Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Asymptotic efficiency of ridge estimator in linear and semiparametric linear models

June Luo

Clemson University, Department of Mathematical Sciences, United States

ARTICLE INFO

Article history:
Received 22 January 2010
Received in revised form 13 January 2011
Accepted 24 August 2011
Available online 1 September 2011

Keywords:
High dimension
Ridge estimator
Differencing sequence
Continuity

ABSTRACT

The linear model with a growing number of predictors arises in many contemporary scientific endeavor. In this article, we consider the commonly used ridge estimator in linear models. We propose analyzing the ridge estimator for a finite sample size n and a growing dimension p. The existence and asymptotic normality of the ridge estimator are established under some regularity conditions when $p \to \infty$. It also occurs that a strictly linear model is inadequate when some of the relations are believed to be of certain linear form while others are not easily parameterized, and thus a semiparametric partial linear model is considered. For these semiparametric partial linear models with $p > n$, we develop a procedure to estimate the linear coefficients as if the nonparametric part is not present. The asymptotic efficiency of the proposed estimator for the linear component is studied for $p \to \infty$. It is shown that the proposed estimator of the linear component asymptotically performs very well.

1. Introduction

High dimensional data have received extensive attention in statistics and biomedical studies. How to efficiently estimate the linear coefficients has been discussed and various estimators have been proposed in the current literature. The asymptotic behavior of those proposed estimators have been obtained to show the asymptotic efficiency. However, almost all the discussions about the asymptotic distributions assume a finite dimension p and a growing sample size n. Given the fact that some of the high dimensional data, such as microarray data, exhibits higher dimension than the sample size, we think that the asymptotic distributions of the estimators should also be addressed assuming a finite sample size n and an infinite dimension p.

Consider the standard multiple regression model for a fixed design

$$Y = X\beta + \epsilon, \tag{1}$$

where X is an $n \times p$ observation matrix, β is an unknown $p \times 1$ vector of regression coefficients, and ϵ is normally distributed with a mean vector $\underline{0}$ and covariance matrix I_n. With $p > n$, the $X'X$ is singular and thus the least squares estimator becomes inappropriate. On the other side, ridge estimator $\hat{\beta} = (X'X + h_p I_p)^{-1}X'Y$ has shown mean square error consistency when the dimension $p \to \infty$ (see Hoerl and Kennard, 1970, Shao and Chow, 2007 and Luo, 2010). In this paper, we will discover the asymptotic distribution of $\hat{\beta}$ and explicitly express the rate of convergence.

Besides the multiple regression models, semiparametric partial linear models have been extensively studied in high dimensional data analysis. In these semiparametric models, some of the relations are assumed to be linear while others are not known. In this paper, we consider the following semiparametric partial linear model

$$y_i = x_i\beta + f(z_i) + \epsilon_i, \quad i = 1, 2, \ldots, n + 1, \tag{2}$$

E-mail address: jluo@clemson.edu.

0167-7152/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.08.019

where $x_i \in R^p$, $z_i \in R^1$, β is an unknown $p \times 1$ vector of parameters, $f(.)$ is an unknown function and ϵ_i's are independent and identically distributed normal random error with mean 0 and variance σ^2. Several approaches, such as the methods in Ahn and Powell (1993) and Liang et al. (1999), have been developed to construct the estimators for β under the assumption $n > p$. Yatchew (1997, 2000) proposed a class of estimators which use appropriate differences that aim to remove the trend in the data that arise from the unknown function f. Such methods do not require an estimator of the function f and are often called difference-based estimators. See also Fan and Huang (2001) and Lam and Fan (2008).

Most recently, Wang et al. (2011) have proposed a difference based estimation method for the semiparametric model (2). Wang et al. used a special class of differences and adopted the least squares estimation method to estimate the linear component after taking the differences. The asymptotic distribution of the difference based least squares estimator was obtained assuming an infinite n and a finite p. Wang et al. (2011) claimed that the correlation structure should not be incorporated so that the optimality results about the estimator are valid. For a model (2) with $p > n$, a new approach to estimate the linear component is unfinished but necessary.

