应用概率统计 第三十一卷
第四期 2015年8月

Chinese Journal of Applied Probability
and Statistics Vol.31 No.4 Aug. 2015

General Shrinkage Rule with Respect to γ-Norms for Bridge Estimation *

GAI YUJIE ZHANG ZHANLI

(*School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 100081*)

ZHANG JUN*

(*Shen Zhen – Hong Kong Joint Research Centre for Applied Statistical Sciences, College of Mathematics and Computational Science, Institute of Statistical Sciences, Shenzhen University, Shenzhen, 518060*)

《应用概率统计》版权所有

Abstract

Bridge regression, a special family of penalized regressions of a penalty function $\sum|\beta_j|^\gamma$ with $\gamma > 0$, has been studied in many literatures. In this paper, we provide some theoretical results of how the shrinkage rule changing with γ under two settings: $\gamma \geq 1$ and $0 < \gamma < 1$, respectively. Simulation results are conducted to evaluate the performance of the proposed method.

Keywords: Bridge estimation, high-dimensional linear model, variable selection, shrinkage rule.

AMS Subject Classification: 62F12, 62G05, 62G08.

§1. Introduction

Consider the linear model

$$Y_i = x_i'\boldsymbol{\beta}_0 + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{1.1}$$

where $Y_i \in \mathbb{R}^1$ is a response variable, $x_i \in \mathbb{R}^p$ is a $p \times 1$ covariate vector, and the ε_i's are i.i.d. model errors. Without loss of generality, we assume that the response Y_i and covariates x_i are centered to have mean zero. This results in a vanished intercept term,

*Gai Yujie's research was supported by the National Natural Science Foundation of China (11201499, 11301309, 71473280). Zhang Zhanli's research was supported by the National Natural Science Foundation of China (11401602). Zhang Jun's research was supported by the National Natural Science Foundation of China (11401391, 11101157), Tianyuan Youth Fund for Mathematics of the National Natural Science Foundation of China (11326179) and Natural Science Foundation of Shenzhen University (801, 00036112).

*Corresponding author, E-mail: zhangjunstat@gmail.com.

Received December 20, 2012. Revised January 29, 2015.

doi: 10.3969/j.issn.1001-4268.2015.04.002

《应用概率统计》版权所有

and we can concentrate on the estimation for the unknown coefficient $\boldsymbol{\beta}_0$. It is known that variable selection is a crucial step in high-dimensional modelling and various powerful penalization methods have been developed for variable selection in parametric models. The usual method for estimating $\boldsymbol{\beta}_0$ is proposed by minimizing the penalized least squares objective function

$$L_n(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - x_i'\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{p}|\beta_j|^\gamma, \tag{1.2}$$

where λ_n is a penalty parameter. For any given $\gamma > 0$, the minimizer $\widehat{\boldsymbol{\beta}}_n$ of (1.2) is called a bridge estimator by Frank and Friedman (1993), who introduced it as a generalization of ridge regression (which occurs for $\gamma = 2$). The special case for $\gamma = 1$ is related to the LASSO estimator proposed by Tibshirani (1996), and the LASSO method is well-known as a variable selection and shrinkage method. For $0 < \gamma \leq 1$, some components of the minimizer $\widehat{\boldsymbol{\beta}}_n$ of (1.2) can be exactly zero if λ_n is sufficiently large (Knight and Fu, 2000). In addition to bridge estimators, other penalization methods have been proposed for the purpose of simultaneous variable selection and shrinkage estimation. Examples include the SCAD penalty (Fan, 1997; Fan and Li, 2001) and the elastic-net (Enet) penalty (Zou and Hastie, 2005). For the SCAD penalty, Fan and Li (2001) studied asymptotic properties of penalized likelihood methods when the number of parameters is finite. Fan and Peng (2004) considered the same problem when the number of parameters diverges and obtain an "oracle" estimator. Here the oracle property means that the resulting estimators perform asymptotically as efficient as if the true model were known.

