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)'

Jun Shao & Sheng Zhang

To cite this article: Jun Shao & Sheng Zhang (2019) A discussion of 'prior-based Bayesian information criterion (PBIC)', Statistical Theory and Related Fields, 3:1, 19-21, DOI: 10.1080/24754269.2019.1583086

To link to this article: https://doi.org/10.1080/24754269.2019.1583086

STATISTICAL THEORY AND RELATED FIELDS
2019, VOL. 3, NO. 1, 19–21
https://doi.org/10.1080/24754269.2019.1583086

Taylor & Francis
Taylor & Francis Group

SHORT COMMUNICATION

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

Jun Shao and Sheng Zhang

Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

ARTICLE HISTORY Received 10 February 2019; Revised 12 February 2019; Accepted 12 February 2019

Professors Bayarri, Berger, Jang, Ray, Pericchi, and Visser deserve a special congratulation for their great work on Bayesian model and variable selection and a pioneering idea of prior-based Bayesian information criterion (PBIC). This work opens the door for contemporary advances in the difficult problem of model and variable selection.

There exist three types of commonly used Bayesian approaches. The first type works on information criterion, such as the well-known BIC. The newly proposed PBIC belongs to this category. The second type includes the indicator model selection (see, e.g., Brown, Vannucci, & Fearn, 1998; Dellaportas, Forster, & Ntzoufras, 1997; George & McCulloch, 1993; Kuo & Mallick, 1998; Yuan & Lin, 2005), the stochastic search method (e.g., O'Hara & Sillanpää, 2009), and the model space method by Green (1995). The third type, which is considered in this discussion, is to apply priors on the regression coefficients that promotes the shrinkage of coefficients towards 0. This last type of approaches is intrinsically connected with frequentist methods in the sense that, first of all, such priors play the same role as the assumption that the coefficients are sparse for the frequentist approach and secondly, in some sense, the Bayesian solution is equivalent to the corresponding frequentist counterpart with a certain penalty parameter. Typical research papers for this type include Griffin and Brown (2009), Park and Casella (2008), and Kyung, Gilly, Ghosh, and Casella (2010).

The shrinkage prior approach may not provide sparse estimates of regression coefficients in general, which could not only complicate the interpretation but also inflate statistical error in analysis. Even without a well-defined variable selection approach, a Bayesian analysis based on a subset of covariates with size considerably less than the original dimensionality, which is referred to as sparse Bayesian analysis, could produce better results than the Bayesian analysis based on all covariates. Several attempts have been made to obtain sparse Bayesian estimates based on shrinkage priors. For instance, Hoti and Sillanpää (2006) proposed a method based on thresholding; however, the method is based on certain approximations, and the choice of the threshold is ad hoc. Another example is the sparse Bayesian learning by Tipping (2001), but it involves complicated nonconvex optimisation and assumes that the variance of the error term is known.

Under the framework of shrinkage priors, we consider a Bayesian variable selection method via a benchmark variable. The benchmark variable serves as a standard to measure the importance of each variable based on the posterior distribution of the corresponding coefficient.

As the first attempt, we focus on linear regression models with normally distributed errors. Let $\boldsymbol{y}$ be an n-dimensional vector of response and, without loss of generality, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ be p centralised n-dimensional vectors of predictors or covariates. Conditional on $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, $\boldsymbol{y}$ is assumed to be multivariate normally distributed as $N(\beta_0\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$, where $\boldsymbol{\beta}$ is a p-dimensional column vector whose jth component is β_j, $\beta_0, \beta_1, \ldots, \beta_p$ are $p+1$ unknown parameters, σ is an unknown positive parameter, $\mathbf{1}$ is a $n-$dimensional vector with all components 1 and $\boldsymbol{I}$ is the identity matrix of order n.

