Journal of Statistical Research *ISSN 0256 - 422 X*
2011, Vol. 45, No. 1, pp. 1-6
Bangladesh

HYPOTHESIS TESTING AND VARIABLE SELECTION IN RIDGE REGRESSION

JUNE LUO

Department of Mathematical Sciences, Clemson University, Clemson SC, 29634, USA
Email: jluo@clemson.edu

SUMMARY

There are discussions about variable selection when sample size $n \to \infty$ and dimension p fixed, but few dealing with n fixed and $p \to \infty$. Recently, high dimensional data, such as Microarray, exhibits very high dimension p and much smaller sample size n. In the present paper, we consider data set with fixed sample size n and growing dimension $p \to \infty$. We also conduct a hypothesis testing for variable selection in ridge regression. We also prove the consistency of the variable selection method for data with fixed sample size and infinite number of variables.

Keywords and phrases: hypothesis testing; variable selection; ridge estimator; shrinkage estimator

AMS Classification: 62F03; 62F12; 34K25; 34D05

1 Introduction

Upon the arrival of microarray data, statisticians found a fatal mistake in the typical assumption of statistical analysis regarding asymptotic properties of various estimators. The typical assumption allows the sample size to go to infinity, but due to the huge expenses of microarray experiments, the sample size in a microarray data is limited. Meanwhile, the number of gene variables could be sufficiently large. The discovery makes the existing statistical methods inappropriate in this situation.

In the current statistical literature, there has been discussions about variable selection using shrinkage techniques (Knight and Fu 2000; Tibshiriani 19996; Zheng and Loh 1997). Almost all the related discussions assume that sample size increases to infinity. However, the condition does not hold in real microarray data, and it is hardly appropriate to consider asymptotic methods for $n \to \infty$ when in fact the sample size n is fairly small. Luo (2010) proposed a variable selection procedure for fixed sample size n and infinite dimension p under the assumption that the random error has a finite moment. In this paper, we will propose a variable screening method for fixed sample size and infinite dimensions without assuming the finite moment on the random error. We will prove that the new variable screening procedure is consistent as well.

It has been proved by Shao and Chew (2007) that ridge estimators are mean square error (MSE) consistent when both the sample size and dimensions go to infinity but with different rates. Considering the limited sample size, Luo (2010) proved that under some regularity conditions, ridge estimator is mean square error consistent, but there was no explicit formula for the rate. Furthermore, how to reduce the assumptions for MSE consistency was one of the concerns in the discussion part in Luo (2010). In the present article, we will make less assumptions for MSE consistency of ridge estimator. Meanwhile, we will provide the explicit expression for the bias and variance of the ridge estimator and thus conduct hypothesis tests for variable selection for high dimensional data in ridge regression.

As a summary, we will provide explicit expressions for the asymptotic behaviors of ridge estimator, conduct hypothesis tests for variable selection and propose a screening method to select significant variables. The addressed questions will be answered under the assumption that sample size is fixed throughout the data and number of predictors is growing to infinity. Those three questions are lacking of study in the literature, so we believe that the results of the paper will fill in a significant void in current statistical theory.

2 Asymptotic properties of ridge estimator

For a high dimensional data with p predictors, we consider the following regression model

$$Y = X\beta + \varepsilon\,,$$

where ε has a multivariate normal distribution with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_p^2 I_n$. Also X is a $n \times p$ matrix and β is a $p \times 1$ unknown regression vector. Apply ridge regression to get the estimator of β (Hoerl and Kennard 1970), i.e.

$$\min_{\beta}(Y - X\beta)'(Y - X\beta) + h_p \sum_{j=1}^{p} \beta_j^2\,,$$

where h_p is a ridge parameter. For estimating the ridge parameter, we refer Kibria (2003) and Muniz and Kibria (2009) among others, so the ridge estimator is formulated as

$$\hat{\beta} = (X'X + h_p I_p)^{-1} X'Y\,.$$

We consider a fixed design in the paper. Since $X'X$ can have at most n positive eigenvalues, without loss of generality, we let λ_{ip} be the i^{th} nonzero eigenvalue of $X'X$ and assume $\lambda_{ip} > 0$ for all $i = 1, 2, \ldots, n$. Throughout the paper, sample size n is finite and dimension $p \to \infty$.

