# The successive raising estimator and its relation with the ridge estimator 

J. García and D.E. Ramirez

July 23, 2016


#### Abstract

The raise estimators are used to reduce collinearity in linear regression models by raising a column in the experimental data matrix which may be nearly linear with the other columns. The raising procedure has two components, namely stretching and rotating which we can analyze separately. We give the relationship between the raise estimators and the classical ridge estimators. Using a case study, we show how to determine the perturbation parameter for the raise estimators by controlling the amount of precision to be retained in the original data.

\section*{J. García}

Department of Applied Economy, Almeria University, Almeria, Spain e-mail jgarcia@ual.es

\section*{D. Ramirez}

Department of Mathematics, University of Virginia, Charlottesville, VA USA e-mail der@virginia.edu


Keywords: Raise estimators, ridge estimators, collinearity, Stein estimators
MSC: 62J05, 62J07

## 1 Introduction

A technique to mitigate the existence of collinearity is constrained estimators such as ridge estimators and Stein estimators. In both cases, the goal is to diminish the Mean Square Error $M S E$ and to mitigate the problems inherent with collinearity. These estimators improve the effect of collinearity by constraining the Mahalanobis distance of the estimator $\widehat{\boldsymbol{\beta}}$ under the constraint $\boldsymbol{\beta}^{\prime} \mathbf{Q} \boldsymbol{\beta}=\delta^{2}$ with $\mathbf{Q}$ a positive definite matrix and $\delta>0$.

From the model $\mathbf{y}=\mathbf{X} \boldsymbol{\beta}+\mathbf{u}$ with uncorrelated, zero-mean, and homoscedastic errors, we know that the $O L S$ estimators $\widehat{\boldsymbol{\beta}}$ are given by

$$
\begin{equation*}
\widehat{\boldsymbol{\beta}}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1} \mathbf{X}^{\prime} \mathbf{y} \tag{1}
\end{equation*}
$$

A direct application of Lagrange multipliers gives the constrained estimator $\widehat{\boldsymbol{\beta}}_{c s}$ with $\widehat{\boldsymbol{\beta}}_{c s}^{\prime} \mathbf{Q} \widehat{\boldsymbol{\beta}}_{c s}=\delta^{2}$ as

$$
\begin{equation*}
\widehat{\boldsymbol{\beta}}_{c s}=\left(\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{Q}\right)^{-1} \mathbf{X}^{\prime} \mathbf{y} \tag{2}
\end{equation*}
$$

where $k$ is the Lagrange parameter chosen to satisfy the constraint $\widehat{\boldsymbol{\beta}}_{c s}^{\prime} \mathbf{Q} \widehat{\boldsymbol{\beta}}_{c s}=$ $\delta^{2}$.

From the Loewner partial ordering $\mathbf{X}^{\prime} \mathbf{X} \preceq{ }_{L} \mathbf{X}^{\prime} \mathbf{X}+k \mathbf{Q}$, the inverses will have the opposite partial order $\left(\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{Q}\right)^{-1} \preceq_{L}\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}$. When $\mathbf{Q}$ and $\mathbf{X}^{\prime} \mathbf{X}$ commute, for example with $\mathbf{Q}=k \mathbf{I}_{p}$ or $\mathbf{Q}=k \mathbf{X}^{\prime} \mathbf{X}$, the two cases considered below, the pair ( $\mathbf{X}^{\prime} \mathbf{X}, \mathbf{X}^{\prime} \mathbf{X}+k \mathbf{Q}$ ) can be simultaneously diagonalized into the pair $\left(\mathbf{I}_{p}, \boldsymbol{\Lambda}\right)$ with $\boldsymbol{\Lambda}$ the diagonal matrix containing the characteristic roots of $\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{Q}$. From the Loewner partial ordering $\mathbf{X}^{\prime} \mathbf{X} \preceq_{\mathbf{L}} \mathbf{X}^{\prime} \mathbf{X}+k \mathbf{Q}$, the Loewner partial ordering $\mathbf{I}_{p} \preceq_{L} \boldsymbol{\Lambda}$ follows, and thus $\mathbf{I}_{p}^{2} \preceq{ }_{L} \boldsymbol{\Lambda}^{2}$ so $\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{2} \preceq_{L}\left(\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{Q}\right)^{2}$, and it follows that $\left\|\widehat{\boldsymbol{\beta}}_{c s}\right\|^{2} \leq\|\widehat{\boldsymbol{\beta}}\|^{2}$ with the constrained estimators shown to have decreased in length. We note that in general, $\mathbf{A} \preceq_{L} \mathbf{B}$ does not imply that $\mathbf{A}^{2} \preceq{ }_{L} \mathbf{B}^{2}$, Puntanen et al. (2011, p. 316). From this idea, we consider two types of constrained estimators:

1. Stein estimators: From $\mathbf{Q}=\mathbf{X}^{\prime} \mathbf{X}$ the expression (2) becomes:

$$
\begin{equation*}
\hat{\boldsymbol{\beta}}_{s}=\left(\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{X}^{\prime} \mathbf{X}\right)^{-1} \mathbf{X}^{\prime} \mathbf{y}=\frac{1}{1+k} \hat{\boldsymbol{\beta}} \tag{3}
\end{equation*}
$$

For a detailed study of the Stein estimators we recommend the original reference of Stein (1960). These estimators were also studied by Mayer and Willke (1973). These estimators, which satisfy expression (3), define a whole class of biased estimators, the so called shrinkage estimators. They are obtained by shrinking the least squares estimator towards the origin, and they satisfy the $M S E$ Admissibility Condition which assures an improvement (decrease) in $M S E$ for some $k \in(0, \infty)$.
2. Ridge estimators: From $\mathbf{Q}=\mathbf{I}_{p}$, the expression (2) becomes:

$$
\begin{equation*}
\hat{\boldsymbol{\beta}}_{r}=\left(\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{I}_{p}\right)^{-1} \mathbf{X}^{\prime} \mathbf{y} \tag{4}
\end{equation*}
$$

where $\mathbf{I}_{p}$ is the matrix identity. Hoerl and Kennard (1970) established that the ridge estimators also satisfy the $M S E$ Admissibility Condition assuring an improvement in $M S E$ for some $k \in(0, \infty)$. The earliest detailed expositions of these estimators are found in Marquardt (1963) and Hoerl and Kennard (1970) with Marquardt (1963) acknowledging that Levenberg (1944) had observed that a perturbation of the diagonal improved convergence in steepest descent algorithms. The history of the early use of matrix diagonal increments in statistical problems is given in the article by Piegorsh and Casella (1989).

Numerous papers have been written justifying the need to find values $0<k_{i}$ $(1 \leq i \leq p)$ to perturb each of the characteristic roots of the matrix $\mathbf{X}^{\prime} \mathbf{X}$ to
improve the condition number of the inverse $\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}$. For a positive definite matrix $\mathbf{A}$, we define the condition number $\kappa$ as the ratio of the largest to smallest of the eigenvalues of $\mathbf{A}$. From this, the generalized ridge estimator is established as

$$
\widehat{\boldsymbol{\beta}}_{r}=\left(\mathbf{X}^{\prime} \mathbf{X}+\mathbf{K}\right)^{-1} \mathbf{X}^{\prime} \mathbf{y}
$$

where $\mathbf{K}$ is the diagonal matrix $\operatorname{diag}\left(k_{1}, \ldots, k_{p}\right)$ containing these perturbations. This regression was briefly commented on by Hoerl and Kennard (1970) and Goldstein and Smith (1974). As the ridge estimator is not equivariant under scaling, the common convention is to scale $\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}$ to correlation form with the explanatory variables centered and scaled to unit length and allowing for a constant term in the linear model. With $\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}$ in correlation form, the estimated variances $\widehat{\operatorname{var}}\left(\beta_{i}\right)$ of the estimates are equal, and the perturbation matrix is usually taken to be $\mathbf{K}=k \mathbf{I}_{p}$ with $k>0$.