Recently, there have been several studies of large sample properties of high-dimensional problems. Fan and Li (2006) provided a review of statistical challenges in high-dimensional problems that arise in many important applications. Huang et al. (2008) studied the asymptotic properties of bridge estimators with $0 < \gamma < 1$ when the number of covariates p may increase to infinity with sample size n. However, all of them as we indicated above focus on the relationship between the consistency of model selection (or parameters) and the tuning parameter λ_n, and also the rule of λ_n. It should be noted that the rule of λ_n usually depends on γ, which also plays a decisive role in the consistency of model selection (or parameters). For any fixed sample size n and γ, the larger value of λ_n is, the smaller number of non-zeros of $\widehat{\boldsymbol{\beta}}_n$ is, and also the smaller the sub-model is. Tibshirani (1996) and Fu (1998) studied how the size of γ makes effect to the estimation shrinkage rule when $\gamma \geq 1$, without some rigorous theoretical results. In this paper, we establish in theory that how the estimation of parameters changes with γ in the high-dimensional sparse models. Once some regularity allows, it will provide a criterion for users to choose a proper γ in application.

The rest of this paper is organized as follows. In Section 2, we obtain the shrinkage regular pattern for the bridge estimation of coefficients $\boldsymbol{\beta}_0$, under the variety of γ for $\gamma \geq 1$ and $0 < \gamma < 1$ respectively. In Section 3, we report the results of a simulation study. All the technical proofs of the asymptotic results are given in the Appendix.

《应用概率统计》版权所有

§2. The Shrinkage Rule of the Components

Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip_n})'$ be the $p_n \times 1$ vector of ith observation of covariates, $i = 1, 2, \ldots, n$. Here the subscript p_n means that the dimension of covariates depends on the sample size n. Assume that the covariates $\boldsymbol{x}_i$ are a fixed design, the response Y_i's are centered and the each component x_{ij} are also standardized, i.e.,

$$\sum_{i=1}^{n} Y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \qquad j = 1, 2, \ldots, p_n.$$

For notation simplicity, we write the true parameter $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}', \boldsymbol{\beta}_{20}')'$, where $\boldsymbol{\beta}_{10} \in \mathbb{R}^{q_n}$, $\boldsymbol{\beta}_{20} \in \mathbb{R}^{m_n}$, $q_n + m_n = p_n$. Considering the sparsity, we suppose that $\boldsymbol{\beta}_{10} \neq \mathbf{0}$ and $\boldsymbol{\beta}_{20} = \mathbf{0}$, where $\mathbf{0}$ is the zero vector of size m_n. Let $\boldsymbol{x}_i = (\boldsymbol{z}_i', \boldsymbol{u}_i')'$. $\boldsymbol{z}_i$ consists of the first q_n covariates corresponding to the nonzero coefficients and $\boldsymbol{u}_i$ consists of the remaining m_n covariates corresponding to the zero coefficients. Moreover, let $\boldsymbol{X}_n = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)'$, $\boldsymbol{X}_{1n} = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)'$, $\boldsymbol{X}_{2n} = (\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n)'$ respectively. Denote

$$\Sigma_n = n^{-1}\boldsymbol{X}_n'\boldsymbol{X}_n \qquad \text{and} \qquad \Sigma_{1n} = n^{-1}\boldsymbol{X}_{1n}'\boldsymbol{X}_{1n}.$$

Without loss of generality, we further assume that the design matrix of the active variables is orthogonal, that is, $\Sigma_{1n} = I$, where I denotes the $q_n \times q_n$ identity matrix. This assumption is mild as q_n is much smaller than sample size n under the sparsity condition. More conveniently, it allows us to study the characteristics of the shrinkage effect under different choices of γ.

The following conditions are needed for the consistency of bridge estimator.

(C1) ε_i, $i = 1, 2, \ldots, n$ are independent and identically distributed random variables with mean zero and variance σ^2, $0 < \sigma^2 < \infty$;

(C2) There exist constants $0 < b_1 < b_2 < \infty$ such that

$$b_1 \leq \min\{|\beta_{10j}|, 1 \leq j \leq q_n\} \leq \max\{|\beta_{10j}|, 1 \leq j \leq q_n\} \leq b_2;$$

(C3) $\rho_{1n} > 0$ for all n;

(C4) $\lambda_n(q_n/n)^{1/2} \to 0$;

(C5) $\lambda_n n^{-\gamma/2}(\rho_{1n}/\sqrt{p_n})^{2-\gamma} \to \infty$.

