Annals of the Institute of Statistical Mathematics (2023) 75:233–250
https://doi.org/10.1007/s10463-022-00841-7

Group least squares regression for linear models with strongly correlated predictor variables

Min Tsao[1]

Received: 3 September 2021 / Revised: 15 February 2022 / Accepted: 1 June 2022 /
Published online: 26 July 2022

Abstract
Traditionally, the main focus of the least squares regression is to study the effects of individual predictor variables, but strongly correlated variables generate multicollinearity which makes it difficult to study their effects. To resolve the multicollinearity issue without abandoning the least squares regression, for situations where predictor variables are in groups with strong within-group correlations but weak between-group correlations, we propose to study the effects of the groups with a group approach to the least squares regression. Using an all positive correlations arrangement of the strongly correlated variables, we first characterize group effects that are meaningful and can be accurately estimated. We then discuss the group approach to the least squares regression through a simulation study and demonstrate that it is an effective method for handling multicollinearity. We also address a common misconception about prediction accuracy of the least squares estimated model.

Keywords Strongly correlated predictor variables · Multicollinearity · Group effects · Linear models · Least squares regression

1 Introduction

Multicollinearity due to strongly correlated predictor variables is a long-standing problem without a satisfactory solution. It arises frequently in observational studies in social sciences and medical research. In this paper, we show that multicollinearity per se is not a problem; the problem is that what we have been trying to do with the strongly correlated variables are misguided and unattainable. We also present a

✉ Min Tsao
mtsao@uvic.ca

[1] Department of Mathematics and Statistics, University of Victoria, Victoria, BC V8W 2Y2, Canada

solution based on appropriate use of such variables. To introduce the problem, consider multiple regression model

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

where $\mathbf{y}$ is an n-vector of response variable values, $\mathbf{X} = [\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_p]$ is a known $n \times (p+1)$ design matrix with $p \geq 2$ and $\mathbf{1}_n$ being the n-vector of 1's, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the unknown vector of regression parameters, and $\boldsymbol{\varepsilon}$ is an n-vector of i.i.d. normal random errors with mean 0 and variance σ^2. Throughout this paper, we work under the low-dimensional setting where $n > p$ and $rank(\mathbf{X}) = p + 1$ so that the least squares estimator for $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2}$$

is available. We assume that the p predictor variables can be partitioned into k groups $\{\mathbf{X}_i\}_{i=1}^k$ such that (*i*) there is at least one group with 2 or more variables, (*ii*) variables in the same group are strongly correlated, and (*iii*) variables from different groups are weakly correlated. Let $\boldsymbol{\beta}_i$ be the parameter vector for variables in group $\mathbf{X}_i$. Model (1) may be written as

$$\mathbf{y} = \beta_0\mathbf{1}_n + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \cdots + \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}.$$

Here, $\boldsymbol{\beta}_i$ reduces to a scalar if there is only 1 variable in group $\mathbf{X}_i$. Let $\hat{\boldsymbol{\beta}}_i$ be the least squares estimator for $\boldsymbol{\beta}_i$. When there are 2 or more variables in $\mathbf{X}_i$, their strong correlations generate multicollinearity which makes variances of elements of $\hat{\boldsymbol{\beta}}_i$ large, rendering $\hat{\boldsymbol{\beta}}_i$ a poor estimator for $\boldsymbol{\beta}_i$.

There is a large body of literature on detecting and handling the multicollinearity problem; see, for example, Draper & Smith (1998), Belsley et al. (2004), Montgomery et al. (2012). Here, we only briefly discuss the main methods for handling the problem. The most well-known methods are the ridge regression (Hoerl & Kennard, 1970) and principal component regression (Jolliffe, 1986). There are also other methods such as latent root regression (Webster et al., 1974) and model respecification by eliminating some predictor variables. There have been a number of studies that evaluate these methods including Hoerl et al. (1975), Gunst et al. (1976), Gunst & Mason (1977) and Lawless (1978). One of the main criteria used for evaluation is the mean squared error of an estimator $\tilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, $E[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})]$. Estimators given by these methods are biased, but they are capable of achieving smaller mean squared errors than the least squares estimator $\hat{\boldsymbol{\beta}}$. However, except for this advantage, these estimators are difficult to use because their sampling properties are in general not available as they depend on the data in complicated ways. The ridge regression estimator, for example, involves a penalty parameter whose value is usually determined by cross-validation. The distribution of the penalty parameter and thus that of the estimator is unavailable. It is also difficult to choose among these methods as extensive comparisons have found no single best overall method; see Montgomery et al. (2012) for more discussion. Further, some authors such as (Conniffe & Stone 1973) are critical of biased estimation methods. Draper & Van Nostrand (1979)

identified two cases where ridge regression may be appropriate but also recommended against the use of biased estimation methods in general. Nevertheless, these methods are still the most used tools for handling multicollinearity.

Is multicollinearity really such an insurmountable problem for the least squares regression that we have to abandon this simple method of regression in favour of complicated alternatives? Traditionally, the focus of regression analyses has been on the impact of individual predictor variables. For example, in estimation, the focus has been on estimating parameters of individual variables; in variable selection, it has been on inclusion or exclusion of individual variables. With this focus on individual variables, multicollinearity has been a problem for the least squares regression as it cannot accurately estimate parameters of the strongly correlated variables which in turn leads to difficulties in variable selection and prediction. Nevertheless, we argue that neither multicollinearity nor the least squares regression is responsible for these problems; the wrong focus on the impact of individual variables is the real culprit. In Remark [b] of Sect. 2.3, we note that estimating the parameter of a variable in a strongly correlated group is a form of extreme extrapolation. That it cannot be done accurately is solely the consequence of extrapolating far beyond the data range. Strongly correlated variables appear naturally in groups. Individual parameters of these variables are not meaningful. Instead of focusing on their individual impact, we should respect their group nature by handling them in groups and focusing on their collective impact on the response variable. To this end, we propose a group approach to the least squares regression which still relies on $\hat{\boldsymbol{\beta}}$ but differs from the traditional least squares regression in three aspects: (*i*) for a group $\mathbf{X}_i$ with 2 or more variables, the group approach will not attempt to estimate or make inference about individual elements of $\boldsymbol{\beta}_i$; instead, it will focus on estimation and inference for those linear combinations of the elements of $\boldsymbol{\beta}_i$ that represent meaningful group effects of $\mathbf{X}_i$; (*ii*) it will perform variable selection at the group level in that variables in a group $\mathbf{X}_i$ are either all in or all out; and (*iii*) it will analyse prediction accuracy of the least squares estimated model through group effects. For a group $\mathbf{X}_i$ with only 1 variable, its group effect is the parameter of the variable, so the group approach will still estimate and make inference of the parameter just like in the traditional least square regression.

Comparing to existing methods for handling multicollinearity, the group approach to the least squares regression has the advantage that it is very simple in computation and its theories for estimation, inference and prediction are already in place as it is still least squares regression with only a change of focus from individual to group effects for strongly correlated variables. In contrast, computation for the ridge regression and principal component regression is more complicated and theories for these methods are convoluted and even intractable. Additional advantages of the group approach include (*i*) it retains the simple least squares estimators $\hat{\beta}_i$; those for variables not strongly correlated with others are good unbiased point estimators of their parameters we can still use; those for strongly correlated variables are only used for estimation and inference of group effects of such variables and making predictions, but they are not used as point estimators as parameters of such variables are not estimated under the group approach; (*ii*) the regression mean

squared error remains a good unbiased estimator for the error variance σ^2; and (*iii*) existing (non-group based) methods of inference, variable selection and model diagnosis for the least squares regression may be adopted with a minor adjustment of handling strongly correlated variables in groups. The ridge regression and principal component regression have none of these advantages.

