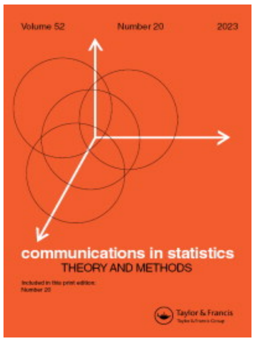

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Two-stage shrunken least squares estimator and its superiority

Quanhong Song, Lichun Wang & Liqun Wang

To cite this article: Quanhong Song, Lichun Wang & Liqun Wang (2023): Two-stage shrunken least squares estimator and its superiority, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2023.2250487

To link to this article: https://doi.org/10.1080/03610926.2023.2250487

Published online: 29 Aug 2023.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2023.2250487

Two-stage shrunken least squares estimator and its superiority

Quanhong Song[a], Lichun Wang[a], and Liqun Wang[b]

[a]Department of Statistics, Beijing Jiaotong University, Beijing, China; [b]Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada

ABSTRACT
In linear regression model, the superiority of ordinary least squares estimator (OLSE) will be failed when there exist multi-collinearity problems. Based on the class of generalized shrunken least squares (GSLS) estimators suggested by Wang (1990), this article proposes a two-stage shrunken least squares estimator and discusses its superiority theoretically, and finally verifies the results by numerical simulations.

ARTICLE HISTORY
Received 31 December 2022
Accepted 16 August 2023

KEYWORDS
Two-stage estimator; generalized shrunken least squares estimator; multi-collinearity

1. Introduction

Considering the following linear regression model:

$$Y = X\beta + e, \quad e \sim (0, \sigma^2 I_n), \tag{1.1}$$

where Y is an $n \times 1$ vector of observations, X is an $n \times p$ design matrix with full column rank, e is an $n \times 1$ random error vector, β is a $p \times 1$ vector of unknown regression coefficients. According to Gauss-Markov Theorem, the OLSE of β is

$$\hat{\beta} = (X'X)^{-1}X'Y, \tag{1.2}$$

which is best linear unbiased estimator. However, with the wide applications of the OLSE, we find that when multi-collinearity problems exist, the OLSE tends to perform poorly. This is because that $X'X$ is close to be a singular matrix when multi-collinearity exists, which comes its eigenvalues (denoted by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$) to be close to be zero and accordingly the mean squares error of the OLSE will be very large.

Stein (1956) proves that the OLSE of a normal mean vector is inadmissible in the case that its dimension is greater than 2, that is, there is another estimator which consistently outperforms the OLSE in some sense. Trenkler (1981) points out that the performance of OLSE $\hat{\beta}$ may become poor when there exist multi-collinearity. In recent decades, many new estimators have been proposed, among which are ridge estimator, principal component estimator, stein estimator, etc. These estimators are all biased estimator, but they have smaller variance compared to the OLSE $\hat{\beta}$. In what follows, we introduce a new estimator class, generalized shrunken least squares (GSLS) estimators, to which many of the commonly used biased estimators belong. Then, the two-stage shrunken least squares (TSLS) estimator in this class will be given, and finally its property will be discussed.

CONTACT Lichun Wang wlc@amss.ac.cn Department of Statistics, Beijing Jiaotong University, Beijing, China.

2. Two-stage shrunken least squares estimator

Definition 2.1. An estimator class of the following form is known as a generalized shrunken least squares estimators (see Wang 1990), that is,

$$\hat{\beta}_{GS}(A) = PAP'\hat{\beta}, \tag{2.1}$$

where $A = \text{diag}(a_1, \ldots, a_p), 0 \leq a_i \leq 1 (i = 1, \ldots, p)$ and P is a $p \times p$ orthogonal matrix such that

$$P'X'XP = \text{diag}(\lambda_1, \ldots, \lambda_p) = \Lambda,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are eigenvalues of $X'X$.

Research based on GSLS estimators has already yielded some results, for example, Zhao (1995) gives a note on GSLS estimators, Guo and Guo (1997) discusses the problem of choosing parameter about GSLS estimators, Sun (1997) discusses the advantages of GSLS estimators and the multiple k-class GSLS estimator suggested by Shi (1999), Duan (1999) gives a new method to choose A, and Sun (1999) proposes a new criterion for selecting A named Q(c). And this estimator class includes a lot of biased estimators, such as:

1) The generalized ridge regression estimator suggested by Hoerl and Kennard (1970):
$$\hat{\beta}_{RR}(K) = (X'X + PKP')^{-1}X'Y = \hat{\beta}_{GS}(\Lambda(\Lambda + K)^{-1}),$$
where $K = \text{diag}(k_1, \ldots, k_p), k_i \geq 0 (i = 1, \ldots, p)$.
2) The principal component estimator suggested by Kendall (1957):
$$\hat{\beta}_{PC}(r) = \sum_{i=1}^{r} \frac{1}{\lambda_i} P_i P_i' X'Y = \hat{\beta}_{GS}\left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}\right),$$
where $P = (P_1, \ldots, P_p)$.
3) Stein estimator:
$$\hat{\beta}_S(c) = (1 - c)\hat{\beta} = \hat{\beta}_{GS}((1 - c)I),$$
where $c \in (0, 1)$.
4) The universal ridge estimator suggested by Yang (1991):
$$\hat{\beta}_U R = (PKP' + X'X)^{-1}PSP'X'Y = \hat{\beta}_{GS}((K + \Lambda)^{-1}S\Lambda),$$
where $K = \text{diag}(\mathrm{k}_1, \ldots, \mathrm{k}_\mathrm{p}), \mathrm{S} = \text{diag}(\mathrm{s}_1, \ldots, \mathrm{s}_\mathrm{p}), \mathrm{k}_\mathrm{i} \geq 0, \mathrm{s}_\mathrm{i} \geq 0, \mathrm{i} = 1, \ldots, \mathrm{p}$.

Since A is arbitrary, we will concentrate on how to choose an A. To this end, the two-stage shrunken least squares (TSLS) estimator is proposed.

