Available online: January 4, 2019

Commun. Fac. Sci. Univ. Ank. Ser. A1 Math. Stat.
Volume 68, Number 1, Pages 1123–1143 (2019)
DOI: 10.31801/cfsuasmas.508223
ISSN 1303–5991 E-ISSN 2618-6470
http://communications.science.ankara.edu.tr/index.php?series=A1

REVIVING SOME GEOMETRIC ASPECTS OF SHRINKAGE ESTIMATION IN LINEAR MODELS

FIKRI AKDENIZ AND FIKRI ÖZTÜRK

Abstract. It is well known that the least squares estimator is the best linear unbiased estimator of the parameter vector in a classical linear model. But, it is 'too long' as a vector and unreliable, confidence intervals are broad for some components especially in the case of multicollinearity. Shrinkage (contraction) type estimators are efficient remedial tools in order to solve problems caused by multicollinearity. In this study, we consider a class of componentwise shrunken estimators with typical members: Mayer and Willke's contraction estimator, Marquardt's principal component estimator, Hoerl and Kennard's ridge estimator, Liu's linear unified estimator and a discrete shrunken estimator. All estimators considered are "shorter" than the least squares estimator with respect to the Euclidean norm, biased, but insensitive to multicollinearity and admissible within the set of linear estimators with respect to unweighted squared error risk. Some behaviors of these estimators are illustrated geometrically by tracing their trajectories as functions of shrinkage factors in a two-dimensional parameter space.

1. Introduction

Consider the classical linear model

$$y = X\beta + \varepsilon \tag{1.1}$$

where the nx1 vector y is an observable random response vector, β is a $p \times 1$ vector of unknown coefficients, X is a non-stochastic $n \times p$ matrix of the explanatory variables (observed or designed) with full column rank, and ε is a non-observable $n \times 1$ random vector satisfying the assumptions $E(\varepsilon) = 0$,$Cov(\varepsilon) = \sigma^2\mathbf{I}_n$. Remind that individual parameters have their own special meanings in the constructed model.

Let σ^2 be a nuisance parameter and let our primary interest be the estimation of the parameter vector β in the parameter space R^p ($\beta \in R^p$). As is well known,

Received by the editors: January 09, 2018; Accepted: April 11, 2018.
2010 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18.
Key words and phrases. Contraction estimator, Liu estimator, principal component estimator, Ridge regression.

from the Gauss-Markov Theorem

$$\hat{\beta} = (X'X)^{-1}X'y$$

is the best linear unbiased (BLU) estimator of β. The BLU estimator of $X\beta$

$$\widehat{X\beta} = X\hat{\beta} = X(X'X)^{-1}X'y$$

is the orthogonal projection of the observed vector y on the estimation space $[X]$ (column space of X), where the projection matrix is $P_{[X]} = X(X'X)^{-1}X'$. On the other side,

$$\arg\min_{\beta \in R^p} (y - X\beta)'(y - X\beta) = \hat{\beta}$$

is the so called Least Squares (LS) estimator of β, denoted by $\hat{\beta}_{LS}$. Additionally, when ε is normally distributed the LS estimator is the Maximum Likelihood estimator, which is the Uniformly Minimum Variance Unbiased estimator of β. This estimator performs poorly in the presence of multicollinearity. When $X'X$ is ill conditioned, that is, when the condition number $k = \lambda_{\max}/\lambda_{\min}$ is large with a small $\lambda_{\min}$ than the risk of $\hat{\beta}_{LS}$

$$\begin{aligned} MSE(\hat{\beta}_{LS}, \beta) &= E((\hat{\beta}_{LS} - \beta)'(\hat{\beta}_{LS} - \beta)) \\ &= E(\left\|\hat{\beta}_{LS} - \beta\right\|^2) \\ &= tr(Cov(\hat{\beta}_{LS})) \\ &= \sum_{i=1}^{p} \frac{\sigma^2}{\lambda_i} \\ &\geq \frac{\sigma^2}{\lambda_{\min}} \end{aligned}$$

becomes relatively large. Here, $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = \lambda_{\min} > 0$ are the eigenvalues of $X'X$ having the spectral decomposition

$$X'X = U\Lambda U',\ \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p),\ U'U = UU' = I_p,\ (X'X)^{-1} = U\Lambda^{-1}U'.$$

Being optimal, i.e. minimum variance in the class of linear unbiased estimators, the $\hat{\beta}_{OLS}$ may fall far away from β, sometimes with wrong sign for an individual component. Also, the expected Euclidean norm of $\hat{\beta}_{OLS}$ is greater than the norm

of β.

$$\begin{aligned} E\left(\left\|\hat{\beta}_{LS}\right\|\right) &= E\left(\sqrt{\hat{\beta}'_{LS}\hat{\beta}_{LS}}\right) \\ &\geq \sqrt{E(\hat{\beta}'_{LS}\hat{\beta}_{LS})} \quad (Jensen's\ Inequality) \\ &= \sqrt{\sigma^2 tr\left((X'X)^{-1}\right) + \beta'\beta} \\ &\geq \sqrt{\beta'\beta} \\ &= \|\beta\| \end{aligned}$$

The LS estimator $\hat{\beta}_{OLS}$ is 'too long' for β. On the other hand, the LS estimator can uniformly be outperformed with respect to some other criterion. For example, the James-Stein's estimator (James and Stein, 1961)

$$\begin{aligned} \hat{\beta}_{JS} &= \left[1 - \frac{(p-2)(y - X\hat{\beta}_{LS})'(y - X\hat{\beta}_{LS})}{(n-p+2)\hat{\beta}'_{LS}X'X\hat{\beta}_{LS}}\right]\hat{\beta}_{LS} \\ &= \left[1 - \frac{\left\|y - X\hat{\beta}_{LS}\right\|^2/(n-p+2)}{\left\|X\hat{\beta}_{LS}\right\|^2/(p-2)}\right]\hat{\beta}_{LS} \end{aligned}$$

is a uniformly better estimator than the LS estimator with respect to weighted mean squared error criterion

$$\begin{aligned} WMSE(\hat{\beta},\beta) &= E\left((\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)\right) \\ &= E\left(\left\|X\hat{\beta} - X\beta\right\|^2\right) \end{aligned}$$

for normally distributed errors and $p \geq 3$ (Gross, 2003). Note that, the James-Stein's estimator is a non-linear biased estimator of β.

In this study the estimation criterion is the Mean Squared Error (MSE) criterion, where

$$\begin{aligned} MSE(\hat{\beta},\beta) &= E((\hat{\beta}-\beta)'(\hat{\beta}-\beta)) = E(\left\|\hat{\beta}-\beta\right\|^2) \\ &= tr(Cov(\hat{\beta})) + Bias(\hat{\beta})'Bias(\hat{\beta}). \end{aligned}$$

is the risk with respect to scalar unweighted squared error loss.

By suitable reparametrization, the original model (1.1) can be reduced to an orthonormal form. The reparametrized model

$$y = V\gamma + \varepsilon \ , \quad V = XU\Lambda^{-1/2}\,,\ \gamma = \Lambda^{1/2}U'\beta \tag{1.2}$$

is in orthogonal form, $V'V = I$. The LS estimator

$$\hat{\gamma}_{OLS} = (V'V)^{-1}V'y = V'y$$

can uniformly be outperformed by the James-Stein's estimator

$$\hat{\gamma}_{JS} = \left[1 - \frac{(p-2)(y - V\hat{\gamma}_{LS})'(y - V\hat{\gamma}_{LS})}{(n-p+2)\hat{\gamma}'_{LS}\hat{\gamma}_{LS}}\right]\hat{\gamma}_{LS}$$

with respect to mean squared error criterion $MSE(\hat{\gamma}, \gamma) = E\left((\hat{\gamma} - \gamma)'(\hat{\gamma} - \gamma)\right)$. Under certain conditions the LS estimator is inadmissible with respect to MSE criterion. There are better estimators.

