Selecting Tuning Parameters for sparse regression models via Generalized Information Criterion

Fumitake Sakaori*
Chuo University, Tokyo, Japan - sakaori@math.chuo-u.ac.jp

Abstract

Sparse regression models, e.g., Lasso, SCAD, and elastic net, have become very popular. These methods estimate model parameters by minimizing penalized or regularized objective function, where the penalty is imposed to the L_1-norm of the parameters, and then simultaneously produce estimators of the parameter and sparseness of variables, that is, variable selection. Hence, prediction accuracy and sparsity for these methods depend on selection of the tuning parameters. Suitable choice of tuning parameters give a prediction optimality and/or a consistency of selecting variables. Information criteria such as AIC and BIC are often used to select the tuning parameters for gaussian sparse regression models. However, these criteria have been derived for only maximum likelihood estimators, and are not appropriate for regularized estimators. For non-MLE estimators, the generalized information criteria (GIC) given by Konishi and Kitagawa (1996) can be used instead of AIC and BIC. GIC is a natural extension of AIC or TIC for non-MLE estimators with differentiable functional form. In this study, we derive the exact GIC for some sparse regression models, and verify its properties.
Keywords: Lasso; GIC.

1. Introduction

Recently, sparse regression models, or regression modeling with L_1 norm regularization, e.g., Lasso, SCAD, and Adaptive Lasso, have been used considerably in many fields of application, and also have had many theoretical results. These methods estimate model parameters by minimizing penalized or regularized objective function, where the penalty is imposed to the L_1-norm of the parameters, and then simultaneously produce estimators of the parameter and sparseness of variables, that is, variable selection.

The performance of these methods heavily depend on selection of the tuning parameter λ, which controls model sparsity and prediction accuracy. Therefore it is essential to select the tuning parameter appropriately.

Cross-validation (CV) are often used due to advantages of its usefulness and useless of any probabilistic assumptions. However, some literature indicates disadvantages of CV, e.g., high computational cost, high variability and tendency to undersmooth (e.g., Araki *et al.*, 2009). Although information criteria such as AIC and BIC are also used, these are derived for the maximum likelihood estimate, and have no theoretical justification for regularized estimators.

We propose to use the generalized information criteria GIC (Konishi and Kitagawa, 1996) for tuning parameter selection. Matsui and Konishi (2011) and Park *et al.* (2012) use GIC for some variants of the sparse regression models, where they consider the quadratic approximation of the L_1 term. In this study, we do not use any approximation of L_1 term.

In Section 2 we define sparse regression models and state the tuning parameter selection problem. In Section 3 we derive GIC for the sparse regression models and give some numerical results. Section 4 contains a summary of this study and some discussion.

2. Sparse Regression Models and tuning parameter selection

Consider a linear regression model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is a response vector, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is a matrix with explanatory variables in its columns, $\boldsymbol{\beta}$ is a regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an error vector having mean $\mathbf{0}$ and variance covariance matrix $\sigma^2 I_n$. Without loss of generality, the response variable is centered and the explanatory variables have been standardized.

Sparse estimators $\hat{\boldsymbol{\beta}}$ is defined as

$$\hat{\boldsymbol{\beta}} = \mathrm{argmin}||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + P(|\boldsymbol{\beta}|),$$

where $P(|\boldsymbol{\beta}|)$ is a penalty term, where the penalty is imposed to the L_1-norm of $\boldsymbol{\beta}$, for instance, $P(|\boldsymbol{\beta}|) = \lambda|\boldsymbol{\beta}|$ for lasso. If the distribution of $\boldsymbol{\varepsilon}$ is a multivariate normal, then the sparse estimators can be written as the penalized gaussian likelihood estimators with L_1-penalty.

There exists some theories of sparse estimators. Lasso is L_2-consistent and variable selection consistent, whereas SCAD and adaptive lasso has the oracle property (Wang *et al.*, 2007). Also, there exists some studies about the choice of the tuning parameters. Meinshausen and Bühlmann (2006) and Leng *et al.* (2006) states the conflict of prediction optimality and consistency of lasso. Wang *et al.* (2007) and Zhang *et al.* (2010) discuss about SCAD, and conclude that BIC is consistent and Generalized cross-validation is not satisfactory.