Note that, the mean squares error of $\hat{\beta}_{GS}(A)$ is given by

$$\begin{aligned}
\text{MSE}(\hat{\beta}_{GS}(A)) &= \text{tr}\left(\text{Cov}\left(\hat{\beta}_{GS}(A)\right)\right) + \left(\text{Bias}\left(\hat{\beta}_{GS}(A)\right)\right)' \left(\text{Bias}\left(\hat{\beta}_{GS}(A)\right)\right) \\
&= \sigma^2 \text{tr}\left(A^2 \Lambda^{-1}\right) + \beta' P(I - A)^2 P' \beta \\
&= \sum_{i=1}^{p} \left[\frac{\sigma^2 a_i^2}{\lambda_i} + \delta_i^2 (1 - a_i)^2\right] \\
&= \sum_{i=1}^{p} D_i(a_i),
\end{aligned}$$

where δ_i is the i^{th} element of the vector $P'\beta$.

Let $\partial \sum_{i=1}^{p} D_i(a_i)/\partial a_i = 0$, we have

$$a_i^{opt} = \frac{\lambda_i \delta_i^2}{\lambda_i \delta_i^2 + \sigma^2}.$$

Note that for $i \neq j$, we have

$$\frac{\partial^2 \sum_{i=1}^{p} D_i(a_i)}{\partial a_i \partial a_j} = 0$$

and

$$\frac{\partial^2 \sum_{i=1}^{p} D_i(a_i)}{\partial a_i \partial a_i} = \frac{2\sigma^2}{\lambda_i} + 2\delta_i^2 > 0,$$

thus its Hessian matrix is positive definite, that is, $(a_1^{opt}, a_2^{opt}, \ldots, a_p^{opt})$ is the point of optimal value that makes $\mathrm{MSE}(\hat{\beta}_{GS}(A))$ reach the minimum value. Usually we do not know the true value of β and σ, so we use the least squares estimator $\hat{\beta}$ and unbiased estimator $\hat{\sigma}^2 = \|Y - X\hat{\beta}\|^2/(n-p)$ to estimate β and σ^2, respectively. Finally we obtain

$$\hat{a}_i = \frac{\lambda_i \hat{\delta}_i^2}{\lambda_i \hat{\delta}_i^2 + \hat{\sigma}^2}.$$

Definition 2.2. The two-stage shrunken least squares estimator is given by

$$\hat{\beta}_{GS}(\hat{A}) = P\hat{A}P'\hat{\beta}, \tag{2.2}$$

where

$$\hat{A} = \mathrm{diag}(\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p),$$

$$\hat{a}_i = \frac{\lambda_i \hat{\delta}_i^2}{\lambda_i \hat{\delta}_i^2 + \hat{\sigma}^2} = \frac{\lambda_i \hat{\beta}' P_i P_i' \hat{\beta}}{\lambda_i \hat{\beta}' P_i P_i' \hat{\beta} + \hat{\sigma}^2},$$

$\hat{\delta}_i = P_i'\hat{\beta}$ is the i^{th} element of the vector $P'\hat{\beta} = (P_1, \cdots, P_p)'\hat{\beta}$.

The advantages of the TSLS estimator are obvious, because it not only belongs to the class of GSLS, and also it solves the problem of choosing the parameter A.

3. The superiority of TSLS estimator

3.1. Expectation

In what follows, we assume $e \sim N(0, \sigma^2 I_n)$ in the model (1.1).

Theorem 3.1. *When σ is sufficiently small, the expectation of the TSLS estimator has the following approximation:*

$$E(P\hat{A}P'\hat{\beta}) = P\mathrm{diag}\Big(1 - \frac{\sigma^2}{\lambda_1 \beta' P_1 P_1' \beta} + o(\sigma^2), \cdots, 1 - \frac{\sigma^2}{\lambda_p \beta' P_p P_p' \beta} + o(\sigma^2)\Big)P'\beta. \tag{3.1}$$

Proof. Let $m = (Y - X\beta)/\sigma$, then $m \sim N_n(0, I)$ and $\hat{\beta} = \beta + \sigma(X'X)^{-1}X'm$, we obtain

$$\hat{\beta}'P_iP_i'\hat{\beta} = [\beta + \sigma(X'X)^{-1}X'm]'P_iP_i'[\beta + \sigma(X'X)^{-1}X'm]$$
$$= \beta'P_iP_i'\beta + 2\sigma\beta'P_iP_i'(X'X)^{-1}X'm + \sigma^2 m'X(X'X)^{-1}P_iP_i'(X'X)^{-1}X'm.$$

Noting that

$$\hat{\sigma}^2 = \frac{\|Y - X\hat{\beta}\|^2}{n-p} = \frac{\|Y - X\hat{\beta}\|^2}{tr(I - P_X)}.$$

Therefore,

$$\hat{\sigma}^2 = \frac{\sigma^2 m'\big[I - X(X'X)^{-1}X'\big]m}{n-p} = \frac{\sigma^2 m'Mm}{n-p},$$

where

$$M = I - X(X'X)^{-1}X'.$$

Denote $\lambda_i\hat{\beta}'P_iP_i'\hat{\beta} = b_1^i + 2\sigma b_2^i + \sigma^2 m' b_3^i m$, where

$$b_1^i = \lambda_i\beta'P_iP_i'\beta,$$
$$b_2^i = \lambda_i\beta'P_iP_i'(X'X)^{-1}X'm,$$
$$b_3^i = \lambda_i X(X'X)^{-1}P_iP_i'(X'X)^{-1}X'.$$

Then we have

$$\hat{\sigma}^2 + \lambda_i\hat{\beta}'P_iP_i'\hat{\beta} = b_1^i + 2\sigma b_2^i + \sigma^2 m'(b_3^i + \frac{M}{n-p})m.$$

Thus

$$(\hat{\sigma}^2 + \lambda_i\hat{\beta}'P_iP_i'\hat{\beta})^{-1} = \frac{1}{b_1^i}\Big[1 + \frac{2\sigma b_2^i + \sigma^2 m'(b_3^i + \frac{M}{n-p})m}{b_1^i}\Big]^{-1}. \tag{3.2}$$

Because there exists $\sigma > 0$ sufficiently small such that

$$\Big|\frac{2\sigma b_2^i + \sigma^2 m'(b_3^i + \frac{M}{n-p})m}{b_1^i}\Big| < 1.$$