In the following section we consider some shrinkage estimators obtained by componentwise shrinking of the LS estimator. All estimators considered are biased with a small total variance then the LS estimator and are admissible in the set of linear estimators. In section 3 the geometry of these estimators is illustrated in the parameter space. Last section contains the concluding remarks.

2. Some Shrunken Estimators

The parameter set of the vector of unknown coefficients in the linear model (1.1) is the p-dimensional Euclidian space R^p ($\beta \in R^p$). Let us consider the family of componentwise shrunken estimators

$$\mathcal{CS}(\hat{\beta}) = \left\{\hat{\beta}_{\mathcal{CS}}(D) : \ \hat{\beta}_{\mathcal{CS}}(D) = D\hat{\beta}_{LS}\,,\ D = diag(d_1, d_2, ..., d_p)\,,\ d_1, d_2, ..., d_p \in [0,1]\right\}.$$

For example, in the 2-dimensional case, the estimates lie in the rectangle defined by the origin and the LS estimate, as opposite corners. All of the estimators are biased, except the $\hat{\beta}_{\mathcal{CS}}(I) = \hat{\beta}_{LS}$. For any estimator $\hat{\beta}_{\mathcal{CS}}(D)$ in $\mathcal{CS}(\hat{\beta})$

$$\left\|\hat{\beta}_{\mathcal{CS}}(D)\right\| \leq \left\|\hat{\beta}_{LS}\right\|.$$

The MSE of $\hat{\beta}_{\mathcal{CS}}(D)$ is

$$\begin{aligned}
MSE(\hat{\beta}_{\mathcal{CS}}(D), \beta) &= E\left((\hat{\beta}_{\mathcal{CS}}(D) - \beta)'(\hat{\beta}_{\mathcal{CS}}(D) - \beta)\right) \\
&= tr\left(Cov(\hat{\beta}_{\mathcal{CS}}(D))\right) + Bias\left(\hat{\beta}_{\mathcal{CS}}(D)\right)' Bias\left(\hat{\beta}_{\mathcal{CS}}(D)\right) \\
&= tr\left(DCov(\hat{\beta}_{LS})D'\right) + (D\beta - \beta)'(D\beta - \beta) \\
&= tr(\sigma^2 DU\Lambda^{-1}U'D') + \beta'(D - I)^2\beta \\
&= \sigma^2 \sum_{j=1}^{p}(d_j^2 \sum_{i=1}^{p} \frac{u_{ji}^2}{\lambda_i}) + \sum_{j=1}^{p}(d_j - 1)^2\beta_j^2
\end{aligned}$$

with a minimum at

$$d_j^{opt} = \frac{1}{1 + \frac{\sigma^2 \sum_{i=1}^{p} \frac{u_{ji}^2}{\lambda_i}}{\beta_j^2}} \quad , \quad j = 1, 2, ..., p$$

values of shrinkage (contraction) factors $d_1, d_2, ..., d_p \in [0,1]$. The optimal MSE estimator is

$$\hat{\beta}_{\mathcal{CS}}(D^{opt}) = D^{opt}\hat{\beta}_{OLS}$$

where $D^{opt} = diag(d_1^{opt}, d_2^{opt}, ..., d_p^{opt})$.

Now, let us write the model (1.1) in the canonical form

$$y = Z\alpha + \varepsilon \ , \ Z = XU \ , \ \alpha = U'\beta \tag{2.1}$$

where, $Z'Z = \Lambda$. This is another reparameterized model. The LS estimator for α is

$$\hat{\alpha}_{LS} = \Lambda^{-1}Z'y$$

and $E\left(\|\hat{\alpha}_{LS}\|\right) \geq \|\alpha\|$. Consider the family of componentwise shrunken estimators

$$\mathcal{CS}(\hat{\alpha}) = \{\hat{\alpha}_{\mathcal{CS}}(B): \ \hat{\alpha}_{\mathcal{CS}}(B) = B\hat{\alpha}_{LS}\,,\, B = diag(b_1, b_2, ..., b_p)\,,\, b_1, b_2, ..., b_p \in [0,1]\}\,.$$

All estimators are biased, except the LS estimator $\hat{\alpha}_{\mathcal{CS}}(I) = \hat{\alpha}_{OLS}$ and for any estimator in this family $\|\hat{\alpha}_{\mathcal{CS}}(B)\| \leq \|\hat{\alpha}_{OLS}\|$. The MSE of $\hat{\alpha}_{\mathcal{CS}}(B)$ is

$$\begin{aligned} MSE(\hat{\alpha}_{\mathcal{CS}}(B), \alpha) &= E\left((\hat{\alpha}_{\mathcal{CS}}(B) - \alpha)'(\hat{\alpha}_{\mathcal{CS}}(B) - \alpha)\right) \\ &= tr\left(Cov(\hat{\alpha}_{\mathcal{CS}}(B))\right) + \left(E(\hat{\alpha}_{\mathcal{CS}}(B)) - \alpha\right)'\left(E(\hat{\alpha}_{\mathcal{CS}}(B)) - \alpha\right) \\ &= tr\left(\sigma^2 B\Lambda^{-1}B'\right) + \alpha'(B-I)^2\alpha \\ &= \sigma^2\sum_{i=1}^{p} b_i^2/\lambda_i + \sum_{i=1}^{p}(b_i-1)^2\alpha_i^2 \end{aligned}$$

with minimum at

$$b_i^{opt} = \frac{1}{1+\frac{\sigma^2/\lambda_i}{\alpha_i^2}} = \frac{\lambda_i}{\lambda_i + \frac{\sigma^2}{\alpha_i^2}} \ , \quad i = 1, 2, ..., p.$$

The optimal MSE estimator in $\mathcal{CS}(\hat{\alpha})$ is $\hat{\alpha}_{\mathcal{CS}}(B^{opt})$, where

$$B^{opt} = diag(b_1^{opt}, b_2^{opt}, ..., b_p^{opt}).$$

Let us notice that the estimation spaces for the models (1.1), (2.2-below) and (2.1) are the same. As a result, the "blue response" $\hat{y}$ (BLU estimator of the response mean $E(y) = X\beta$)

$$\hat{y} = P_{[X]}y = P_{[Z]}y = P_{[V]}y$$

is invariant under these reparametrizations. Note also the following correspondency between linear models (1.1) and (2.1).

