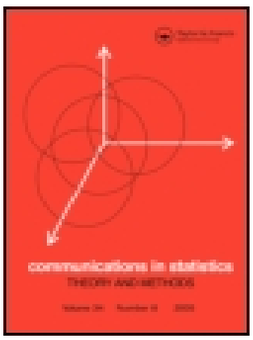

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

# A modified Liu-type estimator with an intercept term under mixture experiments

Ai-Chun Chen & Takeshi Emura

**To cite this article:** Ai-Chun Chen & Takeshi Emura (2017) A modified Liu-type estimator with an intercept term under mixture experiments, Communications in Statistics - Theory and Methods, 46:13, 6645-6667, DOI: 10.1080/03610926.2015.1132327

**To link to this article:** http://dx.doi.org/10.1080/03610926.2015.1132327

Accepted author version posted online: 29 Jun 2016.
Published online: 29 Jun 2016.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2017, VOL. 46, NO. 13, 6645–6667
http://dx.doi.org/10.1080/03610926.2015.1132327

Taylor & Francis
Taylor & Francis Group

# A modified Liu-type estimator with an intercept term under mixture experiments

Ai-Chun Chen and Takeshi Emura

Graduate Institute of Statistics, National Central University, Taoyuan City, Taiwan

**ABSTRACT**

We consider ridge regression with an intercept term under mixture experiments. We propose a new estimator which is shown to be a modified version of the Liu-type estimator. The so-called compound covariate estimator is applied to modify the Liu-type estimator. We then derive a formula of the total mean squared error (TMSE) of the proposed estimator. It is shown that the new estimator improves upon existing estimators in terms of the TMSE, and the performance of the new estimator is invariant under the change of the intercept term. We demonstrate the new estimator using a real dataset on mixture experiments.

**ARTICLE HISTORY**
Received 2 August 2015
Accepted 8 December 2015

**KEYWORDS**
Compound covariate estimator; linear regression; least squares estimator; ridge regression; shrinkage estimator

## 1. Introduction

In a linear regression model, high correlation between regressors is called multicollinearity. When multicollinearity exists, the ordinary least squares (OLS) estimator is less accurate due to large sampling variance. Ridge regression is an alternative estimator derived by Hoerl and Kennard (1970), which becomes one of the most common methods to deal with multicollinearity. Ridge regression aims to reduce the large variance by shrinking the OLS estimator toward the zero vector. It has been theoretically justified that the mean squared error of the ridge estimator is smaller than that of the OLS estimator with appropriate amount of shrinkage (Hoerl and Kennard, 1970; Theobald, 1973).

Multicollinearity arises in mixture experiments, where regressors are proportions of a mixture. Ridge regression has been successfully applied for the Scheffe-type model to overcome the multicollinearity among the proportions (Jang and Anderson-Cook, 2010, 2014). However, the absence of an intercept term in the Scheffe-type model makes it different from the routine practice of linear regression that typically has an intercept term. In the Scheffe-type model, the intercept term is removed and absorbed into the regression coefficients for proportions (Section 2.3). For this removal to be valid, the sum of observed proportions must be 1 (or 100%) for all individuals. Unfortunately, this requirement is not always met, which is indeed the case for our motivating real data example.

In our paper, we consider ridge regression with an intercept term as in the routine practice of linear regression. We particularly propose a new estimator, which is shown to be a modified version of the Liu-type estimator. Here, the so-called compound covariate estimator is applied

**CONTACT** Takeshi Emura takeshiemura@gmail.com Graduate Institute of Statistics, National Central University, No. 300, Zhongda Rd., Zhongli District, Taoyuan City 32001, Taiwan.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

to modify the Liu-type estimator (Liu, 2003). We derive a simple formula of the total mean squared error (TMSE) of the proposed estimator, which allows us to show (i) the TMSE of the proposed method can be always smaller than the TMSE of the OLS estimator and (ii) the optimal value of the shrinkage parameter can be derived and estimated. Simulations show that the new estimator improves upon the OLS estimator and the ridge estimator in terms of the TMSE, and the performance of the new estimator is invariant under the change of the intercept term. We demonstrate the new estimator using a real dataset on mixture experiments.

Section 2 reviews the background. Section 3 introduces the proposed method. Section 4 provides theoretical results. Section 5 performs simulations, and Section 6 analyzes real data. Section 7 concludes.

## 2. Background

This section reviews the background on linear regression, ridge regression, and mixture experiments and also introduces some notations for subsequent discussions. Here, we will emphasize that all our model and notations are tailored for the intercept model (a linear model with an intercept term). This setting is different from the usual literature on ridge regression and mixture experiments in which the model does not have an intercept.

### 2.1. Linear model and ridge regression

A linear model with an intercept term is defined by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where

$$\mathbf{y}_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X}_{n\times(p+1)} = [\,\mathbf{1}_n \quad \mathbf{X}_p\,] = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix},$$

$$\boldsymbol{\beta}_{(p+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where $\mathbf{X}$ is the fixed design matrix with rank $(\mathbf{X}) = p + 1$, $\boldsymbol{\beta} \in R^{p+1}$ is unknown regression coefficients, and $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0},\sigma^2\mathbf{I}_n)$, where $\sigma^2 > 0$ is unknown, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. The ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$ is defined as follows:

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

If $\mathbf{X}$ is standardized, then

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij} = \bar{x}_j = 0, \quad \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = 1 \tag{2}$$

for $j = 1, \ldots, p$. In most textbooks on regression analysis, ridge regression is introduced after standardizing $\mathbf{X}$ to meet Equation (2) (e.g., Ashish and Srivastava, 1990; Draper and Smith, 1998; Hastie et al., 2009). This is because the ridge estimator aims to apply the same scale of shrinkage on all regression coefficients. However, this is not the case for mixture experiments

in which the scales of all regressors are already the same. Thus, if all columns of $\mathbf{X}$ represent proportions of a mixture, then ridge regression is still meaningful.

It is important to emphasize that the model (1) includes an intercept, which is different from the usual literature on (i) ridge regression and (ii) mixture experiments in the following ways:

(i) **Ridge regression**: The intercept term is set to be zero (i.e., $\beta_0 = 0$), and instead, the responses are redefined as $\mathbf{y} - \bar{y}\mathbf{1}_n$, where $\bar{y} = \mathbf{y}^{\mathrm{T}}\mathbf{1}_n/n$ (i.e., responses are centered). This implies that $\beta_0$ is pre-estimated by $\bar{y}$, and the ridge estimator is applied to $p$ regression coefficients (Brown, 1977; Montgomery et al., 2012).

(ii) **Mixture experiment**: The intercept term is set to be zero (i.e., $\beta_0 = 0$), and instead, the regression coefficients are redefined as $\beta_j^* = \beta_0 + \beta_j$, $j = 1, \ldots, p$, under the constraint $\sum_{j=1}^{p} x_{ij} = 0$. This yields the Scheffe's first-order model (Section 2.3 for details).

We still include an intercept in the model (1) with the aim of maintaining the routine practice of linear regression. The ridge estimator introduced by Hoerl and Kennard (1970) takes the form

$$\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}_{(p+1)})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \quad k \geq 0$$

where $k$ is called shrinkage parameter. The equivalent form is as follows:

$$\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k) = \begin{bmatrix} n\bar{y}/(n+k) \\ \hat{\boldsymbol{\beta}}_p^{\mathrm{Ridge}}(k) \end{bmatrix} \equiv \begin{bmatrix} \sum_{i=1}^{n} y_i/(n+k) \\ \left(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p + k\mathbf{I}_p\right)^{-1}\mathbf{X}_p^{\mathrm{T}}\mathbf{y} \end{bmatrix}, \quad k \geq 0$$

As mentioned above, one usually uses ridge regression without an intercept by centering the response. This is because the intercept term should not be shrunken (Brown, 1977).

The alternative versions of the ridge estimator under the intercept model are considered by Brown (1977), Jimichi and Inagaki (1993), Jimichi (2005), and Jimichi (2008), where they suggest a different shrinkage parameter ($k_0$) for the intercept. The Jimichi estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{Jimichi}}(k_0, k) = \begin{bmatrix} n\bar{y}/(n+k_0) \\ \hat{\boldsymbol{\beta}}_p^{\mathrm{Ridge}}(k) \end{bmatrix} \equiv \begin{bmatrix} \sum_{i=1}^{n} y_i/(n+k_0) \\ \left(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p + k\mathbf{I}_{p)}\right)^{-1}\mathbf{X}_p^{\mathrm{T}}\mathbf{y} \end{bmatrix}, \quad k_0, k \geq 0$$

includes the ordinary ridge estimator as a special case, $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k) = \hat{\boldsymbol{\beta}}^{\mathrm{Jimichi}}(k, k)$. They also discuss the optimal choice of $k_0$. The Jimichi estimator includes the Brown estimator by setting $\hat{\boldsymbol{\beta}}^{\mathrm{Brown}}(k) \equiv \hat{\boldsymbol{\beta}}^{\mathrm{Jimichi}}(0, k)$.

Instead of $\hat{\boldsymbol{\beta}}^{\mathrm{Brown}}(k)$ or $\hat{\boldsymbol{\beta}}^{\mathrm{Jimichi}}(k_0, k)$, several authors still have used $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ and have shown its superior performance over $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ (e.g., Sakallıoğlu and Kaçıranlar, 2008; Li and Yang, 2012). This is because ridge regression does not have any restriction about the form of $\mathbf{X}$.

In our paper, we continue to discuss the estimator $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ under the intercept model (1). The major goal of our paper is to propose a new estimator that improves upon $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$, especially under mixture experiments.

### 2.2. Canonical form under an intercept term

In order to study the mean squared error properties of the ridge estimator, the model is usually rewritten in the canonical form. While the canonical form without an intercept is well known, it seems less common to form the canonical form with an intercept. We will follow Jimichi and Inagaki (1993), Jimichi (2005) and Jimichi (2008) to form the canonical form with an intercept.

Let $\lambda_1 \geq \ldots \geq \lambda_p > 0$ be the ordered eigenvalues of the matrix $\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p$, where $\mathbf{X}_p$ is the design matrix without intercept as defined in Equation (1). Let $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ be the $p \times 1$ eigenvectors for $\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p$, corresponding to the ordered eigenvalues. It follows that

$$\boldsymbol{\Gamma}_p^{\mathrm{T}}\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p\boldsymbol{\Gamma}_p = \boldsymbol{\Lambda}_p$$

where $\boldsymbol{\Gamma}_p = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p]$ and $\boldsymbol{\Lambda}_p = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, where $\mathrm{diag}(\boldsymbol{\lambda})$ is the diagonal matrix with $\boldsymbol{\lambda}$ being its diagonals. The model (1) is rewritten in the canonical form $\mathbf{y} = \mathbf{A}\alpha + \boldsymbol{\varepsilon}$, where $\mathbf{A} = \mathbf{X}\boldsymbol{\Gamma}$, $\boldsymbol{\alpha} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\beta}$, and $\boldsymbol{\Gamma}_{(p+1)\times(p+1)}$ are an orthogonal matrix defined as follows:

$$\boldsymbol{\Gamma} = \begin{bmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \boldsymbol{\Gamma}_p \end{bmatrix}$$

It follows that

$$\boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\Gamma} = \mathbf{A}^{\mathrm{T}}\mathbf{A} = \boldsymbol{\Lambda} = \mathrm{diag}(n, \lambda_1, \ldots, \lambda_p), \ \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \mathbf{I}_{(p+1)}$$

If we write $\boldsymbol{\alpha}^{\mathrm{T}} = (\alpha_0, \alpha_1, \ldots, \alpha_p)^{\mathrm{T}}$, then $\alpha_0 = \beta_0$ and $(\alpha_1, \ldots, \alpha_p)^{\mathrm{T}} = \boldsymbol{\Gamma}_p^{\mathrm{T}}(\beta_1, \ldots, \beta_p)^{\mathrm{T}}$.