Cross-validation (CV) are often used due to advantages of its usefulness and useless of any probabilistic assumptions. However, some literature indicates disadvantages of CV, e.g., high computational cost, high variability and tendency to undersmooth (e.g., Araki *et al.*, 2009). Also, information criterion such as AIC and BIC have been used practically to choose the tuning parameters for gaussian sparse regression models. AIC for gaussian sparse regression model can be written as

$$\mathrm{AIC} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2(q+2),$$

where $\hat{\boldsymbol{\theta}}$ is a L_1 regularized estimator of θ, and q is the degrees of freedom of the model which is given as

$$q = \sum_{i=1}^{n} \frac{\mathrm{Cov}(\hat{y}_i, y_i)}{\sigma^2}.$$

On the other hand, BIC for the model can be written as

$$\mathrm{BIC} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + (q+2)\log n.$$

Zou *et al.* (2007) shows that the number of non zero variables is an unbiased estimator of q. Hirose *et al.* (2013) gives an efficient algorithm to calculate the degrees of freedom. We may use AIC and BIC by estimating the degrees of freedom.

However, there is no theoretical justification of AIC and BIC for sparse estimators. AIC is derived from Kullback-Leibler distance for maximum likelihood estimators, and BIC is derived from posterior probability also for maximum likelihood estimators.

3. GIC

In order to overcome the problem of theoretical justification, we propose to use a generalized information criteria GIC (Konishi and Kitagawa, 1996) instead of AIC and BIC. GIC is given by

$$\mathrm{GIC} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + \frac{2}{n}\sum_{i=1}^{n} \mathrm{tr}\left\{\boldsymbol{T}^{(1)}(y_i;\hat{G})\frac{\partial \log f(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{T}(\hat{G})}\right\},$$

where $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ is any estimator of θ and $\boldsymbol{T}^{(1)}(y;G) = \left(T_1^{(1)}(y;G), \ldots, T_p^{(1)}(y;G)\right)^T$ is the influence functions of a p-dimensional functional $\boldsymbol{T}(G)$.

For maximum likelihood estimators, GIC and AIC are equivalent, so GIC is a natural extension of AIC for non-MLE.

GIC is not so common for tuning parameter selection, and only a few studies were done. Matsui and Konishi (2011) considered local quadratic approximation of L_1 term in SCAD for functional regression models. Also, Park *et al.* (2012) considered the same approximation for Lasso-type estimators in robust regression models. In this study we consider GIC with no approximation of L_1 term.

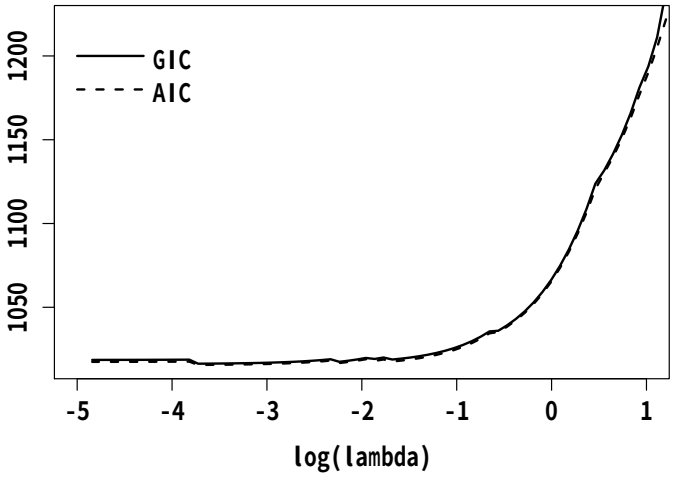

Figure 1: A comparison of GIC and AIC of a simulated data set when $n = 200$

By ignoring non-selected variables, we may calculate the bias correction term in GIC, and GIC can be written as

$$\mathrm{GIC} = -2\sum_{i=1}^{n} \log f(y_i|\hat{\boldsymbol{\theta}}) + 2\mathrm{tr}\left\{R(\boldsymbol{\psi}, \hat{G})^{-1}Q(\boldsymbol{\psi}, \hat{G})\right\},$$

where

$$R(\psi, \hat{G}) = \frac{1}{n\hat{\sigma}^2}\begin{pmatrix} X^TX & \frac{1}{\hat{\sigma}^2}X^TE\mathbf{1} \\ \frac{1}{\hat{\sigma}^2}\mathbf{1}^TEX & \frac{1}{2\hat{\sigma}^2} \end{pmatrix},$$