In the following, we first present Lemma 2.1 obtained by Huang et al. (2008) for the bridge estimators.

Lemma 2.1 Let $\widehat{\boldsymbol{\beta}}_n$ denote the minimizer of (1.2). Under conditions (C1)-(C4), for any $\gamma > 0$, we have $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_p(h_n)$, where $h_n = \min\{h_{1n}, h_{2n}\}$, $h_{1n} = \rho^{-1}(p_n/n)^{1/2}$ and $h_{2n} = [(p_n + \lambda_n q_n)/(n\rho_{1n})]^{1/2}$.

If $\rho_{1n} > \rho_1 > 0$ for all n, Lemma 2.1 yields the rate of convergence $O_p(h_{2n}) = O_p((p_n/n)^{1/2})$ under condition (C4). Moreover, if p_n is further assumed to be finite for all n, then the rate of convergence can be $n^{-1/2}$. However, if $\rho_{1n} \to 0$, the rate of convergence will be slower than $n^{-1/2}$.

The following Lemma 2.2 provides an explicit relationship between the average of bridge estimation and γ.

Lemma 2.2 Let $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{1n}', \widehat{\boldsymbol{\beta}}_{2n}')'$, where $\widehat{\boldsymbol{\beta}}_{1n}$ and $\widehat{\boldsymbol{\beta}}_{2n}$ are estimators of $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\beta}_{20}$, respectively. Under conditions (C1)-(C4), for any $\gamma > 0$ and sufficiently large n, we can have

$$\mathsf{E}(\widehat{\beta}_{1nj}) = \Delta(\gamma) + \frac{1}{2}\frac{\lambda_n}{n}\gamma \mathsf{E} o_p(1) + C_n, \tag{2.1}$$

where

$$\Delta(\gamma) = \beta_{10j} - \frac{1}{2}\frac{\lambda_n}{n}\gamma|\beta_{10j}|^{\gamma-1}\mathrm{sgn}(\beta_{10j}),$$

$\widehat{\beta}_{1nj}$ is the jth component of $\widehat{\boldsymbol{\beta}}_{1n}$, β_{10j} is the jth component of $\boldsymbol{\beta}_{10}$, $j = 1, 2, \ldots, q_n$ and C_n is an expression not relative to γ. Here $\mathrm{sgn}(\cdot)$ stands for the sign function.

2.1 Shrinkage Effect of Bridge Estimation when $\gamma \geq 1$

Firstly considering the case of $\gamma \geq 1$, we aim to find the shrinkage regular pattern for the bridge estimator of $\boldsymbol{\beta}_0$ under the variety of γ. We have the following theorem.

Theorem 2.1 Under conditions (C1)-(C5),

(1) when $0 < \beta_{10j} \leq \mathrm{e}^{-1}$, $\Delta(\gamma)$ is monotonically increasing for $\gamma \in [1, +\infty)$.

(2) when $-\mathrm{e}^{-1} \leq \beta_{10j} < 0$, $\Delta(\gamma)$ is monotonically descending for $\gamma \in [1, +\infty)$.

(3) when $\beta_{10j} \geq 1$, $\Delta(\gamma)$ is monotonically descending for $\gamma \in [1, +\infty)$.

(4) when $\beta_{10j} \leq -1$, $\Delta(\gamma)$ is monotonically increasing for $\gamma \in [1, +\infty)$.