Since $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is unbiased, the total mean squared error (TMSE) of $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is the same as the total variance

$$\mathrm{TMSE}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}) = \mathrm{tr}\{\mathrm{var}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}})\} = \sigma^2 \mathrm{tr}\begin{bmatrix} \frac{1}{n} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & (\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1} \end{bmatrix} = \sigma^2\left(\frac{1}{n} + \sum_{j=1}^{p}\frac{1}{\lambda_j}\right)$$

where "tr" is the trace of a matrix (sum of all the diagonal elements). The TMSE can be large if there is a small eigenvalue of $\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p$, say $\lambda_r \cong 0$, for some $r \in \{1, \cdots, p\}$. If so, we have $\mathbf{X}_p\boldsymbol{\gamma}_r \cong 0$, evidence of multicollinearity as the $r$-th column is linearly dependent with others. This mathematically explains why $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ performs poorly under multicollinearity.

### 2.3. Mixture experiment

Mixture experiment is an experimental design for mixing several ingredients together to form a product (Cornell, 2011). For instance, consider the ingredients of flour, water, and egg to make a cake. If the proportions of the three ingredients are $x_i$, $i = 1, 2, 3$, we have a constraint

$$\sum_{i=1}^{3} x_i = 1, 0 \leq x_i \leq 1, i = 1, 2, 3$$

In this case, the domain of the ingredients forms a three-component simplex region (Figure 1). For instance, the point (0.5, 0, 0.5) in the simplex corresponds to a mixture with 50% flour, 0% water, and 50% egg. However, not every point on the simplex is allowed to occur in the experiments. For example, if we take 5% flour, 90% water, and 5% egg, then we may not complete a cake due to too much water in the mixture. Actually, the proportions are restricted by either a lower bound ($L_i$) or upper bound ($U_i$) so that $0 \leq L_i \leq x_i \leq U_i \leq 1$, $i = 1, 2, 3$.

The Scheffe-type model has been used to perform OLS regression and interpret the results within mixture experiments. The Scheffe's first-order model is defined as follows:

$$y = \beta_1^* x_1 + \cdots + \beta_p^* x_p + \varepsilon \tag{3}$$

which is simply a linear model without intercept. Under the constraint $\sum_{j=1}^{p} x_j = 1$, the intercept model (1) and the Scheffe-type model (3) are related through $\beta_1^* = \beta_0 + \beta_1, \ldots,$

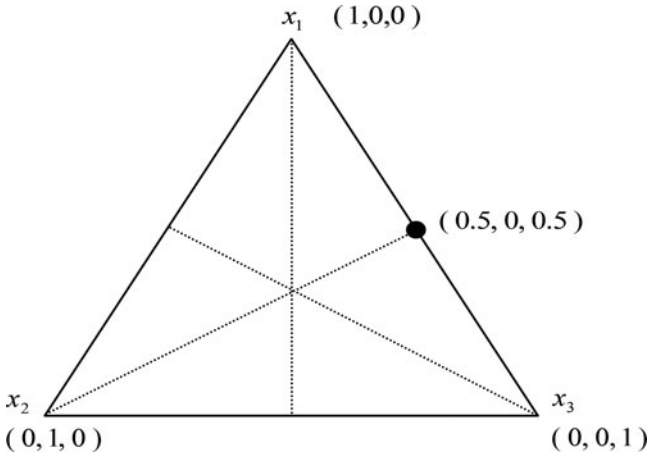


**Figure 1.** The three-component simplex region satisfying the constraint $\sum_{j=1}^{3} x_j = 1$.

$\beta_p^* = \beta_0 + \beta_p$ (Cornell, 2011, p. 27). One cannot use the OLS estimator to the model (3) since one component, say $x_q$, is redundant. Now, we can write $x_q = 1 - \sum_{j \neq q} x_j$. Then, the OLS estimator can be calculated to the re-expressed model as follows:

$$y = \beta_0^{**} + \sum_{j;\, j \neq q} \beta_j^{**} x_j + \varepsilon$$

where the new parameter $\beta_j^{**}$ is the difference $\beta_j - \beta_q$, $j \in \{1, \ldots, p\} \backslash \{q\}$, and the intercept $\beta_0^{**} = \beta_q + \beta_0$ involves the effect of $x_q$(Cornell, 2011, p. 69).

Unfortunately, there are a few reasons that the interpretation of the new parameters is not straightforward to users of regression analysis. Firstly, the interpretation of $\beta_j^{**}$ depends on the chosen coefficient $\beta_q$, which is arbitrary. Second, one does not wish to remove a particular term from a model. Instead of removing $x_q$, Jang and Anderson-Cook (2010, 2014) suggested applying ridge regression directly to the Scheffe-type model (3). Third, both the Scheffe-type model and the re-expressed model are valid only if the constraints $\sum_{j=1}^{p} x_{ij} = 1$ hold for all individuals. In real-world examples of mixture experiments, however, the constraints may not hold exactly. For instance, data may fail to record the minor components in the mixture so that the sum of the proportions is less than 1. This phenomenon will be explained below.

Our motivating example to illustrate mixture experiments is the Portland cement data (Woods et al., 1932). It records the heat (cal/gram), as a response variable ($\mathbf{y}$), evolved during the hardening of Portland cements. The heat depends on the proportion of four compounds in the clinkers, defined as follows:

1. Tricalcium aluminate: $3CaO \cdot Al_2O_3$(for $\mathbf{x}_1$).
2. Tricalcium silicate: $3CaO \cdot SiO_2$(for $\mathbf{x}_2$).
3. Tetracalcium aluminoferrite: $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$(for $\mathbf{x}_3$).
4. $\beta$-dicalcium silicate: $2CaO \cdot SiO_2$(for $\mathbf{x}_4$).

The compounds come from the four original sources (CaO, $SiO_2$, $Al_2O_3$, $Fe_2O_3$), whose amounts are fixed (Table 1; Figure 2). For example, in cement No.1 to No.5, the amount of $SiO_2$ is fixed around 25 (Table 1). Thus, the proportions of $3CaO \cdot SiO_2$ ($\mathbf{x}_2$) and $2CaO \cdot SiO_2$ ($\mathbf{x}_4$) may be complementary. Indeed, in the five cements, the sums of $2CaO \cdot SiO_2$ ($\mathbf{x}_4$) and $3CaO \cdot SiO_2$ ($\mathbf{x}_2$) are 86, 81, 76, 78 and 85%, which are all around 80%($\mathbf{x}_4 + \mathbf{x}_2 \approx 80$), because the source $SiO_2$ becomes these two compounds in chemical reaction. Therefore, the multicollinearity arises when we set these compounds as our explanatory variables. In this example,

**Table 1.** Proportion (%) of the clinkers in the Portland cement data (Woods et al., 1932).

| Cement No. | $SiO_2$ | $Al_2O_3$ | $Fe_2O_3$ | CaO | |
|---|---|---|---|---|---|
| 1 | 27.7 | 3.8 | 2.0 | 65.0 | |
| 2 | 26.0 | 3.5 | 5.1 | 63.1 | |
| 3 | 21.9 | 5.7 | 2.8 | 65.0 | |
| 4 | 24.6 | 5.8 | 2.8 | 64.2 | |
| 5 | 25.0 | 3.9 | 2.1 | 66.6 | |
| Cement No. | $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$ | $3CaO \cdot Al_2O_3$ | $2CaO \cdot SiO_2$ | $3CaO \cdot SiO_2$ | |
| | $\mathbf{x}_3$ | $\mathbf{x}_1$ | $\mathbf{x}_4$ | $\mathbf{x}_2$ | $\sum_{j=1}^{4} \mathbf{x}_j$ |
| 1 | 6 | 7 | 60 | 26 | 99 |
| 2 | 15 | 1 | 52 | 29 | 97 |
| 3 | 8 | 11 | 20 | 56 | 95 |
| 4 | 8 | 11 | 47 | 31 | 97 |
| 5 | 6 | 7 | 33 | 52 | 98 |

NOTE: Five cements are extracted from data loaded from R AICcmodavg package.

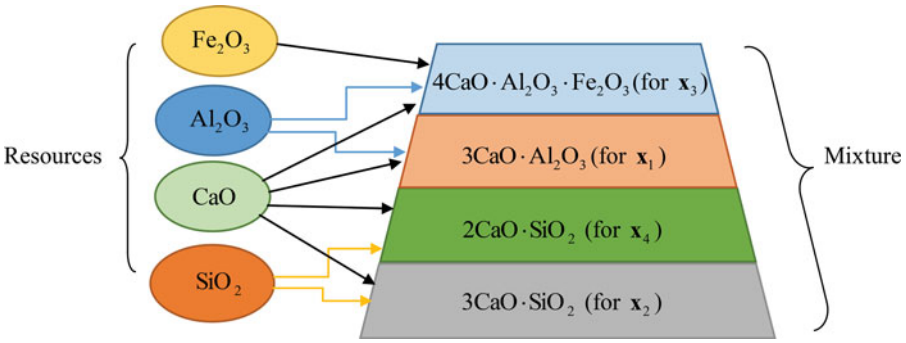


**Figure 2.** Mixture and resources in the Portland cement data (Woods et al., 1932).

the constraints are only approximately satisfied ($\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4 \approx 100$),and in particular, the sum of the proportions is less than 100% (Table 1).

### 2.4. Other ridge-type estimators

To improve the ridge estimator of Hoerl and Kennard (1970), Liu (1993) provided an estimator, called "Liu estimator":

$$\hat{\boldsymbol{\beta}}^{\mathrm{Liu}}(d) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{I})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + d\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}), \quad 0 \le d \le 1$$

Note that $\hat{\boldsymbol{\beta}}^{\mathrm{Liu}}(0) = \hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(1)$ and $\hat{\boldsymbol{\beta}}^{\mathrm{Liu}}(1) = \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$.
Liu (2003) also introduced another estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{Liu}}_{k,d} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} - d\boldsymbol{\beta}^{*}), \quad k > 0, \quad -\infty < d < \infty$$

where $\boldsymbol{\beta}^{*}$ can be any estimator of $\boldsymbol{\beta}$. Liu (2003) showed that $\hat{\boldsymbol{\beta}}^{\mathrm{Liu}}_{k,d}$ is a better estimator over the ridge estimator both on theoretic results and simulations. This estimator is called "Liu-type estimator." Notice the difference between terminologies "Liu estimator" and "Liu-type estimator." The Liu-type estimator has two parameters $d$ and $k$, where $d$ controls bias and $k$ corrects multicollinearity. Note that $\hat{\boldsymbol{\beta}}^{\mathrm{Liu}}_{k,0} = \hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ and $\hat{\boldsymbol{\beta}}^{\mathrm{Liu}}_{0,0} = \hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$. Liu (2003) suggested $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ for $\boldsymbol{\beta}^{*}$.

Sakallıoğlu and Kaçıranlar (2008) introduced a biased estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{SK}}(k, d) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{I})^{-1}\{\mathbf{X}^{\mathrm{T}}\mathbf{y} + d\hat{\boldsymbol{\beta}}(k)^{\mathrm{Ridge}}\}, \quad k > 0, \quad -\infty < d < \infty$$

This is a special case of the Liu estimator, which uses $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ as a substitute of $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$. Sakallıoğlu and Kaçıranlar (2008) discussed some advantage of $\hat{\boldsymbol{\beta}}^{\mathrm{SK}}(k, d)$ by theoretical MSE calculations and simulations.

