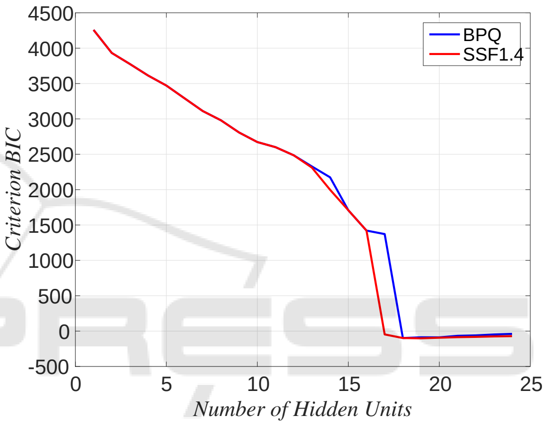

Figure 4: BIC for Data 1.

learning method for each J. AIC of both methods selected $J \geq 24$ as the best model, which is not suitable at all. Figure 4 shows BIC values obtained by each learning method for each J. BIC of BPQ selected J = 18 as the best model, whereas BIC of SSF selected J = 19. Thus, BIC selected a bit smaller models than the true one for this data.

Figure 5 shows WAIC1 and WAIC2 values obtained by each learning method for each J. WAIC1 and WAIC2 of BPQ selected $J \geq 24$ as the best model, which is not suitable. WAIC1 and WAIC2 of SSF selected J = 19 as the best model, which is very close to the true one (J = 20). WAIC1 and WAIC2 selected the same model for each method. Figure 6 shows WBIC values obtained by each learning method for each J. WBIC of BPQ selected $J \geq 24$, which is not suitable, whereas WBIC of SSF selected J = 20, which is right.

Figures 7 to 12 show the results for artificial data 2. Figure 7 shows minimum training error. SSF showed the monotonic decrease, whereas BPQ did not for WBIC. Figure 8 shows test error for $\widehat{\mathbf{w}}$ of the best model obtained by each learning method. BPQ with λ = 0, λ for WAIC, and λ for WBIC got the minimum test error at J = 24, 23, and 9 respectively. SSF with λ = 0, λ for WAIC, and λ for WBIC found the minimum test error at J = 23, 20, and 24 respectively.

Figure 9 shows AIC values obtained by each learning method for each J. AIC of BPQ and SSF selected J = 23 and $J \geq 24$ respectively as the best model, which is not acceptable. Figure 10 shows BIC values obtained by each learning method for each J. BIC of BPQ and SSF selected J = 22 and J = 21 respectively as the best model. BIC selected a bit larger models for this data.

Figure 11 shows WAIC1 and WAIC2 values obtained by each learning method for each J. WAIC1 and WAIC2 of BPQ selected J = 21 as the best model, which is very close to the true model. WAIC1 and WAIC2 of SSF selected J = 20 as the best model, which is exactly the true one. For this data WAIC1 and WAIC2 again selected the same model for each method. Figure 12 shows WBIC values obtained by each learning method for each J. WBIC of BPQ selected J = 10, which is quite unacceptable, whereas

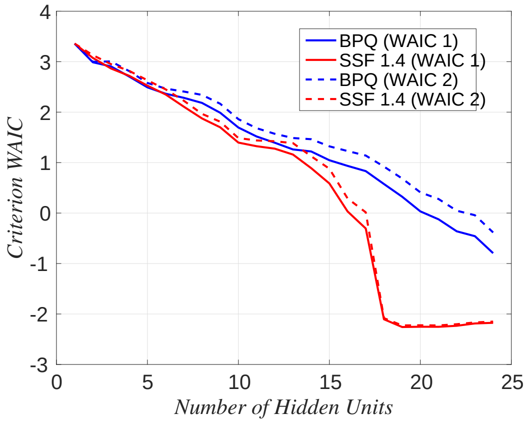

Figure 5: WAIC for Data 1.

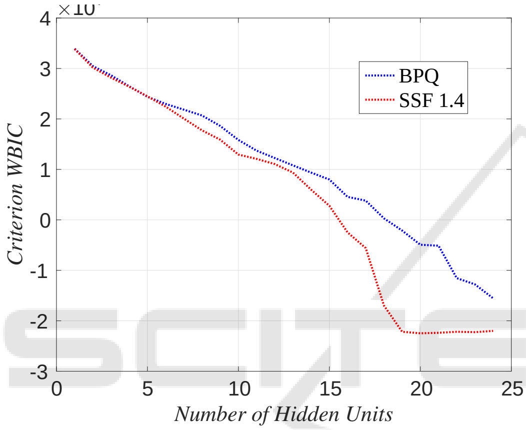

Figure 6: WBIC for Data 1.