Using the Taylor formula to expand Equation (3.2) yielding:

$$(\hat{\sigma}^2 + \lambda_i\hat{\beta}'P_iP_i'\hat{\beta})^{-1} = \frac{1}{b_1^i}\Big[1 - \frac{2\sigma b_2^i + \sigma^2 m'(b_3^i + \frac{M}{n-p})m}{b_1^i}$$
$$+ \Big(\frac{2\sigma b_2^i + \sigma^2 m'(b_3^i + \frac{M}{n-p})m}{b_1^i}\Big)^2$$
$$- \Big(\frac{2\sigma b_2^i + \sigma^2 m'(b_3^i + \frac{M}{n-p})m}{b_1^i}\Big)^3 + o(\sigma^6)\Big]$$
$$= \frac{1}{b_1^i}\Big[1 - \frac{2\sigma b_2^i}{b_1^i} + \sigma^2\Big(\frac{4{b_2^i}^2}{{b_1^i}^2} - \frac{m'(b_3^i + \frac{M}{n-p})m}{b_1^i}\Big)$$
$$+ \sigma^3\Big(\frac{4b_2^i m'(b_3^i + \frac{M}{n-p})m}{{b_1^i}^2} - \frac{8{b_2^i}^3}{{b_1^i}^3}\Big)\Big] + o(\sigma^3).$$

Therefore,

$$\hat{a}_i = \frac{\lambda_i \hat{\beta}' P_i P_i' \hat{\beta}}{\lambda_i \hat{\beta}' P_i P_i' \hat{\beta} + \hat{\sigma}^2} = 1 - \frac{\sigma^2 m' M m}{b_1^i (n-p)} + \frac{2\sigma^3 b_2^i m' M m}{{b_1^i}^2 (n-p)} + o(\sigma^3).$$

Since $E(m'Mm) = n - p$, then

$$E(\hat{a}_i) = 1 - \frac{\sigma^2}{\lambda_i \beta' P_i P_i' \beta} + o(\sigma^2).$$

Noting that $\hat{A} = \mathrm{diag}(\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p)$, which can be seen as a function of $m'Mm$. At the same time, by the fact that

$$(X'X)^{-1} X' M = (X'X)^{-1} X' - (X'X)^{-1} X' X (X'X)^{-1} X' = 0,$$

and from the independence conditions for linear and quadratic forms, we conclude that $m'Mm$ and $(X'X)^{-1}X'm$ are independent. Since $\hat{\beta} = \beta + \sigma (X'X)^{-1} X' m$, so $\hat{A}$ and $\hat{\beta}$ are independent.

Together we come to the conclusion of Theorem 3.1.

The proof of Theorem 3.1 is finished. □

3.2. *The mean squares error*

Theorem 3.2. *When σ is sufficiently small, a sufficient condition for the TSLS estimator to outperform the OLSE in terms of mean squares error criterion is*

$$n > p + 2. \tag{3.3}$$

Proof. Note that $\hat{\beta}_{GS}(\hat{A}) = P\hat{A}P'[\beta + \sigma(X'X)^{-1}X'm]$, and $\hat{\beta}_{GS}(\hat{A}) - \beta = P(\hat{A} - I)P'\beta + \sigma P\hat{A}P'(X'X)^{-1}X'm$, we have

$$\begin{aligned} \mathrm{MSE}(\hat{\beta}_{GS}(\hat{A})) &= E[\beta' P(\hat{A} - I)P'P(\hat{A} - I)P'\beta + \sigma^2 m'X(X'X)^{-1}P\hat{A}P'P\hat{A}P'(X'X)^{-1}X'm \\ &\quad + 2\sigma\beta'P(\hat{A} - I)P'P\hat{A}P'(X'X)^{-1}X'm] \\ &= E[\beta'P(\hat{A} - I)^2P'\beta + \sigma^2 m'X(X'X)^{-1}P\hat{A}^2P'(X'X)^{-1}X'm \\ &\quad + 2\sigma\beta'P(\hat{A} - I)\hat{A}P'(X'X)^{-1}X'm] \\ &= E[\Delta_1 + \Delta_2 + \Delta_3]. \end{aligned}$$

By the fact that

$$P'(X'X)^{-1}X'M = P'(X'X)^{-1}X' - P'(X'X)^{-1}X'X(X'X)^{-1}X' = 0,$$

we know that $\hat{A}$ and $P'(X'X)^{-1}X'm$ are independent, that is, $(\hat{A} - I)\hat{A}$ and $P'(X'X)^{-1}X'm$ are independent. Hence

$$E[\Delta_3] = 2\sigma E[\beta'P(\hat{A} - I)\hat{A}]E[P'(X'X)^{-1}X'm] = 2\sigma E[\beta'P(\hat{A} - I)\hat{A}] \cdot 0 = 0.$$

Then using the facts that

$$\hat{A} - I = \mathrm{diag}\Big(-\frac{\sigma^2 m'Mm}{b_1^1(n-p)} + \frac{2\sigma^3 b_2^1 m'Mm}{{b_1^1}^2(n-p)} + o(\sigma^3), \cdots, -\frac{\sigma^2 m'Mm}{b_1^p(n-p)} + \frac{2\sigma^3 b_2^p m'Mm}{{b_1^p}^2(n-p)} \\ + o(\sigma^3)\Big)$$

and

$$(\hat{A}-I)^2 = \text{diag}\Big(\frac{\sigma^4(m'Mm)^2}{b_1^{1\,2}(n-p)^2}+o(\sigma^4),\cdots,\frac{\sigma^4(m'Mm)^2}{b_1^{p\,2}(n-p)^2}+o(\sigma^4)\Big).$$

Also $E[(m'Mm)^2] = Var(m'Mm)+[E(m'Mm)]^2 = 2tr(M^2)+tr^2(M) = 2(n-p)+(n-p)^2 = (n-p)(n-p+2)$.

