Statistica Sinica Preprint No: SS-2021-0145

Title	Consistency of BIC Model Averaging
Manuscript ID	SS-2021-0145
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0145
Complete List of Authors	Ze Chen, Jianqiang Zhang, Wangli Xu and Yuhong Yang
Corresponding Author	Yuhong Yang
E-mail	yangx374@umn.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Consistency of BIC Model Averaging

Ze Chen[1], Jianqiang Zhang[1], Wangli Xu[1] and Yuhong Yang[2]

[1]*Renmin University of China and* [2]*University of Minnesota*

Abstract: BIC weighting has been frequently applied to high-dimensional linear regression when model averaging is considered to address model selection uncertainty. It also plays a central role in model selection diagnostics. However, little research has been done on its consistency or weak consistency, which are crucial properties for a model averaging method to perform well for various purposes. In addition, previous limited work on model averaging consistency excludes the consideration of categorical covariates. In this note, with both continuous covariates and categorical predictors (with possibly diverging numbers of levels) allowed, we establish both consistency and weak consistency for BIC weighting.

Key words and phrases: BIC-p weighting, Categorical predictors, Consistency, Weak consistency.

1. Introduction

Model averaging is an alternative approach to mitigating model selection uncertainty by weighting estimators across some models. There are various model averaging approaches proposed, see Buckland et al. (1997), Yang

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

(2001), Hjort and Claeskens (2003), Leung and Barron (2006), Hansen (2007), Zhang et al. (2020) and the references therein.

To our knowledge, however, the previous results focus on estimation accuracy and little has been done formally on the consistency of model averaging weighting for general high-dimensional linear modeling. Note that model averaging based on consistent model selection criteria does not necessarily lead to consistent weighting. Lai et al. (2015), as an exception, has derived consistency of the generalized fiducial probabilities for candidate models in the absence of categorical predictors.

What is worth mentioning is that the consistency of weighting plays a central role in some important applications. For instance, it provides a theoretical guarantee for assessing variable selection performance in model selection diagnostics (see Nan and Yang (2014) and Yu et al. (2020)) and measuring variable importance (see Ye et al. (2018)). Thus, it is essential to establish the consistency of weighting for success of model selection diagnostics. With the above background, in this note, focusing on a high-dimensional BIC information criterion (BIC-p) with a sparsity oriented prior on the models, we derive the consistency of BIC-p weighting and provide theoretical support for previous work in the literature. Detailed proofs of the theorems are provided in the Supplementary Material.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2. Main Results

For a linear regression model with both categorical and continuous predictors, we assume, without loss of generality, that among the p predictors $\{X_1, \ldots, X_p\}$, the first q, $\{X_1, \ldots, X_q\}$, are categorical, while the others are continuous. The categorical levels of $\{X_1, \ldots, X_q\}$ are denoted by $\{J_1, \ldots, J_q\}$ respectively. For each categorical variable X_i, we define dummy variables $X_{i,j}$ pertaining to the jth categorical level for $j = 1, \ldots, J_i - 1$, and put $X_{\mathcal{I}_i} = (X_{i,1}, \ldots, X_{i,J_i-1})^{\mathrm{T}}$ with $\mathcal{I}_i \stackrel{\text{def}}{=} \{(i,1), \ldots, (i, J_i - 1)\}$ in the regression. In a similar fashion, put $X_{\mathcal{I}_i} = X_i$ with $\mathcal{I}_i \stackrel{\text{def}}{=} \{i\}$ for each continuous predictor X_i. Given observations $\{y_i, x_i\}_{i=1}^n$ with $x_i = (x_{i,\mathcal{I}_1}^{\mathrm{T}}, \ldots, x_{i,\mathcal{I}_p}^{\mathrm{T}})^{\mathrm{T}}$, where $x_{i,\mathcal{I}_j}$ is the ith sample of $X_{\mathcal{I}_j}$, The linear regression model is written in matrix form as

$$Y = \beta_0 + X\beta + \epsilon, \tag{2.1}$$

where $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$ is an n-dimensional response vector, $X = (x_1, \ldots, x_n)^{\mathrm{T}}$ is a covariate matrix, $\beta = (\beta_{\mathcal{I}_1}^{\mathrm{T}}, \ldots, \beta_{\mathcal{I}_p}^{\mathrm{T}})^{\mathrm{T}}$ is a parameter vector of size $p^* = \sum_{i=1}^q J_i + p - 2q$, $\beta_{\mathcal{I}_i} = (\beta_{i,1}, \ldots, \beta_{i,J_i-1})^{\mathrm{T}}$ for $i = 1, \ldots, q$ and $\beta_{\mathcal{I}_i} = \beta_i$ for $i = q+1, \ldots, p$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}} \sim N(0, \sigma^2 I_n)$, where I_n is the $n \times n$ identity matrix.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