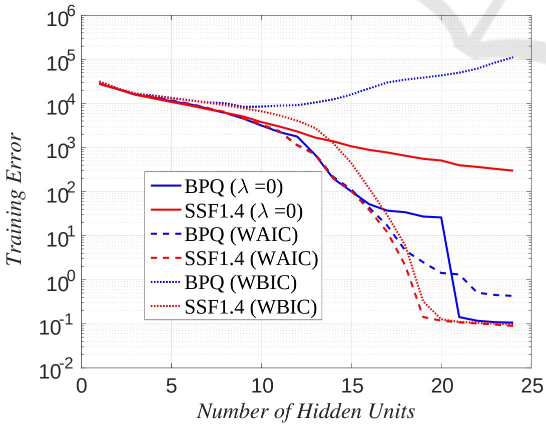

Figure 7: Training Error for Data 2.

WBIC of SSF selected J = 20, which is just the same as the true one.

Tables 1 and 2 summarize our results of model selection using BPQ and SSF for artificial data 1 and 2 respectively.

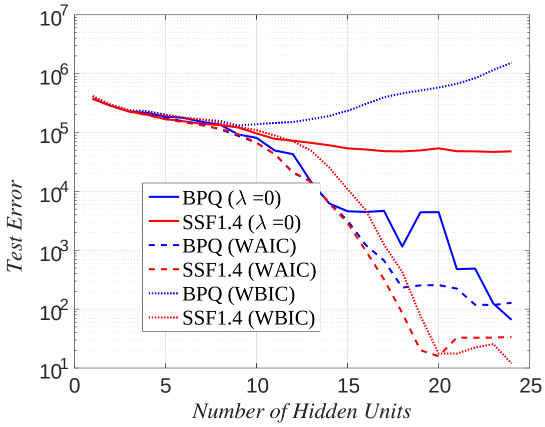

Figure 8: Test Error for Data 2.

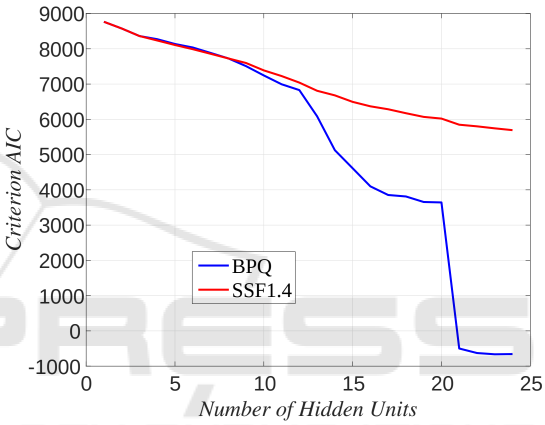

Figure 9: AIC for Data 2.

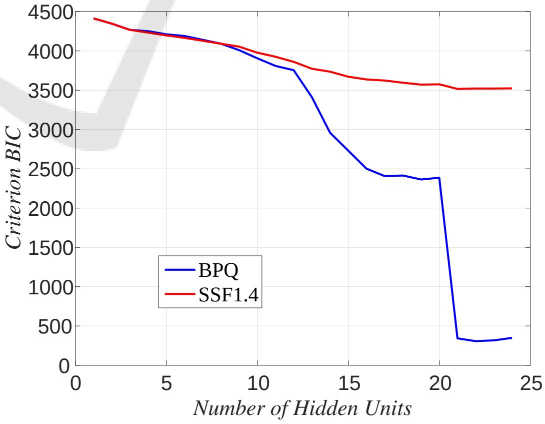

Figure 10: BIC for Data 2.

Considerations:

The results of our experiments may suggest the following. Note, however, that since our experiments are quite limited, more intensive investigation will be needed to make the tendencies more reliable.