Thus, we have

$$\begin{aligned}
E[\Delta_1] &= \beta' PE[(\hat{A}-I)^2]P'\beta \\
&= \beta'(P_1,\cdots,P_p)diag(\frac{\sigma^4(n-p+2)}{b_1^{1\,2}(n-p)},\cdots,\frac{\sigma^4(n-p+2)}{b_1^{p\,2}(n-p)})(P_1,\cdots,P_p)'\beta+o(\sigma^4) \\
&= \sum_{i=1}^{p}\beta' P_i\cdot\frac{\sigma^4(n-p+2)}{n-p}\cdot\frac{1}{(\lambda_i\beta' P_iP_i'\beta)^2}\cdot P_i'\beta+o(\sigma^4) \\
&= \frac{n-p+2}{n-p}\sigma^4\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i^2\beta' P_iP_i'\beta}+o(\sigma^4).
\end{aligned}$$

Because

$$\hat{A}^2 = \text{diag}\Big(1-2\frac{\sigma^2m'Mm}{b_1^1(n-p)}+o(\sigma^2),\cdots,1-2\frac{\sigma^2m'Mm}{b_1^p(n-p)}+o(\sigma^2)\Big),$$

and $P'(X'X)^{-1}P = \Lambda^{-1}$, so

$$\begin{aligned}
E[\Delta_2] &= E[\sigma^2m'XPP'(X'X)^{-1}P\hat{A}^2P'(X'X)^{-1}PP'X'm] \\
&= E[\sigma^2m'XP\Lambda^{-1}\hat{A}^2\Lambda^{-1}P'X'm] \\
&= \sigma^2E[m'X(P_1,\cdots,P_p)\Lambda^{-1}\hat{A}^2\Lambda^{-1}(P_1,\cdots,P_p)'X'm] \\
&= \sigma^2E\left\{m'\Big[\sum_{i=1}^{p}XP_i(\frac{1}{\lambda_i^2}(1-2\frac{\sigma^2m'Mm}{b_1^i(n-p)}))P_i'X'\Big]m\right\}+o(\sigma^4) \\
&= \sigma^2E\Big[\sum_{i=1}^{p}m'\frac{XP_iP_i'X'}{\lambda_i^2}m-2\sigma^2\sum_{i=1}^{p}\Big(m'\frac{XP_iP_i'X'}{\lambda_i^2}m\cdot\frac{m'Mm}{b_1^i(n-p)}\Big)\Big]+o(\sigma^4),
\end{aligned}$$

where we further have

$$\begin{aligned}
E\Big[\sum_{i=1}^{p}m'\frac{XP_iP_i'X'}{\lambda_i^2}m\Big] &= \sum_{i=1}^{p}E\Big[m'\frac{XP_iP_i'X'}{\lambda_i^2}m\Big] \\
&= \sum_{i=1}^{p}\text{tr}\Big(\frac{XP_iP_i'X'}{\lambda_i^2}\Big) = \sum_{i=1}^{p}\frac{\text{tr}(XP_iP_i'X')}{\lambda_i^2} \\
&= \sum_{i=1}^{p}\frac{1}{\lambda_i}.
\end{aligned}$$

Given that $\lambda_i^{-2}MXP_iP_i'X' = 0$, $m'Mm$ and $\lambda_i^{-2}m'XP_iP_i'X'm$ are independent, then we obtain

$$\begin{aligned} E\Big[\sum_{i=1}^{p}\Big(m'\frac{XP_iP_i'X'}{\lambda_i^2}m\cdot\frac{m'Mm}{b_1^i(n-p)}\Big)\Big] &= \sum_{i=1}^{p}\Big[E\Big(m'\frac{XP_iP_i'X'}{\lambda_i^2}m\Big)\cdot E\Big(\frac{m'Mm}{b_1^i(n-p)}\Big)\Big] \\ &= \sum_{i=1}^{p}\Big[\frac{\operatorname{tr}(XP_iP_i'X')}{\lambda_i^2}\cdot\frac{1}{\lambda_i\beta'P_iP_i'\beta}\Big] \\ &= \sum_{i=1}^{p}\frac{1}{\lambda_i^2\beta'P_iP_i'\beta}. \end{aligned}$$

Thus

$$\begin{aligned} E[\Delta_2] &= \sigma^2E\Big[\sum_{i=1}^{p}m'\frac{XP_iP_i'X'}{\lambda_i^2}m - 2\sigma^2\sum_{i=1}^{p}\Big(m'\frac{XP_iP_i'X'}{\lambda_i^2}m\cdot\frac{m'Mm}{b_1^i(n-p)}\Big)\Big] + o(\sigma^4) \\ &= \sigma^2\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i} - 2\sigma^4\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i^2\beta'P_iP_i'\beta} + o(\sigma^4) \\ &= \operatorname{MSE}(\hat\beta) - 2\sigma^4\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i^2\beta'P_iP_i'\beta} + o(\sigma^4). \end{aligned}$$

Together, we obtain

$$\begin{aligned} \operatorname{MSE}(\hat\beta_{GS}(\hat A)) &= \sigma^2\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i} + \frac{n-p+2}{n-p}\sigma^4\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i^2\beta'P_iP_i'\beta} - 2\sigma^4\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i^2\beta'P_iP_i'\beta} + o(\sigma^4) \\ &= \sigma^2\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i} + \Big(\frac{n-p+2}{n-p} - 2\Big)\sigma^4\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i^2\beta'P_iP_i'\beta} + o(\sigma^4). \end{aligned}$$

Note that

$$\operatorname{MSE}(\hat\beta_{GS}(\hat A)) - \operatorname{MSE}(\hat\beta) = \Big(\frac{n-p+2}{n-p} - 2\Big)\sigma^4\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i^2\beta'P_iP_i'\beta} + o(\sigma^4).$$

When σ is sufficiently small, let the above equation be less than 0, then a sufficient condition for $\hat\beta_{GS}(\hat A)$ to be superior to $\hat\beta$ in terms of mean squares error criterion is

$$\frac{n-p+2}{n-p} - 2 < 0,$$

that is,

$$n > p + 2.$$

So, we come to the conclusion of Theorem 3.2.

The proof of Theorem 3.2 is complete. □

3.3. The matrix mean squares error

Theorem 3.3. *When σ is sufficiently small, a sufficient condition for the TSLS estimator to outperform the OLSE in terms of matrix mean squares error criterion is*

$$\frac{n-p+2}{n-p}\sum_{1\le i,j\le p}\frac{P_iP_i'\beta\beta'P_jP_j'}{\lambda_i\lambda_j\beta'P_iP_i'\beta\beta'P_jP_j'\beta}-\sum_{1\le i\le p}\frac{2P_iP_i'}{\lambda_i^2\beta'P_iP_i'\beta} \tag{3.4}$$

is non positive definite.