There is a widely held view that when there is multicollinearity in the data, alternative regression methods in general and the ridge regression in particular give more accurate predictions than the least squares regression. Although there is no proof to support this view, it has appeared in many papers, books and internet sites. Through a group effect-based analysis on the prediction accuracy of the least squares estimated model and a comparison with the ridge regression, we show that this is a misconception arising from comparing prediction accuracy at points where predictions are not meaningful and should not be made. At points where predictions are meaningful, the least squares regression is actually more accurate than the ridge regression.

In Sect. 2, we discuss group effects of strongly correlated variables. We characterize effects that can be accurately estimated and argue that such effects are meaningful but individual parameters of these variables are not meaningful. In Sect. 3, we discuss estimation, variable selection and prediction under the group approach through results of a simulation study and apply this approach to analyse the Hald cement data. The full simulation study is in the Supplementary Material for this paper which also contains extra material for Sects. 2.3 and 3.2. We conclude with a few remarks in Sect. 4.

2 Group effects of strongly correlated predictor variables

Group effects lie at the heart of the group approach to the least squares regression. Tsao (2019) studied estimation of group effects in a theoretical model containing strongly correlated predictor variables with a restrictive uniform correlation structure. We now revisit the estimation problem without imposing any parametric correlation structure on the strongly correlated variables and generalize results in Tsao (2019) to all linear models. For this section, we let $\mathbf{X}_1 = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q]$ and $\mathbf{X}_2 = [\mathbf{x}_{q+1}, \mathbf{x}_{q+2}, \ldots, \mathbf{x}_p]$, and write (1) as

$$\mathbf{y} = \beta_0 \mathbf{1}_n + \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{3}$$

where $2 \le q \le p$, $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_q)^T$, $\boldsymbol{\beta}_2 = (\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p)^T$, and $\mathbf{X}_1$ is a group of strongly correlated variables satisfying (*i*) for $1 \le i,j \le q$, absolute values of $r_{ij} = corr(\mathbf{x}_i, \mathbf{x}_j)$ are all above $\frac{\sqrt{2}}{2}$ (≈ 0.71) and (*ii*) variables in $\mathbf{X}_1$ are not strongly correlated with variables in $\mathbf{X}_2$. Condition (*i*) is needed to ensure that variables in $\mathbf{X}_1$ will all have positive correlations after appropriate sign changes; see Eq. (6). For this section, $\mathbf{X}_2$ holds all variables not in $\mathbf{X}_1$. There may be more strongly correlated groups among variables in $\mathbf{X}_2$, but it suffices to study the group effects of just $\mathbf{X}_1$ as results obtained apply to all such groups. Consider the class of linear combinations of $\beta_1, \beta_2, \ldots, \beta_q$,

$$\Xi = \{\xi(\mathbf{w}) | \xi(\mathbf{w}) = w_1\beta_1 + w_2\beta_2 + \cdots + w_q\beta_q\}, \tag{4}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_q)^T$ is any q-vector satisfying $\sum_{i=1}^{q} |w_i| = 1$. Set Ξ is the class of normalized group effects of variables in $\mathbf{X}_1$. Each $\xi(\mathbf{w})$ in Ξ is a (normalized) group effect defined by its weight vector $\mathbf{w}$. It has an interpretation as the expected change in the response variable y when the q predictor variables in $\mathbf{X}_1$ change by the amount $\mathbf{w}$; that is, $x_1, x_2, \ldots, x_q$ change by the amount $w_1, w_2, \ldots, w_q$, respectively, at the same time. In this sense, $\xi(\mathbf{w})$ represents a collective impact or a group effect of $\mathbf{X}_1$ on y.

Throughout this paper, we say a group effect can be accurately estimated if the variance of its minimum-variance unbiased linear estimator is smaller than or comparable to the error variance σ^2. Not all group effects can be accurately estimated and some group effects are not meaningful. For example, β_1 is a special group effect with $w_1 = 1$ and $w_j = 0$ for $j \neq 1$, but it cannot be accurately estimated. It is also not a meaningful effect (see Remark [b]). We now characterize effects that can be accurately estimated. To this end, we first introduce an all positive correlations arrangement of the strongly correlated variables and then study the limiting properties of their correlation matrix.

2.1 All positive correlations arrangement of strongly correlated variables and limiting properties of their correlation matrix

Let $\mathbf{R}$ be the full rank correlation matrix of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$,

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1q} \\ r_{21} & 1 & \cdots & r_{2q} \\ \cdot & \cdot & \cdots & \cdot \\ r_{q1} & r_{q2} & \cdots & 1 \end{bmatrix}_{q\times q}. \tag{5}$$

Some of the r_{ij} may be negative but since all $|r_{ij}|$ are above $\frac{\sqrt{2}}{2}$, let $\text{sgn}(r_{1j})$ be the sign of $r_{1j} = corr(\mathbf{x}_1, \mathbf{x}_j)$ for $j = 2, 3, \ldots, q$, by Theorem 3.1 in Tsao (2019) the following signed version of the set of q variables

$$\mathbf{x}_1, \text{sgn}(r_{12})\mathbf{x}_2, \ldots, \text{sgn}(r_{1q})\mathbf{x}_q \tag{6}$$

satisfies that all pairwise correlations are positive. We call (6) an all positive correlations (APC) arrangement of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q$. For the rest of this section, we assume that these q variables are already in an APC arrangement so that all r_{ij} in (5) are positive. If they are not in an APC arrangement, we can replace them with their APC version (6); see Sect. 3.2 for an example.

The importance of using the APC arrangement is twofold. Firstly, it makes it easy to identify important and meaningful effects; see Remarks [a] and [b]. Secondly, it makes it easy to measure the level of multicollinearity generated by the q variables and to formulate the question of interest. To see the second point, let $r_M = \min\{r_{ij}\}$. Under the APC arrangement, all r_{ij} satisfy $0 < r_M \leq r_{ij} < 1$, so when r_M goes to 1,

all r_{ij} go to 1 which makes the multicollinearity stronger. In this sense, an increase in r_M represents an increase in the level of multicollinearity, so we will use r_M to measure this level. Our question of interest can now be formulated as that of identifying group effects in (4) that can be accurately estimated when r_M is close to 1.

To answer the above question, we first study the limiting properties of $\mathbf{R}$ and $\mathbf{R}^{-1}$ when r_M goes 1. Since $\mathbf{R}$ is a correlation matrix, it is positive definite, so it has q positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > 0$. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q$ be their corresponding orthonormal eigenvectors, respectively. Let $\mathbf{1}_q$ be the q-vector whose elements are all 1's. We have the following results.

Lemma 1 *Correlation matrix* $\mathbf{R}$ *satisfies*

(i) $\lambda_1 \to q$ *and* $\lambda_i \to 0$ *for* $i = 2, 3, \ldots, q$ *as* $r_M \to 1$; *and*
(ii) $\mathbf{v}_1 \to \frac{1}{\sqrt{q}}\mathbf{1}_q$ *as* $r_M \to 1$.

Lemma 2 *The inverse matrix* $\mathbf{R}^{-1}$ *satisfies*

(i) $\mathbf{v}_1^T \mathbf{R}^{-1} \mathbf{v}_1 > \frac{1}{q}$; *and*
(ii) $\mathbf{v}_1^T \mathbf{R}^{-1} \mathbf{v}_1 \to \frac{1}{q}$ *as* $r_M \to 1$.