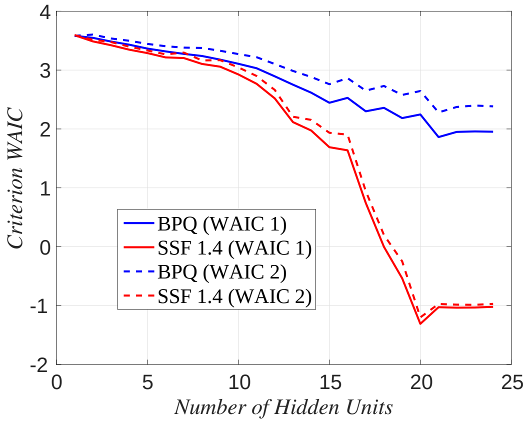

Figure 11: WAIC for Data 2.

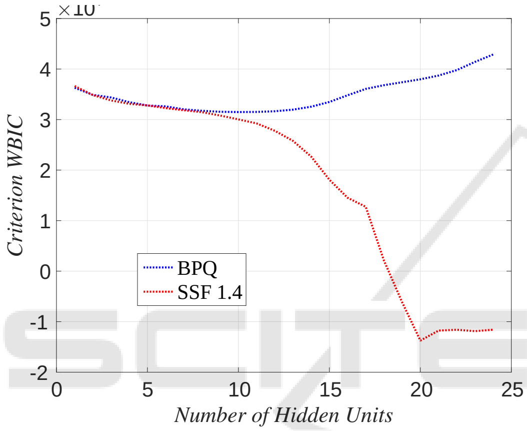

Figure 12: WBIC for Data 2.

Table 1: Comparison of Selected Models for Data 1.

criterion	learning method	
	BPQ	SSF
AIC	≥24	≥24
BIC	18	19
WAIC1	≥24	19
WAIC2	≥24	19
WBIC	≥24	20

(a) Independent learning of BPQ does not guarantee monotonic decrease of training error along with the increase of J, whereas successive learning of SSF does guarantee the monotonic decrease. For MLP model selection, independent learning sometimes did not work well, showing an up-and-down curve of training error and leading to wrong selection, whereas successive learning seems suited for MLP model selection due to the monotonic decrease of training error.

Table 2: Comparison of Selected Models for Data 2.

criterion	learning method	
	BPQ	SSF
AIC	23	≥24
BIC	22	21
WAIC1	21	20
WAIC2	21	20
WBIC	10	20

(b) In our experiments AIC had the tendency to select the largest ($J \geq 24$) among the candidates for any learning method. This is probably because the penalty for model complexity is too small. BIC worked relatively well, having the tendency to select a bit smaller or larger models than the true one.

(c) WAIC and WBIC of SSF worked very well, selecting the true model or models very close to the true one. However, WAIC and WBIC of BPQ sometimes didn't work well. Moreover, there was little difference between WAIC1 and WAIC2 for each learning method in our experiments.

5 CONCLUSION

WAIC and WBIC are new information criteria for singular models. This paper evaluates how they work for MLP model selection using artificial data. We compared them with AIC and BIC using sampling methods. For this sampling, we used independent learning of a conventional learning method BPQ and successive learning of a newly invented SSF. Our experiments showed that WAIC and WBIC of SSF worked very well, selecting the true model or very close models for each data, although WAIC and WBIC of BPQ sometimes did not work well. AIC did not work well selecting larger models, and BIC had the tendency to select a bit smaller or larger models. In the future we plan to do more intensive investigation on WAIC and WBIC.

ACKNOWLEDGEMENT

This work was supported by Grants-in-Aid for Scientific Research (C) 16K00342.

REFERENCES

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

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94:443–458.

Neal, R. (1996). *Bayesian learning for neural networks*. Springer.

Saito, K. and Nakano, R. (1997). Partial BFGS update and efficient step-length calculation for three-layer neural networks. *Neural Comput.*, 9(1):239–257.

Satoh, S. and Nakano, R. (2012). Eigen vector descent and line search for multilayer perceptron. In *IAENG Int. Conf. on Artificial Intelligence & Applications (ICAIA'12)*, volume 1, pages 1–6.

Satoh, S. and Nakano, R. (2013a). Fast and stable learning utilizing singular regions of multilayer perceptron. *Neural Processing Letters*, 38(2):99–115.

Satoh, S. and Nakano, R. (2013b). Multilayer perceptron learning utilizing singular regions and search pruning. In *Proc. Int. Conf. on Machine Learning and Data Analysis*, pages 790–795.

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

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

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

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14:867–897.