Proof. Since $\hat\beta_{GS}(\hat A)-\beta=P(\hat A-I)P'\beta+\sigma P\hat AP'(X'X)^{-1}X'm$, we have

$$\begin{aligned}\mathrm{MMSE}(\hat\beta_{GS}(\hat A))&=E[P(\hat A-I)P'\beta\beta'P(\hat A-I)P'+\sigma^2P\hat AP'(X'X)^{-1}X'mm'X(X'X)^{-1}P\hat AP'\\&\quad+2\sigma P(\hat A-I)P'\beta m'X(X'X)^{-1}P\hat AP']\\&=E[P(\hat A-I)P'\beta\beta'P(\hat A-I)P'+\sigma^2P\hat A\Lambda^{-1}P'X'mm'XP\Lambda^{-1}\hat AP'\\&\quad+2\sigma P(\hat A-I)P'\beta m'XP\Lambda^{-1}\hat AP']\\&\hat{=}E[\Delta^1+\Delta^2+\Delta^3].\end{aligned}$$

Note that

$$\begin{aligned}E[\Delta^3]&=E[2\sigma P(\hat A-I)P'\beta m'XP\Lambda^{-1}\hat AP']\\&=E[2\sigma P(\hat A-I)(P_1,\cdots,P_p)'\beta m'X(P_1,\cdots,P_p)\Lambda^{-1}\hat AP']\\&=2\sigma E\left[P(\hat A-I)\begin{pmatrix}P_1'\beta m'XP_1&\cdots&P_1'\beta m'XP_p\\\vdots&\ddots&\vdots\\P_p'\beta m'XP_1&\cdots&P_p'\beta m'XP_p\end{pmatrix}\Lambda^{-1}\hat AP'\right]\\&=2\sigma E\left[P\begin{pmatrix}\frac{(\hat a_1-1)\hat a_1}{\lambda_1}P_1'\beta m'XP_1&\cdots&\frac{(\hat a_1-1)\hat a_p}{\lambda_p}P_1'\beta m'XP_p\\\vdots&\ddots&\vdots\\\frac{(\hat a_p-1)\hat a_1}{\lambda_1}P_p'\beta m'XP_1&\cdots&\frac{(\hat a_p-1)\hat a_p}{\lambda_p}P_p'\beta m'XP_p\end{pmatrix}P'\right]\\&=2\sigma E\left[\sum_{1\le i,j\le p}\frac{(\hat a_i-1)\hat a_j}{\lambda_j}P_iP_i'\beta m'XP_jP_j'\right].\end{aligned}$$

By the fact that when $1\le i,j\le p$

$$MXP_iP_j=XP_iP_j-X(X'X)^{-1}X'XP_iP_j=0,$$

we know that $(\hat a_1-1)\hat a_1$ and $m'XP_iP_j'$ are independent. Hence

$$E\left[\frac{(\hat a_i-1)\hat a_j}{\lambda_j}P_iP_i'\beta m'XP_jP_j'\right]=E\left[\frac{(\hat a_i-1)\hat a_j}{\lambda_j}P_iP_i'\beta\right]E\left[m'XP_jP_j'\right]=0.$$

So we can obtain $E[\Delta^3]=0$.

Then, using the facts that

$$(\hat a_i-1)(\hat a_j-1)=\frac{\sigma^4(m'Mm)^2}{\lambda_i\lambda_j(n-p)^2\beta'P_iP_i'\beta\beta'P_jP_j'\beta}+o(\sigma^4)$$

and

$$E[(m'Mm)^2] = (n-p)(n-p+2).$$

Thus, we have

$$\begin{aligned}
E[\Delta^1] &= E[P(\hat{A}-I)P'\beta\beta'P(\hat{A}-I)P'] \\
&= E\Big[P(\hat{A}-I)\begin{pmatrix} P_1'\beta\beta'P_1 & \cdots & P_1'\beta\beta'P_p \\ \vdots & \ddots & \vdots \\ P_p'\beta\beta'P_1 & \cdots & P_p'\beta\beta'P_p \end{pmatrix}(\hat{A}-I)P'\Big] \\
&= E\Big[P\begin{pmatrix} (\hat{a}_1-1)(\hat{a}_1-1)P_1'\beta\beta'P_1 & \cdots & (\hat{a}_1-1)(\hat{a}_p-1)P_1'\beta\beta'P_p \\ \vdots & \ddots & \vdots \\ (\hat{p}_1-1)(\hat{1}_1-1)P_p'\beta\beta'P_1 & \cdots & (\hat{a}_p-1)(\hat{a}_p-1)P_p'\beta\beta'P_p \end{pmatrix}P'\Big] \\
&= E\Big[\sum_{1\le i,j\le p}(\hat{a}_i-1)(\hat{a}_j-1)P_iP_i'\beta\beta'P_jP_j'\Big] \\
&= E\Big[\sum_{1\le i,j\le p}\frac{\sigma^4(m'Mm)^2}{\lambda_i\lambda_j(n-p)^2\beta'P_iP_i'\beta\beta'P_jP_j'\beta}P_iP_i'\beta\beta'P_jP_j' + o(\sigma^4)\Big] \\
&= \frac{n-p+2}{n-p}\sum_{1\le i,j\le p}\frac{\sigma^4}{\lambda_i\lambda_j\beta'P_iP_i'\beta\beta'P_jP_j'\beta}P_iP_i'\beta\beta'P_jP_j' + o(\sigma^4).
\end{aligned}$$