The attempts to find the basis of the ridge estimator have been numerous; for example, by using the constrained regression imposing the constraint $\boldsymbol{\beta}^{\prime} \boldsymbol{\beta} \leq \delta^{2}$, Davidov (2006) used the Kuhn-Tucker Theorem to show that the conditioned least squared solution agrees with the ridge solution with constraint $\boldsymbol{\beta}^{\prime} \boldsymbol{\beta}=\delta^{2}$. Some anomalies of the ridge methodology can be found in Jensen and Ramirez (2008), Kapat and Goel (2010), and Jensen and Ramirez (2010a).

Using the notation from García et al. (2015) define the augmented matrix for $\mathbf{X}$ and the augmented vector for $\mathbf{y}$ by $\mathbf{X}_{A}=\left[\mathbf{X}^{\prime}, \sqrt{k} \mathbf{I}_{p}\right]^{\prime}$ and $\mathbf{y}_{A}=\left[\mathbf{y}^{\prime}, \mathbf{0}_{p}^{\prime}\right]^{\prime}$, respectively. The least squared solutions for the constrained model $\left(\mathbf{X}^{\prime} \mathbf{X}\right) \boldsymbol{\beta}=\mathbf{X}^{\prime} \mathbf{y}$ subject to $\boldsymbol{\beta}^{\prime} \boldsymbol{\beta}=\delta^{2}$ and the unconstrained model $\left(\mathbf{X}_{A}^{\prime} \mathbf{X}_{A}\right) \boldsymbol{\beta}=\mathbf{X}_{A}^{\prime} \mathbf{y}_{A}$ both have solution $\widehat{\boldsymbol{\beta}}=\left(\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{I}_{p}\right)^{-1} \mathbf{X}^{\prime} \mathbf{y}$. Augmenting the matrix and the response vector is similar in spirit to imposing identifiability constraints on the parameters as in Seber (1977). Hoerl and Kennard (1970, Eq. 2.3) suggested the relationship between the ridge estimator and the $O L S$ estimator as $\widehat{\boldsymbol{\beta}}_{r}=\left[\mathbf{I}_{p}+k\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}\right]^{-1} \widehat{\boldsymbol{\beta}}$. However, note that since $\widehat{\boldsymbol{\beta}}_{r}$ is constrained to lie on the sphere $\boldsymbol{\beta}^{\prime} \boldsymbol{\beta}=\delta^{2}, \widehat{\boldsymbol{\beta}}_{r}$ has a joint singular distribution in $\mathbb{R}^{n}$ with rank $n-1$ while $\widehat{\boldsymbol{\beta}}$ has a joint non-singular distribution in $\mathbb{R}^{n}$ with rank $n$. These and other anomalies with ridge regression are discussed in Jensen and Ramirez (2010a).

To avoid problems of singularity in the distribution of $\widehat{\boldsymbol{\beta}}_{r}$, one can consider a perturbation of the data $\mathbf{X} \rightarrow \mathbf{X}_{S}$ with $\mathbf{X}_{S}^{\prime} \mathbf{X}_{S}=\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{I}_{p}$ as in Jensen and Ramirez (2008). They call $\mathbf{X}_{S}$ the surrogate matrix, and it is defined using a perturbation of the singular values of $\mathbf{X}$, and yields the surrogate estimator $\widehat{\boldsymbol{\beta}}_{S}$ from $\left(\mathbf{X}_{S}^{\prime} \mathbf{X}_{S}\right) \boldsymbol{\beta}=\mathbf{X}_{S}^{\prime} \mathbf{y}$. Similarly, the raised estimators, discussed below, also improve the ill-conditioning $\mathbf{X} \rightarrow \widetilde{\mathbf{X}}$ with $\widetilde{\mathbf{X}}$ used on both sides of the normal equation $\left(\widetilde{\mathbf{X}}^{\prime} \widetilde{\mathbf{X}}\right) \boldsymbol{\beta}=\widetilde{\mathbf{X}}^{\prime} \mathbf{y}$, an importance difference with ridge estimators. Some authors bypass the constrained optimization foundation of ridge estimators and adopt the view, from numerical analysis given in Riley (1955), that perturbations of the diagonal improves the stability for matrix inverses.

