数理解析研究所講究録
第1621巻 2009年 130-134

Extending the g-prior for Bayesian model selection

丸山 祐造
YUZO MARUYAMA
東京大学・空間情報科学研究センター
CENTER FOR SPATIAL INFORMATION SCIENCE, THE UNIVERSITY OF TOKYO *
Edward, I. George
DEPARTMENT OF STATISTICS, UNIVERSITY OF PENNSYLVANIA †

Suppose the normal linear regression model is used to relate Y to the potential predictors $X_1, \ldots, X_p$,

$$Y \sim N_n(\alpha 1_n + X_F \beta_F, \sigma^2 I_n) \tag{1}$$

where α is an unknown intercept parameter, 1_n is an $n \times 1$ vector each component of which is one, $X_F = (X_1, \ldots, X_p)$ is an $n \times p$ design matrix, β_F is a $p \times 1$ vector of unknown regression coefficients, I_n is an $n \times n$ identity matrix and σ^2 is an unknown positive scalar. We assume that the columns of X_F have been standardized so that for $1 \leq i \leq p$, $\bar{X}_i = 0$ and $X_i' X_i / n \equiv 1$. The subscript $_F$ of X_F and β_F in (1) means the model (1) is the full model. We shall be particularly interested in the variable selection problem where we would like to select an unknown subset of the important predictors. It will be convenient throughout to index each of these 2^p possible subset choices by the vector

$$\gamma = (\gamma_1, \ldots, \gamma_p)$$

where $\gamma_i = 0$ or 1. We use $q_\gamma = \gamma' 1_p$ to denote the size of the γth subset. The problem then becomes that of selecting a submodel of (1) which has a density of the form

$$p_y(Y|\alpha, \beta_\gamma, \sigma^2, \gamma) = \phi_n(y; \alpha 1_n + X_\gamma \beta_\gamma, \sigma^2 I_n) \tag{2}$$

where $\phi_n(y; \mu, \Sigma)$ denotes the n-variate normal density with mean vector μ and covariance matrix Σ. In (2), X_γ is the $n \times q_\gamma$ matrix whose columns corresponds to the γth subset

*maruyama@csis.u-tokyo.ac.jp
†edgeorge@wharton.upenn.edu

of $X_1, \ldots, X_p$, β_γ is a $q_\gamma \times 1$ vector of unknown regression coefficients. The rank of X_γ is assumed to be

$$\text{rank } X_\gamma = \min(n-1, q_\gamma) = r_\gamma. \tag{3}$$

Let $\mathcal{M}_\gamma$ denote the submodel given by (2). In the paper we will omit γ if it does not confuse the readers.

A Bayesian approach to this problem entails the specification of prior distributions on the models $p_\gamma = \Pr(\mathcal{M}_\gamma)$, and on the parameters $p(\alpha, \beta_\gamma, \sigma^2)$ of each model. For each such specification, of key interest is the posterior probability of $\mathcal{M}_\gamma$ given y

$$\Pr(\mathcal{M}_\gamma|y) = \frac{p_\gamma m_\gamma(y)}{\sum_\gamma p_\gamma m_\gamma(y)} = \frac{p_\gamma \text{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]}{\sum_\gamma p_\gamma \text{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]}, \tag{4}$$

where $m_\gamma(y)$ is the marginal density of y under $\mathcal{M}_\gamma$. In (4), $\text{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]$ is so called "null-based Bayes factor" for comparing each of $\mathcal{M}_\gamma$ to the null model $\mathcal{M}_N$ which is defined as

$$\text{BF}[\mathcal{M}_\gamma; \mathcal{M}_N] = \frac{m_\gamma(y)}{m_N(y)},$$

where the null model $\mathcal{M}_N$ is given by

$$Y \sim N_n(\alpha 1_n, \sigma^2 I_n) \tag{5}$$

and $m_N(y)$ is the marginal density of y under the null model. In Bayesian model selection, the Bayes factor is often used as a criterion instead of the marginal density directly. A popular strategy is to select the model for which $\Pr(\mathcal{M}_\gamma|y)$ or $p_\gamma \text{BF}[\mathcal{M}_\gamma; \mathcal{M}_N]$ is largest.

