Electronic Journal of Statistics
Vol. 6 (2012) 656–663
ISSN: 1935-7524
DOI: 10.1214/12-EJS685

Further asymptotic properties of the generalized information criterion*

ChangJiang Xu† and A. Ian McLeod

Department of Statistical & Actuarial Science,
The University of Western Ontario
London, Ontario N6A 5B7, Canada
e-mail: changjiang.xu@mail.mcgill.ca; aim@stats.uwo.ca

Abstract: Asymptotic properties of the generalized information criterion for model selection are examined and new conditions under which this criterion is overfitting, consistent, or underfitting are derived.

AMS 2000 subject classifications: Primary 62J02; secondary 62J12.
Keywords and phrases: Variable selection, model selection, information criterion, consistency.

Received November 2011.

Contents

1. Introduction

By model selection we mean the choice of the best model from a set of candidate models that may have been produced by maximum likelihood estimation (MLE). The best model is usually selected using bootstrapping, cross-validation or an information criterion. The generalized information criterion (GIC) [2, 5, 7, 8] includes many well known information criteria, such as AIC [1] and BIC [6], as well as others. Asymptotic properties of GIC have been investigated [5, 7, 12]. In this article we reexamine these asymptotic properties and we give new conditions under which the GIC is overfitting, consistent, or underfitting.

*This research was supported by an NSERC Discovery Grant awarded to A. I. McLeod.
†Current address: H463, 3755 Cote-Sainte-Catherine, Montreal, Quebec, Canada, H3T 1E2

2. Generalized information criterion

Consider a family of probability distributions, $f(z;\theta)$, where $z \in \Re^{d+1}$ consisting of both response, y, and explanatory variables, $x \in \Re^d$, and $\theta \in \Theta \subset \Re^m$ is a set of parameters. Let $\mathcal{S}$ be a subset of $\{1,2,\ldots,d\}$. Each subset $\mathcal{S}$ represents a class of probability models $\{f(z;\theta) : \theta \in \Theta(\mathcal{S}) \subset \Theta\}$. Let $\{z_i, i = 1,\ldots,n\}$ be a sample of random vector, $Z \sim f(z;\theta_0)$, where θ_0 represents the true model, denoted by $\mathcal{S}_0$.

Let $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K\}$ be a set of candidate model subsets. Model selection is to choose the best model in $\mathfrak{S}$. Let $l_n(\theta) = \sum_i \log f(z_i;\theta)$ be the log-likelihood function, and $\hat{\theta}_n(\mathcal{S}_k)$ be the MLE of $\theta(\mathcal{S}_k)$. We consider the problem of model selection using the generalized information criterion,

$$\text{GIC} = -2l_n(\hat{\theta}_n(\mathcal{S}_k)) + \alpha\kappa(\mathcal{S}_k), \tag{2.1}$$

where $\alpha \geq 0$ is the tuning parameter, and $\kappa(\mathcal{S}_k)$ is the model size defined as the number of elements in $\mathcal{S}_k$. Let $\mathcal{S}_{k_n} \in \mathfrak{S}$ be the model selected by GIC. Model selection is consistent if $\Pr\{\mathcal{S}_{k_n} = \mathcal{S}_0\} \to 1$ as $n \to \infty$. The selected model is overfitted or underfitted according as the selected model size, $\kappa(\mathcal{S}_{k_n})$, is greater than or less than the true model size, $\kappa(\mathcal{S}_0)$, respectively. In this paper, probability (Pr) and expectation (E) statements are all with respect to the true underlying model.

We use the terms overfitting and underfitting as defined in the sense of efficiency by [4]. In this framework, model selection is to find the best model that is either a true model or a model closest to the true model. This definition is appropriate for multicollinear variables. For example, suppose there are three variables x_1, x_2, and x_3, and $x_3 = x_1 + x_2$. Assume the true model consists of the variable x_3. Then the model having the variables x_1 and x_2 may be considered as overfitting. If $\kappa(\mathcal{S}_{k_n}) > \kappa(\mathcal{S}_0)$ then using (2.1), we have $l_n(\hat{\theta}_n(\mathcal{S}_{k_n})) > l_n(\hat{\theta}_n(\mathcal{S}_0))$. So the overfitting implies the model has a larger likelihood value.