$$\begin{aligned}
\beta &= U\alpha \quad , \quad \alpha = U'\beta \quad , \quad \|\beta\| = \|\alpha\| \\
\hat{\beta}_{LS} &= U\hat{\alpha}_{LS} \quad , \quad \hat{\alpha}_{LS} = U'\hat{\beta}_{LS} \quad , \quad \left\|\hat{\beta}_{LS}\right\| = \|\hat{\alpha}_{LS}\| \\
MSE(\hat{\beta}_{LS}, \beta) &= MSE(\hat{\alpha}_{LS}, \alpha) = \sum_{i=1}^{p} \frac{\sigma_2}{\lambda_i} \\
\hat{\beta}_{\mathcal{CS}}(D) &= DUB^{-1}\hat{\alpha}_{\mathcal{CS}}(B) \quad , \quad \hat{\alpha}_{\mathcal{CS}}(B) = BU'D^{-1}\hat{\beta}_{\mathcal{CS}}(D) \\
\hat{\beta}_{\mathcal{CS}}(D) &= DUD^{-1}\hat{\alpha}_{\mathcal{CS}}(D) \quad , \quad \hat{\alpha}_{\mathcal{CS}}(D) = DU'D^{-1}\hat{\beta}_{\mathcal{CS}}(D) \\
\left\|\hat{\beta}_{\mathcal{CS}}(D)\right\| &= \|\hat{\alpha}_{\mathcal{CS}}(D)\| \leq \left\|\hat{\beta}_{LS}\right\| = \|\hat{\alpha}_{LS}\| \\
MSE(\hat{\beta}_{\mathcal{CS}}(D), \beta) &= MSE(\hat{\alpha}_{\mathcal{CS}}(D), \alpha) \\
MSE(\hat{\beta}_{\mathcal{CS}}(D), \beta) &= \sigma^2 \sum_{i=1}^{p} d_i^2/\lambda_i + \sum_{i=1}^{p} (d_i - 1)^2 \beta_i^2 \\
MSE(\hat{\alpha}_{\mathcal{CS}}(B), \alpha) &= \sigma^2 \sum_{i=1}^{p} b_i^2/\lambda_i + \sum_{i=1}^{p} (b_i - 1)^2 \alpha_i^2
\end{aligned}$$

Estimators in $\mathcal{CS}(\hat{\beta})$ and $\mathcal{CS}(\hat{\alpha})$ have smaller total variance then the LS estimator. It is necessary to provide some assurance that the benefit of reduced total variance is not likely to be offset by a large squared bias. The MSE's depend not only on shrinkage factors and the eigenvalues of $X'X$, but also on the unknown parameters of the model, so the optimal estimators cannot be used in practice. The shrinkage factors can be treated as parametric functions, also called shrinkage parameters or biasing (tuning) paremeters. To make the estimators operational some estimates of the shrinkage parameters can be used. It seems reasonable to use plug-in estimates using $\hat{\beta}_{LS}$, $\hat{\alpha}_{LS}$ and $\hat{\sigma}^2$, where

$$\hat{\sigma}_2 = \frac{(y - X\hat{\beta}_{LS})'(y - X\hat{\beta}_{LS})}{n - p} = \frac{\left\|y - X\hat{\beta}_{LS}\right\|^2}{n - p} = \frac{\|y - Z\hat{\alpha}_{LS}\|^2}{n - p}.$$

2.1. **Meyer and Wilke's (MW) Estimator.** Meyer and Wilke (1973) proposed the following contraction estimator.

$$\hat{\beta}_c = c\hat{\beta}_{OLS} \quad , \quad c \in [0, 1] \tag{2.2}$$

$\hat{\beta}_c = \hat{\beta}_{\mathcal{CS}}(cI)$ is the simplest shrinkage estimator in $\mathcal{CS}(\hat{\beta})$.

$$MSE\left(\hat{\beta}_{\mathcal{CS}}(cI)\right) = c^2\sigma^2 tr((X'X)^{-1}) + (c - 1)^2\beta'\beta$$

A necessary and sufficient condition for $\hat{\beta}_{\mathcal{CS}}(cI)$to have a smaller MSE than $\hat{\beta}_{LS}$ is that the following inequality

$$1 - \frac{2\sigma^2 tr((X'X)^{-1})}{\beta'\beta + \sigma^2 tr((X'X)^{-1})} < c < 1$$

is satisfied. The optimal value for $MSE\left(\hat{\beta}_{\mathcal{CS}}(cI)\right)$ is obtained at

$$c^{opt} = \frac{\beta'\beta}{\beta'\beta + \sigma^2 tr((X'X)^{-1})}.$$

The simplest shrinkage estimator in $\mathcal{CS}(\hat{\alpha})$ is $\hat{\alpha}_{\mathcal{CS}}(cI) = U'\hat{\beta}_{\mathcal{CS}}(cI)$. Note that

$$\begin{aligned}
\hat{\beta}_{\mathcal{CS}}(cI) &= c\hat{\beta}_{LS} = cU\hat{\alpha}_{LS} \quad , \quad \hat{\alpha}_{\mathcal{CS}}(cI) = c\hat{\alpha}_{LS} = cU'\hat{\beta}_{LS} \ , \ c \in [0,1] \\
MSE(\hat{\beta}_{\mathcal{CS}}(cI), \beta) &= \sigma^2 c^2 \sum_{i=1}^{p} 1/\lambda_i \ + (c-1)^2 \sum_{i=1}^{p} \beta_i^2 \\
MSE(\hat{\alpha}_{\mathcal{CS}}(cI), \alpha) &= \sigma^2 c^2 \sum_{i=1}^{p} 1/\lambda_i \ + (c-1)^2 \sum_{i=1}^{p} \alpha_i^2 \\
MSE(\hat{\beta}_{\mathcal{CS}}(cI), \beta) &= MSE(\hat{\alpha}_{\mathcal{CS}}(cI), \alpha).
\end{aligned}$$

2.2. **Principal Component (PC) Estimator.** As is known, the principal component (PC) analysis is a widely used dimension reduction technique, with some information being lost. The PC estimation procedure is a dimension reduction for the "cloud of LS estimates" in the parameter space R^p by projecting them orthogonally on the space spanned by the eigenvectors corresponding to the first largest r $(r < p)$ eigenvalues of $X'X$. The PC estimator remove the information which is responsible for the increase of impreciseness in LS estimation.

Let

$$\begin{gathered}
U_1 = [u_1, u_2, ..., u_r], \ U_2 = [u_{r+1}, u_{r+2}, ..., u_p], \ U = [U_1 \, U_2], \\
\Lambda_1 = diag(\lambda_1, \lambda_2, ..., \lambda_r), \ \Lambda_2 = diag(\lambda_{r+1}, \lambda_{r+2}, ..., \lambda_p), \\
\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}.
\end{gathered}$$

The orthogonal projection matrix on the column space $[U_1] = span(u_1, u_2, ..., u_r)$ of U_1 is $P_{[U_1]} = U_1(U_1'U_1)^{-1}U_1' = U_1U_1'$. So, the PC estimator is

$$\begin{aligned}
\hat{\beta}_{PC}(r) &= P_{[U_1]}\hat{\beta}_{LS} \\
&= U_1U_1'(X'X)^{-1}X'y \\
&= U_1U_1'\left(U_1\Lambda_1^{-1}U_1' + U_2\Lambda_2^{-1}U_2'\right)X'y \\
&= U_1\Lambda_1^{-1}U_1'X'y.
\end{aligned}$$

It can be shown that the PC estimator is the restricted LS estimator under the restriction $U_2'\beta = 0$. A necessary and sufficient condition for $\hat{\beta}_{PC}(r)$ to have a smaller MSE than $\hat{\beta}_{LS}$ is that the inequality $\beta'U_2\Lambda_2U_2'\beta \leq \sigma^2$ is satisfied. When

applying the PC estimator, one is faced with the problem of choosing an appropriate r value. In the normal case a pretest estimation procedure can be used.