Li and Yang (2012) introduced the so-called modified Liu estimator, which incorporates some prior knowledge on the Liu estimator. While they showed some remarkable improvement in the MSE over many other ridge-type estimators, they did not discuss the estimation of the shrinkage parameter.

A common feature of the aforementioned ridge-type estimators is that they involve one or more tuning parameters. Selection or estimation of the tuning parameters is required, and its variability needs to be accounted for the MSE evaluation. This is especially the case of the Liu-type estimator in which $\boldsymbol{\beta}^*$ and $k$ are left unspecified.

## 3. Proposed method

This section proposes a new ridge-type estimator of $\boldsymbol{\beta}$ under the intercept model (1). We follow the model and notations in Section 2.

A legitimate concern of $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ under the model (1) is that $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ shrinks the intercept toward zero as $k \to 0$. If the intercept is far from zero, $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ suffers large bias, especially for large $k$. This concern may be removed by shrinking $\hat{\boldsymbol{\beta}}$ toward non zero values, say $\boldsymbol{\beta}^* \neq \mathbf{0}$. Consider a penalized residual sum, defined as follows:

$$(\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\beta) + k(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*), \quad k \geq 0$$

where $\boldsymbol{\beta}^*$can be any estimator of $\boldsymbol{\beta}$. Then, a penalized least squares estimator is obtained as follows:

$$\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k, \boldsymbol{\beta}^*) = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}_{(p+1)})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + k\boldsymbol{\beta}^*)$$

Note that, this estimator is a special case of the Liu-type estimator with the choice $d = -k$, or $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k, \boldsymbol{\beta}^*) = \hat{\boldsymbol{\beta}}^{\mathrm{Liu}}_{k,-k}$. Such a choice has not been particularly discussed in the literature.

For the new estimator to work well, finding a good estimator $\boldsymbol{\beta}^*$ is important. We require:

(A) Estimate $\boldsymbol{\beta}^*$ is a good guess of the true $\boldsymbol{\beta}$ and does not suffer multicollinearity.

(B) Estimate $\boldsymbol{\beta}^*$ does not shrink toward $\mathbf{0}$.

(C) Estimate $\boldsymbol{\beta}^*$ permits a simple formula for the TMSE of $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ as a function of $k$.

While Liu (2003) chose $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ to estimate $\boldsymbol{\beta}^*$, the use of $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ is problematic as it suffers large variance under multicollinearity or under mixture experiments. The choice $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ is also recommended by Liu (2003) for $\boldsymbol{\beta}^*$. However, since $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ shrinks toward $\mathbf{0}$, it does not fit our purpose.

Our novel approach to find $\boldsymbol{\beta}^*$ is inspired by the "compound covariate" method originally advocated by John W. Tukey (Tukey, 1993). It is a method of prediction that aggregates the univariate regression estimates when the scales of all regressors are the same. The method intends to reduce the variance of the regression estimator in a large number of regressors. The compound covariate method has been shown to be a valid prediction method (Radamacher et al., 2002; Matsui, 2006; Emura et al., 2012, 2017; Emura and Chen, 2016) in medical applications, but it has rarely been used as an estimation method. Emura et al. (2012) considered the compound covariate method to form an estimator under high-dimensional regressors in the

Cox proportional hazards model. The resultant estimator is termed "the compound covariate estimator." Following their idea, we propose to find $\boldsymbol{\beta}^*$ as follows:

***Definition*** *(Compound Covariate Estimator):*

*(i) Use the (reduced) univariate model* $y_i = \beta_0^* + \varepsilon_i^*$ *to estimate* $\beta_0^*$ *by* $\bar{y} = \sum_{i=1}^n y_i/n$.

*(ii) After estimating the intercept by* $\hat{\beta}_0^* = \bar{y}$*, use the univariate model for* $j^{\text{th}}$*regressor:* $y_i = \hat{\beta}_0^* + \beta_j^* x_{ij} + \varepsilon_i^*$ *to estimate* $\beta_j^*$, $j = 1, \cdots, p$ *by* $\hat{\beta}_j^* = \sum_{i=1}^n x_{ij} y_i / \sum_{i=1}^n {x_{ij}}^2$.

*The compound covariate estimator is formed by* $\hat{\boldsymbol{\beta}}^* = (\hat{\beta}_0^* \hat{\beta}_1^* \cdots \hat{\beta}_p^*)^{\mathrm{T}}$.

The compound covariate estimator is alternatively derived as the minimizer of

$$RSS^*(\boldsymbol{\beta}) = \sum_{j=1}^{p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_j x_{ij})^2$$

This is the sum of univariate residual sum squares for $j = 1, \cdots, p$. This parallels Emura et al. (2012) who defined the compound covariate estimator as the maximizer of the combined univariate partial likelihoods under the Cox model.

The compound covariate estimator can be expressed in a matrix form as follows:

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \bar{y} \\ \sum_{i=1}^n x_{i1} y_i / \sum_{i=1}^n {x_{i1}}^2 \\ \vdots \\ \sum_{i=1}^n x_{ip} y_i / \sum_{i=1}^n {x_{ip}}^2 \end{bmatrix} = \{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

where $\mathrm{diag}(\Delta)$ is the diagonal matrix with the same diagonal elements as $\Delta$. If $\mathbf{X}$ is standardized to meet Equation (2), then one has a simple expression $\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \mathrm{diag}(n, n-1, \ldots, n-1)$. If $\mathbf{X}$ is not standardized, then the simple expression does apply. In general, our proposed estimator takes the form

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\mathrm{New}}(k) &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}_{(p+1)})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + k\hat{\boldsymbol{\beta}}^*) \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}_{(p+1)})^{-1}[\mathbf{I}_{(p+1)} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\mathrm{T}}\mathbf{y} \end{aligned}$$

The expression is linear in $\mathbf{y}$, which makes it easier to study the sampling distribution of $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$. It is straightforward to see that

$$\lim_{k\to\infty} E\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\} = \{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta$$

Figure 3 explains the behavior of the proposed estimator with a shrinkage toward $\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta$. The proposed shrinkage scheme is different from that of the ridge estimator that shrinks toward $\mathbf{0}$. If $\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta$ is between the true $\boldsymbol{\beta}$ and $\mathbf{0}$, then we expect that our new estimator is superior to the ridge estimator (Figure 3).

## 4. Theory

In this section, we derive a simple formula of the total mean squared error (TMSE) of the proposed estimator. We use this formula to show (I) there exists some $k > 0$ such that the TMSE of the proposed method is smaller than the TMSE of the OLS estimator and (II) the optimal value of $k$ can be derived and estimated.

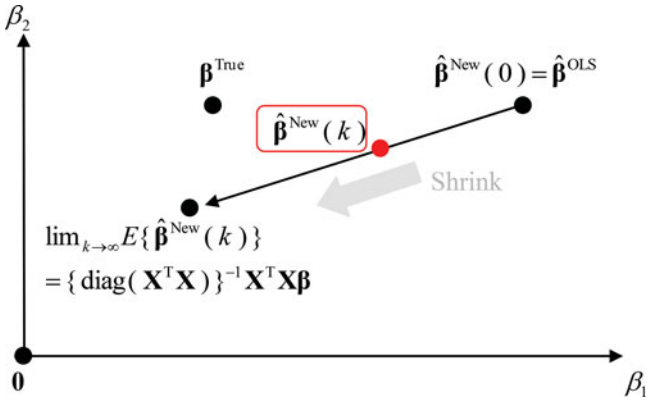


**Figure 3.** The shrinkage scheme for the new estimator when $p = 2$.

### 4.1. Mean squared error calculation

Consider a linear estimator of the form $\tilde{\boldsymbol{\beta}} = C\mathbf{y}$, where $C$ is a fixed $(p+1) \times n$ matrix. Then,

$$\text{TMSE}(\tilde{\boldsymbol{\beta}}) = E(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\text{T}}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \text{bias}(\tilde{\boldsymbol{\beta}})^{\text{T}}\text{bias}(\tilde{\boldsymbol{\beta}}) + \text{v}(\tilde{\boldsymbol{\beta}}) \tag{4}$$

where $\text{bias}(\tilde{\boldsymbol{\beta}}) \equiv E(\tilde{\boldsymbol{\beta}}) - \boldsymbol{\beta} = (C\mathbf{X} - \mathbf{I}_{(p+1)})\boldsymbol{\beta}$ and $\text{v}(\tilde{\boldsymbol{\beta}}) = \sigma^2\text{tr}\{CC^{\text{T}}\}$.

**Lemma 1.** *Suppose that* $\mathbf{X}$ *is standardized to meet Equation (2). Then,*

$$\text{B}(k) \equiv \text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\}^{\text{T}}\text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\} = \sum_{i=1}^{p}\frac{k^2\alpha_i^2(\lambda_i - n + 1)^2}{(\lambda_i + k)^2(n-1)^2}$$

$$\text{V}(k) \equiv \text{v}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\} = \sigma^2\left\{\frac{1}{n} + \sum_{i=1}^{p}\frac{\lambda_i(k+n-1)^2}{(\lambda_i + k)^2(n-1)^2}\right\}$$

**Proof of Lemma 1:** Recall that $\boldsymbol{\Gamma}^{\text{T}}\mathbf{X}^{\text{T}}\mathbf{X}\boldsymbol{\Gamma} = \boldsymbol{\Lambda} = \text{diag}(n, \lambda_1, \ldots, \lambda_p)$ in Section 2.2. Then, the bias is as follows:

$$\begin{aligned}
&\text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\} \\
&= \{(\mathbf{X}^{\text{T}}\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\text{diag}(\mathbf{X}^{\text{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\text{T}}\mathbf{X} - \mathbf{I}\}\boldsymbol{\beta} \\
&= \boldsymbol{\Gamma}\{(\boldsymbol{\Lambda} + k\mathbf{I})^{-1}[\mathbf{I} + k\boldsymbol{\Gamma}^{\text{T}}\{\text{diag}(\mathbf{X}^{\text{T}}\mathbf{X})\}^{-1}\boldsymbol{\Gamma}]\boldsymbol{\Lambda} - \mathbf{I}\}\boldsymbol{\alpha} \\
&= \boldsymbol{\Gamma}\left(\begin{bmatrix} \frac{1}{n+k} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_1+k} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_p+k} \end{bmatrix}\begin{bmatrix} \frac{k+n}{n} & 0 & \cdots & 0 \\ 0 & \frac{k+n-1}{n-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{k+n-1}{n-1} \end{bmatrix}\begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & \lambda_1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} - \mathbf{I}\right)\boldsymbol{\alpha}
\end{aligned}$$

$$= \boldsymbol{\Gamma} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \dfrac{k(\lambda_1 - n + 1)}{(\lambda_1 + k)(n-1)} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{k(\lambda_p - n + 1)}{(\lambda_p + k)(n-1)} \end{bmatrix} \boldsymbol{\alpha}$$

where we simply write $\mathbf{I} = \mathbf{I}_{(p+1)}$. Since $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \mathbf{I}_{(p+1)}$, the bias becomes

$$\begin{aligned} &\mathrm{bias}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\}^{\mathrm{T}}\mathrm{bias}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\} \\ &= \boldsymbol{\alpha}^{\mathrm{T}} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \dfrac{k(\lambda_1 - n + 1)}{(\lambda_1 + k)(n-1)} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{k(\lambda_p - n + 1)}{(\lambda_p + k)(n-1)} \end{bmatrix} \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma} \\ &\quad \times \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \dfrac{k(\lambda_1 - n + 1)}{(\lambda_1 + k)(n-1)} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{k(\lambda_p - n + 1)}{(\lambda_p + k)(n-1)} \end{bmatrix} \boldsymbol{\alpha} \\ &= \sum_{i=1}^{p} \frac{k^2\alpha_i^2(\lambda_i - n + 1)^2}{(\lambda_i + k)^2(n-1)^2} \end{aligned}$$