Consider the following prior density on $\boldsymbol{\beta}$ conditioned on σ^2,

$$p(\boldsymbol{\beta}|\sigma^2) = \prod_{i=1}^{p} \frac{\lambda^{1/\delta}}{2^{1/\delta}\Gamma(1/\delta)\sigma} \exp\left(-2\lambda \left|\frac{\beta_i}{2\sigma}\right|^{\delta}\right) \quad (1)$$

where $\lambda > 0$ and $1 \le \delta \le 2$ are hyper-parameters. When $\delta = 1$, this is the Laplace prior which was considered by Park and Casella (2008) for their Bayesian Lasso. When $\delta = 2$, the prior in (1) is a multivariate normal density and produces the posterior mode of $\boldsymbol{\beta}$ equivalent to the ridge regression estimate. As the Laplace prior is 'sharper' than the Gaussian one, it is expected to yield more sparse predictive models with the potentiality of easier interpretation, which is especially desirable for high-dimensional data with a considerably large amount of noisy

CONTACT Jun Shao shao@stat.wisc.edu

variables. However, the posterior inferences associated with Laplace prior involves relatively intensive computation.

For β_0 and σ^2 that are not involved with variable selection, we consider noninformative priors so that the overall prior for all parameters is

$$p(\beta_0, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2} p(\boldsymbol{\beta}|\sigma^2)$$

If the posterior distribution of β_i is nearly the same as that from a noise variable centred at 0, then it is natural to eliminate $\boldsymbol{x}_i$ as an unimportant covariate. However, the question is how to quantify whether a posterior distribution to be close to that of a noise.

To illustrate our idea, let us first consider an artificial case where a covariate $\boldsymbol{z}$ exists and is known to have no effect on $\boldsymbol{y}$. For example, $\boldsymbol{y}$ is distributed as $N(\boldsymbol{z}\beta_z + \mathbf{1}\beta_0 + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ with $\beta_z = 0$. Although we know $\boldsymbol{z}$ is redundant, we still put a prior on β_z such that β_z and β_i are independently identically distributed conditioning on σ^2. Under this setting, $\boldsymbol{x}_i$ could be treated as an unimportant variable if the posterior of β_i is similar to the posterior of β_z. In other words, the variable $\boldsymbol{z}$ serves as a benchmark in measuring the importance of $\boldsymbol{x}_i$'s.

A benchmark variable should have a posterior distribution centred at 0 and should not affect the Bayesian analysis concerning $\boldsymbol{\beta}$. The question is, how do we find a benchmark variable when we do not have a redundant variable at hand?

We now show that there is a universal solution. Since $\boldsymbol{X}$ is column-wisely centralised, the density of $\boldsymbol{y}$ given $(\boldsymbol{X}, \boldsymbol{z}, \beta_0, \boldsymbol{\beta}, \beta_z, \sigma^2)$ is

$$\begin{aligned} &p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{z}, \beta_0, \boldsymbol{\beta}, \beta_z, \sigma^2) \\ &\quad\propto \frac{1}{\sigma^n} \exp\left(-\frac{||\boldsymbol{y} - \boldsymbol{z}\beta_z - \mathbf{1}\beta_0 - \boldsymbol{X}\boldsymbol{\beta}||^2}{2\sigma^2}\right) \\ &\quad= \frac{1}{\sigma^n} \exp \\ &\qquad\times \left(-\frac{\begin{array}{c}||\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}||^2 + ||\boldsymbol{z} - \bar{z}\mathbf{1}||^2 \beta_z^2 - 2\beta_z \boldsymbol{z}' \\ (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + n(\beta_0 - \bar{y} + \beta_z \bar{z})^2\end{array}}{2\sigma^2}\right) \end{aligned}$$

where $\bar{y}$ is the average of components of $\boldsymbol{y}$, $\bar{z}$ is the average of components of $\boldsymbol{z}$, $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}$, $||\boldsymbol{a}||^2 = \boldsymbol{a}'\boldsymbol{a}$, and $\boldsymbol{a}'$ is the transpose of $\boldsymbol{a}$. Under the previously described prior, the joint conditional posterior distribution $p(\beta_0, \boldsymbol{\beta}, \beta_z|\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{y}, \sigma^2)$ can be obtained. Since the intercept β_0 is not of interest, we integrate it out from $p(\beta_0, \boldsymbol{\beta}, \beta_z|\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{y}, \sigma^2)$ and obtain the conditional posterior density of $(\boldsymbol{\beta}, \beta_z)$ given $(\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{y}, \sigma^2)$ as follows:

$$\begin{aligned} &p\left(\boldsymbol{\beta}, \beta_z|\boldsymbol{X}, \boldsymbol{z}, \boldsymbol{y}, \sigma^2\right) \\ &\quad\propto \frac{1}{\sigma^{n+1}} \\ &\qquad\times \left[p\left(\boldsymbol{\beta}|\sigma^2\right) \exp\left(-\frac{||\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}||^2 + 2\beta_z \boldsymbol{z}'\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}\right)\right] \\ &\qquad\times \left[p\left(\beta_z|\sigma^2\right) \exp\left(-\frac{||\boldsymbol{z} - \bar{z}\mathbf{1}||^2 \beta_z^2 - 2\boldsymbol{z}'\tilde{\boldsymbol{y}}\beta_z}{2\sigma^2}\right)\right] \end{aligned} \tag{2}$$

