Statistical Theory and Related Fields

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

Discussion of 'Prior-based Bayesian Information Criterion (PBIC)'

Bertrand S. Clarke

To cite this article: Bertrand S. Clarke (2019) Discussion of 'Prior-based Bayesian Information Criterion (PBIC)', Statistical Theory and Related Fields, 3:1, 26-29, DOI: 10.1080/24754269.2019.1611143

To link to this article: https://doi.org/10.1080/24754269.2019.1611143

STATISTICAL THEORY AND RELATED FIELDS
2019, VOL. 3, NO. 1, 26–29
https://doi.org/10.1080/24754269.2019.1611143

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

Discussion of 'Prior-based Bayesian Information Criterion (PBIC)'

Bertrand S. Clarke

Statistics Department, University of Nebraska-Lincoln Lincoln, NE, USA

ARTICLE HISTORY Received 16 April 2019; Accepted 22 April 2019

1. Summary

The authors have the basis for a reformulation of the BIC as we think of it now. This problem is both hard and important. In particular, to address it, the authors have put six incisive ideas in sequence. The first is the separation of parameters that are common across models versus those that aren't. The second is the use of an orthogonal (why not orthonormal?) transformation of the Fisher information matrix to get diagonal entries d_i that summarise the parameter-by-parameter efficiency of estimation. The third is using Laplace's method only on the likelihood, i.e. Taylor expanding the log-likelihood and using the MLE rather than centring a Taylor expansion at the posterior mode. (From an estimation standpoint the difference between the MLE and posterior mode is $\mathcal{O}(1/n)$ and can be neglected.) The fourth is the particular prior selection that the third step enables. Since the prior is not approximated by, say $\pi(\hat{\theta})$, the prior can be chosen to have an impact and the only way the prior won't wash out is if it's tails are heavy enough. Fifth is defining an effective sample size n_i^e that differs from parameter to parameter. Finally, sixth, is imposing a relationship between the diagonal elements d_i and the 'unit information' b_i by way of the n_i^e. (All notation and terminology here is the same as in the paper, unless otherwise noted.)

Taken together, the result is a PBIC that arises as an approximation to $-2\log m(x^n)$, where $X^n = (X_1, \ldots, X_n) = (x_1, \ldots, x_n) = x^n$ is the data. This matches the $\mathcal{O}(1)$ asymptotics of the usual BIC.

The main improvement in perspective on the BIC that this paper provides is the observation that different efficiencies for estimating different parameters are important to include in model selection. Intuitively, if a parameter is easier to estimate in one model (larger Fisher information) than the corresponding parameter in another model then *ceteris paribus* the first model should be preferred. (The use of *ceteris paribus* covers a lot of ground, but helps make the point about efficiency.) Neglecting comparitive efficiences of parameters is an important gap to fill in the literature on the BC and model selection more generally.

The focus on the Fisher information $I(\cdot)$ – see Sec. 3.2 in particular – supports this view, however, one must wonder if there is more to be gained from either off-diagonal elements of I or from the orthogonal (orthonormal) matrix $\mathbf{O}$. The constraint $b_i = n_i^e d_i$ is also a little puzzling. It makes sense because b_i is interpretable as something like the Fisher information relative to parameter i. (In this sense it's not clear why it's called the 'unit information'.) The prior selection is very perceptive – and works – but there does not seem to be any unique, general conceptual property that it possesses. Even though it gives an effective result, the prior selection seems a little artificial. The authors may of course counter-argue that one of the reasons to use a prior is precisely that it represents information one has outside the data.

Setting aside such knit-picking, let us turn to the substance of the contribution.

2. Other forms for the BIC?

For comparison, let us try to modify the BIC in three other ways. The first is a refinement of the BIC to identify the constant c in Result 1.1. The second is to look more closely at the contrast between the PBIC the authors propose and a more conventional approach. The third is a discussion of an alternative that starts with an effective sample size rather than bringing it in via the prior.