Next, the total variance is calculated as follows:

$$\begin{aligned} &\mathrm{v}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\} \\ &= \sigma^2\mathrm{tr}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\mathrm{T}}\mathbf{X}[\mathbf{I} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}](\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}\} \\ &= \sigma^2\mathrm{tr}\{\boldsymbol{\Gamma}(\boldsymbol{\Lambda} + k\mathbf{I})^{-1}[\mathbf{I} + k\boldsymbol{\Gamma}^{\mathrm{T}}\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}\boldsymbol{\Gamma}]\boldsymbol{\Lambda}[\mathbf{I} + k\boldsymbol{\Gamma}^{\mathrm{T}}\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}\boldsymbol{\Gamma}](\boldsymbol{\Lambda} + k\mathbf{I})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\} \\ &= \sigma^2\mathrm{tr}\left\{\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\begin{bmatrix} \dfrac{1}{n+k} & 0 & \cdots & 0 \\ 0 & \dfrac{1}{\lambda_1 + k} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{1}{\lambda_p + k} \end{bmatrix}^2 \begin{bmatrix} \dfrac{k+n}{n} & 0 & \cdots & 0 \\ 0 & \dfrac{k+n-1}{n-1} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \dfrac{k+n-1}{n-1} \end{bmatrix}^2 \right. \end{aligned}$$

$$\times \left\{ \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & \lambda_1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} \right\}$$

$$= \sigma^2 \left\{ \frac{1}{n} + \sum_{i=1}^{p} \frac{\lambda_i (k+n-1)^2}{(\lambda_i + k)^2 (n-1)^2} \right\}$$

□

Lemma 1 immediately yields the following theorem:

**Theorem 1.** *Suppose that* $\mathbf{X}$ *is standardized to meet Equation (2). Then,*

$$\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\} = \sum_{i=1}^{p} \frac{k^2 \alpha_i^2 (\lambda_i - n + 1)^2}{(\lambda_i + k)^2 (n-1)^2} + \sigma^2 \left\{ \frac{1}{n} + \sum_{i=1}^{p} \frac{\lambda_i (k+n-1)^2}{(\lambda_i + k)^2 (n-1)^2} \right\}$$

*Furthermore, the preceding formula does not depend on the intercept term* $\beta_0 = \alpha_0$.

### 4.2. Existence theorem

We give a similar result as the existence theorem of the ridge estimator by Hoerl and Kennard (1970). In our proposed estimator, the statement of the existence theorem is the same as Hoerl and Kennard (1970), but the proof is more complicated.

**Theorem 2.** *(Existence theorem of the new estimator) Suppose that* $\mathbf{X}$ *is standardized to meet Equation (2). There always exist some small value of* $k > 0$ *such that the proposed estimator strictly improves upon the OLS estimator in the sense of* $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\} < \mathrm{TMSE}(\hat{\boldsymbol{\beta}}^{\mathrm{OLS}})$.

**Proof of Theorem 2.**

From Theorem 1,

$$\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\} = \sum_{i=1}^{p} \frac{k^2 \alpha_i^2 (\lambda_i - n + 1)^2}{(\lambda_i + k)^2 (n-1)^2} + \sigma^2 \left\{ \frac{1}{n} + \sum_{i=1}^{p} \frac{\lambda_i (k+n-1)^2}{(\lambda_i + k)^2 (n-1)^2} \right\}$$
$$\equiv \mathrm{B}(k) + \mathrm{V}(k)$$

where $\mathrm{B}(k)$ and $\mathrm{V}(k)$ are available in Lemma 1. Accordingly,

$$\begin{cases} \dfrac{d}{dk}\mathrm{B}(k) = \displaystyle\sum_{i=1}^{p} \frac{2k\alpha_i^2 \lambda_i (\lambda_i - n + 1)^2}{(\lambda_i + k)^3 (n-1)^2} & \Rightarrow \lim\limits_{k \to 0^+} \dfrac{d}{dk}\mathrm{B}(k) = 0. \\ \dfrac{d}{dk}\mathrm{V}(k) = \sigma^2 \displaystyle\sum_{i=1}^{p} \frac{2\lambda_i (\lambda_i - n + 1)(k+n-1)}{(\lambda_i + k)^3 (n-1)^2} & \Rightarrow \lim\limits_{k \to 0^+} \dfrac{d}{dk}\mathrm{V}(k) = 2\sigma^2 \displaystyle\sum_{i=1}^{p} \frac{(\lambda_i - n + 1)}{\lambda_i^2 (n-1)} \end{cases}$$

This means that the sign of $\lim_{k \to 0^+} d\mathrm{V}(k)/dk$ determines the slope of the TMSE function at $k = 0^+$. Recall that $\lambda_1 \geq \cdots \geq \lambda_p > 0$. Here, we consider three cases: **Case (i)** $n > \lambda_1 + 1$, **Case (ii)** $\lambda_s + 1 \geq n > \lambda_{s+1} + 1$, $1 \leq s < p$, and **Case (iii)** $\lambda_p + 1 \geq n$ (see Figure 4).

**Case (i):** We have $(\lambda_i - n + 1) < 0$ for all $i = 1, \ldots, p$. Thus, $\lim_{k \to 0^+} d\mathrm{V}(k)/dk < 0$.

**Case (ii):** We divide the summation into two parts.

$$\lim_{k\to 0^+}\frac{d}{dk}\mathrm{V}(k) = 2\sigma^2\sum_{i=1}^{p}\frac{(\lambda_i - n + 1)}{{\lambda_i}^2(n-1)}$$
$$= \underbrace{2\sigma^2\left\{\sum_{i=1}^{s}\frac{(\lambda_i - n + 1)}{{\lambda_i}^2(n-1)}\right\}}_{\boxed{\text{nonnegative}}} + \underbrace{2\sigma^2\left\{\sum_{i=s+1}^{p}\frac{(\lambda_i - n + 1)}{{\lambda_i}^2(n-1)}\right\}}_{\boxed{\text{negative}}}$$

Since $\lambda_i \geq \lambda_s$ for all $i = 1, \ldots, s$, and $\lambda_i \leq \lambda_s$ for all $i = s+1, \ldots, p$,

$$\lim_{k\to 0^+}\frac{d}{dk}\mathrm{V}(k) < 2\sigma^2\left\{\sum_{i=1}^{s}\frac{(\lambda_i - n + 1)}{{\lambda_s}^2(n-1)}\right\} + 2\sigma^2\left\{\sum_{i=s+1}^{p}\frac{(\lambda_i - n + 1)}{{\lambda_s}^2(n-1)}\right\}$$
$$= \frac{2\sigma^2}{{\lambda_s}^2(n-1)}\sum_{i=1}^{p}(\lambda_i - n + 1) = \frac{2\sigma^2}{{\lambda_s}^2(n-1)}\left(\sum_{i=1}^{p}\lambda_i - np + p\right)$$
$$= \frac{2\sigma^2}{{\lambda_s}^2(n-1)}\left\{\mathrm{tr}\left(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p\right) - np + p\right\} = \frac{2\sigma^2}{{\lambda_s}^2(n-1)}\{p(n-1) - np + p\}$$
$$= 0$$

Thus, we obtain $\lim_{k\to 0^+} d\mathrm{V}(k)/dk < 0$.

**Case (iii):** The case of $\lambda_p + 1 \geq n$ corresponds to small $n$. This actually never happens in our model; since $p\lambda_p < \sum_{i=1}^{p}\lambda_i = \mathrm{tr}(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p) = p(n-1)$, we have $\lambda_p < n - 1$.

From the results of **Cases (i)–(iii)**, it holds that $\lim_{k\to 0^+} dT\ \mathrm{MSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\}/dt < 0$. This implies the conclusion of Theorem 2. □

### 4.3. Optimal value of k

There exist many different ways to estimate the shrinkage parameter $k > 0$ for the ridge regression. See Wong and Chiu (2015) that give a comprehensive list of all the available estimators in the literature. These available estimators cannot be directly applied to our new estimator.

Since the estimator $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ is linear in $\mathbf{y}$, the formula of $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\}$ is easily computed by Equation (4). The optimal value for $k$ that minimizes $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\}$ is as follows:

$$k^{\mathrm{New}} = \arg\min_{k\geq 0}[\sigma^2\mathrm{tr}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\mathrm{T}}$$

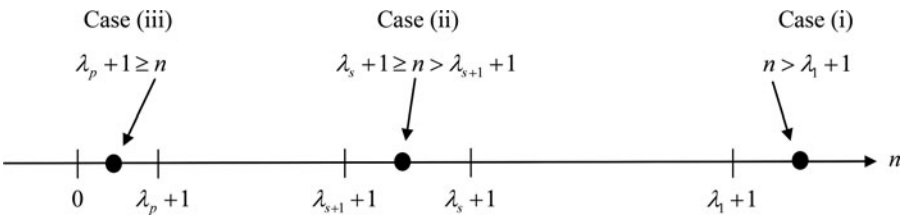


**Figure 4.** Three cases occurring in the proof of Theorem 2.

$$\times \mathbf{X}[\mathbf{I} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}](\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}\}$$

$$+ \|\{(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{I}\}\boldsymbol{\beta}\|^2]$$

If $\mathbf{X}$ is standardized to meet Equation (2), one can also apply the equivalent formula

$$k^{\mathrm{New}} = \arg\min_{k\geq 0}\left[\sum_{i=1}^{p}\frac{k^2\alpha_i^2(\lambda_i - n + 1)^2}{(\lambda_i + k)^2(n-1)^2} + \sigma^2\left\{\frac{1}{n} + \sum_{i=1}^{p}\frac{\lambda_i(k+n-1)^2}{(\lambda_i+k)^2(n-1)^2}\right\}\right]$$

If we use $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\beta}^{\mathrm{OLS}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\beta}^{\mathrm{OLS}})/(n-p-1)$, and $\hat{\boldsymbol{\alpha}}^{\mathrm{OLS}} = \boldsymbol{\Gamma}^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$ in place of $\boldsymbol{\beta}$, $\sigma^2$, and $\boldsymbol{\alpha}$, respectively, then the estimated optimal shrinkage parameter for $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ is as follows:

$$\hat{k}^{\mathrm{New}} = \arg\min_{k\geq 0}[\hat{\sigma}^2\mathrm{tr}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\mathrm{T}}$$

$$\times\mathbf{X}[\mathbf{I} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}](\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}\}$$

$$+ \|\{(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\mathrm{T}}\mathbf{X} - \mathbf{I}\}\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}\|^2]$$

or

$$\hat{k}^{\mathrm{New}} = \arg\min_{k\geq 0}\left[\sum_{i=1}^{p}\frac{k^2\hat{\alpha}_i^{2,\mathrm{OLS}}(\lambda_i - n + 1)^2}{(\lambda_i + k)^2(n-1)^2} + \hat{\sigma}^2\left\{\frac{1}{n} + \sum_{i=1}^{p}\frac{\lambda_i(k+n-1)^2}{(\lambda_i+k)^2(n-1)^2}\right\}\right]$$

The numerical minimization can be done, for example, by R "optimize" routine.

## 5. Simulation

This section examines whether the new estimator improves upon the OLS estimator and the ridge estimator in terms of the TMSE. In addition, we examine the correctness of our theoretical formulas for bias, variance, and TMSE obtained in Lemma 1 and Theorem 1.