Assumption A. Let $h_p \to \infty$ as $p \to \infty$. For sufficiently large p, there exists a constant $\delta > 0$ such that each component of $\beta_{p\times 1}$ is $O(p^{-2-\delta})$.

Theorem 1. *Under the Assumption A, we have bias($\hat{\beta}_j$) $= o(1)$ for all $j = 1, 2, \ldots, p$.*

Proof. When p is large enough, let $\Gamma = (\tau_{ij})_{p\times p}$ be an orthogonal matrix such that

$$\Gamma' X' X \Gamma = \begin{bmatrix} \Lambda_{n\times n} & O_{n\times(p-n)} \\ O_{(p-n)\times n} & O_{(p-n)\times(p-n)} \end{bmatrix}_{p\times p},$$

where $\Lambda_{n\times n}$ is a diagonal matrix with elements $\lambda_{ip}, i = 1, 2, \ldots, n$. Then it follows that

$$\begin{aligned} \text{bias}(\hat{\beta}) &= E(\hat{\beta}) - \beta \\ &= (X'X + h_p I_p)^{-1} X'X\beta - \beta \\ &= -\left(\frac{X'X}{h_p} + I_p\right)^{-1} \beta \\ &= -\Gamma\left(\frac{\Gamma' X' X \Gamma}{h_p} + I_p\right)^{-1} \Gamma'\beta \\ &\hat{=} -\Gamma A \Gamma' \beta\,, \end{aligned}$$

where $A = (\frac{\Gamma' X' X \Gamma}{h_p} + I_p)^{-1}$ is a diagonal matrix with $h_p(h_p + \lambda_{ip})^{-1}$, $i = 1, 2, \ldots, n$ as first n diagonal elements, and the rest $(p-n)$ diagonal elements all equal to 1. Since $h_p(h_p + \lambda_{ip})^{-1} < 1$ for all $i = 1, 2, \ldots, n$ and Γ is an orthogonal matrix, each component of matrix ΓA is $O(1)$.

Under Assumption A, each component of matrix $\Gamma'\beta$ is $O(p^{-2-\delta}p)$, which leads to

$$\text{bias}(\hat{\beta}) = O(p^{-1-\delta}p) = o(1)\ .$$

That completes the proof for Theorem 1.

Assumption B. For simplicity, we choose h_p such that $\lambda_{ip} = o(h_p)$ for all $i = 1, 2, \ldots, n$.

Theorem 2. *Under Assumption B, we claim that*

$$\textit{var}(\hat{\beta}_j) \to \frac{\sigma_p^2}{h_p^2}\ \textit{diag}(X'X)_j \ \textit{for all}\ j = 1, 2, \ldots, p\,,$$

where diag$(X'X)_j$ *means the* j*th diagonal element of* $X'X$.

Proof. For the ridge estimator $\hat{\beta}$, we have the covariance matrix of $\hat{\beta}$

$$\begin{aligned} \text{Cov}\,(\hat{\beta}) &= (X'X + h_p I_p)^{-1} X' \sigma_p^2 X (X'X + h_p I_p)^{-1} \\ &= \frac{\sigma_p^2}{h_p}\left(\frac{X'X}{h_p} + I_p\right)^{-1} \frac{X'X}{h_p}\left(\frac{X'X}{h_p} + I_p\right)^{-1} \\ &= \frac{\sigma_p^2}{h_p}\left[\left(\frac{X'X}{h_p} + I_p\right)^{-1} - \left(\frac{X'X}{h_p} + I_p\right)^{-1}\left(\frac{X'X}{h_p} + I_p\right)^{-1}\right] \\ &= \frac{\sigma_p^2}{h_p}[\Gamma A \Gamma' - \Gamma A \Gamma' \Gamma A \Gamma'] \\ &= \frac{\sigma_p^2}{h_p}[\Gamma(A - A^2)\Gamma']\,, \end{aligned}$$

where Γ and A are defined in Theorem 1. So

$$\text{var}(\hat{\beta}_j) = \frac{\sigma_p^2}{h_p^2} \sum_{i=1}^{n} \tau_{ji}^2 \frac{h_p^2 \lambda_{ip}}{(h_p + \lambda_{ip})^2} \text{ for all } j = 1, 2, \ldots, p \ .$$