The PC estimator has a very simple interpretation in the reparametrized model with canonical form.

$$\begin{aligned}\hat{\alpha}_{PC}(r) &= U'\hat{\beta}_{PS}(r) \\ &= [U_1\, U_2]' U_1 \Lambda_1^{-1} U_1' X' y \\ &= \begin{bmatrix} \Lambda_1^{-1} U_1' X' y \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} I_r \\ 0 \end{bmatrix} \hat{\alpha}_{LS}\end{aligned}$$

PC estimator $\hat{\alpha}_{PC}(r)$ cancel some components of $\hat{\alpha}_{LS}$ having large variances. This is a kind of variable selection, but not the original variables. $\hat{\alpha}_{PC}(r)$ is a componentwise shrunken estimator

$$\hat{\alpha}_{PC}(r) = \hat{\alpha}_{CS}(B) \quad , \quad B = diag(\underbrace{1, 1, ..., 1}_{r}, \underbrace{0, 0, ...0}_{p-r})$$

where the shrinkage factors are equal to either zero or one.

$$MSE(\hat{\alpha}_{PC}(r), \alpha) = \sigma^2 \sum_{i=1}^{r} \lambda_i^{-1} + \sum_{i=r+1}^{p} (u_i'\alpha)^2$$

A sufficient condition for $\hat{\alpha}_{PC}(r)$ to have a smaller MSE than $\hat{\alpha}_{LS}$ is that the inequality $\beta'\beta / \sum_{i=r+1}^{p} \lambda_i^{-1} < \sigma^2$ is satisfied.

The PC estimator takes values in $[U_1]$. A bit flexible estimator, permitting for dispersion in $R^p = [U_1] \oplus [U_2]$ is the Marquardt's estimator

$$\begin{aligned}\hat{\beta}_{MPC}(r, m) &= U \begin{bmatrix} I_r \\ c \end{bmatrix} \hat{\alpha}_{LS} \\ &= U \begin{bmatrix} \Lambda_1^{-1} U_1' X' y \\ c\Lambda_2^{-1} U_2' X' y \end{bmatrix} \\ &= U_1 \Lambda_1^{-1} U_1' X' y + m U_2 \Lambda_2^{-1} U_2' X' y \\ &= P_{[U_1]} \hat{\beta}_{LS} + m P_{[U_2]} \hat{\beta}_{LS}\end{aligned}$$

(Marquardt, 1970, Marquardt and Snee, 1975).

2.3. **Ridge Regression Estimator.** The Ridge regression estimator in its simplest form, called ordinary ridge estimator (OR), is

$$\hat{\beta}_{OR}(k) = (X'X + kI)^{-1} X' \tag{2.3}$$

where k is some nonnegative scalar constant. The generalized ridge estimator (GR) is

$$\hat{\beta}_{GR}(K) = (X'X + K)^{-1}X'y \tag{2.4}$$

where $K = diag(k_1, k_2, ..., k_p)$, $k_i \geq 0$, $i = 1, 2, ..., p$. The general ridge estimator is defined as

$$\hat{\beta}_{GR}(G) = (X'X + G)^{-1}X'y \tag{2.5}$$

where G is a symmetric nonnegative definite matrix.

The generalized ridge estimator in the reparameterized model (2.1)

$$\hat{\alpha}_{GR}(K) = (\Lambda + K)^{-1}\Lambda\hat{\alpha}_{LS} = \hat{\alpha}_{CS}(B)$$

is a componentwise shrunken estimator, where

$$B = diag(\frac{\lambda_1}{\lambda_1 + k_1}, \frac{\lambda_2}{\lambda_2 + k_2}, ..., \frac{\lambda_p}{\lambda_p + k_p})$$

and the OR estimator is also a componentwise shrunken estimator

$$\hat{\alpha}_{OR}(k) = (\Lambda + kI)^{-1}\Lambda\hat{\alpha}_{LS} = \hat{\alpha}_{CS}\left((\Lambda + kI)^{-1}\Lambda\right).$$

The correspondency of ridge estimators between the models (1.1) and (2.1) is as follows.

$$\begin{aligned}
\hat{\alpha}_{GR}(K) &= (\Lambda + K)^{-1}Z'y \\
&= U'(X'X + UKU')^{-1}X'y \\
&= U'\hat{\beta}_{GR}(UKU')
\end{aligned}$$

$$\hat{\beta}_{GR}(K) = U\hat{\alpha}_{GR}(U'KU)$$

$$\hat{\alpha}_{OR}(k) = U'\hat{\beta}_{OR}(k)$$

$$\hat{\beta}_{OR}(k) = U\hat{\alpha}_{OR}(k)$$

The optimal componentwise shrunken estimator of α in the model (2.1) is the generalized ridge estimator $\hat{\alpha}_{GR}(K)$, for $K = diag(k_1, k_2, ..., k_p)$, $k_i = \frac{\sigma^2}{\alpha_i^2}$, $i = 1, 2, ..., p$.

All ridge estimators are admissible in the class of linear estimators with respect to MSE criterion. Hoerl and Kennard (1970a, 1970b) proved that there exist always positive values of k such that

$$MSE(\hat{\beta}_{OR}(k), \beta) < MSE(\hat{\beta}_{LS}, \beta).$$

Theobald (1974) has shown that a sufficient condition for $\hat{\beta}_{OR}(k)$ to be a smaller MSE estimator than $\hat{\beta}_{LS}$ is that k satisfy $0 < k < 2\sigma^2/\beta'\beta$ and a necessary condition is that k satisfy

$$0 < k < \sigma^2 / \max_i \left\{\alpha_i^2\right\}.$$

For the general ridge estimator $\hat{\beta}_{GR}(G)$ the mentioned sufficient condition is $\beta'\left(G^{-1} + (X'X)^{-1}\right)^{-1}\beta \leq \sigma^2$ (Gross, 2003). For the special case $G = kI$ the inequality becomes $\beta'(\frac{2}{k}I + (X'X)^{-1})^{-1}\beta \leq \sigma^2$ which is equivalent to $0 < k < 2\sigma^2/\beta'\beta$.

Ridge estimators are not operational. There are numerous heuristic and theoretically rigorous techniques available for choosing a suitable value for the biasing parameter. Hoerl, Kennard and Baldwin (1975) showed that, for $\hat{k}_{HKB} = p\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$ a significant improvement in the MSE of $\hat{\beta}_{OR}(\hat{k}_{HKB})$ is obtained. Hoerl and Kennard (1976) also give an algorithm for iterative estimation of k, where at every stage $\hat{\beta}$ in $\hat{k}_{HKB} = p\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$ is replaced by $\hat{\beta}_{OR}(\hat{k}_{HKB})$. For the generalized ridge estimator minimum MSE is obtained at $k_i^{opt} = \frac{\sigma^2}{\alpha_i^2}$, $i = 1, 2, ..., p$. Hence $\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$, $i = 1, 2, ..., p$ is the natural choice.