### 5.1. Simulation design

We generate data by mimicking the setting of the Portland cement data that have a four-component mixture ($\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4 \approx 100$) under $n = 13$ and $p = 4$. We consider four cases:

Case 1: $\boldsymbol{\beta} = (50, 1, 1, 1, 1)^{\mathrm{T}}$and $\sigma^2 = 1$,
Case 2: $\boldsymbol{\beta} = (50, 1, 1, 1, 1)^{\mathrm{T}}$and $\sigma^2 = 2$,
Case 3: $\boldsymbol{\beta} = (1, 1, 1, 1, 1)^{\mathrm{T}}$and $\sigma^2 = 1$,
Case 4: $\boldsymbol{\beta} = (1, 1, 1, 1, 1)^{\mathrm{T}}$and $\sigma^2 = 2$.

**Step 1.** Generate the design matrix $\mathbf{X}_p = (\mathbf{x}_1, \ldots, \mathbf{x}_4)$ as follows:

$$\begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix} \sim N_2\left(\begin{bmatrix} 12 \\ 25 \end{bmatrix}, 5 \times \mathbf{I}_2\right), \quad \begin{bmatrix} x_{i3} \\ x_{i4} \end{bmatrix} \sim \begin{bmatrix} -x_{i1} \\ -x_{i2} \end{bmatrix} + N_2\left(\begin{bmatrix} 50 \\ 50 \end{bmatrix}, 5 \times \mathbf{I}_2\right), \; i = 1, \cdots, n$$

In this way, we have the approximate constraint ($\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4 \approx 100$), and the multicollinearity ($\mathbf{x}_2 + \mathbf{x}_4 \approx 50$) and ($\mathbf{x}_1 + \mathbf{x}_3 \approx 50$). After standardization to meet Equation (2),

we get

$$\mathbf{X} = \begin{bmatrix} 1 & -1.015 & -2.184 & 0.660 & 2.082 \\ 1 & -0.020 & 0.940 & -1.170 & -0.927 \\ 1 & -1.272 & -0.154 & 0.563 & -0.165 \\ 1 & 1.713 & -0.127 & -0.759 & 0.485 \\ 1 & 0.159 & 0.771 & 1.141 & -0.291 \\ 1 & -1.253 & 0.656 & 0.873 & -1.083 \\ 1 & 0.353 & 0.443 & 0.169 & -0.936 \\ 1 & 0.661 & 0.748 & -0.428 & -0.409 \\ 1 & 0.461 & 0.620 & -1.427 & -0.028 \\ 1 & -0.621 & -0.042 & 0.160 & -0.150 \\ 1 & 1.610 & -1.973 & -1.387 & 2.006 \\ 1 & 0.233 & 0.468 & -0.132 & -0.174 \\ 1 & -1.008 & -0.165 & 1.738 & -0.410 \end{bmatrix} \equiv [\mathbf{1}, \mathbf{X}_p]$$

The sample correlation form of $\mathbf{X}_p$ is as follows:

$$\text{Sample Corr}(\mathbf{X}_p) = \begin{bmatrix} 1.000 & -0.039 & -0.714 & 0.278 \\ -0.039 & 1.000 & 0.014 & -0.924 \\ -0.714 & 0.014 & 1.000 & -0.227 \\ 0.278 & -0.924 & -0.227 & 1.000 \end{bmatrix}$$

We see that the four regressors are correlated as expected.

**Step 2**. Given $\mathbf{X}$, generate $\boldsymbol{\varepsilon}^{(r)} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and set $\mathbf{y}^{(r)} = \mathbf{X}\beta + \boldsymbol{\varepsilon}^{(r)}$ for $r = 1, \cdots, 100,000$.

Then, $\hat{\boldsymbol{\beta}}^{\text{New}}(k)^{(r)} = (\mathbf{X}^{\text{T}}\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}^{\text{T}}\mathbf{y}^{(r)} + k\hat{\boldsymbol{\beta}}^{*(r)})$ is computed, where $\hat{\boldsymbol{\beta}}^{*(r)} = \{\text{diag}(\mathbf{X}^{\text{T}}\mathbf{X})\}^{-1}\mathbf{X}^{\text{T}}\mathbf{y}^{(r)}$. Compute the approximate bias, variance, and TMSE as follows:

$$\begin{aligned} \text{B}^{\text{New}}(k) &= \{\bar{\hat{\boldsymbol{\beta}}}^{\text{New}}(k)^{(\cdot)} - \boldsymbol{\beta}\}^{\text{T}}\{\bar{\hat{\boldsymbol{\beta}}}^{\text{New}}(k)^{(\cdot)} - \boldsymbol{\beta}\} \\ \text{V}^{\text{New}}(k) &= \frac{1}{100,000}\sum_{r=1}^{100,000}[\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)^{(r)} - \bar{\hat{\boldsymbol{\beta}}}^{\text{New}}(k)^{(\cdot)}\}^{\text{T}}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)^{(r)} - \bar{\hat{\boldsymbol{\beta}}}^{\text{New}}(k)^{(\cdot)}\}] \\ \text{TMSE}^{\text{New}}(k) &= \frac{1}{100,000}\sum_{r=1}^{100,000}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)^{(r)} - \boldsymbol{\beta}\}^{\text{T}}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)^{(r)} - \boldsymbol{\beta}\} \end{aligned}$$

where $\bar{\hat{\boldsymbol{\beta}}}^{\text{New}}(k)^{(\cdot)} = \sum_{r=1}^{100,000}\hat{\boldsymbol{\beta}}^{\text{New}}(k)^{(r)}/100,000$.

**Step 3**. Plot the bias, variance, and TMSE against $k$ and then identify the minimizer of the TMSE, $k^{\text{New}}$.

Perform the same algorithms for $\hat{\boldsymbol{\beta}}^{\text{Ridge}}(k)^{(r)}$ and then identify the minimizer of the TMSE, $k^{\text{Ridge}}$.

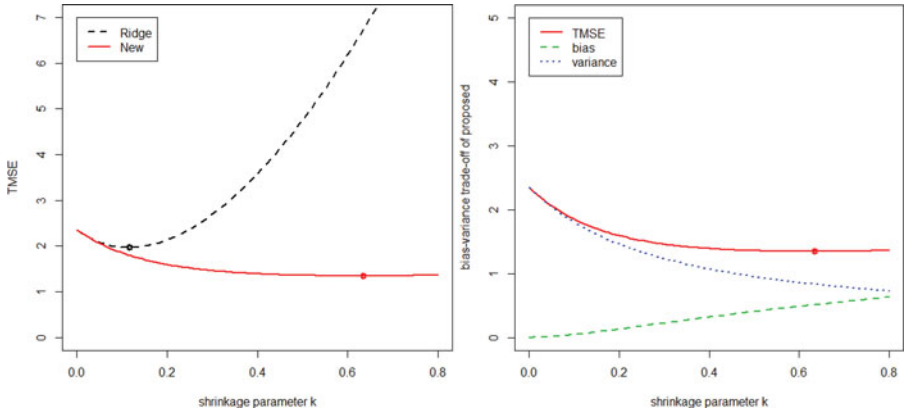


**Figure 5.** Simulation results for the TMSE plot under $\boldsymbol{\beta} = (50, 1, 1, 1, 1)^{\mathrm{T}}$ and $\sigma^2 = 1$ (Case 1). The black point is $k^{\mathrm{Ridge}} = 0.1163$ (left) and the red point is $k^{\mathrm{New}} = 0.6343$ (left and right). The right plot shows the bias–variance trade-off the proposed estimator.

### 5.2. Simulation result (comparison between $\hat{\beta}^{\mathrm{Ridge}}(k)$ and $\hat{\beta}^{\mathrm{New}}(k)$)

Figures 5–8 depict the TMSE plots. Both $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ and $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ show a good amount of improvement in TMSE relative to $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = \hat{\boldsymbol{\beta}}^{\mathrm{New}}(0) = \hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(0)$ in a range of $k > 0$. This implies that both $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ and $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ fix the problem of multicollinearity.

Figures 5 and 6 show that $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ has the smaller TMSE than the ridge estimator $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ when the intercept term is large (when $\boldsymbol{\beta} = (50, 1, 1, 1, 1)^{\mathrm{T}}$). The superior performance of $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ over $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ is more remarkable when $k$ is larger. The worse performance of $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ with large $k$ is due to too much shrinkage of the true intercept term. Then, the superiority of $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ is attributable to the different shrinkage schemes of the intercept term by the compound covariate estimator.

Figures 7 and 8 show that $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ is slightly worse than $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ if the intercept term is small (when $\boldsymbol{\beta} = (1, 1, 1, 1, 1)^{\mathrm{T}}$). However, the difference is very small.

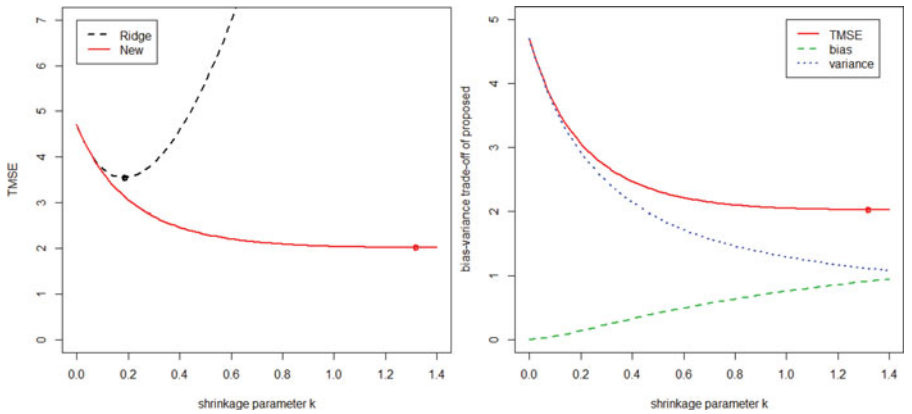


**Figure 6.** Simulation results for the TMSE plot under $\boldsymbol{\beta} = (50, 1, 1, 1, 1)^{\mathrm{T}}$ and $\sigma^2 = 2$ (Case 2). The black point is $k^{\mathrm{Ridge}} = 0.1850$ (left), and the red point is $k^{\mathrm{New}} = 1.3171$ (left and right). The right plot shows the bias–variance trade-off the proposed estimator.

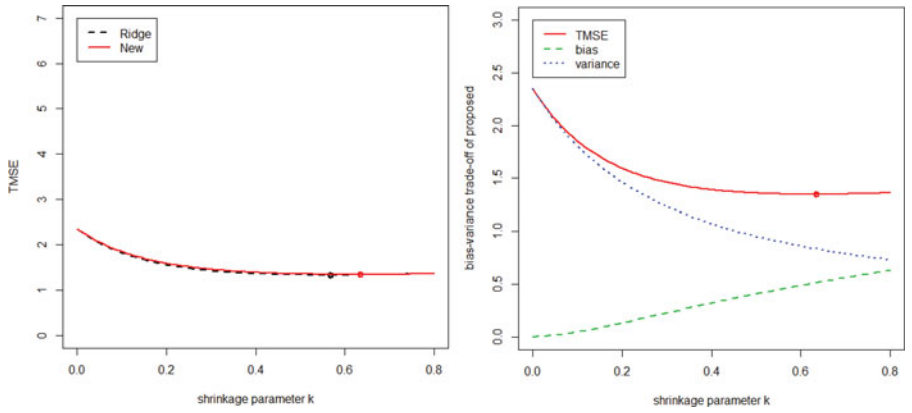


**Figure 7.** Simulation results for the TMSE plot under $\boldsymbol{\beta} = (1, 1, 1, 1, 1)^{\mathrm{T}}$ and $\sigma^2 = 1$ (Case 3). The black point is $k^{\mathrm{Ridge}} = 0.5676$ (left), and the red point is $k^{\mathrm{New}} = 0.6343$ (left and right). The right plot shows the bias–variance trade-off the proposed estimator.