$$Q(\psi, \hat{G}) = \frac{1}{n\hat{\sigma}^2}\begin{pmatrix} \frac{1}{\hat{\sigma}^2}X^TE^2X - \lambda\mathbf{sgn}(\hat{\boldsymbol{\beta}})X^TE\mathbf{1}\mathbf{1}^T & -\frac{X^TE\mathbf{1}}{2\hat{\sigma}^2} + \frac{X^TE^3\mathbf{1}}{2\hat{\sigma}^4} + \frac{n\lambda\mathbf{sgn}(\hat{\boldsymbol{\beta}})}{2} - \frac{\lambda E^2\mathbf{1}}{2\hat{\sigma}^2} \\ -\frac{\mathbf{1}^TEX}{2\hat{\sigma}^2} + \frac{\mathbf{1}^TE^3X}{2\hat{\sigma}^4} & \frac{\mathbf{1}^TE^4\mathbf{1}}{4\hat{\sigma}^6} - \frac{n}{4\hat{\sigma}^2} \end{pmatrix},$$

$E = \mathrm{diag}(y_1 - \boldsymbol{x}_1^T\hat{\boldsymbol{\beta}},\ \ldots, y_n - \boldsymbol{x}_n^T\hat{\boldsymbol{\beta}})$, $\mathbf{1} = (1, \ldots, 1)^T$, and $\mathbf{sgn}(\hat{\boldsymbol{\beta}}) = (0, \mathrm{sgn}(\hat{\beta}_1), \ldots, \mathrm{sgn}(\hat{\beta}_p))^T$.

The difference between GIC and AIC is the bias correction term. AIC includes only the degrees of freedom, whereas GIC includes the estimator of $\boldsymbol{\beta}$ and the error variance σ^2. For the estimation of the error variance, we may consider

$$\hat{\sigma}^2_\lambda = \frac{1}{n}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\lambda)^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_\lambda)$$

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{\mathrm{ML}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{\mathrm{ML}}),$$

or unbiased sample variance of each using the degrees of freedom of the model.

Figure 1 shows a toy example. We generate a dataset by the following settings: $\boldsymbol{\beta} = (3,\ 1.5,\ 0,\ 0,\ 2,\ 0,\ 0,\ 0)^T$, $x_j \overset{iid}{\sim} N(0,1)$, $\mathrm{Cor}(x_j, x_k) = 0.5^{|i-j|}$, $\varepsilon \overset{iid}{\sim} N(0, 3^2)$, $n = 200$. In this case, GIC and AIC are almost the same value, and give the same λ value. As n decreases, the difference betweenn GIC and AIC becomes large.

4. Conclusions

In this study, we propose to use GIC for tuning parameter selection in sparse regression models, which has theoretical justification for sparse estimators. By some examples and simulations, we compare the difference between GIC and AIC. Further study is needed for the variance estimate.

References

Araki, Y., Konishi, S., Kawano, S. & Matsui, H. (2009). Functional regression modeling via regularized Gaussian basis expansions. Annals of Institute of Statistical Mathematics, 61, 811-833.

Hirose, K., Tateishi, S., & Konishi, S. (2013). Tuning parameter selection in sparse regression modeling, Computational Statistics & Data Analysis, 59, 28-40.

Konishi, S. & Kitagawa, G. (1996). Generalised information criteria in model selection. Biometrika, 83, 4, 875-890

Leng, C., Lin, Y. & Wahba, G. (2006). A note on the Lasso and related procedures in model selection, Statistica Sinica, 16, 1273-1284

Matsui H. & Konishi, S. (2011). Variable selection for functional regression models via the L1 regularization, Computational Statistics & Data Analysis, 55, 12, 3304-3310

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso, Ann. Statist., 34, 3, 1436-1462.

Park, H., Sakaori, F. & Konishi, S. (2012). Selection of tuning parameters in robust sparse regression modeling, Proceedings of COMPSTAT2012.

Wang, H., Li, R. & Tsai, C.L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation Method, Biometrika, 94, 3, 553-568.

Zhang, Y., Li, R. & Tsai, C.L. (2010). Regularization Parameter Selections via Generalized Information Criterion, J. Amer. Stat. Assoc., 105. 489, 312-323.

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "degrees of freedom " of the Lasso, Ann. Statist., 35, 5, 2173-2192.