Theorem 2.1 plays an important role in searching the shrinkage rule of γ. Although the second term in (2.1) contains γ, when n is sufficiently large, it will be very small and its effect on $\mathsf{E}(\widehat{\beta}_{1nj})$ could be neglected comparing to $\Delta(\gamma)$. As a result, the monotonicity

《应用概率统计》版权所有

of $\mathsf{E}(\widehat{\beta}_{1nj})$ on γ can be treated the same as that of $\Delta(\gamma)$ asymptotically. Also, Theorem 2.1 tells us that, when $0 < |\beta_{10j}| \le \mathrm{e}^{-1}$, the larger value of γ is, the large value of the bridge estimator $|\widehat{\beta}_{1nj}|$ is; on the contrary, when $|\beta_{10j}| \ge 1$, the larger value of γ is, the smaller value of the bridge estimator $|\widehat{\beta}_{1nj}|$ is, $j = 1, 2, \ldots, q_n$.

Remark 1 Theorem 2.1 does not provide any conclusions when $\mathrm{e}^{-1} < |\beta_{10j}| < 1$, because the shrinkage rule is no longer monotonous in γ in this context. However, what we care most is the case $0 < |\beta_{10j}| \le \mathrm{e}^{-1}$, especially when $|\beta_{10j}|$ is relative small. In such cases, the bridge regression could retain variables with large size of coefficients for $\gamma \ge 1$. Theorem 2.1 shows that larger value of γ tends to retain smaller parameters of $\boldsymbol{\beta}_0$, while smaller value of γ tends to shrink smaller parameters of $\boldsymbol{\beta}_0$ into zero.

Therefore, it implies that if the true model includes many small but nonzero regression parameters, the LASSO will perform poorly but the bridge estimation for large value of γ will perform well. If the true model includes many zero parameters, the LASSO will perform well but the bridge estimation for large value of γ will perform poorly. This phenomenon was also revealed by Tibshirani (1996) and Fu (1998) through simulation studies, without providing some rigorous theoretical results. Generally, if one wants to retain small parameters, a relative large γ can be used. To see the results of Theorem 2.1 more clearly, we plot $\mathsf{E}(\widehat{\beta}_{1nj})$ against $|\beta_{10j}|$ in Figure 1, whose absolute values are plotted in the diagonal. The figures illustrate the same results as the theoretical findings given above.

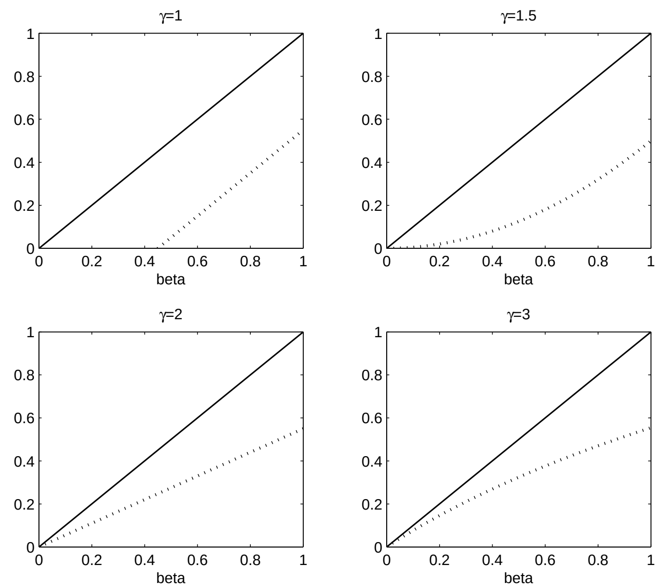

Figure 1 Shrinkage effect of bridge regressions for fixed $\lambda_n > 0$

《应用概率统计》版权所有

2.2 Shrinkage Effect of Bridge Estimation when $0 < \gamma < 1$

For $0 < \gamma < 1$, we also obtain the shrinkage regular pattern for the bridge estimation of $\boldsymbol{\beta}_0$ under the variety of γ. We have the following theorem.

Theorem 2.2 For $0 < \gamma < 1$, when $|\beta_{10j}| \geq \mathrm{e}^{-1}$, $\Delta(\gamma)$ is monotonically descending for $\gamma \in (0, 1)$.