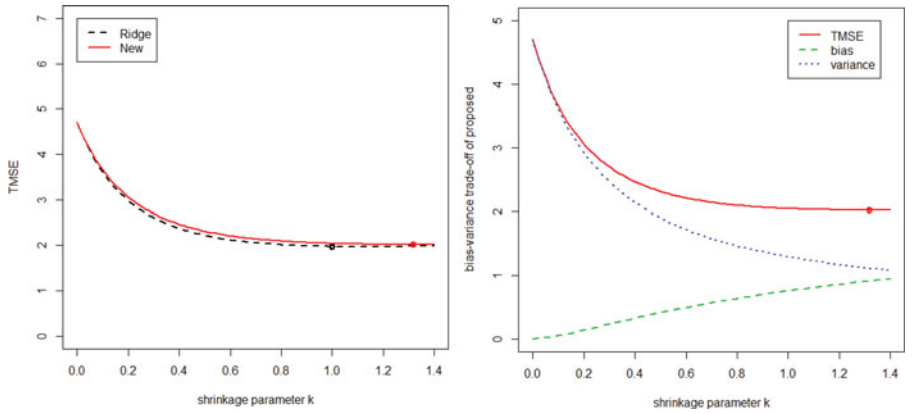


**Figure 8.** Simulation results for the TMSE plot under $\boldsymbol{\beta} = (1, 1, 1, 1, 1)^{\mathrm{T}}$ and $\sigma^2 = 2$ (Case 4). The black point is $k^{\mathrm{Ridge}} = 1.1234$ (left), and the red point is $k^{\mathrm{New}} = 1.3171$ (left and right). The right plot shows the bias–variance trade-off the proposed estimator.

An important result is that the TMSE of $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ is invariant for the change of the intercept term. This property can also be verified theoretically (Theorem 1). Therefore, if the model has a large intercept term, $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ is more desired than $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ as a way to fix multicollinearity.

Figures 5–8 reveal that the bias is flat, and the total variance decreases steeply at $k = 0^{+}$. These results numerically support the conclusion of the existence theorem (Theorem 2).

Table 2 summarizes the conclusion. The new method is more robust against the change of the intercept term than the ridge estimator is.

**Table 2.** The effects of intercept term and $\sigma^2$ on two estimators (Ridge and New).

| | $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ | $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ |
|---|---|---|
| Large intercept term | affected | not affected |
| Large $\sigma^2$ | affected | affected |

**Table 3.** The bias of the proposed estimator $\hat{\boldsymbol{\beta}}^{\text{New}}(k)$, namely $\text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\}^{\text{T}}\text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\}$.

| | | $k = 0.2$ | $k = 0.4$ | $k = 0.6$ | $k = 0.8$ |
|---|---|---|---|---|---|
| $\boldsymbol{\beta} = \begin{bmatrix}50\\1\\1\\1\\1\end{bmatrix}, \sigma^2 = 1$ | Formula (i) | 0.1250 | 0.3104 | 0.4779 | 0.6202 |
| | Formula (ii) | 0.1250 | 0.3104 | 0.4779 | 0.6202 |
| | Monte Carlo | 0.1246 | 0.3099 | 0.4773 | 0.6196 |
| $\boldsymbol{\beta} = \begin{bmatrix}50\\1\\1\\1\\1\end{bmatrix}, \sigma^2 = 2$ | Formula (i) | 0.1250 | 0.3104 | 0.4779 | 0.6202 |
| | Formula (ii) | 0.1250 | 0.3104 | 0.4779 | 0.6202 |
| | Monte Carlo | 0.1244 | 0.3096 | 0.4770 | 0.6194 |
| $\boldsymbol{\beta} = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \sigma^2 = 1$ | Formula (i) | 0.1250 | 0.3104 | 0.4779 | 0.6202 |
| | Formula (ii) | 0.1250 | 0.3104 | 0.4779 | 0.6202 |
| | Monte Carlo | 0.1246 | 0.3099 | 0.4773 | 0.6196 |
| $\boldsymbol{\beta} = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \sigma^2 = 2$ | Formula (i) | 0.1250 | 0.3104 | 0.4779 | 0.6202 |
| | Formula (ii) | 0.1250 | 0.3104 | 0.4779 | 0.6202 |
| | Monte Carlo | 0.1244 | 0.3096 | 0.4770 | 0.6194 |

Formula (i):

$$\text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\}^{\text{T}}\text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\} = \boldsymbol{\beta}^{\text{T}}\{(\mathbf{X}^{\text{T}}\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\text{diag}(\mathbf{X}^{\text{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\text{T}}\mathbf{X} - \mathbf{I}\}^{\text{T}} \times \{(\mathbf{X}^{\text{T}}\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\text{diag}(\mathbf{X}^{\text{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\text{T}}\mathbf{X} - \mathbf{I}\}\boldsymbol{\beta}$$

Formula (ii): $\text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\}^{\text{T}}\text{bias}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\} = \sum_{i=1}^{p} \frac{k^2\alpha_i^2(\lambda_i - n + 1)^2}{(\lambda_i + k)^2(n-1)^2}$

Monte Carlo: $\text{B}^{\text{New}}(k) = \{\bar{\hat{\boldsymbol{\beta}}}^{\text{New}}(k)^{(\cdot)} - \boldsymbol{\beta}\}^{\text{T}}\{\bar{\hat{\boldsymbol{\beta}}}^{\text{New}}(k)^{(\cdot)} - \boldsymbol{\beta}\}$

### 5.3. Result (checking theoretical properties of $\hat{\boldsymbol{\beta}}^{\text{New}}(k)$)

In Sections 5.1 and 5.2, we use the Monte Carlo approximations to examine the bias, total variance, and TMSE. One can alternatively use the exact expressions for the bias, variance, and TMSE (Lemma 1 and Theorem 1), which are free from error due to the finite number of Monte Carlo replications.

We calculate exact values of the bias, total variance, and TMSE of the proposed estimator at $k = 0.2, 0.4, 0.6$, and $0.8$ using the formulas in Lemma 1 and Theorem 1 and compare them with the Monte Carlo approximated values. Tables 3–5 show that the exact values of the bias, total variance, and TMSE (from Lemma 1 and Theorem 1) are very close to the Monte Carlo versions in all simulation settings. Furthermore, the numerical results of Tables 3–5 agree with Figures 5–8. Therefore, we have numerically verified the correctness of our theoretical formulas.

### 5.4. TMSE comparison under estimated parameter k

So far, the performance of the estimators is compared under fixed $k$. In reality, one needs to estimate $k$ by data, which induces some variation. Therefore, we take into account for the

**Table 4.** The total variance of the proposed estimator $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$, namely $\mathrm{v}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\}$.

| | | $k=0.2$ | $k=0.4$ | $k=0.6$ | $k=0.8$ |
|---|---|---|---|---|---|
| $\boldsymbol{\beta}=\begin{bmatrix}50\\1\\1\\1\\1\end{bmatrix}, \sigma^2=1$ | Formula (i) | 1.4767 | 1.0784 | 0.8641 | 0.7332 |
| | Formula (ii) | 1.4767 | 1.0784 | 0.8641 | 0.7332 |
| | Monte Carlo (SE) | 1.4799(0.0048) | 1.0806(0.0032) | 0.8658(0.0023) | 0.7346(0.0019) |
| $\boldsymbol{\beta}=\begin{bmatrix}50\\1\\1\\1\\1\end{bmatrix}, \sigma^2=2$ | Formula (i) | 2.9533 | 2.1568 | 1.7281 | 1.4664 |
| | Formula (ii) | 2.9533 | 2.1568 | 1.7281 | 1.4664 |
| | Monte Carlo (SE) | 2.9597(0.0096) | 2.1612(0.0064) | 1.7315(0.0047) | 1.4691(0.0037) |
| $\boldsymbol{\beta}=\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \sigma^2=1$ | Formula (i) | 1.4767 | 1.0784 | 0.8641 | 0.7332 |
| | Formula (ii) | 1.4767 | 1.0784 | 0.8641 | 0.7332 |
| | Monte Carlo (SE) | 1.4799(0.0048) | 1.0806(0.0032) | 0.8658(0.0023) | 0.7346(0.0019) |
| $\boldsymbol{\beta}=\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \sigma^2=2$ | Formula (i) | 2.9533 | 2.1568 | 1.7281 | 1.4664 |
| | Formula (ii) | 2.9533 | 2.1568 | 1.7281 | 1.4664 |
| | Monte Carlo (SE) | 2.9597(0.0096) | 2.1612(0.0064) | 1.7315(0.0047) | 1.4691(0.0037) |

Formula (i):

$$\sigma^2\mathrm{tr}\{(\mathbf{X}^{\mathrm{T}}\mathbf{X}+k\mathbf{I})^{-1}[\mathbf{I}+k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\mathrm{T}}\mathbf{X}[\mathbf{I}+k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}](\mathbf{X}^{\mathrm{T}}\mathbf{X}+k\mathbf{I})^{-1}\}$$

Formula (ii) : $\mathrm{v}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)\}=\sigma^2\{\frac{1}{n}+\sum_{i=1}^{p}\frac{\lambda_i(k+n-1)^2}{(\lambda_i+k)^2(n-1)^2}\}$

Monte Carlo : $\mathrm{V}^{\mathrm{New}}(k)=\frac{1}{100,000}\sum_{r=1}^{100,000}[\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)^{(r)}-\bar{\hat{\boldsymbol{\beta}}}^{\mathrm{New}}(k)^{(\cdot)}\}^{\mathrm{T}}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)^{(r)}-\bar{\hat{\boldsymbol{\beta}}}^{\mathrm{New}}(k)^{(\cdot)}\}]$

variation of $k$ by changing Step 2 as follows:

$$\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}})\}=\frac{1}{100,000}\sum_{r=1}^{100,000}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}^{(r)}})^{(r)}-\boldsymbol{\beta}\}^{\mathrm{T}}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}^{(r)}})^{(r)}-\boldsymbol{\beta}\}$$

where $\hat{k}^{\mathrm{New}^{(r)}}$ is based on the data $\mathbf{y}^{(r)}=\mathbf{X}\beta+\boldsymbol{\varepsilon}^{(r)}$. Similarly, $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\hat{k}^{\mathrm{Ridge}})\}$ is defined.

A comparison between $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}})\}$ and $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\hat{k}^{\mathrm{Ridge}})\}$ is made in Table 6. Both $\hat{k}^{\mathrm{New}}$ and $\hat{k}^{\mathrm{Ridge}}$ are reasonably good estimators of their true $k^{\mathrm{New}}$ and $k^{\mathrm{Ridge}}$, respectively. Accordingly, $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}})\}$ is smaller than $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\hat{k}^{\mathrm{Ridge}})\}$ in the case of a large intercept (first and second columns). On the other hand, $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}})\}$ is slightly larger than $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\hat{k}^{\mathrm{Ridge}})\}$ in the case of a small intercept (third and fourth columns). Importantly, $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}})\}$ is not affected by the size of the intercept term, while $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\hat{k}^{\mathrm{Ridge}})\}$ is. These results agree with Figures 5–8.