As Wang et al. (2011) strictly require finite and fixed p in model (2), the differencing matrix in Wang et al. (2011) does not apply to remove the nonparametric effect in model (2) for infinite p and finite n. We will concentrate our attention on the estimation of the linear component and study the asymptotic distribution of the proposed estimator. By applying a special matrix, we can ignore the bias induced from the presence of the nonparametric term $f(.)$ and thus estimate the linear coefficients as if the nonparametric part is known. A ridge estimator is suggested after we incorporate the covariance matrix of the random noise. The properties of the resulting estimator followed the asymptotic properties developed for $\hat{\beta}$ in model (1) under certain assumptions.

For the purpose of this paper, we will employ the ridge estimation concept that was proposed to solve the multicollinearity in both linear models and semiparametric partial linear models. The remainder of the paper is organized as follows. In Section 2, we consider linear models and the asymptotic distribution of ridge estimator as $p \to \infty$ is provided. In Section 3, the semiparametric partial linear model and the special matrix to remove the nonparametric effect are introduced. The proposed ridge estimator for the linear component is followed after taking the differences. The asymptotic distribution of the ridge estimator for the linear component is discussed in detail.

2. The asymptotic efficiency of ridge estimator

Our discussion begins with considering the standard multiple regression model $Y = X\beta + \epsilon$, where X is an $n \times p$ observation matrix, β is an unknown vector of regression coefficients, and ϵ is normally distributed with a mean vector $\underline{0}$ and covariance matrix I_n. As $p > n$, the observation matrix X has rank n and thus $X'X$ is singular. So the ridge estimator

$$\hat{\beta} = (X'X + h_p I_p)^{-1} X'Y \tag{3}$$

is more appropriate than the least squares estimator. Since $X'X$ can have at most n positive eigenvalues, without loss of generality, we let λ_{ip} be the ith nonzero eigenvalue of $X'X$ and assume $\lambda_{ip} > 0$ for all $i = 1, 2, \ldots, n$. Throughout the paper, the sample size is finite and the 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 $O\left(\frac{1}{p^{2+\delta}}\right)$. Choose $h_p = o(p^\delta)$ and assume $\lambda_{ip} = o(h_p)$. One such choice is $\delta = 0.2$ and $h_p = p^{0.1}$.

Theorem 1. *Under the Assumption A, for $\hat{\beta}$ given in (3), $h_p(\hat{\beta} - \beta) \to N(\underline{0}, X'X)$ as $p \to \infty$.*

Proof.

$$\begin{aligned} h_p(\hat{\beta} - \beta) &= h_p((X'X + h_p I_p)^{-1} X'(X\beta + \epsilon) - \beta) \\ &= \left(\frac{X'X}{h_p} + I_p\right)^{-1} X'\epsilon - h_p \left(\frac{X'X}{h_p} + I_p\right)^{-1} \beta \\ &\hat{=} M - N, \end{aligned}$$

where $M = \left(\frac{X'X}{h_p} + I_p\right)^{-1} X'\epsilon$ and $N = h_p \left(\frac{X'X}{h_p} + I_p\right)^{-1} \beta$. □

Results in Lemmas 1 and 2 yield $h_p(\hat{\beta} - \beta) \to N(0, X'X)$ as $p \to \infty$ when the random error ϵ is normally distributed with a mean vector $\underline{0}$ and covariance matrix I_n.