Similar to the analysis of Theorem 2.1, Theorem 2.2 tells us that for sufficiently large n and $|\beta_{10j}| \geq \mathrm{e}^{-1}$, the bridge estimator $|\widehat{\beta}_{1nj}|$ shrinks towards $|\beta_{10j}|$ much as γ towards to 1.

Remark 2 Theorem 2.2 does not provide any conclusions for $0 < |\beta_{10j}| < \mathrm{e}^{-1}$, the shrinkage rule in this context is no longer monotonous in γ. For example, if we let $\gamma = -1/\log(|\beta_{10j}|)$ when $0 < |\beta_{10j}| < \mathrm{e}^{-1}$, the shrinkage efficiency will be the largest since γ moves towards to the endpoint 1. In other words, if the true model includes many small but nonzero regression coefficients, we tend to choose γ near to 1 for keeping small coefficients.

For $0 < \gamma < 1$, we also plot $\mathsf{E}(\widehat{\beta}_{1nj})$ against $|\beta_{10j}|$ in Figure 2. All the parameters are designed the same as those used in Figure 1. From Figure 2, we can see clearly that the shrinkage rule coincides with the conclusion obtained from Theorem 2.2.

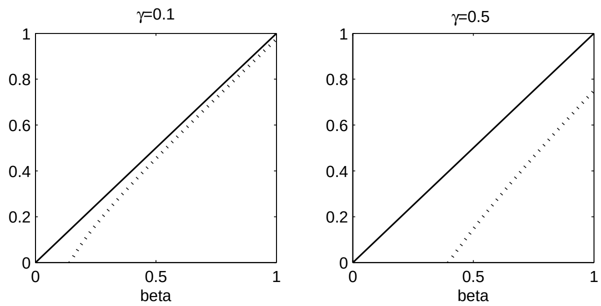

Figure 2 Shrinkage effect of bridge regressions for fixed $\lambda_n > 0$

§3. A Simulation Study

In this section, a simulation study is conducted to examine the performance of bridge estimation under different choices of γ for the model selection. The sample size $n = 200$ and $p = 10$ and 400 are used in the example.

Consider the following linear model

$$y = X'\beta + \varepsilon, \tag{3.1}$$

《应用概率统计》版权所有

《应用概率统计》版权所有

where $\varepsilon \sim \mathrm{N}(0,1)$, the coefficients

$$\beta = (3.0, 3.0, 3.0, 5.0, 5.0, 5.0, 0.2, 0.15, 0.1, 0.08)' \quad \text{for} \quad p = 10,$$

and

$$\beta = (3.0, 3.0, 3.0, 5.0, 5.0, 5.0, 0.5, 0.5, 0.1, 0.08, 0, \dots, 0)' \quad \text{for} \quad p = 400.$$

The covariate vector $X = (x_1, x_2, \dots, x_p)'$ is generated from a multivariate normal distribution with the pairwise correlation between the i^{th} and j^{th} elements as $\rho^{|i-j|}$, $i, j = 1, 2, \dots, p$. $\rho = 0$ denotes all elements of X are independent of each other.

Some tables are provided to investigate the performance of different bridge estimators through γ, in terms of model selection accuracy and coefficients estimation accuracy under different estimation methods. Numerical results are shown in Table 1 and Table 2, giving the median number of predictors being included in the model, denoted as S, and the median of the estimators for the nonzero coefficients, denoted as $\widehat{\beta}_i$'s.

Table 1　Results of coefficient estimations via four variable selection methods when $n > p$

	γ	S	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	$\widehat{\beta}_9$	$\widehat{\beta}_{10}$
$p=10$,	0.9	8	2.9450	2.9527	2.9478	4.9542	4.9498	4.9490	0.1338	0.0814	0.0297	0.0000
$\rho=0$	1	10	2.9531	2.9571	2.9528	4.9503	4.9549	4.9539	0.1541	0.1097	0.0663	0.0395
	2	10	2.9167	2.9311	2.9203	4.8901	4.8944	4.8870	0.1775	0.1332	0.0858	0.0623
	0.1	8	2.9857	2.9950	3.0001	4.9969	4.9944	4.9955	0.1398	0.0813	0.0000	0.0000
$p=10$,	0.9	9	2.9442	2.9525	2.9581	4.9531	4.9543	4.9535	0.1283	0.0864	0.0162	0.0103
$\rho=0.5$	1	10	2.9482	2.9525	2.9577	4.9532	4.9533	4.9539	0.1500	0.1087	0.0655	0.0425
	2	10	1.0486	1.0313	1.0352	1.6804	1.7273	1.7141	0.0525	0.0537	0.0328	0.0179
true	$\boldsymbol{\beta}_{10}$	–	3.00	3.00	3.00	5.00	5.00	5.00	0.20	0.15	0.10	0.08