For $N > 1$, let $\mathcal{M} \stackrel{\text{def}}{=} \{M_i, i = 1, \ldots, N\}$ be a candidate model set, where $M_i = \bigcup_{j \in \mathcal{A}_i} \mathcal{I}_j$, $\mathcal{A}_i \subset \{1, \ldots, p\}$. Let $\| \cdot \|_2$ be the l_2-norm and denote by $|\cdot|$ the number of elements of a set. For the linear regression model (2.1), the index set of true variables is defined as $M_0 \stackrel{\text{def}}{=} \bigcup_{i=1}^p \mathcal{I}_i^0$ with $\mathcal{I}_i^0 = \mathcal{I}_i$ for $\beta_{\mathcal{I}_i} \neq 0$ and $\mathcal{I}_i^0 = \emptyset$ otherwise. Note that under the sparsity assumption, that is, $|M_0| \ll n$, the number of continuous variables in the true model can increase to infinity with n, which is also applicable to the numbers of categorical variables and the number of levels of each categorical variable.

Throughout the note, we assume that $|M_0| \log p^* = o(n)$ and $p^* \to \infty$. Define $\mathcal{M} \stackrel{\text{def}}{=} \{M_i : |M_i| \leq (p^*)^\alpha \wedge (Cn/\log p^*) \text{ and } i \in \{1, \ldots, N\}\}$ for some constants $C > 0$ and $0 < \alpha < 1$ such that $|M_0| = o((p^*)^\alpha)$, where $a \wedge b \stackrel{\text{def}}{=} \min\{a, b\}$. It is worth noting that the condition $|M| \leq (p^*)^\alpha \wedge (Cn/\log p^*)$ for $M \in \mathcal{M}$ is much weaker than the condition $|M| \leq k|M_0|$ for some $k > 1$, which was assumed by Chen and Chen (2008), Luo and Chen (2013), Lai et al. (2015) and others. To ensure that any finitely many categorical variables can be included in our candidate model set, we assume $\max\{J_i : 1 \leq i \leq q \text{ and } \mathcal{I}_i \not\subset M_0\} = o((p^*)^\alpha \wedge (n/\log p^*))$.

For each element M in $\mathcal{M}$, we calculate the corresponding weight w_M below according to the BIC-p weighting method. Let $RSS_M \stackrel{\text{def}}{=} \|Y - \hat{\beta}_0 - X_M \hat{\beta}_M\|_2^2$ be the residual sum of squares of the model M, where X_M denotes

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

an $n \times |M|$ submatrix of the design matrix X, and $\hat{\beta}_0$ and $\hat{\beta}_M$ are corresponding least squares estimators. Let $I_M \stackrel{\text{def}}{=} n\log(RSS_M) + |M|\log n - n\log n$. Following Nan and Yang (2014), the BIC-p weight w_M is defined as

$$w_M \stackrel{\text{def}}{=} \exp\Big(-\frac{I_M}{2} - \psi C_M\Big) / \sum_{M' \in \mathcal{M}} \exp\Big(-\frac{I_{M'}}{2} - \psi C_{M'}\Big), \tag{2.2}$$

where $C_M = |M|\log(e \cdot p^*/|M|) + 2\log(|M|+2)$ and $\psi > 0$ is a constant.

For ease of notation, let $w_i \triangleq w_{M_i}$ for $M_i \in \mathcal{M}$. Given the candidate models $\mathcal{M}$ and a weighting $w = \{w_i, i = 1, ..., N\}$, we define *weight concentration index* (WCI) as $WCI(w) = \sum_{i=1}^N w_i|M_i \nabla M_0|$, where ∇ denotes the symmetric difference of two sets. Clearly, when WCI is close to zero, the weights of the candidate models are concentrated well around the true model. Based on $WCI(w)$, we give the definition of consistency and weak consistency as follows.