Under Assumption B,

$$\lim_{p \to \infty} \frac{h_p^2}{(h_p + \lambda_{ip})^2} = 1 \text{ for all } i = 1, 2, \ldots, n \ .$$

Recall that n is finite and the $\sum_{i=1}^{n} \tau_{ji}^2 \lambda_{ip}$ is the jth diagonal element of $X'X$, i.e.

$$\text{var}(\hat{\beta}_j) \to \frac{\sigma_p^2}{h_p^2} \text{diag}(X'X)_j$$

for all $j = 1, 2, \ldots, p$ as $p \to \infty$, where $\text{diag}(X'X)_j$ means the jth diagonal element of $X'X$. That finishes the proof for Theorem 2.

3 Hypothesis testing and variable selection

Since the discovery of microarray, as a primary goal in high dimensional data analysis, variable selection has received extensive attention in statistics. We will propose two consistent methods to eliminate insignificant variables in the ridge regression.

Assumption C. Choose $\sigma_p = o(h_p)$ and $p^{-\delta} h_p = o(\sigma_p)$.

Assumption C guarantees that the bias part of ridge estimator goes to 0 faster than the standard deviation of the ridge estimator. Because the random error ε is multivariate normal, $\hat{\beta}$ has a multivariate normal. Therefore we have the following result.

Theorem 3. *Under the Assumption A, B and C, for sufficiently large p, consider the following hypothesis testing:*

$$H_0 : \beta_j = 0 \quad vs \quad H_a : \beta_j \neq 0.$$

The p-value of the test is

$$2\Phi \left(- \frac{|\hat{\beta}_j|}{\frac{\sigma_p}{h_p} \sqrt{diag(X'X)_j}} \right) ,$$

where function Φ is the standard normal distribution function.

Theorem 3 can serve as a variable selection method. We will also propose a screening method and prove the consistency of the screening method. Let a_p be a sequence of positive numbers satisfying $a_p = o(1)$. For each p value, we screen out the jth gene if and only if $|\hat{\beta}_j| \leq a_p$. Therefore, after applying the screening out procedure, only genes associated with $|\hat{\beta}_j| > a_p$ are kept in the model as predictors. The sequence a_p acts as a filter in the process and eliminates genes with relatively small coefficients.

Theorem 4. *Under Assumption A, B and C, when a_p is chosen so that*

$$\frac{h_p^2 a_p^2}{\sigma_p^2 \log(p)} \to \infty \text{ as } p \to \infty\,,$$

the variable screening method is consistent in the sense that

$$\lim_{p\to\infty} P(|\hat{\beta}_j| > a_p) = 1 \text{ for any j with } \beta_j \neq 0 \tag{3.1}$$

and

$$\lim_{p\to\infty} P(|\hat{\beta}_j| \leq a_p \text{ for all j with } \beta_j = 0) = 1\,. \tag{3.2}$$

Proof. Assumption A and B guarantee that $\hat{\beta}_j$ is mean square error consistent for β_j, which implies for any $\eta > 0$,

$$\lim_{p\to\infty} P(|\hat{\beta}_j - \beta_j| > \eta) = 0\,.$$

When $\beta_j \neq 0$, since $a_p = o(1)$, we have

$$\lim_{p\to\infty} P(|\hat{\beta}_j| > a_p) = 1\,,$$

and this finishes the proof for (3.1).