Also

$$\hat{a}_i\hat{a}_j = 1 - \frac{\sigma^2 m'Mm}{n-p}\Big(\frac{\lambda_i\beta'P_iP_i'\beta + \lambda_j\beta'P_jP_j'\beta}{\lambda_i\lambda_j\beta'P_iP_i'\beta\beta'P_jP_j'\beta}\Big) + o(\sigma^2),$$

so

$$\begin{aligned}
E[\Delta^2] &= \sigma^2 P\hat{A}\Lambda^{-1}P'X'mm'XP\Lambda^{-1}\hat{A}P' \\
&= \sigma^2 E\Big[P\hat{A}\Lambda^{-1}\begin{pmatrix} P_1'X'mm'XP_1 & \cdots & P_1'X'mm'XP_p \\ \vdots & \ddots & \vdots \\ P_p'X'mm'XP_1 & \cdots & P_p'X'mm'XP_p \end{pmatrix}\Lambda^{-1}\hat{A}P'\Big] \\
&= \sigma^2 E\Big[P\begin{pmatrix} \frac{\hat{a}_1\hat{a}_1}{\lambda_1\lambda_1}P_1'X'mm'XP_1 & \cdots & \frac{\hat{a}_1\hat{a}_p}{\lambda_1\lambda_p}P_1'X'mm'XP_p \\ \vdots & \ddots & \vdots \\ \frac{\hat{a}_p\hat{a}_1}{\lambda_p\lambda_1}P_p'X'mm'XP_1 & \cdots & \frac{\hat{a}_p\hat{a}_p}{\lambda_p\lambda_p}P_p'X'mm'XP_p \end{pmatrix}P'\Big] \\
&= \sigma^2 E\Big[\sum_{1\le i,j\le p}\frac{\hat{a}_i\hat{a}_j}{\lambda_i\lambda_j}P_iP_i'X'mm'XP_jP_j'\Big].
\end{aligned}$$

Since

$$MX = X - X(X'X)^{-1}X'X = 0,$$

we conclude that $\hat{a}_i$ and $m'X$ are independent, accordingly, $\hat{a}_i$ and $X'mm'X$ are independent. Hence

$$
\begin{aligned}
E[\Delta^2] &= \sigma^2 E\Big[\sum_{1\le i,j\le p} \frac{\hat{a}_i\hat{a}_j}{\lambda_i\lambda_j} P_iP_i'X'mm'XP_jP_j'\Big] \\
&= \sigma^2 E\Big[\sum_{1\le i,j\le p} \frac{P_iP_i'X'XP_jP_j'}{\lambda_i\lambda_j} \\
&\quad - \sum_{1\le i,j\le p} \frac{P_iP_i'X'XP_jP_j'}{\lambda_i\lambda_j}\frac{\sigma^2 m'Mm}{n-p}\Big(\frac{\lambda_i\beta'P_iP_i'\beta + \lambda_j\beta'P_jP_j'\beta}{\lambda_i\lambda_j\beta'P_iP_i'\beta\beta'P_jP_j'\beta}\Big)\Big] + o(\sigma^4) \\
&= \mathrm{MMSE}(\hat{\beta}) - \sigma^4 \sum_{1\le i,j\le p} \frac{P_iP_i'X'XP_jP_j'}{\lambda_i\lambda_j}\Big(\frac{\lambda_i\beta'P_iP_i'\beta + \lambda_j\beta'P_jP_j'\beta}{\lambda_i\lambda_j\beta'P_iP_i'\beta\beta'P_jP_j'\beta}\Big)\Big] + o(\sigma^4) \\
&= \mathrm{MMSE}(\hat{\beta}) - \sigma^4 \sum_{1\le i\le p} \frac{P_iP_i'}{\lambda_i}\Big(\frac{2}{\lambda_i\beta'P_iP_i'\beta}\Big) + o(\sigma^4).
\end{aligned}
$$

Together, we obtain

$$
\begin{aligned}
\mathrm{MMSE}(\hat{\beta}_{GS}(\hat{A})) - \mathrm{MMSE}(\hat{\beta}) = \sigma^4\Big[&\frac{n-p+2}{n-p}\sum_{1\le i,j\le p}\frac{P_iP_i'\beta\beta'P_jP_j'}{\lambda_i\lambda_j\beta'P_iP_i'\beta\beta'P_jP_j'\beta} \\
&- \sum_{1\le i\le p}\frac{2P_iP_i'}{\lambda_i^2\beta'P_iP_i'\beta}\Big] + o(\sigma^4).
\end{aligned}
$$

When σ is sufficiently small, let the above equation be less than 0, then a sufficient condition for $\hat{\beta}_{GS}(\hat{A})$ to be superior to $\hat{\beta}$ in terms of MMSE criterion is

$$
\frac{n-p+2}{n-p}\sum_{1\le i,j\le p}\frac{P_iP_i'\beta\beta'P_jP_j'}{\lambda_i\lambda_j\beta'P_iP_i'\beta\beta'P_jP_j'\beta} - \sum_{1\le i\le p}\frac{2P_iP_i'}{\lambda_i^2\beta'P_iP_i'\beta} \le 0.
$$

□

The condition of the above theorem has a simpler form. The condition $n > p + 2$ is actually implied here.

Note that

$$
\begin{aligned}
&\frac{n-p+2}{n-p}\sum_{1\le i,j\le p}\frac{P_iP_i'\beta\beta'P_jP_j'}{\lambda_i\lambda_j\beta'P_iP_i'\beta\beta'P_jP_j'\beta} \\
&= \frac{n-p+2}{n-p}P\begin{pmatrix}
\frac{1}{\lambda_1^2\beta'P_1P_1'\beta} & \cdots & \frac{P_1'\beta\beta'P_p}{\lambda_1\lambda_p\beta'P_1P_1'\beta\beta'P_pP_p'\beta} \\
\vdots & \ddots & \vdots \\
\frac{P_p'\beta\beta'P_1}{\lambda_p\lambda_1\beta'P_pP_p'\beta\beta'P_1P_1'\beta} & \cdots & \frac{1}{\lambda_p^2\beta'P_pP_p'\beta}
\end{pmatrix}P'
\end{aligned}
$$

and

$$2\sum_{1\le i\le p}\frac{P_iP_i'}{\lambda_i^2\beta'P_iP_i'\beta} = 2P\begin{pmatrix}\frac{1}{\lambda_1^2\beta'P_1P_1'\beta} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \frac{1}{\lambda_p^2\beta'P_pP_p'\beta}\end{pmatrix}P',$$

so the condition of Theorem 3.3 is equivalent to

$$\begin{pmatrix}\left(\frac{n-p+2}{n-p}-2\right)\frac{1}{\lambda_1^2\beta'P_1P_1'\beta} & \cdots & \frac{n-p+2}{n-p}\frac{P_1'\beta\beta'P_p}{\lambda_1\lambda_p\beta'P_1P_1'\beta\beta'P_pP_p'\beta}\\ \vdots & \ddots & \vdots\\ \frac{n-p+2}{n-p}\frac{P_p'\beta\beta'P_1}{\lambda_p\lambda_1\beta'P_pP_p'\beta\beta'P_1P_1'\beta} & \cdots & \left(\frac{n-p+2}{n-p}-2\right)\frac{1}{\lambda_p^2\beta'P_pP_p'\beta}\end{pmatrix}\le 0. \quad (3.5)$$