First observe that the conventional expression for the BIC is actually only accurate to $\mathcal{O}_P(1)$ not $o_P(1)$. However, the constant term can be identified. Let x^n be IID P_θ. Staring at Result 1.1 and using standard Lapace's method analysis of $m(x^n)$ gives that

$$\left| \log \frac{p(X^n \mid \hat{\theta})\pi(\hat{\theta})}{m(X^n)} - \frac{p}{2}\log\frac{n}{2\pi} - \frac{1}{2}\log\det \hat{I}(\hat{\theta}) \right| \to 0, \tag{1}$$

CONTACT Bertrand S. Clarke bclarke3@unl.edu

in P_θ-probability; see Clarke and Barron (1988). So, a more refined version of the BIC expression, which approximates the posterior mode, is

$$\begin{aligned}\mathrm{BIC}_{\mathrm{better}} &= -2\ell(\hat\theta) + p\log n = -2\log m(x^n) + p\log 2\pi \\ &\quad - \log\det \hat I(\hat\theta) + 2\log\pi(\hat\theta) + o_P(1). \qquad (2)\end{aligned}$$

Using (2) may largely address Problem 1 as identified by the authors. Minimising $\mathrm{BIC}_{\mathrm{better}}$ over candidate models is loosely like maximising $m(x^n)$ subject to a penalty term in p and I, i.e. loosely like finding the model that achieves the maximal penalised maximum likelihood if the mixture density were taken as the likelihood. Expression (2) can be re-arranged to give an expression for $m(x^n)$. Indeed, one can plausibly argue that maximising $m(x^n)$ over models (and priors) under some restrictions should be a useful statistic for model selection.

This is intuitively reasonable ... until you want to take the intuition of the authors into account, *viz.* that different θ_j's in $\theta = (\theta_1, \ldots, \theta_p)$ require different sample sizes to estimate equally well or correspond to different effective sample sizes. One expects this effect to be greater as more and more models are under consideration. It is therefore natural to focus on the parameters that distinguish the models from each other rather than the common parameters. So, for ease of exposition we assume that $\theta = \theta_{(1)}$ i.e. that $\theta_{(2)}$ does not appear. (In simple examples like linear regression $\theta_{(2)}$ often corresponds to the intercept and can be removed by centring the data.)

So, second, let us look at the Laplace's method applied to $m(x^n)$. Being informal about a second order Taylor expansion and using standard notation gives

$$\begin{aligned}m(x^n) &= \int p(x^n \mid \theta)\pi(\theta)\mathrm{d}\theta = p(x^n \mid \hat\theta) \\ &\quad \times \int \mathrm{e}^{-(n/2)(\theta-\hat\theta)^{\mathrm{T}}\hat I(\tilde\theta)(\theta-\hat\theta)}\pi(\theta)\mathrm{d}\theta.\end{aligned}$$

(The domain of intgration is $\mathbb{R}^p$ but this can be cut down to a ball $B(\hat\theta, \epsilon)$ by allowing error terms of order $\mathcal{O}(\mathrm{e}^{-n\gamma})$ for suitable $\gamma > 0$. Then, the Taylor expansion can be used. Finally, one can go back to the original domain of integration again by adding an exponentialy small error term.) Standard conditions (see e.g. Clarke & Barron, 1988) give that the $\tilde\theta$ can be replaced by θ_T and the empirical Fisher information, $\hat I$, can be replaced by the actual Fisher information. Thus:

$$m(x^n) \approx p(x^n \mid \hat\theta)\int \mathrm{e}^{-(n/2)(\theta-\hat\theta)^{\mathrm{T}} I(\theta_T)(\theta-\hat\theta)}\pi(\theta)\mathrm{d}\theta.$$

The integrand is a normal density that can be integrated in closed form, apart from the π. By another approximation (that seems to be asymptotically tight up to $\mathcal{O}_P(\epsilon n^{p/2})$ factor where ϵ can be arbitrarily small) we get:

$$m(x^n) \approx p(x^n \mid \hat\theta)\int \mathrm{e}^{-(n/2)(\theta-\theta_T)^{\mathrm{T}} I(\theta_T)(\theta-\theta_T)}\pi(\theta)\mathrm{d}\theta. \qquad (3)$$

So far, this is standard. It becomes more interesting when the technique of the authors is invoked. Essentially, they diagonalise $I(\theta_T)$. For this, the p eigenvalues must be strictly positive, but that is not usually a difficult assumption to satisfy. Write $D = O^t I(\theta_T) O$ where O is an orthonormal matrix, i.e., a rotation, and $D = \mathrm{diag}(d_1, \ldots, d_p)$. (The authors use an orthogonal matrix, but an orthonormal matrix seems to give cleaner results.) Now, consider the transformation $\xi = O^T(\theta - \theta_T)$ so that $\mathrm{d}\xi = \mathrm{d}\theta$ by the orthonormality of O. Note that the transformation has been simplified since the argument of O is θ_T. Now, the integral in right hand side of expression (3) is

$$\begin{aligned}&\int \mathrm{e}^{-(n/2)\xi^T O^T I(\theta_T) O\xi}\pi(O(\theta_T)\xi + \theta_T)\mathrm{d}\theta \\ &\quad = \int \mathrm{e}^{-(n/2)\xi^T D\xi}\pi(O(\theta_T)\xi + \theta_T)\mathrm{d}\theta. \qquad (4)\end{aligned}$$

At this point the authors, rather than using Laplace's method on the integral, choose π as a product of individual π_i's for each ξ_i. Each factor in that product has hyperparameters λ_i, d_i, and b_i and the resulting p-dimensional integral in (4) has a closed form as given at the end of Sec. 2.