Lemma 1. *Under the Assumption A, each component of vector N is $o(1)$.*

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

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

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

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

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

$$\text{each component of } N = O(h_p p^{-1-\delta} p) = o(1).$$

That finishes the proof of Lemma 1. □

Lemma 2. *Under the Assumption A, $M \to N(\underline{0}, X'X)$ as $p \to \infty$.*

Proof. Note that ϵ is normally distributed, so $M = \left(\frac{X'X}{h_p} + I_p\right)^{-1} X'\epsilon$ is normally distributed. We also know that $E(\epsilon) = \underline{0}$, so M has a mean of $\underline{0}$. The covariance matrix of M is

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

where matrix A was defined in the proof of Lemma 1. It follows that $h_p(A - A^2)$ is a diagonal matrix with $\frac{\lambda_{ip} h_p^2}{(h_p+\lambda_{ip})^2},\ i = 1, 2, \ldots, n$ as first n diagonal elements, and the rest $(p - n)$ diagonal elements all equal to 0. So the ijth component of covariance matrix $\mathrm{Cov}(M)$ is

$$\sum_{q=1}^{n} \tau_{iq} \frac{\lambda_{qp} h_p^2}{(h_p + \lambda_{qp})^2} \tau_{jq}.$$

Since $\lambda_{ip} = o(h_p)$, it follows $\frac{h_p^2}{(h_p+\lambda_{qp})^2} \to 1$ as $p \to \infty$.

So

$$\begin{aligned} \sum_{q=1}^{n} \tau_{iq} \frac{\lambda_{qp} h_p^2}{(h_p + \lambda_{qp})^2} \tau_{jq} &\to \sum_{q=1}^{n} \tau_{iq} \lambda_{qp} \tau_{jq} \\ &= \text{ the } ij\text{th component of matrix } X'X. \end{aligned}$$

That finishes the proof of Lemma 2. □

Corollary 1. *For a linear model $Y = X\beta + \epsilon$ in (1), when ϵ is normally distributed with a mean vector $\underline{0}$ and covariance matrix $\sigma^2 I_n$, under the Assumption A, the ridge estimator $\hat{\beta} = (X'X + h_p I_p)^{-1} X'Y$ has the asymptotic distribution*

$$h_p(\hat{\beta} - \beta) \to N(\underline{0}, \sigma^2 X'X).$$

In general, for a model $Y = X\beta + \epsilon$, where ϵ is normally distributed with a mean vector $\underline{0}$ and covariance matrix Σ. To take the covariance structure of the random noise ϵ into account, we suggest performing a transformation before applying the ridge estimator. Since $\Sigma^{-\frac{1}{2}}\epsilon$ has the covariance matrix I_n, we replace X with $\Sigma^{-\frac{1}{2}}X$ and thus we conclude

$$h_p(\hat{\beta} - \beta) \to N(\underline{0}, X'\Sigma^{-1}X), \tag{4}$$

where $\hat{\beta} = (X'\Sigma^{-1}X + h_pI_p)^{-1}X'\Sigma^{-1}Y$ defined as the transformed ridge estimator.

3. Ridge estimator in semiparametric partial linear models

Partial linear models are more flexible than standard linear models since they have a parametric and a nonparametric component. They can be a suitable choice when one suspects that the response y linearly depends on x, but y is nonlinearly related to z. Consider a fix design version of the semiparametric partial linear model (2) where z_i is a function of p. Model (2) has the matrix form

$$Y = X\beta + f(Z) + \epsilon,$$

where $Y = (y_1, y_2, \ldots, y_{n+1})'$, X is an $(n+1) \times p$ matrix whose ith row is given by x_i, and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_{n+1})'$ is normally distributed with a mean vector $\underline{0}$ and covariance matrix $\sigma^2 I_{n+1}$. Wang et al. (2011) used a higher order sequence of differences and studied the asymptotic properties of least squares estimator when sample size $n \to \infty$. As the differencing method and the least squares estimator in Wang et al. (2011) is not applicable for $p > n$, we propose the following method to estimate the linear component without an estimator of the function $f(.)$.

Assumption B. The sequence $\{z_i\} \to c_0$ as $p \to \infty$, for all $i = 1, 2, \ldots, n+1$, where c_0 is a finite constant. The unknown function f is continuous at point c_0. One such example is $z_i = i/p$, $c_0 = 0$ and f is any function that is continuous at 0.