If the inequality (3.5) holds, we conclude

$$\left(\frac{n-p+2}{n-p}-2\right)\frac{1}{\lambda_1^2\beta'P_1P_1'\beta}\le 0,$$

which is just the condition of Theorem 3.2.

4. Numerical simulations

In this section, some numerical simulations are conducted to demonstrate the performances of the TSLS estimator and test the superiority condition, which are designed as follows:

Let the model be

$$Y = 1 + 2X_1 + 3X_2 + 5X_3 + 4X_4 + e,$$

where $e \sim N(0,\sigma)$, $X_1 \sim N(0,\sigma_0)$, $X_2 \sim N(1,\sigma_0)$, $X_3 \sim N(2,\sigma_0)$, $X_4 = 0.5X_1 + 3X_2 + e_1$, $e_1 \sim N(0,\Omega)$.

Among the above parameters, Ω is used to control the degree of multi-collinearity that exists among X_4, X_1, X_2, which becomes greater as Ω gets smaller. And n is the number of observations and σ and σ_0 denotes the random error.

Experiment 1: To compare the performances of TSLS estimator with some estimator at different levels of multi-collinearity.

Set $n = 500$, $\sigma = 1$ and number of simulations $iter = 100$, the sample mean squares error $\sum_{i=1}^{iter}\|\hat{\beta}^i - \beta\|^2/(iter-5)$ will be used in this experiment as a proxy for the mean squares error. where $\hat{\beta}^i$ denotes the estimator obtained from the i-th simulation, and then vary the size of Ω to observe the performance of the estimators.

Experiment 2: To verify TSLS estimator is superior when the σ is sufficiently small.

Set $n = 500$, $\Omega = 0.001$ and number of simulations $iter = 100$, and then vary the size of σ to compare the mean squares error of the estimators.

Table 1. $\sigma = 1$.

Ω	MSE_TSLS	MSE_OLSE	MSE_Stein	MSE_Ridge
1	0.0482	0.0470	0.6144	0.0470
0.1	1.8181	1.9303	2.1829	1.9296
0.01	84.1752	171.6564	138.7557	164.8265
0.001	10670.4812	22875.7016	18536.0625	2475.2325

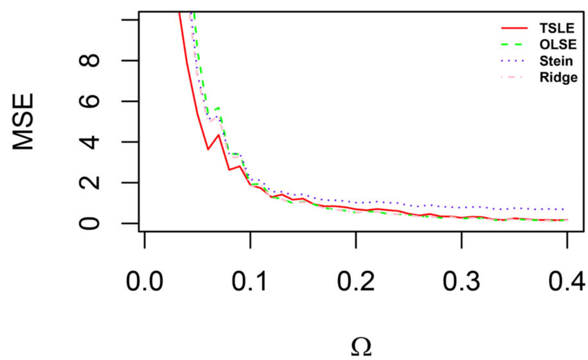

Figure 1. MSE while $\sigma = 1$.

Table 2. $\Omega = 0.001$.

	MSE_TSLS	MSE_OLS	MSE_Stein	MSE_Ridge
0.01	1.8339	2.0351	2.2318	1.9338
0.1	73.9481	174.3933	142.1233	20.5975
1	13418.4613	25012.8790	20261.1433	2640.4777
10	947655.5705	2105286.5463	1705291.1071	225292.6000

Finally two experiments produce the following results:

From Table 1, we can see that as Ω decreases and the multi-collinearity increases, both the mean squares errors of the TSLS estimator and the other estimators increase significantly, while the advantage of the TSLS estimator becomes more obvious, and this is also visualized in Figure 1. Table 2 then verifies the conclusion that the TSLS estimator performs the best when σ is sufficiently small, that is, in the case that σ is 0.01 or less in this experiment. For different cases, the requirements for "sufficiently small" will be different, but for most models, it is basically guaranteed "σ is sufficiently small".

5. Real data analysis

This section uses real data to illustrate the superiority of TSLS estimator in comparison to other estimators.

5.1. Boston housing data

The data used in this subsection is named Boston Housing (see Harrison and Rubinfeld 1978), which is a set of 506 rows and 14 columns of data on Boston house prices and some basic social information. We can find the data form package "MASS" in the R software. This data set contains the following columns: crim (per capita crime rate by town), zn (proportion of

Table 3. Comparisons under Boston housing data.

TSLS	OLSE	Stein	Ridge
3.194	3.355	3.275	3.337

Table 4. Head of Hartnagel.

	Year	tfr	Partic	Degrees	fconvict	ftheft	mconvict	mtheft
1	1931	3200	234	12.4000	77.1000		778.7000	
2	1932	3084	234	12.9000	92.9000		745.7000	
3	1933	2864	235	13.9000	98.3000		768.3000	
4	1934	2803	237	13.6000	88.1000		733.6000	
5	1935	2755	238	13.2000	79.4000	20.4000	765.7000	247.1000
6	1936	2696	240	13.2000	91.0000	22.1000	816.5000	254.9000

residential land zoned for lots over 25,000 sq.ft.), indus (proportion of non retail business acres per town.), chas (Charles River dummy variable (= 1 if tract bounds river; 0 otherwise), nox (nitrogen oxides concentration (parts per 10 million)), rm (average number of rooms per dwelling), age (proportion of owner-occupied units built prior to 1940), dis (weighted mean of distances to five Boston employment centers), rad (index of accessibility to radial highways), tax (full-value property-tax rate per \$10,000), ptratio (pupil-teacher ratio by town), black $(1000(Bk - 0.63)^2$, where Bk is the proportion of blacks by town), lstat (lower status of the population (percent)), and medv (median value of owner-occupied homes in \$1000s). The dependent variable is medv, and the independent variables are crim, zn, indus, chas, nox, rm, age, dis, rad, tax, ptratio, black, and lstat.