An alternative is the more conventional approach of recognising that as $n \to \infty$ the integrand converges to unit mass at $\xi = 0$. Using this gives that $m(x^n)$ is approximately

$$\begin{aligned}&p(x^n \mid \hat\theta) w(\theta_T)(2\pi)^{p/2}\det(nD)^{-1/2} \\ &\qquad \times \frac{\int \mathrm{e}^{-(1/2)\xi((nD)^{-1})^{-1}\xi}\mathrm{d}\xi}{(2\pi)^{p/2}\det(nD)^{-1/2}} \\ &\quad = p(x^n \mid \hat\theta) w(\theta_T)(2\pi)^{p/2}\frac{1}{n^{p/2}((\Pi_{i=1}^p d_i)^{1/p})^{p/2}} \\ &\quad = p(x^n \mid \hat\theta) w(\theta_T)(2\pi)^{p/2}\frac{1}{(ns)^{p/2}},\end{aligned}$$

where s is the geometric mean of the d_i's. The geometric mean is the side length of a p-dimensional cuboid whith volume equal do $\Pi_{i=1}^p d_i$. Thus, s plays the role of a sort of average Fisher information for the collection of ξ_i's. This sequence of approximations gives

$$\log m(x^n) = \ell(\hat\theta) + \log w(\theta_T) + \frac{p}{2}\log(2\pi) - \frac{p}{2}\log ns.$$

This leads to a form of the BIC as

$$\begin{aligned}BIC_S &= -2\ell(\hat\theta) + p\log n = -2\log m(x^n) \\ &\quad + 2\log w(\theta_T) - p\log(2\pi) - p\log s + o(1).\end{aligned} \qquad (5)$$

Comparing (5) and (2), the only difference is that the Fisher information is summarised by s, a sort of average efficiency that in effect puts all parameters on the

same scale. Roughly, $p \log s$ and $\log \det I(\theta)$ correspond to the term $\sum_{i=1}^{p} \log(1 + n_i^e)$ in the PBIC. The extra term in the PBIC, $-2\sum_{i=1}^{p} \log((1 - e^{v_i})/\sqrt{2}v_i)$, seems to correspond to the log prior density term.

As a third way to look at the BIC, oberve that neither (5) nor (2) have any clear analog to n_i^e apart from the treatment of Fisher information and its interpretation as an efficiency. So, two natural questions are what the effective sample sizes mean and what they are doing. In the PBIC they are introduced as hyperparameters and are restricted to linear models. For instance, in Example 3.3, effective sample sizes are average precisions divided by the maximal precision even though it is unclear why this expression has a claim to be an effective sample size.

On the other hand, in Sec. 3.2 a general definition of n_j^e in terms of entries of $I(\theta)$ is given for each $j = 1, \ldots p$. This is a valid generalisation of sample size because the n_j^e's reduce to n. Indeed, in the IID case with large n, $I_{jj}^*(\hat{\theta}) \approx nI_{jj}(\theta_T)$ and $w_{ij} \approx 1/n$. So, $\sum_{i=1}^{n} w_{ij} I_{ijj}^*(\theta_T) \approx (1/n) \sum_{i=1}^{n} I_{ijj}^*(\theta_T) \approx (1/n)(nI_{jj}(\theta_T)) = I_{jj}(\theta_T)$. This gives $n_i^e \approx nI_{jj}(\theta_T)^*/I_{jj}(\theta_T) = n$. In this generalisation, each n_j^e is closely related to the Fisher information and hence to the relative efficiency of estimating different parameters. Indeed, n_i^e is, roughly, the total Fisher information for θ_i (over the sample) as a fraction of the convex combination of Fisher informations for the θ_j's over the data.