Define a matrix

$$D = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & & & & & \\ 0 & \cdots & 0 & 0 & 1 & -1 \end{bmatrix}_{n\times(n+1)}. \tag{5}$$

We now consider the model $DY = DX\beta + Df(Z) + D\epsilon$, which is

$$y_{i+1} - y_i = (x_{i+1} - x_i)\beta + f(z_{i+1}) - f(z_i) + \epsilon_{i+1} - \epsilon_i, \quad i = 1, 2, \ldots, n. \tag{6}$$

Since $\{z_i\}$ has the same limit c_0 for all $i = 1, 2, \ldots, n+1$, for any $\psi > 0$, there exists a large enough p such that

$$\max_{1\le i\le n+1} |z_i - c_0| < \psi.$$

Function f is continuous at point c_0, so for a large enough p, we have

$$\max_{1\le i\le n+1} |f(z_i) - f(c_0)| < \psi,$$

which implies that

$$\max_{1\le i\le n} |f(z_{i+1}) - f(z_i)| = o(1).$$

Therefore, the application of the matrix D removes the nonparametric effect when p is sufficiently large. That allows us to ignore the presence of nonparametric part in model (6) and estimate the coefficient as if the nonparametric term is not present. Thus, we may rewrite (6) as

$$DY = DX\beta + D\epsilon, \tag{7}$$

where matrix D is given in (5). Let Ψ be the $n \times n$ covariance matrix of $D\epsilon$, so $\Psi = \sigma^2 DD'$. As followed by the results in Section 2, the transformed ridge estimator $\hat{\beta} = (X'D'\Psi^{-1}DX + h_pI_p)^{-1}X'D'\Psi^{-1}DY$ has the asymptotic distribution

$$h_p(\hat{\beta} - \beta) \to N(\underline{0}, X'D'\Psi^{-1}DX). \tag{8}$$

4. Conclusions and discussions

The use of ridge estimator is a crucial part for constructing asymptotic distributions for the estimator. Both linear models and semiparametric partial linear models are considered in the paper. In general, the regression function between Y and X may be not strictly linear or even partial linear. When there exists a kernel transformation function $k(.)$ so that it is linear or partial linear for Y and $K(X)$ in the feature space, our results for both linear models and partial linear models can be extended to the feature space for new variables under the specified assumptions. The established asymptotic distribution for the ridge estimator implies that the ridge estimator is mean square error consistent for the parameter β when dimension $p \to \infty$.

Acknowledgments

I want to deliver my appreciation to Professor Yijun Zuo and Professor Hira L. Koul at Michigan State University. They both offered a tremendous amount of help during my dissertation.

References

Ahn, H., Powell, J., 1993. Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics 58, 3–29.

Fan, J., Huang, L., 2001. Goodness-of-fit tests for parametric regression models. Journal of the American Statistical Association 96 (454), 640–652.

Hoerl, A., Kennard, R., 1970. Ridge regression biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Lam, C., Fan, J., 2008. Profile-kernel likelihood inference with diverging number of parameters. The Annals of Statistics 36 (5), 2232–2260.

Liang, H., Hardle, W., Carroll, R., 1999. Estimation in semiparametric partially linear errors-in-variables model. The Annals of Statistics 27, 1519–1535.

Luo, J., 2010. The discovery of mean square error consistency of ridge estimator. Statistics and Probability Letters 80, 343–347.

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

Wang, L., Brown, L., Cai, T., 2011. A difference based approach to semiparametric partial linear model. Electronic Journal of Statistics 5, 619–641.

Yatchew, A., 1997. An elementary estimator of the partial linear model. Economics Letters 57, 135–143.

Yatchew, A., 2000. Scale economies in electricity distribution: a semiparametric analysis. Journal of Applied Economics 15 (2), 187–210.