Then we use the first 406 rows of the data as the regression group and the last 100 rows as the test group. The TSLS estimator and the other estimators are calculated using the data of regression group. The estimated medv is then obtained by substituting the data from the test group, summing the square of the difference with the true value, and comparing the magnitude of the values. The results are given in Table 3.

From the results, we can see that the TSLS estimator is superior to the OLSE, which corresponds to the conclusion of this article. And compared to other estimators the TSLS estimator will be a good choice.

5.2. Hartnagel data

So far, we have done a real data analysis that illustrates the superiority of TSLS estimator. And here, we will choose a small data set with a comparatively large multi-collinearity problem to see how TSLS performs. The data used here is named Hartnagel (see Fox and Hartnagel 1979) and is derived from the data package "car" in the R software and is structured as follows in Table 4.

The data consists of a total of 38 rows and 7 columns of time series data on crime rates in Canada from 1931 to 1968, with a few missing data. This data set contains the following columns: year (1931–1968), tfr (Total fertility rate per 1000 women), partic (Women's labor-force participation rate per 1000), degrees (Women's post-secondary degree rate per 10,000), fconvict (Female indictable-offense conviction rate per 100,000), ftheft (Female theft conviction rate per 100,000), mconvict (Male indictable-offense conviction rate per 100,000),

Table 5. Comparisons under Hartnagel data.

TSLS	OLSE	Stein	Ridge
62510.91	78330.01	41106.06	78331.32

and mtheft (Male theft conviction rate per 100,000). The dependent variable is tfr, and the independent variables are partic, degrees, fconvict, ftheft, mconvict, and mtheft.

Here, the missing data from the first four years are removed, and then the remaining 34 years of data are used as the regression group for the first 30 years and as the test group for the last four years. The OLSE and the TSLS estimator are calculated using the data from the first 30 years. The estimated tfr is then obtained by substituting the data from the test group, summing the square of the difference with the true value, and comparing the magnitude of the values. The results are given in Table 5.

From the results, we can see that the TSLS estimator performs better than the OLSE in the situation of small data set, and even compared to other biased estimators it is also an option worth considering.

6. Conclusions

In this article, a theoretical derivation gives a condition for the superiority of the newly introduced two-stage shrunken least squares estimator in terms of mean squares error (MSE) criterion, that is, when σ is sufficiently small, if $n > p + 2$ then the two-stage shrunken least squares estimator is superior to the ordinary least squares estimator (OLSE). This conclusion suggests that the sufficient conditions for the two-stage shrunken least squares estimator to outperform the OLSE under the MSE criterion are only related to n and p when σ is sufficiently small, which is fairly easy to determine in practice. Under the assumptions of this article, the above sufficient conditions that "$n > p + 2$" and "σ is sufficiently small" are basically satisfied. In other words, for the multiple linear model the two-stage shrunken least squares estimator is superior to the OLSE in terms of MSE criterion in most situations. Also we discuss the superiority of the two-stage shrunken least squares estimator under the matrix mean squares error (MMSE) criterion. Compared to the MSE criterion, the corresponding conditions are a bit more complex, but it also requires the condition that "$n > p + 2$". With respect to the MMSE criterion, the readers can refer to the conclusion in Wang (1990), which states that $\mathrm{MMSE}(\hat{\beta}_{GS}(\hat{A})) \leq \mathrm{MMSE}(\hat{\beta})$ if and only if $\beta' P(I - \hat{A})(I + \hat{A})^{-1} P' \beta \leq \sigma^4$. The proposed sufficient conditions are fairly easy to determine and have a little Bayesian flavor in practice. At the same time, the obtained results enrich and extend the study on biased estimation for the multiple linear models.

Acknowledgments

The authors thank the editor and the two reviewers for their comments and suggestions which led to substantial improvement of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Supported by NNSF of China (11371051).

References

Duan, Q. T. 1999. A method of choosing parameters of the generalized shrunken least squares estimate. *Journal of Zhengzhou Institute of Light Industry* 14 (4):75–7.

Fox, J., and T. F. Hartnagel. 2008. Changing social roles and female crime in Canada: A time series analysis*. *Canadian Review of Sociology/revue Canadienne De Sociologie* 16 (1):96–104. doi:10.1111/j.1755-618X.1979.tb01012.x.

Guo, J. L., and F. X. Guo. 1997. The Q(c) criterion of choosing parameter about the generalized shrunken least squares estimate. *Journal of Fuzhou University (Natural Science)* 25 (3):1–7.

Harrison, D., and D. L. Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 5 (1):81–102. doi:10.1016/0095-0696(78)90006-2.

Hoerl, A. E., and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12 (1):55–67. doi:10.1080/00401706.1970.10488634.

Kendall, M. G. 1957. *A course in multivariate analysis*. Royal Statistical Society: Series A (General) 121 (4):480–1.

Shi, J. H. 1999. The multiple k-class generalized shrunken least squares estimator. *Journal of Shanxi Teacher's University Natural Science Edition* 13 (4):8–11.

Stein, C. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1:197–206.

Sun, X. Q. 1997. Advantages of generalized shrunken least square estimator. *Journal of Suzhou Railway Teachers College* 14 (4):5–9.

Sun, X. Q. 1999. Multivariate generalized shrunken least square estimate. *Journal of Foshan University* 14 (2):17–22.

Trenkler, G. 1981. *Biased estimators in the linear regression model*. Cambridge, Massachusetts: Oelgeschlager, Gunn & Hail Publisher, Inc.

Wang, L. Q. 1990. Generalized shrunken least squares estimators. *Chinese Journal of Applied Probablilty and Statistics* 6 (3):225–32.

Yang, H. 1991. Universal ridge estimates on the regression coefficient. *Journal of Chongqing Jiaotong Institute* 10 (3):42–8.

Zhao, Z. M. 1995. A note on the generalized shrunken least squares estimator. *Methematica Applicata* 8 (1):90–5.