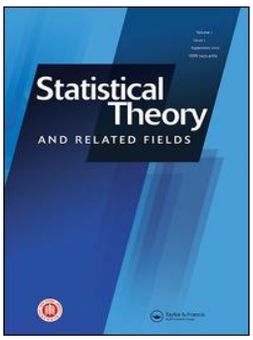

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

# A discussion of prior-based Bayesian information criterion (PBIC)

Jiming Jiang & Thuan Nguyen

**To cite this article:** Jiming Jiang & Thuan Nguyen (2019): A discussion of prior-based Bayesian information criterion (PBIC), Statistical Theory and Related Fields, DOI: 10.1080/24754269.2019.1583631

**To link to this article:** https://doi.org/10.1080/24754269.2019.1583631

Published online: 13 Mar 2019.

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2019.1583631

Taylor & Francis
Taylor & Francis Group



# A discussion of prior-based Bayesian information criterion (PBIC)

Jiming Jiang[a,b] and Thuan Nguyen[a,b]

[a]Department of Statistics, University of California, Davis, Davis, CA, USA; [b]Department of Public Health and Preventive Medicine, Oregon Health and Science University, Portland, OR, USA

**ARTICLE HISTORY** Received 12 February 2019; Accepted 13 February 2019

Professor Bayarri and coauthors' paper (hereafter, PBIC) offers a stimulating and welcomed addition to the already extensive and yet still rapid expanding literature on model selection and related topics. In a 2013 review on model selection in linear mixed models by Müller, Scealy, and Welsh (2013), the authors classified main approaches in mixed model selection into three categories, the information criteria, the shrinkage methods and the fence methods. The current paper is not specifically regarding mixed model selection problems; however, as one shall see, there is a connection in various ways.

The paper focuses on a special case of the information criteria, namely, the Bayesian information criterion (BIC) and its extensions. In this regard, two other references may be mentioned, in addition to those cited by the authors. One is the $\delta$-BIC method of Broman and Speed (2002), in which a tuning constant, $\delta$, is multiplied to the logarithm penalty to improve finite-sample performance; the other is an extended BIC proposed by Chen and Chen (2008), which allows the number of covariates to increase with the sample size.

The current paper has noted a number of problems with general use of BIC. Some similar notes were made regarding not just BIC but the information criteria in general by Jiang, Rao, Gu, and Nguyen (2008) in the context of mixed model selection. Among the problems mentioned in both papers is the so-called effective sample size (ESS). The issue was naturally raised in Jiang et al. (2008) because the latter authors were concerned with correlated observations. Intuitively, when the data are correlated, the ESS is smaller than the total number of observations due to the 'redundancy' in the data that each data point does not bring as much new information as an independent data point. Take a look at an extreme case where $n$ data points are so correlated that they are identical; obviously, in this case the ESS should be 1, rather than $n$. Another example, given in Jiang et al. (2008) (also see Jiang & Nguyen, 2015), is a linear mixed model, which may be viewed as a two-way extension of the group mean model discussed extensively in PBIC. In the linear mixed model, the observations, $y_{ij}$, satisfy $y_{ij} = x'_{ij}\beta + u_i + v_j + e_{ij}$, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, where $x_{ij}$ is a vector of known covariates, $\beta$ is a vector of unknown regression coefficients (the fixed effects), $u_i$, $v_j$ are random effects, and $e_{ij}$ is an additional error. It is assumed that $u_i$'s, $v_j$'s and $e_{ij}$'s are independent such that $u_i \sim N(0, \sigma_u^2)$, $v_j \sim N(0, \sigma_v^2)$, $e_{ij} \sim N(0, \sigma_e^2)$. It is well-known (e.g., Hartley & Rao, 1967; Harville, 1977; Miller, 1977) that, in this case, the ESS for estimating $\sigma_u^2$ and $\sigma_v^2$ is not the total sample size, $n = m_1 m_2$, but $m_1$ and $m_2$, respectively. Now suppose that one wishes to select the fixed covariates, which are components of $x_{ij}$, under the assumed model structure using BIC. It is not clear what should be in place of $n$ in the $\log(n)$ penalty (it does not make sense to let $n = m_1 m_2$). Note that the $m_1, m_2$ as the ESS for estimating $\sigma_u^2, \sigma_v^2$, respectively, can be interpreted intuitively – they are the numbers of appearance of the $u_i$s and $v_j$s, respectively, in the model. In general, the ESS for correlated data is somewhere between 1 and $n$, the sample size (this is also noted in PBIC), but exact quantification of ESS is difficult. In PBIC, the authors consider independent, rather than dependent data; still, they show that the ESS issue arises, when it comes to estimating different parameters. More importantly, the authors are able to quantify the ESS, in a certain way. I wonder if the quantification has some general, intuitive explanation, as in the special examples discussed above. By the way, the notation $n_i^e$, used to denote the ESS for estimating the $i$th group mean in the group mean model, might cause some confusion as being the $e$th power of $n_i$; perhaps, $n_{e,i}$ is a better notation?