The proofs of these lemmas are in the Appendix.

2.2 The eigen-effect of strongly correlated predictor variables

In this section, we identify one group effect for the standardized version of (3) that can be very accurately estimated at high levels of multicollinearity. It will be used to identify other effects that can be accurately estimated.

Let $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ni})^T$, $\bar{x}_i = \frac{1}{n}\sum_{j=1}^n x_{ji}$ and $s_i^2 = \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2$ which is $(n-1)$ times the sample variance of $\mathbf{x}_i$. We call

$$\mathbf{x}_i' = \frac{\mathbf{x}_i - \bar{x}_i \mathbf{1}_n}{s_i} \tag{7}$$

the standardized variable which has mean zero and length one. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, $\bar{y} = \frac{1}{n}\sum_{j=1}^n y_j$ and $\mathbf{y}' = \mathbf{y} - \bar{y}$. We can write (3) as

$$\mathbf{y}' = \mathbf{X}_1' \boldsymbol{\beta}_1' + \mathbf{X}_2' \boldsymbol{\beta}_2' + \boldsymbol{\varepsilon}, \tag{8}$$

where $\mathbf{X}_1' = [\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_q']$, $\mathbf{X}_2' = [\mathbf{x}_{q+1}', \mathbf{x}_{q+2}', \ldots, \mathbf{x}_p']$, $\boldsymbol{\beta}_1' = (\beta_1', \beta_2', \ldots, \beta_q')^T$, and $\boldsymbol{\beta}_2' = (\beta_{q+1}', \beta_{q+2}', \ldots, \beta_p')^T$. We call model (8) the standardized model. The relationship between parameters in models (8) and (3) is

$$\beta_0 = \bar{y} - \sum_{i=1}^{p} \bar{x}_i \beta_i'/s_i \text{ and } \beta_i = \beta_i'/s_i \text{ for } i = 1, 2, \ldots, p. \tag{9}$$

Let $\mathbf{X}' = [\mathbf{X}_1', \mathbf{X}_2']$. Then, $\mathbf{X}'^T\mathbf{X}' = [r_{ij}] \in \mathbb{R}^{p\times p}$ is the correlation matrix of the p predictor variables in models (8) or (3) where $r_{ij} = corr(\mathbf{x}_i', \mathbf{x}_j') = corr(\mathbf{x}_i, \mathbf{x}_j)$. Partition this correlation matrix as follows:

$$\mathbf{X}'^T\mathbf{X}' = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}_{p\times p}, \tag{10}$$

where $\mathbf{R}_{11} = \mathbf{R} \in \mathbb{R}^{q\times q}$ is the correlation matrix (5) of the q variables in $\mathbf{X}_1'$, and $\mathbf{R}_{12}$ is the between-group correlation matrix of $\mathbf{X}_1'$ and $\mathbf{X}_2'$. By (10),

$$[\mathbf{X}'^T\mathbf{X}']^{-1} = \begin{bmatrix} [\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}]^{-1} & \mathbf{R}_{11}^{-1}\mathbf{R}_{12}[\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} - \mathbf{R}_{22}]^{-1} \\ [\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} - \mathbf{R}_{22}]^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1} & [\mathbf{R}_{22} - \mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}]^{-1} \end{bmatrix}. \tag{11}$$

Let $\mathbf{R}^* = [\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}]$. Then, $\mathbf{R}^*$ is a symmetric positive definite matrix as $\mathbf{R}^{*-1}$ is a diagonal block of the positive definite matrix $[\mathbf{X}'^T\mathbf{X}']^{-1}$ in (11). Let λ_1^* be its largest eigenvalue and $\mathbf{v}_1^* = (v_{11}^*, v_{12}^*, \ldots, v_{1q}^*)^T$ be the corresponding orthonormal eigenvector. We call linear combination

$$\xi_E = \mathbf{v}_1^{*T}\boldsymbol{\beta}_1' = v_{11}^*\beta_1' + v_{12}^*\beta_2' + \cdots + v_{1q}^*\beta_q' \tag{12}$$

the eigen-effect. Since $\|\mathbf{v}_1^*\| = 1$, $1 \le \sum_{i=1}^{q} |v_{1i}^*| \le \sqrt{q}$ and so ξ_E may not be a normalized effect. Nevertheless, for technical convenience, we will first study ξ_E and will give a simple normalized representation of ξ_E later.

Let $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1', \hat{\beta}_2', \ldots, \hat{\beta}_p')^T$ be the least squares estimator for $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1'^T, \boldsymbol{\beta}_2'^T)^T$. The minimum-variance unbiased linear estimator for ξ_E is

$$\hat{\xi}_E = \mathbf{v}_1^{*T}\hat{\boldsymbol{\beta}}_1' = v_{11}^*\hat{\beta}_1' + v_{12}^*\hat{\beta}_2' + \cdots + v_{1q}^*\hat{\beta}_q'. \tag{13}$$

Since $\hat{\xi}_E$ is an unbiased estimator for ξ_E, it is accurate if $var(\hat{\xi}_E)$ is small. Although none of the β_i' in (12) is accurately estimated by $\hat{\beta}_i'$ in (13) when r_M is high, the following theorem shows ξ_E is accurately estimated by $\hat{\xi}_E$.

Theorem 1 *For the group of strongly correlated variables in* $\mathbf{X}_1'$ *in* (8),

(i) *if they are uncorrelated with variables in* $\mathbf{X}_2'$, *then* (i_1) $var(\hat{\xi}_E) > \sigma^2/q$ *and* (i_2) $var(\hat{\xi}_E) \to \sigma^2/q$ *as* $r_M \to 1$; *and*

(ii) *if they are correlated with variables in* $\mathbf{X}_2'$ *but the between-group correlation matrix* $\mathbf{R}_{12} \to \mathbf{0}$ *as* $r_M \to 1$, *then* $var(\hat{\xi}_E) \to \sigma^2/q$ *as* $r_M \to 1$.[1]

To interpret Theorem 1, when variables in $\mathbf{X}_1'$ are uncorrelated with those in $\mathbf{X}_2'$, result (i_1) gives a lower bound on $var(\hat{\xi}_E)$ and result (i_2) shows $var(\hat{\xi}_E)$ approaches this lower bound as r_M approaches its upper bound 1. Thus, ξ_E is more accurately estimated by $\hat{\xi}_E$ at higher levels of multicollinearity. Result (ii) gives the asymptotic behaviour of $var(\hat{\xi}_E)$ when r_M goes to 1 and correlations between variables in $\mathbf{X}_1'$ and $\mathbf{X}_2'$ go to zero ($\mathbf{R}_{12} \to \mathbf{0}$). It implies that when such correlations are weak and the level of multicollinearity is high, $var(\hat{\xi}_E)$ is approximately σ^2/q. The proof of Theorem 1 is in the Appendix.

Theorem 1 does not cover the case where some variables in $\mathbf{X}_1'$ are strongly correlated with some variables in $\mathbf{X}_2'$. We are not interested in this case as it weakens the notion of $\mathbf{X}_1'$ being a (stand-alone) group of strongly correlated variables which renders its group effects not meaningful. Turning now to other effects defined by unit vectors that may be accurately estimated when r_M is high, the following result shows where such effects may be found.