For the augmented model $\mathbf{y}_{A}=\mathbf{X}_{A} \boldsymbol{\beta}+\mathbf{u}$ the error terms cannot be centered. This could be understood as a critique to ridge estimator obtained from the
augmented model. Minimizing the sum of the squared errors $e_{i}=y_{i}-\widehat{y}_{i}$ for the augmented rows $\sum_{i=1}^{p} e_{i}^{2}=k \widehat{\boldsymbol{\beta}}^{\prime} \widehat{\boldsymbol{\beta}}$ replaces the constraint $\boldsymbol{\beta}^{\prime} \boldsymbol{\beta}=\delta^{2}$. Case studies in García et al. (2015) show that the variance inflation factors VIF $(1 \leq i \leq p)$ computed for the surrogate model $\mathbf{y}=\mathbf{X}_{S} \boldsymbol{\beta}+\mathbf{u}$ and for the augmented model $\mathbf{y}_{A}=\mathbf{X}_{A} \boldsymbol{\beta}+\mathbf{u}_{A}$ are asymptotically equal $(n \rightarrow \infty)$.

To conclude, by using the words of Marshall and Olkin (1979) it seems that the ridge estimator can find a justification in the result provided by Riley (1955) where it is shown that the matrix $\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{I}_{p}$ is better conditioned than the matrix $\mathbf{X}^{\prime} \mathbf{X}$.

In this paper, we present some properties of the raised estimators originally presented by García et al. (2011) and its varieties: the stretched and the rotated estimators. We focus on finding an explication about the basis of the ridge and Stein estimators. Raised regression is similar to ridge estimator. One difference is that the ridge procedure seeks stability of $\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}$ on the left-hand side of the normal equations $\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1} \widehat{\boldsymbol{\beta}}=\mathbf{X}^{\prime} \mathbf{y}$, transforming the normal equations into the ridge equations $\left(\mathbf{X}^{\prime} \mathbf{X}+\mathbf{k I}_{\mathbf{p}}\right)^{-1} \widehat{\boldsymbol{\beta}}=\mathbf{X}^{\prime} \mathbf{y}$. Our procedure is in the spirit of the surrogate estimators of Jensen and Ramirez (2010b) where $\mathbf{X}$ is modified $\mathbf{X} \rightarrow \widetilde{\mathbf{X}}$ and the modified matrix is used on both sides of the normal equations. Thus the raised procedure allows the user to visualize the perturbed design matrix $\widetilde{\mathbf{X}}$, a feature not available with ridge regression.

The paper is organized as follows: Section 2 presents the generalized raised estimator for two variables. Section 3 presents the Stretched and the Rotated estimators as varieties of the raised estimator. Section 4 studies the raised estimators in the case of two variables and shows the way to obtain $\left(\mathbf{X}^{\prime} \mathbf{X}+\mathbf{k} \mathbf{I}_{\mathbf{p}}\right)$ by using the successive raising procedure. Section 5 studies the variance inflation factors and the metric number. Section 6 gives the relation between the raise and ridge estimators. Our case study is presented in Section 7. Finally, Section 8 contains the conclusions. Some technical results are shown in the two appendices.

## 2 The generalized raise method for two variables

We consider the linear model $\mathbf{y}=\mathbf{X} \boldsymbol{\beta}+\mathbf{u}$ where $E(\mathbf{u})=\mathbf{0}, E\left(\mathbf{u u ^ { \prime }}\right)=\sigma^{2} \mathbf{I}_{n}$ and $\mathbf{X}$ is a full rank matrix $n \times p$. For convenience, we have taken $\sigma^{2}=1$. We assume that the variables are centered and standardized, that is, $\mathbf{X}^{\prime} \mathbf{X}$ is in correlation form.