Table 2　Results of coefficient estimations via four variable selection methods when $n < p$

	γ	S	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	$\widehat{\beta}_9$	$\widehat{\beta}_{10}$
	0.1	6	2.9704	2.9654	2.9547	4.9733	4.9733	4.9804	0.0000	0.0000	0.0000	0.0000
$p=400$,	0.9	8	2.7837	2.7754	2.7704	4.7819	4.7810	4.7910	0.2016	0.2034	0.0000	0.0000
$\rho=0$	1	9	2.9099	2.8980	2.8960	4.8950	4.8975	4.9034	0.3986	0.3977	0.0000	0.0000
	2	10	0.8161	0.8074	0.8202	1.3694	1.3678	1.3732	0.1165	0.1244	0.0203	0.0214
	0.1	6	2.9594	2.9501	2.9558	4.9770	4.9795	4.9702	0.0000	0.0000	0.0000	0.0000
$p=400$,	0.9	8	2.7729	2.7705	2.7676	4.7875	4.7933	4.7851	0.1952	0.2048	0.0000	0.0000
$\rho=0.5$	1	9	2.9119	2.9111	2.9073	4.9033	4.9093	4.9061	0.4068	0.4034	0.0056	0.0000
	2	10	0.6428	0.6258	0.6332	1.0454	1.0488	1.0626	0.0807	0.0790	0.0125	0.0241
true	$\boldsymbol{\beta}_{10}$	–	3.00	3.00	3.00	5.00	5.00	5.00	0.50	0.50	0.10	0.08

From the numerical results in Tables 1-2, it is clearly that those larger coefficients are shrunk largely as γ grows, for those small coefficients, samller γ $(0<\gamma<1)$ will shrink smaller coefficient largely, whenever $p\leq n$ or $p>n$. This result coincides with the theoretical analysis in Theorem 2.1 and Theorem 2.2. Moreover, in Table 2, we can see that ridge estimator with $\gamma=2$ collapses in the context of $p>n$. It implies that $\gamma\leq 1$ can be used to deal with the "$p\gg n$" problems in practice.

Appendix

Proof of Lemma 2.2 Under conditions (C1) and (C3), $\widehat{\boldsymbol{\beta}}_n$ is consistent by Lemma 2.1. By condition (C2), each component of $\widehat{\boldsymbol{\beta}}_{1n}$ stays away from zero for n sufficiently large. Thus, when n is large enough, the derivative $(\partial/\partial\boldsymbol{\beta}_1)L_n(\widehat{\boldsymbol{\beta}}_{1n},\widehat{\boldsymbol{\beta}}_{2n})$ exists. That is

$$-2\sum_{i=1}^{n}(Y_i-\boldsymbol{z}_i'\widehat{\boldsymbol{\beta}}_{1n}-\boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}_{2n})\boldsymbol{z}_i+\lambda_n\gamma|\widehat{\boldsymbol{\beta}}_{1n}|^{\gamma-1}\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_{1n})=0.$$

Put $\boldsymbol{\beta}_{20}=0$ and $\varepsilon_i=Y_i-\boldsymbol{z}_i'\boldsymbol{\beta}_{10}$ into above equation, we can get

$$-2\sum_{i=1}^{n}(\varepsilon_i-\boldsymbol{z}_i'(\widehat{\boldsymbol{\beta}}_{1n}-\boldsymbol{\beta}_{10})-\boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}_{2n})\boldsymbol{z}_i+\lambda_n\gamma|\widehat{\boldsymbol{\beta}}_{1n}|^{\gamma-1}\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_{1n})=0.$$