Our main focus in this paper is to propose and study specifications for the parameter prior $p(\alpha, \beta, \sigma^2)$. Although we will not dwell on the model prior specifications, we note in passing that it may be a good idea to avoid a prior which puts equal weights all the models. A default choice that has emerged is the independent Bernoulli prior

$$p_\gamma(\gamma|w) = w^{q_\gamma}(1-w)^{p-q_\gamma},$$

which is controlled by a single hyperparameter $w \in (0,1)$. Under this prior, each predictor is independently included in the model with the same probability w. This prior includes the uniform distribution over models, $w = 1/2$, which is considered by many as the natural "non-informative" choice. We will observe that such unequal weights model space priors may play an important role, especially in many predictors case.

We turn to prior density of the unknown parameters of the submodel $\mathcal{M}_\gamma$. In particular, the joint density we consider has a form

$$p(\alpha, \beta, \sigma^2) = p(\alpha)p(\sigma^2)p(\beta|\sigma^2) = p(\alpha)p(\sigma^2)\int_0^\infty p(\beta|\sigma^2, g)p(g)dg,$$

where $p(\alpha) = 1$ and $p(\sigma^2) = \sigma^{-2}$. The validness of non-informative but improper priors for α and σ^2 is clear.

The most tractable prior distribution of β is normal conjugate since we consider the normal linear regression model. Among a class of normal conjugate priors, so-called Zellner's (1986) g-prior

$$p(\beta|\sigma^2, g) = N_q(0, g\sigma^2(X'X)^{-1}), \tag{6}$$

is often used (George and Foster (2000), Fernandez et.al. (2001), Liang et.al. (2008)). Since (6) includes the inverse of $X'X$, it is applicable only for the traditional situation $p \leq n-1$, which means $q \leq n-1$ for any submodel. George and Foster (2000) show that the marginal density of y given g and σ^2 under $\mathcal{M}_\gamma$, which is denoted by $m_\gamma(y|g, \sigma^2)$, is given by

$$m_\gamma(y|g, \sigma^2) \propto \exp\left(\frac{g}{g+1}\left\{\max_{\alpha,\beta} \log p(Y|\alpha, \beta, \sigma^2) - qH(g)\right\}\right) \tag{7}$$

where $H(g) = (2g)^{-1}(g+1)\log(g+1)$ under the priors $p(\alpha) = 1$ and (6). Hence if the variance σ^2 is known and we choose g independently of y which satisfies $H(g) = 2$ or $\log n$ the Bayesian strategy using (7) exactly corresponds to AIC by Akaike (1974) or BIC by Schwarz (1978). (See George and Foster (2000) and Liang et.al. (2008) for the detail of a wide variety of related choices for penalty term $H(g)$.) This equivalence is the essential reason that g-prior of β is considered useful for selecting the best model in the Bayesian framework. Another point of the g-prior is that it makes the marginal density or the Bayes factor as a function of the important statistics from the frequentist point of view, like the maximum log likelihood or the residual sum of squares (RSS).

However there are some unsatisfactory points in the last paragraph. When the maximization of (7) is considered, the variance σ^2 is assumed to be known whereas it is usually unknown in the real situations. George and Foster (2000) insert the unbiased estimator of variance based on the full model given by (1) or the submodel $\mathcal{M}_\gamma$ after deriving the criterion (7). In rigorous Bayesian point of view, however, if the variance is unknown, the prior distribution should be given. Furthermore we would like to assume the prior distribution of g instead of estimating a fixed g by empirical Bayes method. In such full Bayes methods, even if inverse-gamma conjugate prior for σ^2 are also used, there still remains an integral with respect to a hyperparameter g, which is usually calculated by MCMC or the Laplace approximation. We believe that a more analytical result in full Bayes setting is desirable for the theoretical and practical point of view. This is one of our motivation of this paper.