For the $n \times k$ matrix $\mathbf{A}=\left[\mathbf{a}_{1}, \mathbf{a}_{2}, \ldots, \mathbf{a}_{p}\right]$, the columns span is denoted by $S p(\mathbf{A})$, with $\mathbf{A}_{(j)}$ denoting the $j^{t h}$ column vector $\mathbf{a}_{j}$, and $\mathbf{A}_{[j]}$ denoting the $n \times(p-1)$ matrix formed by deleting $\mathbf{A}_{(j)}$ from $\mathbf{A}$. For the linear model $\mathbf{y}=\mathbf{X} \boldsymbol{\beta}+\mathbf{u}$, central to a study of collinearity is the relationship between $\mathbf{X}_{(j)}$ and $S p\left(\mathbf{X}_{[j]}\right)$. Indeed, there is a monotone relationship between the collinearity indices $\kappa_{j}$ of Stewart (1987), the variance inflation factors $V I F_{j}$, and the angle between $\left(\mathbf{X}_{(j)}, S p\left(\mathbf{X}_{[j]}\right)\right)$, see Jensen and Ramirez (2013, Theorem 4).

Let $\mathbf{P}_{[j]}=\mathbf{X}_{[j]}\left(\mathbf{X}_{[j]}^{\prime} \mathbf{X}_{[j]}\right)^{-1} \mathbf{X}_{[j]}^{\prime}$ be the projection operator onto the subspace $S p\left(\mathbf{X}_{[j]}\right) \subset \mathbb{R}^{n}$ spanned by the columns of the reduced (or relaxed) matrix $\mathbf{X}_{[j]}$. From the geometry of the right triangle formed by $\left(\mathbf{X}_{(j)}, \mathbf{P}_{[j]} \mathbf{X}_{(j)}\right)$, the squared lengths satisfy

$$
\left\|\mathbf{X}_{(j)}\right\|^{2}=\left\|\mathbf{P}_{[j]} \mathbf{X}_{(j)}\right\|^{2}+\left\|\mathbf{X}_{(j)}-\mathbf{P}_{[j]} \mathbf{X}_{(j)}\right\|^{2}
$$

where $\left\|\mathbf{X}_{(j)}-\mathbf{P}_{[j]} \mathbf{X}_{(j)}\right\|^{2}=R S S_{j}$ is the residual sum of squares from the regression $\mathbf{X}_{(j)}=\mathbf{X}_{[j]} \boldsymbol{\alpha}+\mathbf{v}$ with $\widehat{\boldsymbol{\alpha}}=\mathbf{P}_{[j]} \mathbf{X}_{(j)}$. The angle $\theta_{j}$ between $\mathbf{X}_{(j)}$ and $S p\left(\mathbf{X}_{[j]}\right)$ is given by the angle between $\mathbf{X}_{(j)}$ and $\mathbf{P}_{[j]} \mathbf{X}_{(j)}$ and satisfies

$$
\cos \left(\theta_{j}\right)=\frac{\mathbf{X}_{(j)}^{\prime} \mathbf{P}_{[j]} \mathbf{X}_{(j)}}{\left\|\mathbf{X}_{(j)}\right\|\left\|\mathbf{P}_{[j]} \mathbf{X}_{(j)}\right\|}=\frac{\left\|\mathbf{P}_{[j]} \mathbf{X}_{(j)}\right\|}{\left\|\mathbf{X}_{(j)}\right\|}
$$

We show the details leading to the raise method with $p=2$.
In the case of two variables, $\mathbf{X}=\left[\mathbf{x}_{1}, \mathbf{x}_{2}\right]$, the model is given by $\mathbf{y}=\beta_{1} \mathbf{x}_{1}+$ $\beta_{2} \mathbf{x}_{2}+\mathbf{u}$ where the matrix $\mathbf{X}^{\prime} \mathbf{X}$ is defined as $\mathbf{X}^{\prime} \mathbf{X}=\left(\begin{array}{cc}1 & \sum x_{2 i} x_{1 i} \\ \sum x_{2 i} x_{1 i} & 1\end{array}\right)=$ $\left(\begin{array}{ll}1 & \rho \\ \rho & 1\end{array}\right)$ with $\rho$ the correlation coefficient between the exogenous variables.

