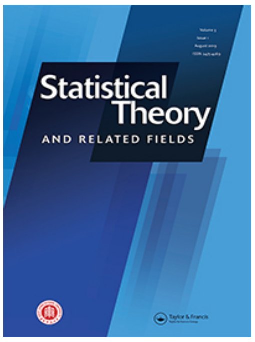

Statistical Theory and Related Fields

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

A discussion of ‘prior-based Bayesian information criterion’

Jiahua Chen & Zeny Feng

STATISTICAL THEORY AND RELATED FIELDS
2019, VOL. 3, NO. 1, 14–16
https://doi.org/10.1080/24754269.2019.1583628

Taylor & Francis
Taylor & Francis Group

A discussion of 'prior-based Bayesian information criterion'

Jiahua Chen[a] and Zeny Feng[b]

[a]Department of Statistics, University of British Columbia, Vancouver, BC, Canada; [b]Department of Mathematics and Statistics, University of Guelph, Guelph, ON, Canada

ARTICLE HISTORY Received 11 February 2019; Revised 13 February 2019; Accepted 13 February 2019

We would like to thank the authors (Bayarri et al., 2018) for their interesting and provoking paper, and we wish to discuss some issues related to sample size in general and the number of covariates in the context of linear regression model when using the Bayesian information criteria (BIC) for model selection. Schwarz (1978) was the first to develop tools for estimating the dimension of parameters among distributions in exponential family and consequently, introduce the BIC to serve as an approximation to the Bayesian posterior probability of a given model. The BIC has been used in a broad context and has been widely adapted for model selection despite that there are situations where the BIC might not be appropriate. Returning to its root as in this discussion paper is essential when the model and the data structure markedly deviate from the original context.

The original BIC criterion targets models arose from distributions belonging to an exponential family which permits a neat and simple analytical form after Laplace approximation. The neatness of this form is a blessing, but unfortunately, can be a curse as well. When the data is deprived of the independent and identically distributed (iid) structure, a blind application of BIC will not survive a close scrutiny. As discussed in Bayarri et al. (2018), the sample size in BIC becomes problematic. The prior-based BIC (PBIC) proposed in Bayarri et al. (2018) is essential to overcome these issues. This paper timely draws our attention to many unsettling issues related to the use of BIC in non-standard situations.

1. The classical and mutated BICs

Suppose we have a statistical model, referring to a specific family of distributions as usual, denoted as

$$M = \{f(x;\theta) : \theta \in \Theta \subset \mathcal{R}^p\}.$$

The density function $f(x;\theta)$ under consideration is usually chosen to have nice mathematical properties such as being regular. The dimension of the parameter θ remains the same within a model. This assumption is not obviously seen in the above presentation.

When the above model M is chosen for a population and a random sample is provided, statistical analysis is to infer the θ value of the population. In the context of Bayesian analysis, the θ value is regarded as uncertain and the level of uncertainty is specified by a prior density, say $\pi(\theta)$. The combination of the prior on the θ and the data sampled from the population lead to the posterior distribution which is the basis of the statistical decision.

When there are many competing models, say $M_1, M_2, \ldots, M_J$, a prior probability should be decided for each of these M_j's. Let α_j denote the prior probability for $M_j, j = 1, \ldots, J$. For notational simplicity, we use $f_j(x;\theta_j)$ for the density function in model M_j, p_j for the dimension of Θ_j which is the parameter space of M_j, and we also use some obvious conventions such as M, p and θ as some generic versions.

Let x_n be a sample of size n from a distribution which is a member of model M. By Bayes formula, the posterior probability that this M is M_j, is proportional to

$$\text{post}(M_j) = \alpha_j \int f_j(x_n;\theta_j)\pi_j(\theta_j)\,\mathrm{d}\theta_j. \tag{1}$$

Equation (1) precisely corresponds to $S(\mathbf{Y}, n, j)$ of Schwarz (1978, p. 462). The development in Schwarz (1978) is restricted to exponential family and under the assumption that the density is a function of x_n through $\mathbf{Y} = \mathbf{Y}(x_n)$. Other than the factor α_j, our (1) duplicates the function $m(x)$ of Bayarri et al. (2018). The subindex j in our expression highlights the fact that the form of θ_j depends on model M_j.