If $\beta_j = 0$, following the result is Theorem 3,

$$P(|\hat{\beta}_j| > a_p) = \Phi\left(\frac{\text{bias}(\hat{\beta}_j) - a_p}{\text{sd}(\hat{\beta}_j)}\right) + \Phi\left(\frac{-\text{bias}(\hat{\beta}_j) - a_p}{\text{sd}(\hat{\beta}_j)}\right)\,,$$

where Φ is the standard normal distribution function. It is proven that $\text{bias}(\hat{\beta}_j) = O(p^{-\delta})$. Under the assumption $h_p^2 a_p^2/(\sigma_p^2 \log(p)) \to \infty$, we have $\sigma_p = o(h_p a_p)$ and thus

$$\text{bias}(\hat{\beta}_j) = O(p^{-\delta}) = o(\frac{\sigma_p}{h_p}) = o(a_p) \text{for all } j = 1, 2, \ldots, p.$$

In Theorem 2, we proved $\text{sd}(\hat{\beta}_j) \to f_j$ where $f_j = \frac{\sigma_p}{h_p}\sqrt{\text{diag}(X'X)_j}$ for all $j = 1, 2, \ldots, p$, which tells us

$$\frac{\pm\text{bias}(\hat{\beta}_j) - a_p}{\text{sd}(\hat{\beta}_j)} / (\frac{a_p}{f_j}) \to -1 \text{ as } p \to \infty\,,$$

so for a large enough p value, there exists a positive constant $\mu \in (0, 1)$ such that

$$P(|\hat{\beta}_j| > a_p) \leq 2\Phi\left(-\mu a_p / f_j\right)\,.$$

Recall the fact that $\text{diag}(X'X)$ are all finite constants. For a sufficiently large p, we have the following

$$\frac{\mu\, a_p}{f_j} = \frac{\mu}{\text{diag}(X'X)_j}\frac{h_p a_p}{\sigma_p} \geq \sqrt{2q\log(p)} \geq 1 \quad \text{for a } \; q > 1\,.$$

Now we apply the inequality $2\Phi(-x) \leq e^{-x^2/2}$ for any $x \geq 1$. The probability of interest becomes

$$P(|\hat{\beta}_i| > a_p) \leq 2\Phi(-\mu a_p / f_j) \leq 2\Phi(-\sqrt{2q\log(p)}) \leq e^{-q\log(p)} = p^{-q}\,.$$

Then it follows

$$P(|\hat{\beta}_j| > a_p \text{ for at least one } j \text{ with } \beta_j = 0) \leq \sum_{\text{all } j \text{ with } \beta_j = 0} P(|\hat{\beta}_j| > a_p) \leq p^{1-q} \to 0 \text{ as } p \to \infty\,,$$

and thus

$$\lim_{p\to\infty} P(|\hat{\beta}_j| \leq a_p \text{ for all j with } \beta_j = 0) = 1\ .$$

That finishes the proof for (3.2).

Acknowledgments

This is a continuation of my dissertation work. I want to thank Professor Yijun Zuo and Professor Hira L. Koul at Michigan State University. I also like to thank the referees for thoughtful comments.

References

[1] Hoerl, A. and Kennard, R. (1970). Ridge regression biased estimation for nonorthogonal problems. *Technometrics,* **12***, 55-67.*

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

[3] Kibria, B.M.G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computations,* **32***, 419-435.*

[4] Luo, J. (2010). The discovery of mean square error consistency of ridge estimator. *Statistics and Probability Letters,* **80***, 343-347.*

[5] Muniz, G. and Kibria, B.M.G. (2009). On some ridge regression estimators: An Empirical Comparisons. *Communications in Statistics-Simulation and Computations,* **38:3***, 621-630.*

[6] Shao, J. and Chow, S. (2007). Variable screening in predicting clinical outcome with high-dimensional microarrays. *Journal of Multivariate Analysis,* **98 (8)***, 1529-1538.*

[7] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society,* **58***, 267-288.*

[8] Zheng, X. and Loh, W. (1997). A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica,* **7***, 311-325.*