There are many different interpretations of ridge estimators, like data augmented estimator, restricted estimator, Bayesian estimator etc. The Bayesian interpretation says that, under the normal prior $N(0, \frac{\sigma^2}{k}I)$ for the parameter vector, the Bayes estimator with respect to squared error loss is the OR estimator (2.3), and under the normal prior $\beta \sim N(0, \sigma^2 K^{-1})$ the Bayes estimator is the GR estimator (2.4). Under the prior information $\beta \sim (0, \sigma^2 G^{-1})$ the Bayes estimator in the set of linear homogeneous estimators $\left\{\hat{\beta} \, : \, \hat{\beta} = Ay \, , \, A \in R^{p\times n}\right\}$ is defined as

$$\hat{\beta}_{LB}(G) = \left(\arg\min_A MSE^B\left(A, \sigma^2, G\right)\right) y$$

Rao (1976). The resulting estimator is called Linear Bayes (LB) estimator, and can be obtained as follows.

$$\begin{aligned} MSE^B(A, \sigma^2, G) &= E_\beta\left(MSE(Ay, \beta)\right) \\ &= E_\beta\left(\sigma^2 tr(AA') + tr\left((AX - I)\beta\beta'(AX - I)'\right)\right) \\ &= \sigma^2 tr(AA') + tr\left((AX - I)E_\beta(\beta\beta')(AX - I)'\right) \\ &= \sigma^2 tr(AA') + \sigma^2 tr\left((AX - I)G^{-1}(AX - I)'\right) \end{aligned}$$

$$\arg\min_A MSE^B\left(A, \sigma^2, G\right) = \left(X'X + G\right)^{-1} X'$$

and

$$\hat{\beta}_{LB}(G) = (X'X + G)^{-1}X'y$$

(Gross, 2003). So, under the prior information $\beta \sim (0, \sigma^2 G^{-1})$ the LB estimator is equal to GR estimator. Under the prior information $\beta \sim (0, \frac{\sigma^2}{k}I)$ we have

$$\hat{\beta}_{LB}(k) = \hat{\beta}_{OR}(k).$$

Although being formally the same, these estimators are conceptionally different. A statistician employing the Bayes estimator uses the sample information with an extra prior information. A statistician employing the ridge estimator only uses the sample information and has to estimate the shrinkage parameters in order to make the estimator operational. The resulting estimator is a nonlinear function of the sample.

The OR estimator $\hat{\beta}_{OR}(k)$ is the shortest estimator in an equivalence class of estimators with the same residual sum of squares. In the case of normally distributed

errors, $\hat{\beta}_{OR}(k)$is the shortest β vector for given likelihood. The OR estimator $\hat{\beta}_R(k)$ has a minimum residual sum of squares in an equivalence class of estimators of the same length. In the case of normally distributed errors, $\hat{\beta}_R(k)$ is the most-likely β vector of the same length (Hoerl and Kennard, 1970a).

The family of OR estimators

$$\hat{\beta}_R(k) = (X'X + kI)^{-1}X'y \quad , \quad k \in [0, \infty)$$

traces a curved path through the parameter space from $\hat{\beta}_R(0) = \hat{\beta}_{LS}$ to 0(origin), as k increases from zero to infinity, in such a way that $\left\|\hat{\beta}_R(k)\right\| < \left\|\hat{\beta}_{LS}\right\|$ for $k \in (0, \infty)$. Hoerl and Kennard (1970a) proved that the expected distance between $\hat{\beta}_R(k)$ and β must initially decrease as k increases (illustrated in Section 3).

The name of ridge regression comes from ridge analysis developed by Hoerl (1959) for examining higher-dimensional quadratic response surfaces used in experimental design. Hoerl (1962) noted that the residual sum of squares in regression could be written as a quadratic function of the coefficients. So, ridge analysis, used in the maximization of the response function, could be used to calculate the coefficients as one moved along the minimum ridge of the residual sum of squares from the overall minimum (least squares) to some more stable solution closer to the origin (Hoerl, 1985).

Ridge regression is a popular technique to deal with multicollinearity. The properties, optimality conditions and problems associated with Ridge estimators have been investigated in hundreds of papers. An old list, including small abstracts are the Hoerl and Kennard's (1979, 1981) works, for the early years publications about ridge regression. In the last years there appeared a lot of papers on combining the Ridge and LS estimators, under the name (r-k) class estimator.

2.4. **Linear Unified Estimator.** Kejian Liu (Liu, 1993) proposed the following estimator.

$$\hat{\beta}_{Liu}(d) = (X'X + I)^{-1}(X'y + d\,\hat{\beta}_{LS}) \quad , \quad 0 < d < 1 \tag{2.6}$$

This estimator is a linear combination of the ridge estimator with ridge parameter k=1 and the least squares estimator. It is also a linear transformation of $\hat{\beta}_{LS}$.

$$\begin{aligned} \hat{\beta}_{Liu}(d) &= (X'X + I)^{-1}(X'y + d\,\hat{\beta}_{LS}) \\ &= (X'X + I)^{-1}(X'X)(X'X)^{-1}(X'y + d\,\hat{\beta}_{LS}) \\ &= (X'X + I)^{-1}(X'X + dI)\hat{\beta}_{LS} \end{aligned}$$

This estimator is named as **Li**near **U**nified (abbreviated as Liu) estimator by Akdeniz and Kaçıranlar (1995) and was accepted by the related community. Liu (2003) mention that, when there exists severe collinearity, the shrinkage parameter selected by existing methods for ridge regression may not fully address the ill conditioning problem. In order to solve this problem, he proposes a new two-parameter estimator and shows, using both theoretical results and simulation study, that the

proposed estimator has two advantages over ridge regression. First, the estimator has less MSE, second, the estimator can better address the ill-conditioning problem. The Liu-type estimation procedure is a popular technique in the last two decade. The new fashion is combining PC, Ridge and Liu estimators. You can make (r-d) or ((r-k)-d) combinations (see for example Alheety and Kibria (2013)). The two parameter Liu (2003) estimator is the (k-d) combination.

Liu (1993) showed that the estimator $\hat{\beta}_{Liu}(d)$, given above (2.6), is superior to the $\hat{\beta}_{LS}$ in the MSE sense. The optimal value for d is

$$d_{opt} = 1 - \frac{\sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i(1+\lambda_i)}}{\sum_{i=1}^{p} \frac{\alpha_i^2}{(1+\lambda_i)}}.$$

The estimator $\hat{\beta}_{Liu}(d)$is not operational. The plug-in estimator of d_{opt} is

$$\hat{d}_{opt} = 1 - \frac{\hat{\sigma}^2 \sum_{i=1}^{p} \frac{1}{\lambda_i(1+\lambda_i)}}{\sum_{i=1}^{p} \frac{\hat{\alpha}_i^2}{(1+\lambda_i)}}.$$

In Section-3 we used,

$$\widehat{d} = \begin{cases} 0 & if \quad \widehat{d}^{opt} < 0 \\ \widehat{d}^{opt} & if \quad 0 \leq \widehat{d}^{opt} \leq 1 \\ 1 & if \quad \widehat{d}^{opt} > 1. \end{cases}$$