3. Asymptotic properties

The asymptotic properties of GIC are derived under the following conditions:

C1. The true model $\mathcal{S}_0$ is identifiable, that is, each model $\mathcal{S}$ with $\kappa(\mathcal{S}) \leq \kappa(\mathcal{S}_0)$ satisfies the condition:

$$\sup_{\theta\in\Theta(\mathcal{S}_0)} E\{\log f(Z;\theta)\} - \sup_{\theta\in\Theta(\mathcal{S})} E\{\log f(Z;\theta)\} \geq \Delta,$$

where $\Delta > 0$ is a constant.

C2. $\mathcal{S}_0 \in \mathfrak{S}$, $\mathcal{S}_0 \subset \mathcal{S}_K$, and the largest model size, $\kappa(\mathcal{S}_K)$, is fixed or bounded.

C3. $\hat{\theta}_n(\mathcal{S}_k)$ converges to a point, $\theta^*(\mathcal{S}_k)$, almost surely, and Wald consistency conditions [10] hold.

For linear models the identifiable condition C1 is the same as used by [7]. If the largest model size, $\kappa(\mathcal{S}_K)$, grows up with the sample size n, then further

assumptions are required as Wang, Li and Leng [11]. If $\mathcal{S}_0 \notin \mathfrak{S}$, the asymptotic properties may be considered for selecting the best model, which is the most parsimonious and closest to the true model.

Without loss of generality, we assume $\kappa(\mathcal{S}_1) \le \kappa(\mathcal{S}_2) \le \cdots \le \kappa(\mathcal{S}_K)$, and $\kappa(\mathcal{S}_1) < \kappa(\mathcal{S}_K)$. For notational simplicity, let $l_{n,k} = l_n(\hat{\theta}_n(\mathcal{S}_k))$ and $\kappa_k = \kappa(\mathcal{S}_k)$. Similarly, let $l_{n,0} = l_n(\hat{\theta}_n(\mathcal{S}_0))$ and $\kappa_0 = \kappa(\mathcal{S}_0)$. Define

$$\gamma = \lim_{n\to\infty} \min_{\kappa_k < \kappa_0} \frac{n^{-1}\{2l_{n,0} - 2l_{n,k}\}}{\kappa_0 - \kappa_k} \tag{3.1}$$

If $\kappa_0 = 1$, $\gamma = \infty$. It is shown in Lemma 4.1 that under condition C3,

$$\lim_{n\to\infty} n^{-1} l_{n,k} = \lim_{n\to\infty} n^{-1} l_n(\theta^*(\mathcal{S}_k)) = E\{\log f(Z;\theta^*(\mathcal{S}_k))\}.$$

Then the limit in (3.1) exists, and along with condition C1,

$$\gamma = \min_{\kappa_k < \kappa_0} \frac{2E\{\log f(Z;\theta_0) - \log f(Z;\theta^*(\mathcal{S}_k))\}}{\kappa_0 - \kappa_k} \ge 2\Delta/(\kappa_0 - 1). \tag{3.2}$$

The γ may also be defined as a limit inferior if the limit in (3.1) doesn't exist.

Let $\kappa(\alpha)$ be the model size selected by GIC with tuning parameter α. To emphasize the tuning parameter depends on the sample size n, we also use the notation α_n instead of α.

Theorem 3.1. *Let $n^{-1}\alpha_n \to r$.*

(1) If $\alpha_n < \infty$, then as $n \to \infty$,

$$\begin{aligned} \Pr\{\kappa(\alpha) = \kappa_0\} &\le \Pr\{\chi^2_{\kappa_K - \kappa_0} \le \alpha(\kappa_K - \kappa_0)\}, \\ \Pr\{\kappa(\alpha) > \kappa_0\} &= 1 - \Pr\{\kappa(\alpha) = \kappa_0\}. \end{aligned}$$

(2) If $\alpha_n \to \infty$ and $r < \gamma$, $\Pr\{\kappa(\alpha) = \kappa_0\} \to 1$ as $n \to \infty$.
(3) If $\alpha_n \to \infty$ and $r > \gamma$, $\Pr\{\kappa(\alpha) < \kappa_0\} \to 1$ as $n \to \infty$.
(4) If $\alpha_n \to \infty$ and $r = \gamma$, $\Pr\{\kappa(\alpha) \le \kappa_0\} \to 1$ as $n \to \infty$.