Theorem 2 *For* $\delta > 0$, *define a neighbourhood of* $\mathbf{v}_1^*$ *on the unit sphere*

$$\mathcal{N}_\delta = \{\mathbf{v} \in \mathbb{R}^q : \|\mathbf{v}\| = 1 \text{ and } \sqrt{1-\delta} < \mathbf{v} \cdot \mathbf{v}_1^* \le 1\}. \tag{14}$$

Suppose the between-group correlation matrix $\mathbf{R}_{12} \to \mathbf{0}$ *as* $r_M \to 1$. *If a unit vector* $\mathbf{v} \notin \mathcal{N}_\delta$, *then* $var(\mathbf{v}^T \hat{\boldsymbol{\beta}}_1') \to \infty$ *as* $r_M \to 1$.

2.3 Characterization of group effects that can be accurately estimated

Theorem 2 implies that all $\mathbf{v}^T \boldsymbol{\beta}_1'$ that can be accurately estimated at high r_M levels are given by $\mathbf{v} \in \mathcal{N}_\delta$. Let $s(\mathbf{v})$ be the sum of absolute values of elements of $\mathbf{v}$. Then, $1 \le s(\mathbf{v}) \le \sqrt{q}$ and $\mathbf{w} = \mathbf{v}/s(\mathbf{v})$ is a bijection that maps $\mathcal{N}_\delta$ into a small open neighbourhood of the normalized eigenvector $\mathbf{v}_1^*/s(\mathbf{v}_1^*)$ on the simplex $\sum_{i=1}^q w_i = 1$. Weight $\mathbf{w}$ of group effects that can be accurately estimated are in this open neighbourhood. In this sense, such effects are in a neighbourhood of the normalized eigen-effect $\xi_E^* = \xi_E/s(\mathbf{v}_1^*)$.

To identify a simpler effect to represent ξ_E^* and its neighbourhood, when variables in $\mathbf{X}_1'$ are uncorrelated with variables in $\mathbf{X}_2'$, $\mathbf{R}_{12} = \mathbf{0}$ and $\mathbf{R}^* = \mathbf{R}$, so $\lambda_1^* = \lambda_1$ and $\mathbf{v}_1^* = \mathbf{v}_1$. By Lemma 1, $\mathbf{v}_1 \to \frac{1}{\sqrt{q}}\mathbf{1}_q$ as $r_M \to 1$, which implies $s(\mathbf{v}_1) \to \sqrt{q}$ and $\mathbf{v}_1/s(\mathbf{v}_1) \to \frac{1}{q}\mathbf{1}_q$. When variables in $\mathbf{X}_1'$ and $\mathbf{X}_2'$ are correlated, $\mathbf{v}_1^* \to \frac{1}{\sqrt{q}}\mathbf{1}_q$ and thus

[1] $\mathbf{R}_{12} \to \mathbf{0}$ denotes element-wise convergence of $\mathbf{R}_{12}$ to zero. It implies $\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} \to \mathbf{0}$ under general conditions such as $\|\mathbf{R}_{22}^{-1}\|_{max}$ is bounded or $(\|\mathbf{R}_{12}\|_{max})^2(\|\mathbf{R}_{22}^{-1}\|_{max}) = o(1)$. This observation will be used in the proof of (ii) which requires $\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} \to \mathbf{0}$.

$\mathbf{v}_1^*/s(\mathbf{v}_1^*) \to \frac{1}{q}\mathbf{1}_q$ also hold under general conditions (see proof of Theorem 1(*ii*)). Thus, $\xi_E^* \to \xi_A$ as $r_M \to 1$ where

$$\xi_A = \frac{1}{q}\mathbf{1}_q^T\boldsymbol{\beta}_1' = \frac{1}{q}(\beta_1' + \beta_2' + \cdots + \beta_q'). \tag{15}$$

We call ξ_A the average group effect of the q strongly correlated variables in $\mathbf{X}_1'$. The minimum-variance unbiased linear estimator for ξ_A is

$$\hat{\xi}_A = \frac{1}{q}\mathbf{1}_q^T\hat{\boldsymbol{\beta}}_1' = \frac{1}{q}\left(\hat{\beta}_1' + \hat{\beta}_2' + \cdots + \hat{\beta}_q'\right). \tag{16}$$

When r_M is close to 1, $\hat{\xi}_A \approx \hat{\xi}_E^*$ and so $var(\hat{\xi}_A) \approx var(\hat{\xi}_E^*) = var(\hat{\xi}_E)/[s(\mathbf{v}_1^*)]^2$. Theorem 1 and $s(\mathbf{v}_1^*) \to \sqrt{q}$ then imply that $var(\hat{\xi}_A) \approx \sigma^2/q^2$. On the other hand, when all variables are uncorrelated, $var(\hat{\xi}_A) = \sigma^2/q$. This shows that the estimation of ξ_A benefits from a high level of multicollinearity in that it makes $var(\hat{\xi}_A)$ approximately q times smaller. Our subsequent discussions will be centred on ξ_A as it has simpler expression and interpretation than ξ_E^*.

For the unstandardised model (3) where $\beta_1, \beta_2, \cdots, \beta_q$ are parameters of the strongly correlated variables in $\mathbf{X}_1$, let $\mathbf{w}^* = (w_1^*, w_2^*, \ldots, w_q^*)^T$ where

$$w_i^* = \frac{s_i}{\sum_{j=1}^q s_j} \tag{17}$$

for $i = 1, 2, \ldots, q$. We call the following weighted average

$$\xi_W = w_1^*\beta_1 + w_2^*\beta_2 + \cdots + w_p^*\beta_q \tag{18}$$

the variability weighted average effect of the variables in $\mathbf{X}_1$ as w_i^* is proportional to the variability (measured by s_i) of $\mathbf{x}_i$. Using the least squares estimator in (2), the minimum-variance unbiased linear estimator for ξ_W is

$$\hat{\xi}_W = w_1^*\hat{\beta}_1 + w_2^*\hat{\beta}_2 + \cdots + w_p^*\hat{\beta}_q. \tag{19}$$

Noting that relationship (9) between the coefficients of the original and standardized models also applies to their respective least squares estimates, $\hat{\xi}_W$ can be expressed in terms of $\hat{\xi}_A$ as

$$\hat{\xi}_W = \frac{1}{\sum_{j=1}^q s_j}\sum_{i=1}^q s_i\hat{\beta}_i = \frac{1}{\sum_{j=1}^q s_j}\left(\sum_{i=1}^q \hat{\beta}_i'\right) = \frac{q}{\sum_{j=1}^q s_j}\hat{\xi}_A. \tag{20}$$

When r_M is close to 1, since $var(\hat{\xi}_A)$ is approximately σ^2/q^2, (20) implies

$$var(\hat{\xi}_W) = \left(\frac{q}{\sum_{j=1}^q s_j}\right)^2 var(\hat{\xi}_A) \approx \frac{\sigma^2}{\left(\sum_{i=1}^q s_i\right)^2}.$$

In practice, $(\sum_{i=1}^{q} s_i)^2$ is usually large, so $var(\hat{\xi}_W)$ is much smaller than σ^2. Using ξ_A and ξ_W as reference points, we now characterize the set of effects that are meaningful and can be accurately estimated in the following remarks. The Supplementary Material has an expanded version of these remarks.