2.5. **A Discrete Shrunken (DS) Estimator.** Let us try to solve the normal equation

$$X'X\beta = X'y$$

numerically by the iteration formula

$$\beta^{(n)} = (I - hX'X)\beta^{(n-1)} + hX'y \quad , \quad n = 1, 2, 3, ... \ , \ \beta^{(0)} = 0$$

obtained from an algorithm based on fixed point theorem, applied to the equation

$$\beta = \beta - hX'X\beta + hX'y.$$

It can be shown that $\beta^{(n)}$ converges to the solution for $0 < h < 2/\lambda_{\max}$ and for any initial value $\beta^{(0)} \in R^p$. The sequence of estimators

$$\hat{\beta}^{(n)}(h, n) = (I - hX'X)\hat{\beta}^{(n-1)} + hX'y \quad , \quad n = 1, 2, 3, ... \ , \ \hat{\beta}^{(0)} = 0$$

converge to $\hat{\beta}_{LS}$ uniformly. For a chosen value of h the set $\left\{\hat{\beta}^{(n)}(h, n) : n = 1, 2, 3, ...\right\}$ is a discrete family of n. A suitable element in the set can be taken as an estimator of β. Based on this idea the estimator,

$$\hat{\beta}_{DS}(h, n) = \left(I - (I - hX'X)^n\right)\hat{\beta}_{LS} \quad , \quad 0 < h < 1/\lambda_{\max}.$$

is defined (Öztürk, 1984). The corresponding estimator of α

$$\hat{\alpha}_{DS}(h,n) = U\hat{\beta}^{(n)} = \left(I - (I - h\Lambda)^n\right)\hat{\alpha}_{LS} \quad , \quad 0 < h < 1/\lambda_{\max}$$

is the componentwise shrunken estimator $\hat{\alpha}_{DS}(h,n) = \hat{\alpha}_{CS}(B)$, $B = I-(I-h\Lambda)^n$. The MSE of $\hat{\alpha}_{DS}(h,n)$ is

$$MSE(\hat{\alpha}_{DS}(h,n)) = \sigma^2 \sum_{i=1}^{n} \frac{(1-(1-h\lambda_i)^n)^2}{\lambda_i} + \sum_{i=1}^{n} (1-h\lambda_i)^{2n}\alpha_i^2.$$

A sufficient condition for $\hat{\alpha}_{DS}(h,n)$ to have a smaller MSE than $\hat{\alpha}_{LS}$ is that the inequality $(1-h\lambda_{\min})^n < \frac{\sigma^2}{\sigma^2+\lambda_{\max}\alpha_{\max}^2}$ is satisfied. The discrete shrunken estimator $\hat{\beta}_{DS}(h,n)$ is similar to the estimator defined by Trenkler (1978). Trenkler's motivation is based on a serial expansion for the Moore-Penrose generalized inverse of X.

Let us mention that we can look at the normal equation $X'X\beta = X'y$ as an inversion problem. Then we encounter the problem of ill-conditioning for bad coefficient matrix $X'X$. The problem of ill-conditioning is a widely investigated topic in applied mathematics since 1960. Some of the solution methods are regularization method, selection method and replacement method. It is very interesting that generalized ridge estimation mimics the regularization method, principal component estimation mimic the selection method and ordinary ridge estimation mimic the replacement method (Öztürk and Akdeniz, 2000).

3. Geometrical Visualization of Estimation Traces in the Parameter Space

When the design matrix X has nearly collinear columns, then the LS estimator is unstable in the sense that different samples can produce dramatically different estimates. This is the problem of multicollinearity.

Consider the following simulation framework (Case II) from (Swindel, 1974).

$$X = \left[\underline{x_1}, \underline{x_2}\right] = \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & 0.8 \\ 0 & 0 \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \epsilon \sim N(0, I).$$

$r_{X_1,X_2} = 0.8846$ indicates a severe collinearity.

$$\begin{aligned} X'X &= \begin{bmatrix} 1.00 & 0.96 \\ 0.96 & 1.00 \end{bmatrix} \\ X'X &= U\Lambda U' \\ U &= \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 0.04 & 0 \\ 0 & 1.96 \end{bmatrix} \end{aligned}$$

The simulated 10 samples for ϵ are:

-1.276	-1.218	-0.453	-0.350	0.723	0.676	-1.099	-0.314	0.394	-0.633
-0.318	-0.799	-1.664	1.391	0.382	0.733	0.653	0.219	-0.681	1.129
-1.377	-1.257	0.495	-0.139	-0.854	0.428	-1.322	-0.315	-0.732	-1.348

10 samples for

$$y = \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & 0.8 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \epsilon$$

are:

0.124	0.182	0.947	1.050	2.123	2.076	0.301	1.086	1.006	0.767
1.082	0.601	-0.264	2.791	1.782	2.133	2.053	1.619	0.719	2.529
-1.377	-1.257	0.495	-0.139	-0.854	0.428	-1.322	-0.315	-0.732	-1.348

and LS estimates for the simulated samples are as below.

$\widehat{\beta}_{LS}$:	-1.964	-0.768	3.271	-2.981	2.247	1.361	-3.539	-0.366	1.334	-3.228
	2.826	1.327	-2.784	5.724	0.547	1.646	5.221	2.299	-0.101	5.582

Estimates for the individual components of β are unstable, even having different signs for different samples. Swindel (1974) illustrates this situation geometrically in the estimation space spanned by the columns of $X = [x_1\ x_2]$ (Figure-1). Imagine the coefficients in the linear combination $c_1x_1+c_2x_2$ represented by the 10 points for simulated 10 data sets. They are the LS estimates of β. Do the same imagination for Case I. Do you feel that they have less variation than in Case II? They are more "stable". They do not suffer from the problem of multicollinearity. Indeed, the estimates are just below (Swindel, 1974):

Case I

$\hat{\beta}_{LS}$:	-0.276	-0.218	0.547	0.650	1.723	1.676	-0.099	0.686	0.606	0.367
	0.682	0.201	-0.664	2.391	1.382	1.733	1.653	1.219	0.319	2.129

The true parameter is $\beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. The estimates seem to be far away. This is because of the relatively big variance of the simulated error term in the model. Thanks to Swindel, to be not afraid of big variance, otherwise the deficiency of LS could not be exhibited so well. Let us also mention that, when the explanatory variables individually obey small variation, the parameter estimates become unstable, have big variances, like in the situation of near collinearity. Near collinearity, geometrically means a small area of the parallelogram defined by x_1 and x_2(volume for p>2). Swindel defeated the small variation in the explanatory variables by taking them normalized. In reality this is not possible, except in design of experiments. For example, A-optimality in design of experiments means optimal total variance,

design with smallest $tr(X'X)$.

Now, imagine the unknown "truth", the point with coordinates (1,1) in the booth (CaseI and CaseII) coordinate systems with basis vectors x_1 and x_2 (the point $X \begin{bmatrix} 1 \\ 1 \end{bmatrix} = x_1 + x_2$). All ten points, randomly fallen on the estimation space want to be close to the "truth". Do you also feel that the shrinking of the vectors represented by the points (vectors beginning from the origin of the coordinate systems and ending at the LS points) will move the points a bit towards the "truth", but not for all of them. This is the direct (Stein-type) shrinking, as the Meyer and Wilke's estimator does in the parameter space. The planes in Figure-1 and Figure-2 represent the estimation space, not the parameter space as in the other figures below. Notice that the scatters are the same for both CaseI and CaseII. The "randomness" is the same. Simulated errors are the same. Swindel

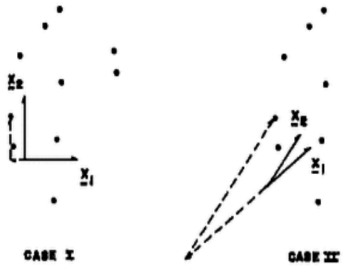

FIGURE 1. A figure from Swindel's (1974) paper.

(1981) also illustrates the geometry of Ridge traces (Figure-2). The advantages of the OR estimator are easily seen. There are some "good" points on the curves, that is, there are some points closer to the point $1x_1 + 1x_2$ identified by an asterisk, rather than the LS estimates.

Using the Swindel's simulated data, we illustrate the behavior of LS, MW, OR, Liu, PC and DS estimators (defined in Section 2) geometrically in the parameter space, instead of the estimation space as in Swindel's works. In all of the figures below, the true value of β is shown as a red colored star, LS estimates are red pentagrams; optimal estimates are circles, operational estimates are squares and MW estimates are colored magenta, PC estimates are colored cyan, OR estimates are colored blue, Liu estimates are colored green, DS estimates are colored black.