Corollary 3.1. *Let $n^{-1}\alpha_n \to r \ne \gamma$. Then asymptotically:*
(1) If $\alpha_n < \infty$, GIC *selects either the true model or an overfitted model.*
(2) If $\alpha_n \to \infty$ and $r < \gamma$, GIC *is consistent.*
(3) If $\alpha_n \to \infty$ and $r > \gamma$, GIC *is underfitting.*

The proof of Theorem 3.1 is given in Section 4. Corollary 3.1 is directly derived from Theorem 3.1. From Theorem 3.1 and Corollary 3.1, the GIC with α_n bounded has the same asymptotic property as AIC. The GIC with α_n unbounded and $\lim n^{-1}\alpha_n < \gamma$, has the same asymptotic property as BIC. For $\alpha_n = \log\log n$, $\log n$, or n^ν, $0 < \nu < 1$, the GIC is consistent since $n^{-1}\alpha_n \to 0$. In the existing results, the consistency of GIC was derived under the condition of $n^{-1}\alpha_n \to 0$ [7, 12]. However, from Corollary 3.1, the condition $n^{-1}\alpha_n \to 0$ is sufficient but not necessary for the consistency.

In the case $\mathcal{S}_0 = \mathcal{S}_K$, from Lemma 4.1 and condition C1, for $k < K$ we have,

$$\lim_{n\to\infty} n^{-1}[l_{n,K} - l_{n,k}] = E\{\log f(Z;\theta_0)\} - E\{\log f(Z;\theta^*(\mathcal{S}_k))\} > 0.$$

So for $\alpha < \infty$, as $n \to \infty$,

$$n^{-1}[\text{GIC }(\mathcal{S}_K) - \text{GIC }(\mathcal{S}_k)] = -2n^{-1}[l_{n,K} - l_{n,k}] + n^{-1}\alpha(\kappa_K - \kappa_k) < 0.$$

Hence the GIC with bounded α almost surely selects the true model, $\mathcal{S}_0$. Then from Theorem 3.1 and Corollary 3.1, we have the following corollary,

Corollary 3.2. *Let $n^{-1}\alpha_n \to r \neq \gamma$. If $\mathcal{S}_0 = \mathcal{S}_K$, then asymptotically*
(1) if $\alpha_n \to \infty$ and $r < \gamma$, GIC *is consistent;*
(2) if $\alpha_n \to \infty$ and $r > \gamma$, GIC *is underfitting.*

3.1. An example of linear regression

Consider a linear model $Y = X^T\beta + \varepsilon$, where $X = (X_1, \ldots, X_d)^T \in \Re^d$, and ε is assumed to be independent normal with mean zero and variance σ^2. $(\beta, \sigma) \in \Theta \subset \Re^d \times [\sigma_0, \infty)$, where $\sigma_0 > 0$. With the assumption that Θ is a closed subset and $E\{|X_jX_k|\} < \infty$, we next show that the condition C3 holds, and the γ, defined in (3.1), is closely related to the signal-to-noise ratio.

The density function

$$\log f(y;\beta|x) = -\frac{1}{2\sigma^2}(y - x^T\beta)^2 - \frac{1}{2}\log\pi\sigma^2. \tag{3.3}$$

Let $\beta_0 \in \Re^d$ with κ_0 elements of nonzero be the true parameter. From (3.3),

$$E\{|\log f(Y;\beta|X)|\} \le \frac{1}{2\sigma^2}(\beta - \beta_0)^T E\{XX^T\}(\beta - \beta_0) + \frac{1}{2}|\log\pi\sigma^2| < \infty.$$

Then Wald Assumption 6 [10] holds. Similarly, it can be directly checked from (3.3) that other Wald assumptions hold.

For independent and identically distributed samples, by strong law of large number, we have $\sum_{i=1}^n X_{ij}X_{ik}/n \to E\{X_jX_k\}$ and $\sum_{i=1}^n X_{ij}Y_i/n \to E\{X_jX^T\}\beta_0$ almost surely as $n \to \infty$. Then $\hat{\beta}(\mathcal{S}_k) \to \beta^*(\mathcal{S}_k)$ almost surely, where $\beta^*(\mathcal{S}_k) = E\{X(\mathcal{S}_k)X(\mathcal{S}_k)^T\}^{-1}E\{X(\mathcal{S}_k)X^T\}\beta_0$, assuming the inverse of matrix above exists. So the condition C3 holds. From (3.2),

$$\gamma = \min_{\kappa_k<\kappa_0} \frac{E\{|X(\mathcal{S}_0)^T\beta_0(\mathcal{S}_0) - X(\mathcal{S}_k)^T\beta^*(\mathcal{S}_k)|^2\}}{(\kappa_0 - \kappa_k)\sigma^2},$$

and then γ may be viewed as an average signal-to-noise ratio.