Additionally, in modern statistics, treating (very) many predictors case ($p \geq n$) becomes more and more important. Note that since RSS is zero in the case where $q \geq n-1$, neither naive AIC nor BIC methods do work. Since the covariance matrix in Zellner's g-prior does

not exist for $q \geq n$, no Bayesian criteria based on the original g-prior including George and Foster (2000) and Liang et.al. (2008) is available. If we use not the original g-prior but a typical prior $\beta \sim N(0, \lambda\sigma^2 I_q)$, the Bayes factor is well-defined in many predictors case. However, it is no longer a function of important statistics from the viewpoint of frequentist unlike the Bayes factor with the g-prior. We would like to extend the goodness of the g-prior to the many predictors case naturally, which is also our motivation in this paper.

In this paper, we find a special variant of Zellner's g-prior which enables us to not only calculate the marginal density based on full Bayes method analytically but also treat many predictors case. Eventually we propose a following analytical Full Bayes Factor (FBF) which is a function of fundamental aggregated information of data:

$$\mathrm{FBF}[\mathcal{M}_\gamma] \tag{8}$$
$$= \begin{cases} \left\{ \overline{\mathrm{sv}}[X] \times \|\hat{\beta}_{LSE}^{MP}\| \right\}^{-n+1} & \text{if } q \geq n-1, \\ \dfrac{\{\mathrm{min.sv}[X]\}^q}{\overline{\mathrm{sv}}[X]^q} \left\{ 1 - R^2 + \{\mathrm{min.sv}[X]\}^2 \|\hat{\beta}_{LSE}\|^2 \right\}^{-1/4-q/2} & \\ \quad \times \left(1-R^2\right)^{-(n-q)/2+3/4} \dfrac{B(q/2+1/4, (n-q)/2-3/4)}{B(1/4, (n-q)/2-3/4)} & \\ \qquad \text{if } q \leq n-2. & \end{cases}$$

In (8), $\|\cdot\|$ denotes the Euclid norm, R^2 is the R-squares under $\mathcal{M}_\gamma$, $\overline{\mathrm{sv}}[X]$ and $\mathrm{min.sv}[X]$ are the geometric average and minimum value of the singular values of X, respectively. Also $\hat{\beta}_{LSE}$ for $q \leq n-2$ is the usual least squares estimator, $\hat{\beta}_{LSE}^{MP}$ for $q \geq n-1$ is the least squares estimator using the Moore-Pennrose inverse matrix, where the response variable is standardized as $(y - \bar{y}1_n)/\|y - \bar{y}1_n\|$. Our criterion $\mathrm{FBF}[\mathcal{M}_\gamma]$ has a reasonable interpretation.

参 考 文 献

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control AC-19*, 716–723. System identification and time-series analysis.

[2] Casella, G. (1980). Minimax ridge regression estimation. *Ann. Statist.* **8**, 5, 1036–1056.

[3] Casella, G. (1985). Condition numbers and minimax ridge regression estimators. *J. Amer. Statist. Assoc.* **80**, 391, 753–758.

[4] Chipman, H., George, E. I., and McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. In *Model selection.* IMS Lecture Notes Monogr. Ser., Vol. **38**. Inst. Math. Statist., Beachwood, OH, 65–134.

[5] FERNÁNDEZ, C., LEY, E., AND STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100**, 2, 381–427.

[6] GEORGE, E. I. AND FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 4, 731–747.

[7] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. (2001). *The elements of statistical learning.* Springer Series in Statistics. Springer-Verlag, New York. Data mining, inference, and prediction.

[8] HURVICH, C. M. AND TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 2, 297–307.

[9] KNIGHT, K. AND FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 5, 1356–1378.

[10] LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A., AND BERGER, J. O. (2008). Mixtures of g-priors for bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 481, 410–423.

[11] MARUYAMA, Y. AND STRAWDERMAN, W. E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *Ann. Statist.* **33**, 4, 1753–1770.

[12] RAFTERY, A. E., MADIGAN, D., AND HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92**, 437, 179–191.

[13] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 2, 461–464.

[14] STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42**, 1, 385–388.

[15] ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques.* Stud. Bayesian Econometrics Statist., Vol. **6**. North-Holland, Amsterdam, 233–243.

[16] ZELLNER, A. AND SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Eds. University of Valencia, 585–603.

[17] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 476, 1418–1429.