Note that marginalisation over β_0 is equivalent to centralising the response $\boldsymbol{y}$. After the integration, it could be regarded that the posterior inferences are drawn from the centralised response $\tilde{\boldsymbol{y}}$ instead of the original $\boldsymbol{y}$. The reason that we introduce β_0 in the model and then integrate it out, instead of eliminating it at the very beginning and directly building a linear regression model as $\tilde{\boldsymbol{y}} = \boldsymbol{z}\beta_z + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, is mainly for the mathematical rigorousness, as $\tilde{\boldsymbol{y}}$ is not of full rank and has a degenerate distribution.

The conditional posterior density in (2) implies that given $(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{z}, \sigma^2)$, $\boldsymbol{\beta}$ and β_z are independent if and only if $\boldsymbol{z}'\boldsymbol{X} = 0$. In other words, $\boldsymbol{z}$ does not affect the posterior of $\boldsymbol{\beta}$ if and only if $\boldsymbol{z}$ is orthogonal to all $\boldsymbol{x}_i$'s, $i = 1, \ldots, p$. Meanwhile, the posterior of β_z is centred at 0 if and only if $\boldsymbol{z}'\tilde{\boldsymbol{y}} = 0$. Is there a $\boldsymbol{z}$ orthogonal to $(\boldsymbol{X}, \tilde{\boldsymbol{y}})$? Clearly, $\boldsymbol{z} = \mathbf{1}$, the column vector of ones, is a direct solution and could be used as a benchmark to assess the importance of $\boldsymbol{x}_i$'s. When $\boldsymbol{z} = \mathbf{1}$, the posterior density of β_z remains the same as its prior, and the posterior density of $(\boldsymbol{\beta}, \beta_z, \sigma^2)$ is simplified to

$$\begin{aligned} &p\left(\boldsymbol{\beta}, \beta_z, \sigma^2|\boldsymbol{X}, \boldsymbol{y}\right) \\ &\quad\propto \frac{1}{\sigma^{n+1}} p\left(\beta_z|\sigma^2\right) p\left(\boldsymbol{\beta}|\sigma^2\right) \exp\left(-\frac{||\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}||^2}{2\sigma^2}\right) \end{aligned} \tag{3}$$

The benchmark serves as a measure to assess the importance of each covariate, and therefore provide guidance on variable selection. How to carry out variable selection using posterior (3) or extend the ideal to more general settings requires more research. In the rest of this discussion, we consider a real data example.

The prostate cancer data originally came from a research conducted by Stamey et al. (1989), and it was studied by Tibshirani (1996) and Zou and Hastie (2005). The goal of the research was to explore the relation between the level of prostate specific antigen and several clinical measures in men before their hospitalisation for radical prostatectomy. The data frame contains 97 observations and 9 variables. The response is the logarithm of prostate-specific antigen (lpsa), while the 8 covariates are the logarithm of cancer volume (lcavol), logarithm of prostate weight (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), the logarithm of capsular penetration (lcp), Gleason score (gleason) and percentage Gleason score 4 or 5 (pgg45).