Therefore,

$$\Sigma_{1n}(\widehat{\boldsymbol{\beta}}_{1n}-\boldsymbol{\beta}_{10})=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{z}_i-\frac{1}{2n}\gamma\lambda_n|\widehat{\boldsymbol{\beta}}_{1n}|^{\gamma-1}\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_{1n})-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}_{2n}\boldsymbol{z}_i.$$

It follows that,

$$\begin{aligned}e_{nj}'(\widehat{\boldsymbol{\beta}}_{1n}-\boldsymbol{\beta}_{10})=&\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i e_{nj}\Sigma_{1n}^{-1}\boldsymbol{z}_i-\frac{1}{2}\frac{\lambda_n}{n}\gamma e_{nj}'\Sigma_{1n}^{-1}|\widehat{\boldsymbol{\beta}}_{1n}|^{\gamma-1}\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_{1n})\\&-\frac{1}{n}\sum_{i=1}^{n}e_{nj}\Sigma_{1n}^{-1}\boldsymbol{z}_i\boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}_{2n},\end{aligned}\tag{4.1}$$

where e_{nj} is a $q_n\times 1$ vector whose other components are all zeros except the jth one is 1, $j=1,2,\ldots,q_n$. Note that $\Sigma_{1n}=I_{q_n\times q_n}$, then the equation (4.1) is simplified to

$$\begin{aligned}\widehat{\beta}_{1nj}-\beta_{10j}=&\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i e_{nj}\boldsymbol{z}_i-\frac{1}{2}\frac{\lambda_n}{n}\gamma|\widehat{\beta}_{1nj}|^{\gamma-1}\mathrm{sgn}(\widehat{\beta}_{1nj})\\&-\frac{1}{n}\sum_{i=1}^{n}e_{nj}\Sigma_{1n}^{-1}\boldsymbol{z}_i\boldsymbol{u}_i'\widehat{\boldsymbol{\beta}}_{2n}.\end{aligned}\tag{4.2}$$

Taking expectation on both sides of (4.2), we have

$$\mathsf{E}(\widehat{\beta}_{1nj})=\beta_{10j}-\frac{1}{2}\frac{\lambda_n}{n}\gamma\mathsf{E}[|\widehat{\beta}_{1nj}|^{\gamma-1}\mathrm{sgn}(\widehat{\beta}_{1nj})]-\frac{1}{n}\sum_{i=1}^{n}e_{nj}\Sigma_{1n}^{-1}\mathsf{E}(\boldsymbol{z}_i\boldsymbol{u}_i')\widehat{\boldsymbol{\beta}}_{2n}.\tag{4.3}$$

《应用概率统计》版权所有

《应用概率统计》版权所有

By Lemma 2.1, we know that

$$\widehat{\beta}_{1nj} \xrightarrow{\mathsf{P}} \beta_{10j},$$

thus

$$|\widehat{\beta}_{1nj}|^{\gamma-1}\text{sgn}(\widehat{\beta}_{1nj}) \xrightarrow{\mathsf{P}} |\beta_{10j}|^{\gamma-1}\text{sgn}(\beta_{10j}).$$

It follows that

$$\mathsf{E}[|\widehat{\beta}_{1nj}|^{\gamma-1}\text{sgn}(\widehat{\beta}_{1nj})] = |\beta_{10j}|^{\gamma-1}\text{sgn}(\beta_{10j}) + \mathsf{E}o_p(1). \tag{4.4}$$

Plugging (4.4) into (4.3), and we denote

$$C_n = \frac{1}{n}\sum_{i=1}^{n} e_{nj}\Sigma_{1n}^{-1}\mathsf{E}(\boldsymbol{z}_i\boldsymbol{u}_i')\widehat{\boldsymbol{\beta}}_{2n},$$

then we obtain that

$$\mathsf{E}(\widehat{\beta}_{1nj}) = \beta_{10j} - \frac{1}{2}\frac{\lambda_n}{n}\gamma|\beta_{10j}|^{\gamma-1}\text{sgn}(\beta_{10j}) - \frac{1}{2}\frac{\lambda_n}{n}\gamma\mathsf{E}o_p(1) + C_n. \tag{4.5}$$