Remark [a] For the q variables in APC arrangement in $\mathbf{X}_1'$ of the standardized model (8), effects $\xi'(\mathbf{w}) = w_1\beta_1' + w_2\beta_2' + \cdots + w_q\beta_q'$ that can be accurately estimated at a given high r_M level are in a small neighbourhood of ξ_A,

$$\mathcal{N}_A = \{\xi'(\mathbf{w}) : ||\mathbf{w} - \mathbf{w}_a|| < \delta_1\} \tag{21}$$

where δ_1 is a small positive constant that depends on r_M and $\mathbf{w}_a = \frac{1}{q}\mathbf{1}_q$ is the weight vector of ξ_A. Similarly, for the q variables in APC arrangement in $\mathbf{X}_1$ of the unstandardised model (3), group effects $\xi(\mathbf{w}) = w_1\beta_1 + w_2\beta_2 + \cdots + w_q\beta_q$ that can be accurately estimated are in a neighbourhood of ξ_W,

$$\mathcal{N}_W = \{\xi(\mathbf{w}) : ||\mathbf{w} - \mathbf{w}^*|| < \delta_2\}, \tag{22}$$

where δ_2 is a small positive constant. An alternative characterization of $\mathcal{N}_W$ is $\mathcal{N}_W = \{\xi(\mathbf{w}) : \xi(\mathbf{w}) \text{ such that the corresponding } \xi'(\mathbf{w}') \in \mathcal{N}_A\}$.

Remark [b] Set $\mathcal{N}_A$ in (21) is also the set of practically important and meaningful group effects for variables in $\mathbf{X}_1'$ in that $\mathbf{w}$ values in the neighbourhood of $\mathbf{w}_a$ represent the most probable changes of the variables in $\mathbf{X}_1'$. Two extreme examples illustrate this point. (*i*) Effect $\beta_1' \notin \mathcal{N}_A$ as its weight vector is $\mathbf{w}_1 = (1, 0, \ldots, 0)$. It represents the group impact on response when x_1' increases by 1 unit but the other variables do not change. (*ii*) Effect ξ_A has $\mathbf{w}_a = \frac{1}{q}\mathbf{1}_q$, so $\xi_A \in \mathcal{N}_A$. It represents the group impact when all variables increase by $(1/q)$th of a unit. With strong positive correlations and in standardized units, the variables are likely to increase at the same time and in similar amounts. So ξ_A is practically important and meaningful, whereas β_1' is not. In fact, estimating β_1' alone amounts to extreme extrapolation and β_1' by itself is neither meaningful nor interpretable as one cannot just increase x_1' by 1 unit while holding other variables constant under strong correlations among variables. Another example showing individual parameters are not meaningful is the extreme case of perfect correlation with $x_1' = \cdots = x_q' = x'$. Let $c = \beta_1' + \cdots + \beta_q'$. Then, the collective impact of these q variables on the response is cx'. There are infinitely many sets of β_i' that sum up to c. The data $(\mathbf{X}', \mathbf{y}')$ contain no information on which set is in the true model. In this sense, the data contain no information about the individual β_i'. Similarly, the data contain little information about the individual β_i' when the level of multicollinearity is high. The large variances of the least squares estimators for β_i' are warnings for this lack of information which is always a problem regardless the method of regression used. With this understanding, we should focus on estimating c, or equivalently $\xi_A = c/q$, and group effects in $\mathcal{N}_A$. For the strongly correlated variables in $\mathbf{X}_1$ in the unstandardised model, a group effect is meaningful if and only if the corresponding effect in the standardized model is meaningful. Thus, $\mathcal{N}_W$ is the set of meaningful group effects for these variables.

Table 1 Correlations of original Hald cement data (left) and renamed data (right)

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$		$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$
$\mathbf{x}_1$	1.00	0.22	−0.82	−0.24	$\mathbf{x}_1$	1.00	0.82	0.22	0.24
$\mathbf{x}_2$	0.22	1.00	−0.13	−0.97	$\mathbf{x}_2$	0.82	1.00	0.13	0.02
$\mathbf{x}_3$	−0.82	−0.13	1.00	0.02	$\mathbf{x}_3$	0.22	0.13	1.00	0.97
$\mathbf{x}_4$	−0.24	−0.97	0.02	1.00	$\mathbf{x}_4$	0.24	0.02	0.97	1.00

3 Group approach to the least squares regression

The performance and advantages of the group approach to the least squares regression are demonstrated in a small simulation study in the Supplementary Material. In this section, we summarize this simulation study. We also apply the group approach to analyse the Hald cement data and illustrate the APC arrangement as well as Theorem 1(*ii*) with this example.

3.1 Estimation, variable selection and prediction under the group approach

We examined three aspects of the group approach, estimation, variable selection and prediction, in the simulation study. Our results are as follows.

(a) *Estimation.* For strongly correctly variables, we demonstrated that group effects in the neighbourhood of the variability weighted average effect (22) are accurately estimated, confirming theoretical results in Sect. 2. Group effects not in this neighbourhood are poorly estimated, but these are not meaningful effects. Parameters of variables not strongly correlated with other variables and linear combinations of these parameters are accurately estimated, showing that the impact of multicollinearity due to strongly correlated variables is only limited to parameters and group effects of such variables. We also gave a t statistic for the group effect.

(b) *Variable selection.* Under the group approach, strongly correlated variables are all in or all out at the same time in the variable selection process. For all subsets regression, the number of models needed to be examined under the group approach is much smaller than that under the traditional non-group-based approach. Simulation results showed that the group approach is more accurate and more stable than the traditional approach.

(c) *Prediction.* Using the group approach, we obtained a more precise characterization of the region in the predictor variable space over which the least squares estimated model gives accurate predictions. We call this region the feasible prediction region and denote it by $\mathcal{R}_{FP}$. We argued that only predictions made over $\mathcal{R}_{FP}$ are meaningful. Further, simulation results showed that the least squares predictor is more accurate than the ridge regression predictor over $\mathcal{R}_{FP}$, demonstrating that the commonly held view that the ridge regression predictor is more accurate is a misconception.

Table 2 Estimated parameter values and average group effects for the standardized model (8); ξ_A^1 is the average group effect for group $\{x_1', x_2'\}$ and ξ_A^2 is that for $\{x_3', x_4'\}$

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
β_1'	31.607	14.308	2.209	0.055
β_2'	−2.261	15.788	−0.143	0.889
β_3'	27.500	36.784	0.748	0.473
β_4'	8.353	38.762	0.215	0.834
ξ_A^1	14.673	1.456	10.072	0.000
ξ_A^2	17.927	1.571	11.409	0.000

3.2 Application to Hald cement data

The Hald cement data have been widely used in the literature to illustrate multicollinearity; see, for example, Draper & Smith (1998). The data set contains 13 observations on 4 predictor variables and 1 response: heat evolved in calories per gram of cement (y), amount of tricalcium aluminate (x_1), amount of tricalcium silicate (x_2), amount of tetracalcium alumino ferrite (x_3), and amount of dicalcium silicate (x_4). The data set is given in the Supplementary Material which also contains a more detailed analysis of this data.

We first illustrate the APC arrangement of a group of strongly correlated variables with these data. In Table 1, the correlation matrix on the left is that of the 4 predictor variables. It shows that there are 2 strongly correlated groups $\{x_1, x_3\}$ and $\{x_2, x_4\}$ with negative correlation within each group, so $\{x_1, -x_3\}$ and $\{x_2, -x_4\}$ are their APC arrangements. For convenience, we rename the variables so that x_1 is still the same, but the old $-x_3$ is now called x_2, the old x_2 now called x_3, and the old $-x_4$ now called x_4. The correlation matrix of the renamed variables is on the right of Table 1. The strongly correlated groups are now $\{x_1, x_2\}$ and $\{x_3, x_4\}$, both in APC arrangement, and there are no strong correlations between variables from different groups.