Definition 1. The weighting w is consistent if

$$WCI(w) \xrightarrow{P} 0, \quad \text{as } n \to \infty$$

Statistica Sinica

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

and the weighting is weakly consistent if

$$\frac{WCI(w)}{|M_0|} \xrightarrow{P} 0, \quad \text{as } n \to \infty.$$

For the theorems below, $\mathcal{M}$ is assumed to contain the true model and it can be up to the collection of all subset models. The following Conditions 1−3 are required for consistency.

Condition 1. All levels of each categorical variable are observed and the ratio of the most frequent levels to the least frequent levels is bounded by some constant.

Condition 2. $\min_{i\in\{1,\dots,p\}}\{\|\beta_{\mathcal{I}_i^0}\|_2^2 : \mathcal{I}_i^0 \neq \emptyset\} \geq c_1 \left(|M_0| \log{(p^*)}/n\right)^{\kappa}$, where $c_1 > 0$, $\kappa = 1 - \varepsilon$ and ε is an arbitrarily small positive constant.

Condition 3. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and the largest eigenvalues, respectively. Then for all M such that $|M| \leq k|M_0|$,

$$0 < c_{\min} \leq \lambda_{\min}\left(\frac{1}{n}X_M^{\mathrm{T}}X_M\right) \leq \lambda_{\max}\left(\frac{1}{n}X_M^{\mathrm{T}}X_M\right) \leq c_{\max} < \infty,$$

for some fixed $k > 1$.

Condition 1 excludes the case of an extremely unbalanced design matrix. Condition 2 requires that the minimum of the l_2-norms of the coeffi-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

cients of both the grouped dummy variables and the continuous variables in the true model are not too small. It should be noted that we impose a restriction on the ith group effect of $\beta_{\mathcal{I}_i^0}$ rather than the individual contribution of $\beta_{i,j}, j = 1, \ldots, J_i - 1$. Therefore, for the true grouped dummy variables, some or even most individual effects can be very small. Condition 3 is the sparse Riesz condition, which is a commonly used regularity condition for $p \gg n$ (see Zhang and Huang (2008); Lai et al. (2015)).

Theorem 1. *Under Conditions 1–3,* $\log(|M_0|)/\log p^* \leq \delta < \alpha$ *and* $\log n / \log p^* \leq \eta$ *for some positive constants* δ, α *and* η, *if* $\psi > (2C(k-1)((\alpha \wedge \eta) - 1))^{-1} k \log(1 - 4C(1+(\alpha \wedge \eta))) + (k/(k-1) - (\alpha \wedge \eta)/2)/(1 - (\alpha \wedge \eta))$ *for some* $C \in (0, 1/(4(1 + (\alpha \wedge \eta))))$, *we have*

$$\max_{M \in \mathcal{M}, M \neq M_0} \frac{w_M}{w_{M_0}} \xrightarrow{P} 0, \qquad \textit{as } n \to \infty. \tag{2.3}$$

Furthermore, the weighting is consistent, i.e.,

$$WCI(w) \xrightarrow{P} 0, \qquad \textit{as } n \to \infty. \tag{2.4}$$

Note that the lower bound on ψ in Theorem 1 is always a positive constant, since $\alpha \in (0, 1)$ and $C \in (0, 1/(4(1 + (\alpha \wedge \eta))))$. Theorem 1 states

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

that the weight w_{M_0} of the true model will tend to one as $n \to \infty$.

Typically there may be some relatively small coefficients in the true (or best) model that violate Condition 2. For $i = 1, \ldots, p$ and given arbitrary constant $c_2 > 0$, we define

$$\mathcal{I}_i^S \stackrel{\text{def}}{=} \begin{cases} \mathcal{I}_i^0 & \text{if } \mathcal{I}_i^0 \neq \emptyset \text{ and } \|\beta_{\mathcal{I}_i^0}\|_2^2/|\mathcal{I}_i^0| < c_2|M_0|\log(p^*)/n, \\ \emptyset & \text{otherwise.} \end{cases}$$

Let $M_0^S \stackrel{\text{def}}{=} \bigcup_{i=1}^p \mathcal{I}_i^S$ denote the set with indices of smaller coefficients. Note that we allow the l_2-norms of the coefficients of the variables in the set M_0^S to be arbitrarily small, but the number of these variables should be limited. Thus, a condition required for the weak consistency of BIC-p weighting is stated as follows:

Condition 4. $|M_0^S|/|M_0| \leq \xi_n$, where $\{\xi_n\}$ is a nonnegative sequence converging to zero as $n \to \infty$.