Another problem, noted both in PBIC and in Jiang et al. (2008), is how to reasonably count the number of (free) parameters, or the degrees of freedom associated with the parameters. In this regard, Ye (1998) introduced the generalised degrees of freedom, which, in particular, is not necessarily an integer. This is similar to ESS, which can also be a non-integer.

As noted, there is an extensive literature in model selection, even if one focuses attention on BIC extensions. Furthermore, even though most of these extensions are proven to be consistent, finite-sample

**CONTACT** Jiming Jiang jimjiang@ucdavis.edu

performance can differ substantially. For example, the $\delta$-BIC (Broman & Speed, 2002) corresponds to a class of criteria with different values of $\delta$, and the finite-sample performance of the criterion depends heavily on the choice of $\delta$. A question about which BIC extension is the best is a difficult one to answer, if it can be answered at all. An alternative is to let the data speak (assuming that the data know the answer but not how to speak without help). A natural way of doing this is via the fence methods (e.g., Jiang & Nguyen, 2015). The idea consists of constructing a statistical fence to carefully isolate a subset of candidate models, known as the correct models. Once the fence is built, the optimal model can be selected from those within the fence based on a criterion of optimality that can incorporate practical considerations. A standard criterion of optimality is parsimony, that is, choosing the model within the fence that is the simplest, e.g., in terms of dimensionality. In a mathematical expression, the fence is constructed via the inequality

$$Q(M) - Q(M_*) \leq c, \tag{1}$$

where $M$ denotes a candidate model, $Q(\cdot)$ is a measure of lack-of-fit, $M_*$ is a candidate model that has the minimum $Q$ [so that $Q(M_*)$ is the basedline measure], and $c$ is a tuning constant. Note that, essentially, all of the model selection strategies, including the information criteria, amount to balancing model fitting and model complexity. The fitting part is controlled by the fence inequality, (1); the complexity part is controlled by the parsimony criterion, if the latter is used to select the optimal model within the fence. Thus, for example, the penalty for model complexity, which corresponds to the expressions other than $-2l(\hat{\theta})$ in PBIC or PBIC* [$-2l(\hat{\theta})$ is the $Q$ in this case], or $\delta \log(n)$ in $\delta$-BIC, are not needed. The final question comes down to the choice of $c$ in (1), which may be viewed as a cut-off. This is where the data have something to say. Typically, a lack-of-fit and complexity measures go opposite directions in a way much like the Type-I and Type-II errors in hypothesis testing. Thus, (1) might be viewed as the standard strategy of controlling the probability of Type-I error, but there is a major difference. Instead of using a given cut-off, such as $\alpha = 0.05$ in hypothesis testing, the $c$ in (1) is chosen in a data-driven manner by maximising the 'posterior' probability that a candidate model is selected, leading to the *adaptive fence* (e.g., Jiang & Nguyen, 2015, ch. 3).

Finally, consistency in model selection has been widely used as *the* standard asymptotic property in model selection, but, it is not very useful in comparing different model selection criteria that are all consistent. Although there has been further asymptotic properties, such as the oracle property (Fan & Li, 2001), much of the issue still exists, that is, virtually every new model selection procedure that is proposed is consistent, and has the oracle property. What is really needed, when it comes to asymptotic comparison of different model selection procedures, is a similar property to efficiency in parameter estimation. So far, such a property has not been established, and widely accepted.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## References

Broman, K. W., & Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experiemental crosses. *Journal of the Royal Statistical Society, Series B*, *64*, 641–656.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*, 759–771.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Hartley, H. O., & Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, *54*, 93–108.

Harville, D. A. (1977). maximum likelihood approaches to variance components estimation and related problems. *Journal of the American Statistical Association*, *72*, 320–340.

Jiang, J., & Nguyen, T. (2015). *The fence methods*. Singapore: World Scientific.

Jiang, J., Rao, J. S., Gu, Z., & Nguyen, T. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, *36*, 1669–1692.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of analysis of variance. *The Annals of Statistics*, *5*, 746–762.

Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, *28*, 135–167.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, *93*, 120–131.