For model (8) with the standardized renamed variables, the matrix $\mathbf{X}'^T\mathbf{X}'$ in (10) is just the correlation matrix on the right of Table 1. Matrix $\mathbf{R}_{11}$ in (10) is the upper-left quarter of this correlation matrix, $\mathbf{R}_{22}$ is the lower-right quarter, and $\mathbf{R}_{12}$ is the upper-right quarter. For $i \neq j$, the r_{ij} in $\mathbf{R}_{11}$ are close to 1, and the r_{ij} in $\mathbf{R}_{12} = \mathbf{R}_{21}^T$ are small. The latter leads to, as an example illustrating footnote 1 for Theorem 1(*ii*), small elements in

$$\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} = \begin{pmatrix} 0.06 & -0.01 \\ -0.01 & 0.22 \end{pmatrix}.$$

Thus, Theorem 1(*ii*) applies to group $\{x_1', x_2'\}$ in that $var(\hat{\xi}_E) \approx \sigma^2/2$ and consequently $var(\hat{\xi}_A) \approx \sigma^2/2^2$. Similarly, it also applies to group $\{x_3', x_4'\}$.

Table 3 Predicted values and their estimated variances at 5 points

	Predicted value	Estimated variance
$\boldsymbol{x}_1$	95.423	0.460
$\boldsymbol{x}_2$	100.496	3.706
$\boldsymbol{x}_3$	94.381	7.359
$\boldsymbol{x}_4$	95.742	5.285
$\boldsymbol{x}_5$	116.827	1689.129

Table 2 shows the estimated values of the 4 parameters β_i' and the 2 average group effects ξ_A^i in (15). The β_i' are poorly estimated with large standard errors due to multicollinearity generated by the two groups of strongly correlated variables. The t-test shows they are not significantly different from zero at the 5% level. The average group effects, on the other hand, are very accurately estimated with small standard errors and are highly significant. The estimated error variance is $\hat{\sigma}^2 = 2.306^2$, so the estimated standard errors of the two average group effects based on Theorem 1(*ii*) are $\hat{\sigma}/2 = 1.153$. It can be seen from Table 2 that the standard errors of the two estimated group effects are indeed close to this value. We write the least squares estimated model (8) as

$$\hat{y}' = (31.607x_1' - 2.261x_2')_G + (27.500x_3' + 8.353x_4')_G, \tag{23}$$

where the $(\dots)_G$ notation indicates that variables inside each $(\dots)_G$ are strongly correlated. Individual estimated parameter values such as 31.607 and −2.261 inside such brackets should not be used as point estimates as the underlying parameters are not meaningful and thus not estimated; they should only be used to estimate or make inference on meaningful group effects, such as ξ_A^1 and ξ_A^2, or make predictions over the feasible prediction region.

Finally, we demonstrate that the least squares estimated model gives accurate predictions over the feasible prediction region $\mathcal{R}_{FP}$, and accurate extrapolation is also possible with this estimated model. Consider 5 points

$$\begin{aligned}
\boldsymbol{x}_1 &= (7.46153, -11.76923, 48.15385, -30.00000),\\
\boldsymbol{x}_2 &= (3.18232, -15.98495, 64.86423, -10.86569),\\
\boldsymbol{x}_3 &= (7.25776, -11.10359, 46.53671, -28.84034),\\
\boldsymbol{x}_4 &= (-4.76478, -25.08204, 75.10608, -1.00862),\\
\boldsymbol{x}_5 &= (13.57470, -18.42563, 75.10608, -47.39482).
\end{aligned}$$

In the more detailed Hald cement data analysis in the Supplementary Material, we showed that $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$ are in $\mathcal{R}_{FP}$, but $\boldsymbol{x}_5$ is not. Plotting (x_1, x_2) of the 5 points and the 13 points in the Hald cement data in Fig. 1 finds $\boldsymbol{x}_4$ and $\boldsymbol{x}_5$ outside the data hull of the 13 points, so making predictions at $\boldsymbol{x}_4$ and $\boldsymbol{x}_5$ is extrapolation. Table 3 gives the predicted values and their estimated variances at the 5 points. The predictions at $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are accurate with small variances as these points are in both the data hull and $\mathcal{R}_{FP}$. Point $\boldsymbol{x}_5$ is not in $\mathcal{R}_{FP}$, so extrapolation at $\boldsymbol{x}_5$ is inaccurate with a

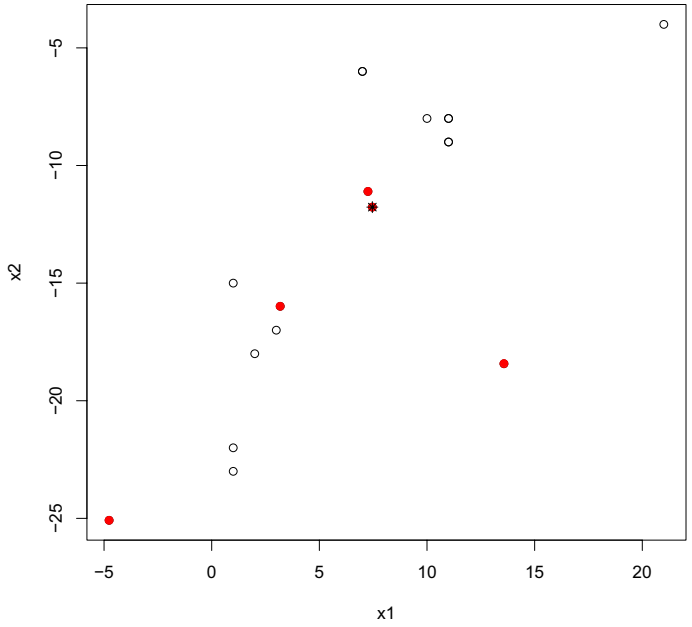

Fig. 1 Points representing (x_1, x_2) of the 13 observations in the Hald cement data are in circles. The "$\star$" symbol represents the mean of the 13 points. Points representing the 5 prediction points are in red dots. Points $\boldsymbol{x}_4$ and $\boldsymbol{x}_5$ are the two red dots outside the circle data hull, and $\boldsymbol{x}_4$ is the one in the lower left corner which is still inside the feasible prediction region. A plot of (x_3, x_4) of these points (not included) gives similar observations

large variance. In contrast, extrapolation at $\boldsymbol{x}_4$ is accurate as $\boldsymbol{x}_4$ is in $\mathcal{R}_{FP}$. This shows accurate extrapolation with the least squares estimated model is still possible when there is multicollinearity, provided it is done within $\mathcal{R}_{FP}$.

4 Concluding remarks

Multicollinearity due to strongly correlated predictor variables manifests in two ways. Numerically, it manifests through the ill-conditioning of the $\mathbf{X}^T\mathbf{X}$ matrix and ultimately the large variances of the least squares estimators for parameters of the strongly correlated variables. Geometrically, it manifests as a tight spatial constraint on the strongly correlated variables in that their data points are clustered tightly around a line.[2] Making predictions outside a narrow band around this line, including estimating parameters of these variables, is extreme extrapolation that may be meaningless and highly inaccurate.

[2] For unstandardised variables and/or variables not in an APC arrangement, this line is difficult to characterize. But for standardized variables in APC arrangement, this line is easy to describe; e.g. for the q variable in $\mathbf{X}_1'$ of (8), this line is $x_1' = x_2' = \cdots = x_q'$.