4. Proofs

Before giving the proof of Theorem 3.1, we introduce some lemmas.

Lemma 4.1. *Let $l_n(\theta) = \sum_i \log f(z_i;\theta)$ be the log-likelihood function. Let $\hat{\theta}_n$ be the* MLE *of θ. Assume $\hat{\theta}_n \to \theta$ almost surely. Under Wald conditions, we have*

$$\lim_{n\to\infty} n^{-1} l_n(\hat{\theta}_n) = \lim_{n\to\infty} n^{-1} l_n(\theta) = E\{\log f(Z;\theta)\}.$$

Proof of Lemma 4.1. Define

$$f(z;\theta,\rho) = \sup_{\|\theta'-\theta\|\le\rho} f(z;\theta'),$$

and $l_n(\theta,\rho) = \sum_i \log f(z_i;\theta,\rho)$. Under Wald conditions, we have $E\{|\log f(Z;\theta)|\} < \infty$ and $E\{|\log f(Z;\theta,\rho)|\} < \infty$. By the strong law of large numbers,

$$\lim_{n\to\infty} n^{-1} l_n(\theta) = E\{\log f(Z;\theta)\},$$

$$\lim_{n\to\infty} n^{-1} l_n(\theta,\rho) = E\{\log f(Z;\theta,\rho)\}.$$

Since $l_n(\hat{\theta}_n) \ge l_n(\theta)$,

$$\liminf_{n\to\infty} n^{-1} l_n(\hat{\theta}_n) \ge \lim_{n\to\infty} n^{-1} l_n(\theta) = E\{\log f(Z;\theta)\}. \tag{4.1}$$

Since $\hat{\theta}_n \to \theta$ almost surely, there exists N_ρ such that $\| \hat{\theta}_n - \theta \| \le \rho$ almost surely for $n > N_\rho$. Then for $n > N_\rho$, $\log f(z;\hat{\theta}_n) \le \log f(z;\theta,\rho)$ and $l_n(\hat{\theta}_n) \le l_n(\theta,\rho)$ almost surely. Hence

$$\limsup_{n\to\infty} n^{-1} l_n(\hat{\theta}_n) \le \lim_{n\to\infty} n^{-1} l_n(\theta,\rho) = E\{\log f(Z;\theta,\rho)\}. \tag{4.2}$$

From (4.1) and (4.2),

$$E\{\log f(Z;\theta)\} \le \liminf_{n\to\infty} n^{-1} l_n(\hat{\theta}_n) \le \limsup_{n\to\infty} n^{-1} l_n(\hat{\theta}_n) \le E\{\log f(Z;\theta,\rho)\}.$$

From Lemma 2 in Wald [10], $\lim_{\rho\to 0} E\{\log f(Z;\theta,\rho)\} = E\{\log f(Z;\theta)\}$. So

$$\lim_{n\to\infty} n^{-1} l_n(\hat{\theta}_n) = \lim_{\rho\to 0} E\{\log f(Z;\theta,\rho)\} = E\{\log f(Z;\theta)\}.$$

□

Lemma 4.2. GIC *can select the model $\mathcal{S}_k$ if and only if $l_{n,k} = \max_{\{j,\kappa_j=\kappa_k\}} l_{n,j}$ and $A_{k,1} \le \alpha \le A_{k,2}$, where*

$$A_{k,1} = \max_{\kappa_j>\kappa_k} 2\{l_{n,j} - l_{n,k}\}/(\kappa_j - \kappa_k),$$

$$A_{k,2} = \min_{\kappa_j<\kappa_k} 2\{l_{n,j} - l_{n,k}\}/(\kappa_j - \kappa_k).$$

Here we define $A_{K,1} = 0$ and $A_{1,2} = \infty$.