In the fifth repetition of the simulation study

$$\epsilon = \begin{bmatrix} 0.7230 & 0.3820 & -0.8540 \end{bmatrix}'$$

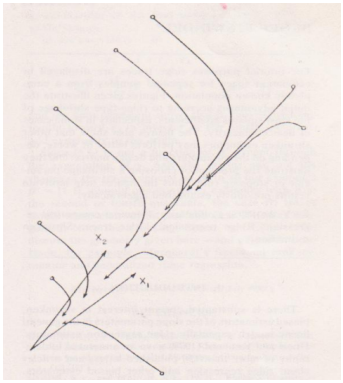

FIGURE 2. A figure from Swindel's (1981) paper

for CaseII. The corresponding "observed" (simulated) response is

$$y = \begin{bmatrix} 2.123 \\ 1.782 \\ -0.854 \end{bmatrix}$$

and the estimates are below (with explanations necessary for the figures).

$$\beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{red colored, asterisk} \qquad \widehat{\beta}_{LS} = \begin{bmatrix} 2.2471 \\ 0.5421 \end{bmatrix} \quad \text{red colored, pentagram}$$

$$\widehat{\beta}_{MW}\left(c^{opt}\right) = \begin{bmatrix} 0.1634 \\ 0.0394 \end{bmatrix} \quad \text{magenta, circle} \qquad \widehat{\beta}_{MW}\left(\widehat{c}^{opt}\right) = \begin{bmatrix} 0.5014 \\ 0.1210 \end{bmatrix} \quad \text{magenta, square}$$

$$\widehat{\beta}_{PC}\left(r\right) = \begin{bmatrix} 1.3946 \\ 1.3946 \end{bmatrix} \quad \text{cyan, circle} \qquad \widehat{\beta}_{MPC}\left(r,m\right) = \begin{bmatrix} 1.6504 \\ 1.1389 \end{bmatrix} \quad \text{cyan, square},$$

$$\widehat{\beta}_{OR}\left(k^{opt}\right) = \begin{bmatrix} 0.9563 \\ 0.8907 \end{bmatrix} \quad \text{blue, circle} \qquad \widehat{\beta}_{OR}\left(\widehat{k}^{opt}\right) = \begin{bmatrix} 1.3331 \\ 1.1152 \end{bmatrix} \quad \text{blue, square}$$

$$\widehat{\beta}_{Liu}\left(d^{opt}\right) = \begin{bmatrix} 1.0019 \\ 0.8784 \end{bmatrix} \quad \text{green, circle} \qquad \widehat{\beta}_{Liu}\left(\widehat{d}\right) = \begin{bmatrix} 0.9563 \\ 0.8907 \end{bmatrix} \quad \text{green, square}$$

$$\widehat{\beta}_{DS}\left(h,n\right) = \begin{bmatrix} 1.3984 \\ 1.3052 \end{bmatrix} \quad \text{black colored square}$$

where,

$$\begin{aligned} c^{opt} &= 0.0727, \hat{c}^{opt} = 0.2232, r = 1, m = 0.3, k^{opt} = 1, \\ \hat{k}^{opt} &= 0.0874, d^{opt} = 0.0353, \hat{d} = 0, \quad h = 0.2, n = 7. \end{aligned}$$

All of the estimators give better estimates then LS estimator, except MW, according to Euclidean distance. The situation can also be seen easily on the graph.

We evaluate the estimates according to their distance to the true parameter (unknown, "big red star"). In Figure-3 the traces of OR, Liu and DS family of

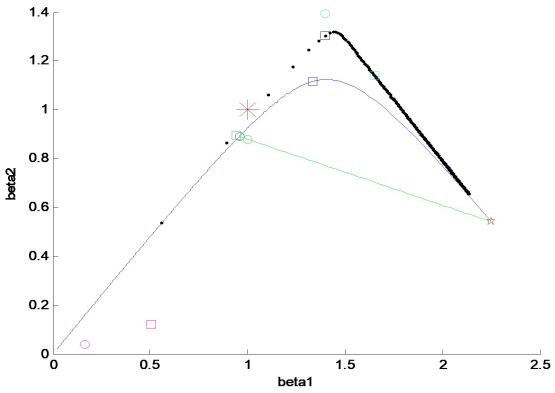

FIGURE 3. OR-trace(blue), Liu-trace(green), DS-trace(black) and LS, MW, PC, OR, Liu, DS estimates

estimates are illustrated, with optimal and operational estimates lying on them. Observe that the Ridge-trace (blue) is a continuous curved path through the parameter space from$\hat{\beta}_{LS}$ to the origin, as k increases from zero to infinity, initially converging to β. Liu estimator traces a straight line (green) from $\hat{\beta}_{LS}$ to $\hat{\beta}_{OR}(1)$. The DS-trace (black) is a discrete curved path from the origin to the$\hat{\beta}_{LS}$.

The OR, Liu and DS traces for all the simulated 10 samples are as in Figure-4. Observe that there are better estimates lying on the traces than the LS estimates. The ingenuity is to choose a "good point" on the trace when "the star isn't shining". The statistical problem here is the estimation of the shrinking (biasing) parameter. It is a serious problem, because the LS estimates (small stars) are far away in the sense of Euclidean distance.

In Figure-5 the performance of MW, PC, R, Liu estimators are compared to the performance of the LS estimator. As is seen, the optimal shrinkage estimators give estimates, which are almost covering the true β. But, it's not possible in reality. For the operational estimators, there is a spread around the unknown parameter vector β.

In Figure-6 the performance of the Ridge estimator is illustrated versus MW, PC, PCM, Liu and DS estimators.

In Figure-7 the performances of Ridge and Liu estimators are illustrated for both of optimal and operational estimates. Optimal estimators are better then operational ones. But, it's difficult to decide which works better, Ridge or Liu.

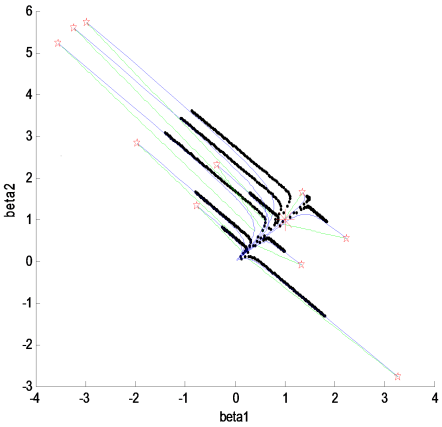

FIGURE 4. Traces in the parameter space for OR, Liu and DS estimators.

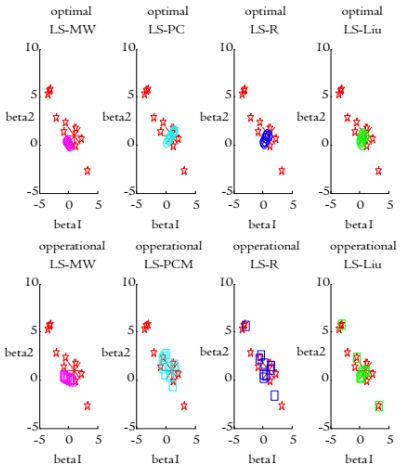

FIGURE 5. LS estimator versus MW, PC, R, Liu estimators.