Existing methods for dealing with multicollinearity such as ridge regression and principal component regression all focus on overcoming the numerical ill-conditioning aspect of multicollinearity in order to produce more accurate estimators for parameters of the strongly correlated variables. They overlooked the geometric implication of multicollinearity which renders these parameters meaningless (see Remark [b] of Sect. 2.3). They may produce estimators with smaller variances than the least squares estimators, but this does not make the parameters they are trying to estimate more meaningful. Indeed, trying to accurately estimate parameters of strongly correlated variables is misguided. It also cannot be done in general as strongly correlated data contain little information about the individual parameters. With the misconception of their having more accurate predictions dispelled, there is little reason for abandoning the simple least squares regression in favour of these methods.

The group approach to the least squares regression respects the group nature of the strongly correlated predictor variables. It studies their group impact and is free of the multicollinearity problem. With the aid of the APC arrangement, it works effectively in estimation, inference, variable selection and prediction. We did not discuss model checking, but on this point, the group approach also has a clear advantage over the ridge regression and principal component regression as various residuals and residual plots for the least squares regression can be directly employed by the group approach with well-understood usages and interpretations, whereas the same cannot be said about the ridge regression and principal component regression. To conclude, we recommend the group approach to the least squares regression over existing methods for handling multicollinearity because of its simplicity and effectiveness.

Appendix

Proof of Lemma 1 Let $\mathbf{A}$ be the $q \times q$ matrix whose elements are all 1. Then, $\mathbf{A}$ has two distinct eigenvalues, $\lambda_1^A = q$ and $\lambda_2^A = 0$. Eigenvalue λ_1^A has multiplicity 1, and λ_2^A has multiplicity $(q-1)$. The orthonormal eigenvector of λ_1^A is $\frac{1}{\sqrt{q}}\mathbf{1}_q$. Here, we ignore the other orthonormal eigenvector of λ_1^A, $-\frac{1}{\sqrt{q}}\mathbf{1}_q$, which differs only in sign from $\frac{1}{\sqrt{q}}\mathbf{1}_q$.

Let $\mathbf{P} = [p_{ij}]$ be a perturbation matrix of $\mathbf{A}$ defined by

$$\mathbf{P} = \mathbf{A} - \mathbf{R}. \tag{24}$$

Then, $\mathbf{P}$ is real and symmetric and $p_{ij} = 1 - r_{ij}$. When $r_M \to 1$, since $p_{ij} = (1 - r_{ij}) \to 0$, we have $\|\mathbf{P}\|_2 \to 0$. It follows from this and $\mathbf{R} = \mathbf{A} - \mathbf{P}$ (so $\mathbf{R}$ is a perturbed version of $\mathbf{A}$) that $\lambda_1 \to \lambda_1^A = q$ and $\lambda_i \to \lambda_2^A = 0$ for $i = 2, 3, \ldots, q$ as $r_M \to 1$ (Horn and Johnson, 1985; page 367).

To show that $\mathbf{v}_1 \to \frac{1}{\sqrt{q}}\mathbf{1}_q$ as $r_M \to 1$, since $\mathbf{R}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$, we have

$$r_{i1}v_{11} + r_{i2}v_{12} + \cdots + r_{iq}v_{1q} = \lambda_1 v_{1i} \tag{25}$$

for $i = 1, 2, \ldots, q$, where $(r_{i1}, r_{i2}, \ldots, r_{iq})$ is the ith row of $\mathbf{R}$ and v_{1i} is the ith element of $\mathbf{v}_1$. All v_{1i} are bounded between -1 and 1 since $v_{1i}^2 \le \|\mathbf{v}_1\|^2 = 1$. When $r_M \to 1$, all $r_{ij} \to 1$, so $(r_{ij}v_{1j} - v_{1j}) \to 0$ for $j = 1, 2, \ldots, q$. Thus,

$$(r_{i1}v_{11} + r_{i2}v_{12} + \cdots + r_{iq}v_{1q}) - (v_{11} + v_{12} + \cdots + v_{1q}) \to 0 \tag{26}$$

as $r_M \to 1$. By (25) and (26), $\lambda_1 v_{1i} - (v_{11} + v_{12} + \cdots + v_{1q}) \to 0$ which implies $\lambda_1^2 v_{1i}^2 - (v_{11} + v_{12} + \cdots + v_{1q})^2 \to 0$ for $i = 1, 2, \ldots, q$. It follows that

$$\lambda_1^2(v_{11}^2 + v_{12}^2 + \cdots + v_{1q}^2) - q(v_{11} + v_{12} + \cdots + v_{1q})^2 \to 0. \tag{27}$$

Since $v_{11}^2 + v_{12}^2 + \cdots + v_{1q}^2 = \|\mathbf{v}_1\|^2 = 1$ and $\lambda_1 \to q$, (27) implies that $(v_{11} + v_{12} + \cdots + v_{1q}) \to \sqrt{q}$. This and (26) imply that

$$(r_{i1}v_{11} + r_{i2}v_{12} + \cdots + r_{iq}v_{1q}) \to \sqrt{q}$$

for $i = 1, 2, \ldots, q$. By (25), we also have $\lambda_1 v_{1i} \to \sqrt{q}$. This and $\lambda_1 \to q$ imply that $v_{1i} \to 1/\sqrt{q}$ for $i = 1, 2, \ldots, q$, that is, $\mathbf{v}_1 \to \frac{1}{\sqrt{q}}\mathbf{1}_q$. □

Proof of Lemma 2 Since $\mathbf{R}$ is positive definite, $\mathbf{R}^{-1}$ is also positive definite. Let $\lambda_1' \ge \lambda_2' \ge \cdots \ge \lambda_q' > 0$ be the eigenvalues of $\mathbf{R}^{-1}$. Then, $\lambda_i' = \lambda_{q-i+1}^{-1}$ and its eigenvector is $\mathbf{v}_i' = \mathbf{v}_{q-i+1}$ for $i = 1, 2, \ldots, q$. In particular, $\lambda_q' = \lambda_1^{-1}$ and $\mathbf{v}_q' = \mathbf{v}_1$. Since all $\lambda_i > 0$ and $trace(\mathbf{R}) = q = \sum_{i=1}^q \lambda_i$, we have $0 < \lambda_1 < q$. Also, $\mathbf{v}_1^T\mathbf{v}_1 = 1$ as $\mathbf{v}_1$ is orthonormal. It follows from these that

$$\mathbf{v}_1^T\mathbf{R}^{-1}\mathbf{v}_1 = \mathbf{v}_q'^T\mathbf{R}^{-1}\mathbf{v}_q' = \mathbf{v}_q'^T\lambda_q'\mathbf{v}_q' = \frac{\mathbf{v}_1^T\mathbf{v}_1}{\lambda_1} = \frac{1}{\lambda_1} > \frac{1}{q}, \tag{28}$$

which proves (i). By Lemma 1, $\lambda_1 \to q$ as $r_M \to 1$. Thus, by (28)

$$\mathbf{v}_1^T\mathbf{R}^{-1}\mathbf{v}_1 = \frac{1}{\lambda_1} \to \frac{1}{q},$$

as $r_M \to 1$, which proves (ii). □

Proof of Theorem 1 For any constant vector $\mathbf{c} \in \mathbb{R}^p$, we have

$$var(\mathbf{c}^T\hat{\boldsymbol{\beta}}') = \sigma^2\mathbf{c}^T[\mathbf{X}'^T\mathbf{X}']^{-1}\mathbf{c}. \tag{29}$$