Proof of Lemma 4.2. GIC selects model $\mathcal{S}_k$ if and only if $-2\{l_{n,j}-l_{n,k}\}+\alpha(\kappa_j-\kappa_k)\geq 0$, that is,

$$\begin{array}{ll} l_{n,j}\leq l_{n,k}, & \kappa_j=\kappa_k;\\ \alpha\leq 2\{l_{n,j}-l_{n,k}\}/(\kappa_j-\kappa_k), & \kappa_j<\kappa_k,\\ \alpha\geq 2\{l_{n,j}-l_{n,k}\}/(\kappa_j-\kappa_k), & \kappa_j>\kappa_k, \end{array}$$

or $A_{k,1}\leq\alpha\leq A_{k,2}$. □

Lemma 4.3. *Asymptotically*

$$\begin{array}{rcl} \Pr\{\kappa(\alpha)=\kappa_0\} & = & \Pr\{A_{0,1}\leq\alpha\leq A_{0,2}\},\\ \Pr\{\kappa(\alpha)>\kappa_0\} & = & \Pr\{A_{0,1}\geq\alpha\},\\ \Pr\{\kappa(\alpha)<\kappa_0\} & = & \Pr\{A_{0,2}\leq\alpha\}. \end{array}$$

Proof of Lemma 4.3. From C1, for $\kappa_k\leq\kappa_0$, $n^{-1}(l_{n,0}-l_{n,k})\geq\Delta$ as n is large enough, and then $l_{n,0}=\max_{\{j,\kappa_j=\kappa_0\}}l_{n,j}$. Therefore, from Lemma 4.2, as n is large enough, $\kappa(\alpha)=\kappa_0$ if and only if $A_{0,1}\leq\alpha\leq A_{0,2}$. That is,

$$\begin{array}{rcl} \Pr\{\kappa(\alpha)=\kappa_0\} & = & \Pr\{l_{n,0}=\max_{\{j,\kappa_j=\kappa_0\}}l_{n,j}, A_{0,1}\leq\alpha\leq A_{0,2}\}\\ & = & \Pr\{A_{0,1}\leq\alpha\leq A_{0,2}\}. \end{array}$$

From Lemma 4.1,

$$\lim n^{-1}l_{n,K}=E\{\log f(Z;\theta^*(\mathcal{S}_K))\}\leq E\{\log f(Z;\theta_0)\}.$$

On the other hand, from the condition C2, $l_{n,K}\geq l_{n,0}$, and then

$$\lim n^{-1}l_{n,K}\geq\lim n^{-1}l_{n,0}=E\{\log f(Z;\theta_0)\}.$$

So $\lim n^{-1}l_{n,K}=E\{\log f(Z;\theta_0)\}$, and $n^{-1}(l_{n,K}-l_{n,0})\to 0$. Hence

$$n^{-1}A_{0,1}\to\ 0. \tag{4.3}$$

From C1, we also have

$$n^{-1}A_{0,2}\geq\Delta\min_{\kappa_j<\kappa_0}\{2/(\kappa_0-\kappa_j)\}. \tag{4.4}$$

From (4.3) and (4.4), $A_{0,1}\leq A_{0,2}$ as n is large enough.

If the selected model size, $\kappa(\alpha)=\kappa_k$, is greater than the true model size, κ_0, then $l_{n,k}>l_{n,0}$, and from the definition, we have

$$A_{k,2}\leq 2\{l_{n,0}-l_{n,k}\}/(\kappa_0-\kappa_k)\leq A_{0,1}.$$

Hence, from Lemma 4.2, if $\kappa(\alpha)>\kappa_0$, then $\alpha\leq A_{0,1}$.

Similarly, if $\kappa(\alpha)<\kappa_0$, we have $\alpha\geq A_{0,2}$. Note that if $\kappa(\alpha)=\kappa_0$, then $A_{0,1}\leq\alpha\leq A_{0,2}$. Hence, Lemma 4.3 follows. □