If an accurate computation of post(M_j) is cheap, then we would select the model M_j that maximises the posterior probability. In most cases, use some computationally feasible approximation is more realistic but may lead to complications.

Suppose x_n consists of n independent and identically distributed observations. Let $\hat{\theta}_j$ be the maximum

CONTACT Jiahua Chen ✉ jhchen@stat.ubc.ca Department of Statistics, University of British Columbia, Vancouver, BC, Canada

likelihood estimator (MLE) of θ_j under model M_j and $\ell_n(\cdot)$ be generic log likelihood function suitable for all $M_1, M_2, \ldots, M_J$. Then under reasonable conditions, Laplace approximation leads to the authentic BIC:

$$\begin{aligned}\mathrm{aBIC}(M_j) &= -2\log\left\{\alpha_j \int f_j(x_n;\theta_j)\pi_j(\theta_j)\,\mathrm{d}\theta_j\right\}\\ &= -2\ell_n(\hat{\theta}_j) + p_j \log n + c_j + o_p(1). \quad (2)\end{aligned}$$

Note that c_j crucially depends on M_j through at least α_j, π_j and f_j as we understand that f_j and M_j are two names for the same notion. When n is very large, the c_j and $o_p(1)$ can be wrapped up to $O_p(1)$ and the aBIC(M_j) arrives at the classical BIC:

$$\mathrm{BIC}(M_j) = -2\ell_n(\hat{\theta}_j) + p_j \log n. \quad (3)$$

However, unless n is very large, the size of c_j is not negligible. In other words, the aBIC in (2) and BIC in (3) can be very different and so that the BIC would no longer be a good approximation to the aBIC. Should c_j be taken into consideration? Bayarri et al. (2018) gives a positive answer to this question.

Taking this in mind, let us look into details for a data set of sample size n in the classical BIC and the prior-based BIC of the authors. Recalled that the BIC in Schwarz (1978) is derived when there are n iid observations from an exponential family and having these n iid observations of some dimensions but not necessarily the dimension of the parameters. The dimension of parameters p in the BIC refers to the dimension of $Y(x)$ where $Y(x)$ is a vector of statistics, not that of x in the exponential family model. Once we leave the comfort zone of iid and exponential family, direct application of BIC in (3) is questionable though it is now a common practice. Consider an extreme case where we have n iid observations from a distribution in an exponential family, but each observation is duplicated exactly twice in the data set. The apparent sample size is therefore $2n$ but the (correct) likelihood is not affected by the duplication. Applying classical BIC merely in formality leads to the wrongful BIC:

$$\mathrm{BIC}_w = -2\sum_{i=1}^{n} \log f(x_i;\hat{\theta}_j) + p_j \log(2n).$$

Yet its difference from the rightful BIC is merely a constant $p_j \log(2)$, which may well be regarded as part of c_j in aBIC. We suggest from this analysis that if omitting terms of $O_p(1)$ in BIC is acceptable, the precise definition of effective sample size is not so crucial.

Suppose that θ is a vector. When θ in a small neighbourhood of the truth, and therefore it is also in a small neighbourhood of $\hat{\theta}$ when the MLE is consistent, we have

$$\ell_n(\theta) \approx \ell_n(\hat{\theta}) - \tfrac{1}{2}(\theta - \hat{\theta})^\tau \mathbb{I}_n(\theta - \hat{\theta}),$$

where $\mathbb{I}_n = \mathbb{I}_n(\hat{\theta})$ is the Fisher information matrix at $\hat{\theta}$. The faithful Laplace approximation would lead to

$$\mathrm{aBIC}(M_j) = -2\ell_n(\hat{\theta}_j) + \log\det\{\mathbb{I}_n(\hat{\theta}_j)\} + c_j + o_p(1) \quad (4)$$

assuming $\log\det\{\mathbb{I}_n(\theta_j)\} \to \infty$ as $n \to \infty$. By this, we realise that c_j remains dependent on α_j, p_j and π_j, but the dependence of aBIC on f_j has been accommodated in the Fisher information. In common applications, we may choose to omit $O_p(1)$ constants related to α_j and π_j in BIC. After which, we seem to have defined the effective sample sizes via $\det\{\mathbb{I}_n(\hat{\theta}_j)\}$.