Figure 1 visualises the posteriors with Laplace prior ($\delta = 1$). Results with normal prior ($\delta = 2$) are similar

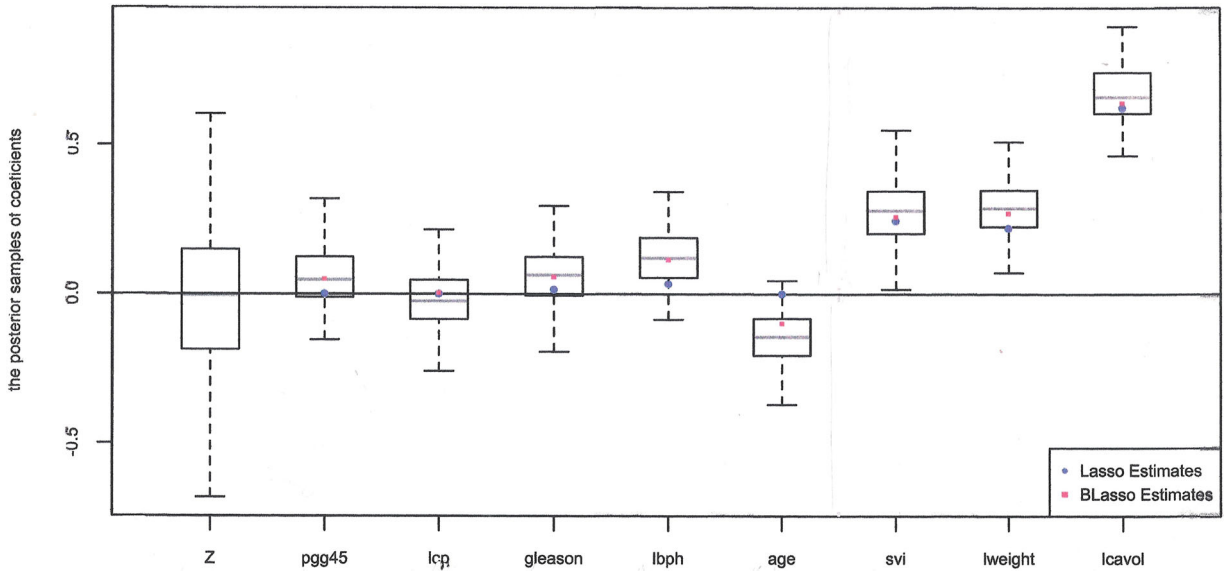

Figure 1. Posterior plots on the prostate cancer data.

and omitted. In Figure 1, the leftmost boxplot is based on the posterior samples of the coefficient for the benchmark $z = \mathbf{1}$. It is distributed almost symmetrically around 0 as expected. Other box plots represent the posterior distributions of the coefficients associated with 8 covariates. It can be seen that the three posteriors plotted in the far right of Figure 1 are clearly different from the posterior of the benchmark and, hence, we conclude that the corresponding three covariates, svi, lweight, and lcavol, are useful for the response. On the other hand, the posteriors of three covariates next to the benchmark in Figure 1 are not different from the benchmark posterior and, hence, the covariates pgg45, lcp, and gleason are not useful. The posteriors of lbph and age are just marginally different from that of the benchmark, and we still consider them to be not useful covariates.

Figure 1 also includes Lasso and Bayesian Lasso estimates of each coefficients, marked as circles and squares in the figure. The Lasso estimates are zero for pgg45, lcp, and age, nonzero for the other 5 covariates. Thus, the Lasso approach agrees with our approach for covariates pgg45, lcp, age, svi, lweight, and lcavol, but does not agree on gleason and lbph. Since the magnitudes of Lasso estimates for gleason and lbph are small, another thresholding added to Lasso will result in the same conclusion with ours. Meanwhile, the Bayesian Lasso evaluates all the coefficients to be nonzero as it doesn't select variables to promote model sparsity.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Brown, P. J., Vannucci, M., & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, *60*, 627–641.

Dellaportas, P., Forster, J. J., & I., Ntzoufras (1997). *On Bayesian model and variable selection using MCMC* (Technical Report). Athens: Department of Statistics, Athens University of Economics and Business.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *85*, 398–409.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Griffin, J. E., & Brown, P. J. (2009). *Inference with Normal-Gamma prior distributions in regression problems* (Technical Report). Institute of Mathematics, Statistics and Actuarial Science, University of Kent.

Hoti, F., & Sillanpää, M. J. (2006). Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity*, *97*, 4–18.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhya Series B*, *60*, 65–81.

Kyung, M., Gilly, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, *5*, 369–412.

O'Hara, R. B., & Sillanpää, M. J. (2009). Review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*, 85–118.

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*, 681–686.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients. *Journal of Urology*, *141*, 1076–1083.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning*, *1*, 211–244.

Yuan, M., & Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, *100*, 1215–1225.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*, 301–320.