Let $\mathbf{c}_E = (\mathbf{v}_1^{*T}, 0, \ldots, 0)^T$. Then, $\xi_E = \mathbf{c}_E^T\boldsymbol{\beta}'$ and $\hat{\xi}_E = \mathbf{c}_E^T\hat{\boldsymbol{\beta}}'$. By (11) and (29),

$$var(\hat{\xi}_E) = \sigma^2 \mathbf{v}_1^{*T}[\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}]^{-1}\mathbf{v}_1^* = \sigma^2 \mathbf{v}_1^{*T}\mathbf{R}^{*-1}\mathbf{v}_1^*. \tag{30}$$

To show (*i*), when variables in $\mathbf{X}_1'$ are uncorrelated with variables in $\mathbf{X}_2'$, $\mathbf{R}_{12} = \mathbf{0}$ and so $\mathbf{R}^* = \mathbf{R}$ and $\mathbf{v}_1^* = \mathbf{v}_1$. By (30),

$$var(\hat{\xi}_E) = \sigma^2 \mathbf{v}_1^T \mathbf{R}^{-1}\mathbf{v}_1. \tag{31}$$

Applying Lemma 2 to the right-hand side of (31), we obtain (i_1) and (i_2).

To show (*ii*), for simplicity, we assume general conditions discussed in footnote 1 hold so that $\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} \to \mathbf{0}$ when $\mathbf{R}_{12} \to \mathbf{0}$. It follows from this and conditions in Theorem 1(*ii*) that $\mathbf{R}_{11}$ and $\mathbf{R}^*$ will both converge to matrix $\mathbf{A}$ in (24). We again define a perturbation matrix of $\mathbf{A}$ as

$$\mathbf{P}^* = \mathbf{A} - \mathbf{R}^*$$

like what we did in (24). By following steps similar to those in the proofs of Lemma 1 and Lemma 2, we can show that $\mathbf{R}^*$ also has the two properties in Lemma 1 and property (*ii*) in Lemma 2. The latter and (30) imply (*ii*). □

Proof of Theorem 2 Since $\mathbf{v} \cdot \mathbf{v}_1^* = \|\mathbf{v}\|\|\mathbf{v}_1^*\| cos(\theta) = cos(\theta)$ where θ is the angle between $\mathbf{v}$ and $\mathbf{v}_1^*$, $\sqrt{1-\delta} < \mathbf{v} \cdot \mathbf{v}_1^* \le 1$ is equivalent to $\sqrt{1-\delta} < cos(\theta) \le 1$ or $0 \le \theta < \theta_\delta$ for some small fixed $\theta_\delta > 0$. Thus, $\mathcal{N}_\delta$ in (14) represents a small open circular region centred on $\mathbf{v}_1^*$ on the surface of the unit sphere.

Similar to $var(\hat{\xi}_E)$ in (30), $var(\mathbf{v}^T\hat{\boldsymbol{\beta}}_1') = \sigma^2 \mathbf{v}^T \mathbf{R}^{*-1}\mathbf{v}$. Since $\mathbf{R}^{*-1}$ is real symmetric positive definite, it has eigendecomposition $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$ where $\mathbf{Q}$ is the matrix of orthonormal eigenvectors including $\mathbf{v}_1^*$ and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. The smallest eigenvalue of $\mathbf{R}^{*-1}$ is $1/\lambda_1^*$ which converges to $1/q$ under the condition of Theorem 2 as r_M goes to 1. The other eigenvalues of $\mathbf{R}^{*-1}$ all go to infinity as r_M goes to 1. For any unit vector $\mathbf{v}$,

$$1 = \mathbf{v}^T\mathbf{v} = \mathbf{v}^T\mathbf{Q}\mathbf{Q}^T\mathbf{v} = \mathbf{v}^T[\tilde{\mathbf{Q}}, \mathbf{v}_1^*][\tilde{\mathbf{Q}}, \mathbf{v}_1^*]^T\mathbf{v} = \mathbf{v}^T\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^T\mathbf{v} + (\mathbf{v}^T\mathbf{v}_1^*)^2 \tag{32}$$

where $\tilde{\mathbf{Q}}$ is the matrix containing all columns of $\mathbf{Q}$ but $\mathbf{v}_1^*$. If $\mathbf{v} \notin \mathcal{N}_\delta$, then $(\mathbf{v}^T\mathbf{v}_1^*)^2 \le 1-\delta$. This and (32) imply that $1 \le \mathbf{v}^T\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^T\mathbf{v} + (1-\delta)$, that is, $\mathbf{v}^T\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}^T\mathbf{v} \ge \delta$. This leads to the following lower bound on $var(\mathbf{v}^T\hat{\boldsymbol{\beta}}_1')$,

$$var(\mathbf{v}^T\hat{\boldsymbol{\beta}}_1') = \sigma^2\mathbf{v}^T\mathbf{R}^{*-1}\mathbf{v} = \sigma^2\mathbf{v}^T\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T\mathbf{v} \ge \sigma^2\mathbf{v}^T\tilde{\mathbf{Q}}\tilde{\boldsymbol{\Lambda}}\tilde{\mathbf{Q}}^T\mathbf{v} \ge \frac{\sigma^2\delta}{\lambda_2^*}, \tag{33}$$

where $\tilde{\boldsymbol{\Lambda}}$ is the diagonal matrix of all eigenvalues of $\mathbf{R}^{*-1}$ except the smallest one $1/\lambda_1^*$, and $1/\lambda_2^*$ is the second smallest eigenvalue of $\mathbf{R}^{*-1}$. Since $1/\lambda_2^* \to \infty$ as $r_M \to 1$, (33) implies that $var(\mathbf{v}^T\hat{\boldsymbol{\beta}}_1') \to \infty$ as $r_M \to 1$ if $\mathbf{v} \notin \mathcal{N}_\delta$. □

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-022-00841-7.

Acknowledgements We would like to thank two anonymous reviewers and an Associate Editor for their helpful comments which have led to many improvements in this paper. This work is supported by the Natural Sciences and Engineering Research Council of Canada.

References

Belsley, D. A., Kuh, E., Welsch, R. E. (2004). *Regression diagnostics*: *Identifying influential data and sources of collinearity*. New York: Wiley & Sons.

Conniffe, D., Stone, J. (1973). A critical view of ridge regression. *American Statistician, 22*, 181–187.

Draper, N. R., Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: Wiley.

Draper, N. R., Van Nostrand, R. C. (1979). Ridge regression and James-Stein estimators: Review and comments. *Technometrics, 21,* 451–466.

Gunst, R. F., Mason, R. L. (1977). Biased estimation in regression: An evaluation using mean squared error. *Journal of the American Statistical Association, 72*, 616–628.

Gunst, R. F., Webster, J. T., Mason, R. L. (1976). A comparison of least squares and latent root regression estimators. *Technometrics, 18*, 75–83.

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

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

Horn, R. A., Johnson, C. A. (1985). *Matrix analysis*. Cambridge: Cambridge University Press.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Lawless, J. F. (1978). Ridge and related estimation procedures: Theory and practice. *Communications in Statistics: Theory and Methods, 7*, 135–164.

Montgomery, D. C., Peck, E. A., Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). New York: Wiley.

Tsao, M. (2019). Estimable group effects for strongly correlated variables in linear models. *Journal of Statistical Inference and Planning, 198*, 29–42.

Webster, J. T., Gunst, R. F., Mason, R. L. (1974). Latent root regression analysis. *Technometrics, 16*, 513–522.