Consider the Example 1.3 of Bayarri et al. (2018) where $n/2$ observations are iid from $N(\theta, 1)$ and another $n/2$ observations are iid from $N(\theta, 1000)$. The Fisher information for θ is given by $\mathbb{I}(\theta) = (1 + 1/1000)(n/2) \approx n/2$. Our understanding is therefore in good agreement with Bayarri et al. (2018). In Example 1.4 of Bayarri et al. (2018), the Fisher information for θ is given by $\sum_{i=1}^{n} i^{-1} \approx \log n$. Hence, using the above suggested approximate aBIC, we would have get

$$\mathrm{aBIC}(\theta \neq 0) = -2\ell_n(\hat{\theta}_j) + \log\log(n).$$

Our suggestion on sample size is also found reasonable when applied to Example 1.2 of Bayarri et al. (2018) and therefore in good agreement with the Prior-based BIC of the authors.

2. The role of parameter dimension *p*

Our view on the role of the dimension of the parameter p in BIC differs from Bayarri et al. (2018). Our starting point is that part of c_j omitted as an $O_p(1)$ term in aBIC to arrive at BIC is related to p. When p is very large, the resulting approximation may lead to a nonsensical model selection criterion. We use the extended BIC (EBIC) as an example which is proposed by Chen and Chen (2008) that are suitable for small-n-large-p problems. Consider the classical linear regression model when n independent observations are obtained and the dimension of the explanatory variable is q. We use q instead of p to avoid potential confusion. In the era of big data, the number of explanatory variables q can be much larger than n. Let M_j be the collection of models where the expectation of the response is a linear combination of exactly j explanatory variables. One generally regards that each specific set of j explanatory variables makes up a model of its own right. Let M_{jk}, $k = 1, 2, \ldots, \binom{q}{j}$ be these models. From this angle, the cardinality of M_j is $\binom{q}{j}$. When aBIC is used, a prior probability α_{jk} is required for every M_{jk} and they lead to a total prior probability for M_j.

Suppose one puts $\alpha_{jk} \propto 1$ as it is clearly the default choice in BIC, we have

$$\alpha_j \propto \sum_{k=1}^{\binom{p}{j}} \alpha_{jk} = \binom{q}{j}$$

for model set M_j. When $q = 1{,}000{,}000$, we have $\alpha_2 = 50,000\alpha_1$. In small-$n$-large-$p$ problems, this implies that a linear model with two explanatory variables is 50,000 times more likely to be selected than a model with one explanatory variable if BIC is applied without any modifications. This is apparently controversial and leads to inconsistent model selection when n has a lower order than q.

To fix this problem, Chen and Chen (2008) suggest to put

$$\alpha_{jk} \propto \binom{q}{j}^{-\gamma}$$

for some $\gamma \in [0, 1]$. In applications, one would put an upper bound J not depending on n or q the number of explanatory variables allowed. Applying the Laplace approximation, the Extended BIC is obtained:

$$\text{EBIC}(M_j) = -2\ell_n(\hat{\theta}_j) + j\log(n) + 2j\gamma\log(q).$$

Although the choice of $\gamma = 1$ is most natural, their simulation results suggest that the choice of $\gamma = 0.5$ is a better trade-off between model complexity and parsimony. When q is very large, EBIC demands stronger evidence in order to accommodate a model with another explanatory variable.

The development of EBIC largely overlooks other Bayesian aspects of BIC. Refinements along the line of PBIC can be fruitful.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada [566065 and 587391].

Notes on contributors

Professor Jiahua Chen is Canada Research Chair, Tier I at the University of British Columbia.

Professor Zeny Feng is an associate professor at the University of Guelph.

References

Bayarri, M., Berger, J. O., Jang, W., Ray, S., Pericchi, L. R., & Visser, I. (2018). Prior-based Bayesian information criterion. *Statistical Theory and Related Fields*.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.