Finally let us look at the estimated MSE values, calculated as simple averages. The estimated value of $MSE(\hat{\beta}_{LS}, \beta)$ is

$$\widehat{MSE}(\hat{\beta}_{LS}, \beta) = \frac{\sum_{s=1}^{10} \left((\hat{\beta}^s_{LS} - \beta)'(\hat{\beta}^s_{LS} - \beta) \right)}{10} = 15.7473$$

and the others are in Table-1, calculated in the same way.

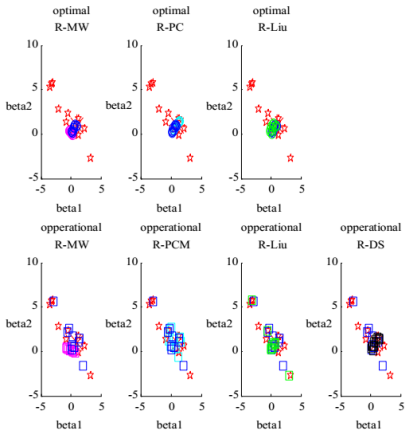

FIGURE 6. Ridge estimator versus MW, PC, Liu, DS estimators

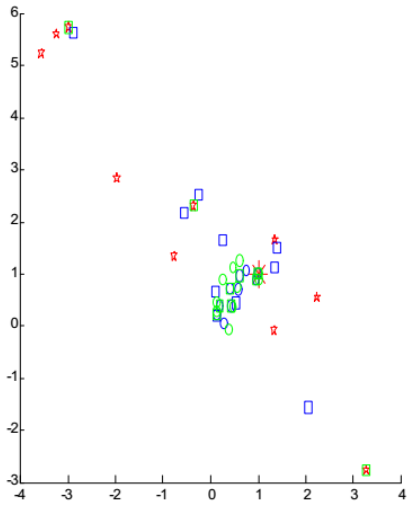

FIGURE 7. Performance of Ridge versus Liu.

Table-1 Observed MSE values

Estimator	Meyer and Wilke's	Principal Component	Ridge	Liu	Discrete Shrinking
optimal shrinkage	1.8347	0.4334 (p=2,r=1)	0.5462	0.5931	
operational shrinkage	2.0726	1.8117 (Marquardt's estimator, c=0.3)	5.6314	6.4884	0.4689 (n=7,h=0.2)

4. Conclusions

Once more we tried to elaborate upon the geometrical aspect of linear models, specially of shrunken estimators. The algebra we used, indeed, is a geometrical speech. Besides the statistical theory, Linear Model teaching needs matrix algebra in a high level. But intuition comes up easily when geometry is used. Once again listen to the Swindel's words.

"I have found that students of statistics do not readily appreciate the foregoing ideas given only the algebraic justification of them prevalent in the current literature. But, they respond enthusiastically to a geometrical explanation. . . " (Swindel, 1974).

". . . The figures also show that other shrunken estimators may perform better or worse, depending on the parameters and design matrix; and they illustrate the problem of choosing a shrinkage parameter or stopping rule. Thus, the figures help motivate results previously established algebraically." (Swindel, 1981).

Let us end by saying that the primary aim of this study was to call to remember some old works, not to make work. Every new generation must learn and rediscover the knowledge of old generations.

References

[1] Akdeniz F. and Kaçıranlar, S., On the almost unbiased generalized Liu estimator and unbiased estimation of the Bias and MSE, *Communications in Statistics- Theory and Methods*, (1995) 24(7): 1789-1797.

[2] Alheety, M.I. and Kibria, B.M.G., Modified Liu-Type Estimator Based on (r-k) Class Estimator, *Communications in Statistics—Theory and Methods*, (2013) 42: 304–319.

[3] Gross, J. Linear Regression, Springer, .2003.

[4] Hoerl, A. E., Optimum solution to many variables equations. Application of ridge analysis to regression problems, *Chemical Engineering Progress*, (1959) 55, 69-78.

[5] Hoerl, A. E., Application of ridge analysis to regression problems, *Chemical Engineering Progress*, (1962) 58, 54-59.

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

[7] Hoerl, A. E. and Kennard, R. W., Ridge regression: applications to nonorthogonal problems, *Technometrics* (1970) 12, 69-82.

[8] Hoerl, A. E. Kennard, R. W. and Baldwin, F. K., Ridge regression: Some simulations, *Communications in Statistics–Theory and Methods*, (1975) 4, 105-123.

[9] Hoerl, A. E. and Kennard, R. W., Ridge regression: Iterative estimation of the biasing parameter, *Communications in Statistics–Theory and Methods*, (1976) A5(1), 77-78.

[10] Hoerl, A. E. and Kennard, R. W., Ridge regression-1979. *Mimeo document -Bibliography,* (1979) 12-77.

[11] Hoerl, A. E. and Kennard, R.W., Ridge regression-1980. Advances, algorithms and applications, *American Journal of Mathematical and Management Sciences*, (1981) 1(1), 5-83.

[12] Hoerl, A. E., Ridge Analysis 25 Years Later, *The American Statistician*, (1985) 39(3),186-192.

[13] James, W. and Stein, C., Estimation with quadratic loss, *Proceedings of the Fourth Berkeley Symposium, University of California Press*, (1961) 361-379.

[14] Liu, K. J., A new class of biased estimate in linear regression, *Communications in Statistics–Theory and Methods*, (1993), 22, 393-402.

[15] Liu, K.J., Using Liu type estimator to combat multicollinearity, *Communications in Statistics–Theory and Methods*, (2003) 32(5), 1009-1020.
[16] Marquardt, D.W., Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, *Technometrics* (1970) 12, 591-612.
[17] Marquardt, D.W. and Snee, R.D., Ridge regression in practice. *The American Statistician,* (1975) 29, 3-20.
[18] Meyer, L.S. and Wilke, T.A., On biased estimation in linear models, *Technometrics* (1973) 15, 497-508.
[19] Öztürk, F., A discrete shrinking method as alternative to least squares, *Communications de la Faculté des Sciences de l'University d'Ankara*, (1984), 33, 179-185.
[20] Öztürk, F. and Akdeniz, F., Ill-conditioning and multicollinearity, *Linear Algebra and its Applications*, (2000) 321, 295-305.
[21] Rao, C.R., Estimation of parameters in a linear model*, The Annals of Statistics*, (1976) 4, 1023-1037.
[22] Swindel, B.F., Instability of regression coefficients illustrted , *The American Statistician*, (1974) 28(2), 63-65.
[23] Swindel, B.F., Good ridge estimators based on prior informatin, *Communications in Statistics–Theory and Methods,* (1976) A5(11), 1065-1075.
[24] Swindel, B.F., Geometry of ridge regression illustrated, *The American Statistician*, (1981) 35(1), 12-15.
[25] Theobald, CM, Generalization of Mean Square Error Applied to Ridge Regression, *Journal of the Royal Statistical Society Series B*, (1974) 36(1), 103-106.
[26] Trenkler, G., An iteration estimator for the linear model, *Computational Statistics, Physica Verlag,* (1978) 125-131.

Current address: Fikri Akdeniz: Çağ University, Tarsus, Turkey.
E-mail address: fikriakdeniz@gmail.com
ORCID Address: http://orcid.org/0000-0002-8427-8653
Current address: Fikri Öztürk: Ankara University, Ankara, Turkey.
E-mail address: ozturkf@ankara.edu.tr
ORCID Address: http://orcid.org/0000-0002-7175-7372