4.1. Proof of Theorem

Proof of Theorem 3.1 (1). Since α is bounded, from Lemma 4.3 and (4.4), asymptotically $\Pr\{\kappa(\alpha) < \kappa_0\} = \Pr\{A_{0,2} \le \alpha\} = 0$. So

$$\Pr\{\kappa(\alpha) = \kappa_0\} = 1 - \Pr\{\kappa(\alpha) > \kappa_0\}.$$

From Lemma 4.3,

$$\begin{aligned}\Pr\{\kappa(\alpha) = \kappa_0\} &= \Pr\{A_{0,1} \le \alpha \le A_{0,2}\}\\ &\le \Pr\{A_{0,1} \le \alpha\}\\ &\le \Pr\{2\{l_{n,K} - l_{n,0}\}/(\kappa_K - \kappa_0) \le \alpha\}\\ &= \Pr\{\chi^2_{\kappa_K - \kappa_0} \le \alpha(\kappa_K - \kappa_0)\}.\end{aligned}$$

□

Proof of Theorem 3.1 (2). Since $r < \gamma$, $n^{-1}\alpha_n < n^{-1}A_{0,2}$ as n is larger. From Lemma 4.3, and $l_{n,K} \ge l_{n,k}$, as n is large enough, we have

$$\begin{aligned}\Pr\{\kappa(\alpha) = \kappa_0\} &= \Pr\{A_{0,1} \le \alpha_n \le A_{0,2}\}\\ &= \Pr\{A_{0,1} \le \alpha_n, \alpha_n/n \le A_{0,2}/n\}\\ &= \Pr\{A_{0,1} \le \alpha_n\}\\ &\ge \Pr\{\max_{\kappa_j > \kappa_0} 2\{l_{n,K} - l_{n,0}\}/(\kappa_j - \kappa_0) \le \alpha_n\}\\ &\ge \Pr\{2\{l_{n,K} - l_{n,0}\} \le \alpha_n\}\\ &= \Pr\{\chi^2_{\kappa_K - \kappa_0} \le \alpha_n\} \to 1.\end{aligned}$$

□

Proof of Theorem 3.1 (3). Since $r > \gamma$, $n^{-1}\alpha_n > n^{-1}A_{0,2}$ as n is larger. Then, from Lemma 4.3, $\Pr\{\kappa(\alpha) < \kappa_0\} = \Pr\{A_{0,2} \le \alpha_n\} = 1$. □

Proof of Theorem 3.1 (4). Since $n^{-1}\alpha_n \to r = \gamma > 0$, from (4.3), $n^{-1}A_{0,1} < n^{-1}\alpha_n$ for n large enough. So, from Lemma 4.3, $\Pr\{\kappa(\alpha) > \kappa_0\} = \Pr\{A_{0,1} \ge \alpha_n\} = 0$. □

5. Conclusions

We have reexamined the asymptotic properties of GIC for model selection and have derived conditions on the tuning parameter, under which the GIC is asymptotically overfitting, consistent or underfitting. These new results elucidate the performance of the GIC. The proofs of the asymptotic properties employed the strong law of large numbers. If the weak law of large number is used, then the type of convergence is in probability, and the asymptotic properties are the weak rather than strong properties.

Recently, Zhang, Li and Tsai [14] considered the asymptotic properties of GIC in the case that the candidate models are produced by non-concave penalized likelihood methods, such as least absolute shrinkage and selection operator (LASSO) [9], smoothly clipped absolute deviation (SCAD) [3] and minimax concave penalty (MCP) [13]. It would be interesting to extend our derived asymptotic properties to this case where the largest model size could also increase with the sample size.

Acknowledgement

The authors would like to thank two referees and an Associate Editor for their insightful comments and suggestions.

References

[1] Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19** 716–723. MR0423716
[2] Akaike, H. (1979). A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. *Biometrika* **66** 237–242. MR0548189
[3] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581
[4] McQuarrie, A. D. R. and Tsai, C. L. (1998). *Regression and Time Series Model Selection*. World Scientific Publishing Company, Singapore. MR1641582
[5] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics* **12** 758–765. MR0740928
[6] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics* **6** 461–464. MR0468014
[7] Shao, J. (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica* **7** 221–262. MR1466682
[8] Shibata, R. (1984). Approximate Efficiency of a Selection Procedure for the Number of Regression Variables. *Biometrika* **71** 43-49. MR0738324
[9] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288. MR1379242
[10] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **20** 595–601. MR0032169
[11] Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B* **71** 671–683. MR2749913
[12] Yang, Y. (2005). Can the Strengths of AIC and BIC be Shared? *Biometrika* **92** 937–950. MR2234196
[13] Zhang, C.-H. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *The Annals of Statistics* **38** 894–942. MR2604701
[14] Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105** 312–323. MR2656055