Theorem 2. *Under Conditions 1 and 3–4,* $\log(|M_0|)/\log p^* \leq \delta < \alpha$ *and* $\log n/\log p^* \leq \eta$ *for some positive constants* δ, α *and* η, *if* $\psi > (2C(k-1)((\alpha\wedge\eta)-1))^{-1}k\log(1-4C(1+(\alpha\wedge\eta)))+(k/(k-1)-(\alpha\wedge\eta)/2)/(1-(\alpha\wedge\eta))$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

for some $C \in (0, 1/(4(1+(\alpha \wedge \eta))))$, *then* w *is weakly consistent, i.e.,*

$$\frac{WCI(w)}{|M_0|} \xrightarrow{P} 0, \qquad \text{as } n \to \infty. \tag{2.5}$$

Not surprisingly, the weak consistency requires milder conditions that are much more realistic in applications.

Supplementary Material

The proofs of Theorems 1 and Theorem 2 are provided in the Supplementary Material document.

Acknowledgements

We sincerely thank Co-Editor Rong Chen for his very constructive suggestions on improving this paper. The work of Wangli Xu, Ze Chen and Jianqiang Zhang was supported by the Beijing Natural Science Foundation (Z20001) and National Natural Science Foundation of China (11971478).

REFERENCE

Buckland, S.T., Augustin, N.H., and Burnham, K.P. (1997). "Model selection - an integral part of inference." *Biometrics*, **53**, 603–618.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Chen, J. and Chen, Z. (2008). "Extended bayesian information criterion for model selection with large model space." *Biometrika*, **94**, 759–771.

Hansen, B. (2007). "Least squares model averaging." *Econometrica*, **75**, 1175–1189.

Hjort, N.L. and Claeskens, G. (2003). "Frequentist model average estimators." *Journal of the American Statistical Association*, **98**, 879–899.

Lai, R.C., Hanning, J., and Lee, T.C. (2015). "Generalized fiducial inference for ultrahigh-dimensional regression." *Journal of the American Statistical Association*, **110**, 760–772.

Leung, G. and Barron, A.R. (2006). "Information theory and mixing least-squares regressions." *IEEE Transactions on Information Theory*, **52**, 3396–3410.

Luo, S. and Chen, Z. (2013). "Extended BIC for linear regression models with diverging number of relevant features and high or ultrahigh feature spaces." *Journal of Statistical Planning and Inference*, **143**, 494–504.

Nan, Y. and Yang, Y. (2014). "Variable selection diagnostics measures for high-dimensional regression." *Journal of Computational and Graphical Statistics*, **23**, 636–656.

Yang, Y. (2001). "Adaptive regression by mixing." *Journal of the American Statistical Association*, **96**, 574–588.

Ye, C., Yang, Y., and Yang, Y. (2018). "Sparsity oriented importance learning for high-dimensional linear regression." *Journal of the American Statistical Association*, **113**, 1797–1812.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Yu, Y., Yi, Y., and Yang, Y. (2020). "Performance assessment of high-dimensional variable identification." *Statistica Sinica, to appear*.

Zhang, C. and Huang, J. (2008). "The sparsity and bias of the lasso selection in high-dimensional linear regression." *The Annals of Statistics*, **36**, 1567–1594.

Zhang, X., Yu, D., Zou, G., Liang, H., and Carroll, R. (2020). "Parsimonious model averaging with a diverging number of parameters." *Journal of the American Statistical Association*, **115**, 972–984.

Ze Chen, Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China.

E-mail: chze96@163.com

Jianqiang Zhang, Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China.

E-mail: zhangjqs@163.com

Wangli Xu, Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China.

E-mail: wlxu@ruc.edu.cn

Yuhong Yang, School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: yangx374@umn.edu
Statistica Sinica