Now, it may make sense to use the definition of n_j^e in Sec. 3.2 to generalise the BIC directly, i.e. find the n_j^e's first, since they depend only on the Fisher informations and on x^n, and use them to propose a new BIC. For instance, consider

$$\mathrm{BIC}_{\mathrm{TESS}} = -2\ell(\hat{\theta}) + \sum_{i=1}^{p} \log n_i^e. \tag{6}$$

In (6), the concept of effective sample size is used to account for the different efficiencies of estimating different parameters, making it valid to compare them. Note that (6) levels the playing field for the $f_i(\cdot\theta)$'s in the log-likelihood so that they do not need to be modified. Thus, effective sample sizes have a meaning something like the sample size required to make the estimation of one parameter (to a prescribed accuracy) close to the sample size required to estimate another parameter (to the same accuracy), a parallel to the appearance of the geometric mean in (5).

At this point, one can go back to (3) and (4) and seek ways to justify using n_i^e in place of n. Because (4) is nearly a product of univariate integrals it may be possible to regard the elements on the diagonal of D as a form of the Fisher information that permits replacement of n with n_i^e. Similarly, the geometric mean used in (5) may be related (by, say, log) to the ratios of sums of Fisher informations used to define n_i^e in Sec. 3.2 thereby relating (5) and (6). Finally, (6) is not obviously related to $m(x^n)$ but one can hope that a suitably reformulated Laplace's method on (3) and (4) may lead to a compatible expression for it.

One interesting query the authors are well-placed to answer is whether the results of Sec. 5.5 hold if the PBIC is replaced by (6). After all, there should be reasonable conditions under which all the n_i^e's from Sec. 3.2 increase fast enough with n, e.g. for all n, $0 < \eta < \min_j I_{jj}^* \leq I_{jj}^* < \max_j I_{jj}^* < B < \infty$.

3. Where to from here?

The authors have a very promising general definition in Sec. 3.2. Establishing a relationship between n_j^e and the effective sample size formulae proposed for linear models would be useful, but more fundamentally, the question is whether the n_j^e from Sec. 3.2 makes sense in such simpler contexts. If it does, then the fact that it differs from 'TESS' may not be very important. We strongly agree with the authors who write, á propos of n_j^e, that it should 'be viewed primarily as a starting point for future investigations of effective sample size'. (They actually limit this point to nonlinear models, but for the sake of a satisfying overall theory it should apply to linear models as well.)

Another tack is to be overtly information-theoretic by defining an effective sample size in terms of codelength. One form of the relative entropy, see Clarke and Barron (1988), is implicit in (2). However, one can use an analogous formulation to convert a putative sample of size n to an effective sample. Use a nonparametric estimator to form $h(x; x^n)$, an estimate of the density of X. Then, choose a 'distortion rate', r and find z^m for the smallest value of m that satisfies $D(h(\cdot; x^n) \| h(\cdot; z^m)) \leq r$, where $D(\cdot\|\cdot)$ is the relative entropy. This is the effective sample and sample size since it recreates the empirical density with a tolerable level of distortion. The larger r is, the more distortion is allowed and the smaller m will be. Information-theoretically, this is the same as approximating a Shannon code based on $h(\cdot; x^n)$ by a Shannon code based on $h(\cdot; z^m)$ in terms of small redundancy, in say, bits. This definition for effective sample size requires choosing r, but D is in bits so it would make sense for r to be some function of bits per symbol e.g. $\epsilon(\sigma n)$, where $\mathrm{Var}(X) = \sigma^2$, for some $c \in (0, 1]$, with $\epsilon = 1/2$ as a default.

Another way to look at this procedure for finding an effective sample size is via data compression. In this context, the rate distortion function is a well-studied quantity, see Cover and Thomas (2006), Chap. 10. The problem is that it's not obvious how to obtain an effective sample size from the rate distortion function, or, in the parlance of information theory, a set of lower dimensional canonical representatives that achieve the rate distortion function lower bound. On the other hand, this can be done in practice and further study may yield good solutions.

Finally, the rate distortion function is the result of an operation performed on a Shannon mutual information that, for parameteric families, usually has an expression in terms of the Fisher information. Likewise, it is well known that certain relative entropies can be expressed in terms of Fisher information. So, the definitions of effective sample size from an information theory perspective (via rate distortion) and form Sec. 3.2 (via efficiency) may ultimately coincide.

Disclosure statement

No potential conflict of interest was reported by the author.

References

Clarke, B., & Barron, A.1988). *Information-theoretic asymptotics of Bayes methods* (Technical Report #26). Stat. Dept., Univ. Illinois.

Cover, T., & Thomas, J. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: John Wiley and Sons.