We complete the proof of Lemma 2.2. □

Proof of Theorem 2.1 By Lemma 2.2, let $\partial\Delta(\gamma)/\partial\gamma = 0$, we have

$$|\beta_{10j}|^{\gamma-1}\text{sgn}(\beta_{10j})(1 + \gamma\log(|\beta_{10j}|)) = 0.$$

Solving the equation of γ, we get

$$\gamma_0 = -\frac{1}{\log(|\beta_{10j}|)}.$$

When $0 < \beta_{10j} \leq \mathrm{e}^{-1}$, $\gamma_0 \leq 1$, it is easy to verify that $\partial^2\Delta(\gamma)/\partial\gamma_0^2 > 0$. As a result, γ_0 is the minimum point of the curve function $\Delta(\gamma)$. It follows that, when $0 < \beta_{10j} \leq \mathrm{e}^{-1}$, $\Delta(\gamma)$ is monotonically increasing for $\gamma \geq 1$. When $-\mathrm{e}^{-1} \leq \beta_{10j} < 0$, $\Delta(\gamma)$ is monotonically descending for $\gamma \geq 1$. As such, the conclusion of Theorem 2.1 (1) is obtained. The conclusion of Theorem 2.1 (2) can be obtained similarly, we omit the details. □

Proof of Theorem 2.2 Proof of Theorem 2.2 is similar to that of Theorem 2.1. So we omit it here. □

References

[1] Frank, I.E. and Friedman, J.H., A statistical view of some chemometrics regression tools (with discussion), *Technometrics*, **35(2)**(1993), 109–148.

[2] Tibshirani, R., Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society: Series B* (*Methodological*), **58(1)**(1996), 267–288.

[3] Knight, K. and Fu, W.J., Asymptotics for Lasso-type estimators, *The Annals of Statistics*, **28(5)** (2000), 1356–1378.

[4] Fan, J., Comments on "Wavelets in statistics: a review" by A. Antoniadis, *Journal of the Italian Statistical Society*, **6(2)**(1997), 131–138.

[5] Fan, J. and Li, R., Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96(456)**(2001), 1348–1360.

[6] Zou, H. and Hastie, T., Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), **67(2)**(2005), 301–320.

[7] Fan, J. and Peng, H., Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics*, **32(3)**(2004), 928–961.

[8] Fan, J. and Li, R., Statistical challenges with high dimensionality: feature selection in knowledge discovery, *Proceedings of the International Congress of Mathematicians, Vol. 3* (Autores: Sanz-Solé, M., Soria, J., Varona, J.L. and Verdera, J.), European Mathematical Society, Zürich, 595–622, 2006.

[9] Huang, J., Horowitz, J.L. and Ma, S., Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *The Annals of Statistics*, **36(2)**(2008), 587–613.

[10] Fu, W.J., Penalized regressions: the bridge versus the Lasso, *Journal of Computational and Graphical Statistics*, **7(3)**(1998), 397–416.

桥回归的系数估计对应于γ-范数的压缩规律

盖玉洁 张占利

(中央财经大学统计与数学学院, 北京, 100081)

张 君

(深圳大学深港联合应用统计科学研究中心, 数学与计算科学学院, 统计科学研究所, 深圳, 518060)

惩罚函数为$\sum|\beta_j|^\gamma$, $\gamma>0$的桥回归, 作为一类特殊的惩罚回归方法, 已经被很多文献研究过. 本文分别给出了估计系数的压缩大小与γ在$\gamma\geq 1$和$0<\gamma<1$两种不同情况下取值的一些理论结果, 并通过模拟来验证压缩估计的估计效果.

关键词: 桥回归, 高维线性模型, 变量选择, 压缩规律.

学科分类号: O212.

《应用概率统计》版权所有