## 6. Real data analysis

We analyze the dataset on Portland cement (Woods et al., 1932). As described in Section 2.3, these data are obtained from a four-component mixture experiment that does not exactly meet the constraint $(\mathbf{x}_1+\mathbf{x}_2+\mathbf{x}_3+\mathbf{x}_4=1)$; in particular, the sum of the four proportions is

**Table 5.** The TMSE of the proposed estimator $\hat{\boldsymbol{\beta}}^{\text{New}}(k)$, namely $\text{TMSE}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\}$.

| | | $k = 0.2$ | $k = 0.4$ | $k = 0.6$ | $k = 0.8$ |
|---|---|---|---|---|---|
| $\boldsymbol{\beta} = \begin{bmatrix}50\\1\\1\\1\\1\end{bmatrix}, \sigma^2 = 1$ | Formula (i) | 1.6017 | 1.3888 | 1.3419 | 1.3534 |
| | Formula (ii) | 1.6017 | 1.3888 | 1.3419 | 1.3534 |
| | Monte Carlo (SE) | 1.6044(0.0053) | 1.3905(0.0042) | 1.3430(0.0037) | 1.3542(0.0034) |
| $\boldsymbol{\beta} = \begin{bmatrix}50\\1\\1\\1\\1\end{bmatrix}, \sigma^2 = 2$ | Formula (i) | 3.0783 | 2.4673 | 2.2060 | 2.0866 |
| | Formula (ii) | 3.0783 | 2.4673 | 2.2060 | 2.0866 |
| | Monte Carlo (SE) | 3.0841(0.0101) | 2.4708(0.0075) | 2.2085(0.0062) | 2.0885(0.0055) |
| $\boldsymbol{\beta} = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \sigma^2 = 1$ | Formula (i) | 1.6017 | 1.3888 | 1.3419 | 1.3534 |
| | Formula (ii) | 1.6017 | 1.3888 | 1.3419 | 1.3534 |
| | Monte Carlo (SE) | 1.6044(0.0053) | 1.3905(0.0042) | 1.3430(0.0037) | 1.3542(0.0034) |
| $\boldsymbol{\beta} = \begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \sigma^2 = 2$ | Formula (i) | 3.0783 | 2.4673 | 2.2060 | 2.0866 |
| | Formula (ii) | 3.0783 | 2.4673 | 2.2060 | 2.0866 |
| | Monte Carlo (SE) | 3.0841(0.0101) | 2.4708(0.0075) | 2.2085(0.0062) | 2.0885(0.0055) |

Formula (i):

$$\sigma^2 \text{tr}\{(\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\text{diag}(\mathbf{X}^\top\mathbf{X})\}^{-1}]\mathbf{X}^\top\mathbf{X}[\mathbf{I} + k\{\text{diag}(\mathbf{X}^\top\mathbf{X})\}^{-1}](\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}\}$$
$$+ \|\{(\mathbf{X}^\top\mathbf{X} + k\mathbf{I})^{-1}[\mathbf{I} + k\{\text{diag}(\mathbf{X}^\top\mathbf{X})\}^{-1}]\mathbf{X}^\top\mathbf{X} - \mathbf{I}\}\boldsymbol{\beta}\|^2$$

Formula (ii) : $\text{TMSE}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)\} = \sum_{i=1}^{p} \frac{k^2\alpha_i^2(\lambda_i - n + 1)^2}{(\lambda_i + k)^2(n-1)^2} + \sigma^2\{\frac{1}{n} + \sum_{i=1}^{p} \frac{\lambda_i(k+n-1)^2}{(\lambda_i+k)^2(n-1)^2}\}$

Monte Carlo : $\text{TMSE}^{\text{New}}(k) = \frac{1}{100,000}\sum_{r=1}^{100,000}\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)^{(r)} - \boldsymbol{\beta}\}^\top\{\hat{\boldsymbol{\beta}}^{\text{New}}(k)^{(r)} - \boldsymbol{\beta}\}$

slightly less than 1 for all individuals. Since the requirement of the Scheffe-type model is not met ($\sum_{j=1}^{4} x_{ij} = 1$ does not hold for all individuals), one cannot apply the Scheffe-type model (Section 2.3). For this reason, we work on the intercept model (1).

We load the Portland cement data from R AICcmodavg package (Mazerolle, 2014). The sample correlation between the columns of $\mathbf{X}_p$ is as follows:

$$\text{Sample Corr}(\mathbf{X}_p) = \begin{array}{c} \begin{matrix}\mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4\end{matrix} \\ \begin{bmatrix} 1.000 & 0.229 & -0.824 & -0.245 \\ 0.229 & 1.000 & -0.139 & -0.973 \\ -0.824 & -0.139 & 1.000 & 0.030 \\ -0.245 & -0.973 & 0.030 & 1.000 \end{bmatrix} \end{array}$$

where the columns $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ indicate the proportions of $3\text{CaO} \cdot \text{Al}_2\text{O}_3$, $3\text{CaO} \cdot \text{SiO}_2$, $4\text{CaO} \cdot \text{Al}_2\text{O}_3 \cdot \text{Fe}_2\text{O}_3$, and $2\text{CaO} \cdot \text{SiO}_2$, respectively. We find that $\mathbf{x}_1$and $\mathbf{x}_3$ are negatively correlated. Also, $\mathbf{x}_2$and $\mathbf{x}_4$ are negatively correlated. Thus, we find necessity to adjust for multicollinearity. Hereafter, we divide our analysis into two cases according to whether $\mathbf{X}$ is standardized or not.

**Table 6.** Simulation results for examining the performance of $\hat{k}^{\mathrm{New}}$, $\hat{k}^{\mathrm{Ridge}}$, $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}})$, and $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\hat{k}^{\mathrm{Ridge}})$ based on 100,000 repetitions.

| | $\boldsymbol{\beta}=\begin{bmatrix}50\\1\\1\\1\\1\end{bmatrix}, \sigma^2=1$ | $\boldsymbol{\beta}=\begin{bmatrix}50\\1\\1\\1\\1\end{bmatrix}, \sigma^2=2$ | $\boldsymbol{\beta}=\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \sigma^2=1$ | $\boldsymbol{\beta}=\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \sigma^2=2$ |
|---|---|---|---|---|
| True $k^{\mathrm{New}}$ | 0.6343 | 1.3171 | 0.6343 | 1.3171 |
| True $k^{\mathrm{Ridge}}$ | 0.1163 | 0.1850 | 0.5676 | 1.1234 |
| $E(\hat{k}^{\mathrm{New}})$ | 0.9131 | 1.4647 | 0.9131 | 1.4647 |
| $E(\hat{k}^{\mathrm{Ridge}})$ | 0.1025 | 0.1548 | 0.7116 | 1.2851 |
| $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}})\}$ | 1.9454 | 3.3734 | 1.9454 | 3.3734 |
| $\mathrm{TMSE}\{\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(\hat{k}^{\mathrm{Ridge}})\}$ | 2.1858 | 4.0983 | 1.9401 | 3.3460 |

**Case (A)**: $\mathbf{X}$ is not standardized.

This is the standard way to perform regression on mixture experiments. Previously, the same data are analyzed in this way by a number of authors, including Sakallıoğlu and Kaçıranlar (2008) and Li and Yang (2012).

Under the un-standardized $\mathbf{X}$, we obtain

$$\hat{\boldsymbol{\beta}}^{\mathrm{OLS}} = (62.41 \quad 1.55 \quad 0.510 \quad 0.102 \quad -0.144)^{\mathrm{T}}, \quad \hat{\sigma}^2 = 5.983$$

In real data analysis, we do not know the true parameters. Sakallıoğlu and Kaçıranlar (2008) regard $\boldsymbol{\beta} = (62.41 \quad 1.55 \quad 0.510 \quad 0.102 \quad -0.144)^{\mathrm{T}}$ and $\sigma^2 = 5.983$ as the true values when calculating the TMSE of the several ridge-type estimators. Following their approach, we compare

$$\hat{\boldsymbol{\beta}} = C^{\mathrm{OLS}}\mathbf{y}, \quad \hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k) = C^{\mathrm{Ridge}}(k)\mathbf{y}, \quad \hat{\boldsymbol{\beta}}^{\mathrm{Liu}}(d) = C^{\mathrm{Liu}}(d)\mathbf{y}, \quad \hat{\boldsymbol{\beta}}^{\mathrm{Liu}}_{k,d} = C^{\mathrm{Liu}}_{k,d}\mathbf{y},$$
$$\hat{\boldsymbol{\beta}}^{\mathrm{SK}}(k,d) = C^{\mathrm{SK}}(k,d)\mathbf{y}, \quad \hat{\boldsymbol{\beta}}^{\mathrm{New}} = C^{\mathrm{New}}(k)\mathbf{y},$$

where

$$\begin{aligned}
C^{\mathrm{OLS}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}} \\
C^{\mathrm{Ridge}}(k) &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}_{(p+1)})^{-1}\mathbf{X}^{\mathrm{T}} \\
C^{\mathrm{Liu}}(d) &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{I}_{(p+1)})^{-1}\{\mathbf{I}_{(p+1)} + d(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\}\mathbf{X}^{\mathrm{T}} \\
C^{\mathrm{Liu}}_{k,d} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}_{(p+1)})^{-1}\{\mathbf{X}^{\mathrm{T}} - dC^{\mathrm{Ridge}}(k)\} \\
C^{\mathrm{SK}}(k,d) &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{I}_{(p+1)})^{-1}\{\mathbf{X}^{\mathrm{T}} + dC^{\mathrm{Ridge}}(k)\} \\
C^{\mathrm{New}}(k) &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + k\mathbf{I}_{(p+1)})^{-1}[\mathbf{I}_{(p+1)} + k\{\mathrm{diag}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}^{-1}]\mathbf{X}^{\mathrm{T}}
\end{aligned}$$

Table 7 summarizes the results. All the ridge-type estimators exhibit some improvement in the TMSE over the OLS estimator. This is the effect of correcting multicollinearity. In particular, the proposed estimator $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(\hat{k}^{\mathrm{New}})$ achieves the smallest TMSE among all the estimators. Except for the new estimator, all other estimators show nearly identical TMSE values as those reported by Sakallıoğlu and Kaçıranlar (2008). Figure 9 shows that the TMSE plot of $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ is uniformly smaller than that of $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$.

where $\lambda_1 \geq \ldots \geq \lambda_{p+1}$ are the eigenvalues of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

The black point is $\hat{k}^{\mathrm{Ridge}} = 0.001521$, and the red point is $\hat{k}^{\mathrm{NEW}} = 0.005192$.

**Table 7.** Analysis of the Portland cement data (Woods et al., 1932) based on the un-standardized design matrix.

| $\hat{\boldsymbol{\beta}}$ | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | Bias | Var | TMSE |
|---|---|---|---|---|---|---|---|---|
| $\hat{\boldsymbol{\beta}}^{\text{OLS}}$ | 62.41 | 1.55 | 0.51 | 0.10 | − 0.14 | 0 | 4912.09 | 4912.09 |
| $\hat{\boldsymbol{\beta}}^{\text{Ridge}}(k), k = \hat{k}_{\text{HK}}$ | 27.63 | 1.91 | 0.87 | 0.47 | 0.21 | 1209.55 | 961.42 | 2170.55 |
| $\hat{\boldsymbol{\beta}}^{\text{Ridge}}(k), k = \hat{k}^{\text{Ridge}}$ | 27.78 | 1.91 | 0.87 | 0.47 | 0.21 | 1199.62 | 971.40 | 2170.62 |
| $\hat{\boldsymbol{\beta}}^{\text{Liu}}(d), d = \hat{d}_{\text{opt}}$ | 62.25 | 1.55 | 0.51 | 0.10 | − 0.14 | 0.02 | 4887.28 | 4887.30 |
| $\hat{\boldsymbol{\beta}}^{\text{SK}}(k,d), k = \hat{k}_{\text{HK}}, d = \hat{d}_{\text{opt}}$ | 27.61 | 1.91 | 0.87 | 0.47 | 0.21 | 1211.46 | 959.50 | 2170.96 |
| $\hat{\boldsymbol{\beta}}^{\text{New}}(k), k = \hat{k}_{\text{HK}}$ | 80.71 | 1.36 | 0.32 | − 0.09 | − 0.33 | 335.20 | 961.78 | 1296.20 |
| $\hat{\boldsymbol{\beta}}^{\text{New}}(k), k = \hat{k}^{\text{New}}$ | 88.99 | 1.28 | 0.24 | − 0.18 | − 0.41 | 707.30 | 177.86 | 884.30 |

NOTES: $\hat{k}_{\text{HK}} = \hat{\sigma}^2/\{(\hat{\boldsymbol{\beta}}^{\text{OLS}})^{\text{T}}(\hat{\boldsymbol{\beta}}^{\text{OLS}})\} = 0.001535$, $\hat{k}^{\text{Ridge}} = 0.001521$, $\hat{k}^{\text{NEW}} = 0.005192$, $\hat{\sigma}^2 = 5.983$, and $\hat{d}_{\text{opt}} = \sum_{i=1}^{p+1} \frac{\lambda_i(\hat{\alpha}_i^2 - \hat{\sigma}^2)}{(\lambda_i+1)^2(\lambda_i+\hat{k}_{\text{HK}})} / \sum_{i=1}^{p+1} \frac{\lambda_i(\lambda_i\hat{\alpha}_i^2+\hat{\sigma}^2)}{(\lambda_i+1)^2(\lambda_i+\hat{k}_{\text{HK}})^2} = 0.997$,

**Case (B)**: $\mathbf{X}$ is standardized as in Equation (2).

This may not be the standard way to perform regression on mixture experiments. However, this is the usual way to apply ridge regression in other applications. In addition, since we have developed theoretical results under the standardized $\mathbf{X}$, it is of our interest to examine this setting.

Under the standardized $\mathbf{X}$, we obtain

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = (95.43 \quad 9.12 \quad 7.94 \quad 0.65 \quad -2.41)^{\text{T}}, \quad \hat{\sigma}^2 = 5.983$$

In real data analysis, we do not know the true parameters. Similar to Case (A), we regard $\boldsymbol{\beta} = (95.43 \quad 9.12 \quad 7.94 \quad 0.65 \quad -2.41)^{\text{T}}$ and $\sigma^2 = 5.983$ as the true values when we calculate the TMSE of $\hat{\boldsymbol{\beta}}^{\text{Ridge}}(k)$ and $\hat{\boldsymbol{\beta}}^{\text{New}}(k)$.

The results of evaluating the TMSE are displayed in Figure 10. In a range of $k > 0$, both $\hat{\boldsymbol{\beta}}^{\text{Ridge}}(k)$ and $\hat{\boldsymbol{\beta}}^{\text{New}}(k)$ perform much better than $\hat{\boldsymbol{\beta}}^{\text{OLS}}$. Among them, $\hat{\boldsymbol{\beta}}^{\text{New}}(k)$ achieves the smallest TMSE with the optimal value $\hat{k}^{\text{New}} = 0.1826$.

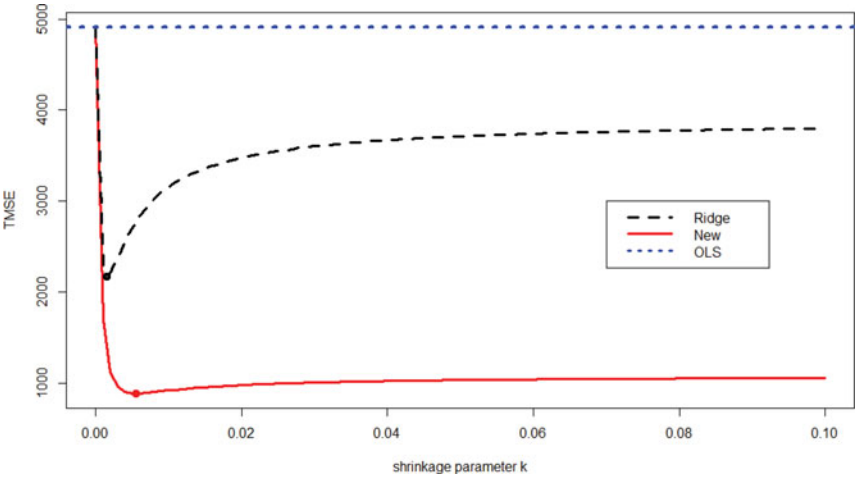


**Figure 9.** The TMSE plots of $\hat{\boldsymbol{\beta}}^{\text{OLS}}$, $\hat{\boldsymbol{\beta}}^{\text{Ridge}}(k)$ and $\hat{\boldsymbol{\beta}}^{\text{New}}(k)$ based on the Portland cement data under the un-standardized design matrix.

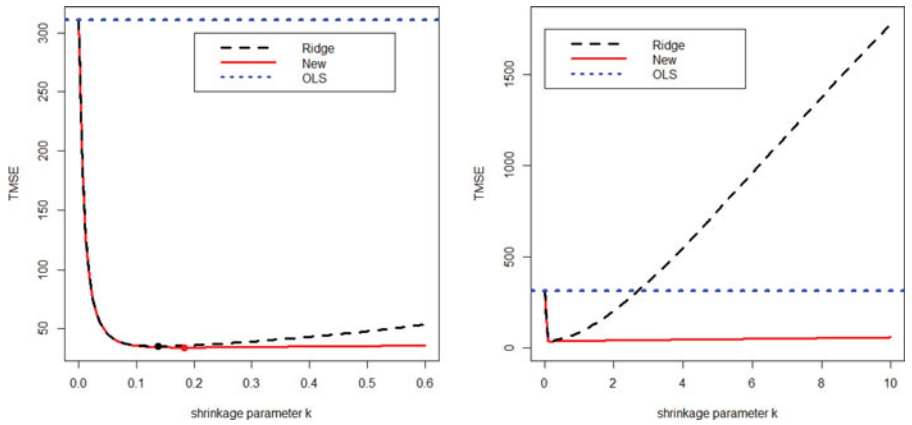


**Figure 10.** The TMSE plots for $\hat{\boldsymbol{\beta}}^{\mathrm{OLS}}$, $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}(k)$ and $\hat{\boldsymbol{\beta}}^{\mathrm{New}}(k)$ based on the Portland cement data under the standardized design matrix. The black point is $\hat{k}^{\mathrm{Ridge}} = 0.1373$ (left), and the red point is $\hat{k}^{\mathrm{New}} = 0.1826$ (left). The range of $k$ is different between the right and left figures.

## 7. Conclusion and discussion

In the Scheffe-type model for mixture experiments, the model does not include an intercept term under the key assumption that the observed proportion of a mixture is the unity (Section 2.3). We have pointed out that this assumption is occasionally not hold, in particular, for the Portland cement data. With this problem, a direct application of ridge regression to the Scheffe-type model can be misleading.

As an appropriate way to modify the ridge estimator, we have proposed a modification to the Liu-type estimator (Liu, 2003) of an intercept and regression coefficients under multicollinearity. An important property of the proposed estimator is that the TMSE is invariant for the size of the intercept, which is verified by both theoretically (Section 4) and numerically (Section 5). This invariance is exactly the consequence of using the compound covariate estimator to modify the Liu-type estimator. As a result, the performance of the proposed estimator is superior to the OLS estimator and the ridge estimator when the intercept is far from zero. If the intercept is near zero, then the proposed estimator is only slightly inferior to the ordinary ridge estimator.

While the present paper focuses on mixture experiments, the proposed idea can be applied to other experimental designs, where there exist some constraints on the design region. For instance, the uniform design over general input domains in computer experiments is considered by Chuang and Hung (2010). This design is applied to study a system which has several queues in parallel. Another interesting instance is the higher-order designs, such as the second-order mixture design.

## Acknowledgments

We thank the reviewer and the associate editor for their helpful comments that improve the manuscript. We are also thankful to Professor Ying-Chao Hung and Professor Hsiang-Ling Hsu for their comments on an earlier version of our paper.

## Funding

This work is supported by the research grant funded by the Ministry of Science and Technology of Taiwan (MOST 103-2118-M-008-MY2).

## References


Ashish, S., Srivastava, M. (1990). *Regression Analysis. Theory, Methods and Applications*. New York: Springer.

Brown, P.J. (1977). Centering and scaling in ridge regression. *Technometrics* 19:35–36.

Chuang, S.C., Hung, Y.C. (2010). Uniform design over general input domains with applications to target region estimation in computer experiments. *Comput. Stat. Data Anal.* 54(1):219–232.

Cornell, J.A. (2011). *A Primer on Experiments with Mixture*. Hoboken, N. J.: Wiley.

Draper, N.R., Smith, H. (1998). *Applied Regression Analysis*, 3rd ed. New York: John Wiley & Sons.

Emura, T., Chen, Y.H., Chen, H.Y. (2012). Survival prediction based on compound covariate under cox proportional hazard models. *PLoS One* 7. doi:10.1371/journal.pone.0047627

Emura, T., Chen, Y.-H. (2016). Gene selection for survival data under dependent censoring, a copula-based approach. *Stat. Methods Med. Res.* 25(6):2840–2857.

Emura, T., Nakatochi, M., Matsui, S., Michimae, H., Rondeau, V. (2017). Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model. *Stat. Methods in Med. Res*. doi:10.1177/0962280216688032.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer-Verlag.

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

Jang, D.-H., Anderson-Cook, C.M. (2010). Fraction of design space plots for evaluating ridge estimators in mixture experiments. *Qual. Reliab. Eng. Int.* 27:27–34.

Jang, D.-H., Anderson-Cook, C.M. (2014). Visualization approaches for evaluating ridge regression estimators in mixture and mixture-process experiments. *Qual. Reliab. Eng. Int.* doi:10.1002/qre.1683

Jimichi, M., Inagaki, N. (1993). Centering and scaling in ridge regression. *Stat. Sci. Data Anal.* 3:77–86.

Jimichi, M. (2005). Improvement of regression estimators by shrinkage under multicollinearity and its feasibility. Ph.D. Thesis. Osaka University, Japan.

Jimichi, M. (2008). Exact moments of feasible generalized ridge regression estimator and numerical evaluations. *J. Jpn. Soc. Comp. Stat.* 21:1–20.

Li, Y., Yang, H. (2012). A new Liu-type estimator in linear regression model. *Stat. Pap.* 53:427–437.

Liu, K. (1993). A new class of biased estimate in linear regression. *Commun. Stat. Theory Methods* 22:393–402.

Liu, K. (2003). Using Liu-type estimator to combat collinearity. *Commun. Stat. Theory Methods* 32:1009–1020.

Matsui, S. (2006). Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. *BMC Bioinf.* 7:156.

Mazerolle, M.J. (2014). AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). R package version 2.0–1, *CRAN*.

Montgomery, D.C., Peck, E.A., Vining, G.G. (2012). *Introduction to Linear Regression Analysis*. New Jersey: Wiley.

Radamacher, M.D., Mcshane, L.M., Simon, R.A. (2002). Paradigm for class prediction using gene expression profiles. *J. Comput. Biol.* 9:505–511.

Sakallıoğlu, S., Kaçıranlar, S. (2008). A new biased estimator on ridge estimation. *Stat. Pap.* 49:669–689.

Tukey, J.W. (1993). Tightening the clinical trial. *Controlled Clin. Trials* 14:266–285.

Theobald, C.M. (1973). Generalizations of mean square error applied to ridge regression. *J. Royal Stat. Soc. Ser. B (Methodological)* 36:103–106.

Wong, K.Y., Chiu, S.N. (2015). An iterative approach to minimize the mean squared error in ridge regression. *Comput. Stat.* 30(2):625–639.

Woods, H., Steinour, H.H., Starke, H.R. (1932). Effect of composition of Portland cement on heat evolved during hardening. *Ind. Eng. Chem.* 24:1207–1214.