On the other hand, the matrix $\mathbf{X}^{\prime} \mathbf{y}$ is given by $\mathbf{X}^{\prime} \mathbf{y}=\left(\sum x_{1 i} y_{i}, \sum x_{2 i} y_{i}\right)^{\prime}$. The collinearity problem occurs when the vectors $\mathbf{x}_{1}$ and $\mathbf{x}_{2}$ are very close; that is, when the angle between the vectors is very small.

To correct this problem, we will try to separate both vectors. The raise method begins with the regression between $\mathbf{x}_{1}$ and $\mathbf{x}_{2}$ in the following way. Pick one of the columns of $\mathbf{X}$, say $\mathbf{X}_{(1)}$, and regress this column using the remaining column(s):

$$
\begin{aligned}
\mathbf{X}_{(1)} & =\mathbf{X}_{[1]} \boldsymbol{\alpha}+\mathbf{v} \\
\mathbf{x}_{1} & =\alpha \mathbf{x}_{2}+\mathbf{v}
\end{aligned}
$$

The matrix $\mathbf{X}_{(2)}$ is $\mathbf{X}_{(2)}^{\prime}=\left(x_{21}, x_{22,}, \ldots, x_{2 n}\right)$, with $\mathbf{X}_{(2)}^{\prime} \mathbf{X}_{(2)}=\sum x_{2 i}^{2}=1$ and $\mathbf{X}_{[1]}^{\prime} \mathbf{X}_{(1)}=\mathbf{X}_{(2)}^{\prime} \mathbf{X}_{(1)}=\sum x_{2 i} x_{1 i}=\rho$. Thus,

$$
\begin{equation*}
\hat{\alpha}=\left(\mathbf{X}_{[1]}^{\prime} \mathbf{X}_{[1]}\right)^{-1} \mathbf{X}_{[1]}^{\prime} \mathbf{X}_{(1)}=\left[\mathbf{X}_{(2)}^{\prime} \mathbf{X}_{(2)}\right]^{-1}\left[\mathbf{X}_{(2)}^{\prime} \mathbf{X}_{(1)}\right]=\rho \tag{5}
\end{equation*}
$$

Summarizing, with $\left\|\mathbf{x}_{1}\right\|^{2}=1=\left\|\mathbf{x}_{2}\right\|^{2}$, we can write that $\mathbf{x}_{1}=\rho \mathbf{x}_{2}+\mathbf{e}_{1}$ with the estimated errors $\mathbf{e}_{1}$ orthogonal to $S p\left(\mathbf{X}_{[1]}\right)$; and, in particular, $\mathbf{e}_{1} \perp \mathbf{x}_{2}$ with $\mathbf{e}_{1}=\mathbf{x}_{1}-\rho \mathbf{x}_{2}$ with

$$
\begin{aligned}
\left\|\mathbf{x}_{1}\right\|^{2} & =\left\|\mathbf{P}_{[1]} \mathbf{x}_{1}\right\|^{2}+\left\|\mathbf{e}_{1}\right\|^{2} \\
\left\|\mathbf{x}_{1}\right\|^{2} & =\left\|\rho \mathbf{x}_{2}\right\|^{2}+\left\|\mathbf{e}_{1}\right\|^{2}
\end{aligned}
$$

As $\left\|\mathbf{x}_{1}\right\|^{2}=\left\|\mathbf{x}_{2}\right\|^{2}=1$, the raise method transforms $\mathbf{x}_{1} \rightarrow \tilde{\mathbf{x}}_{1}$ by the rule

$$
\tilde{\mathbf{x}}_{1}=\mathbf{x}_{1}+\lambda \mathbf{e}_{1}=\mathbf{x}_{1}+\lambda\left(\mathbf{x}_{1}-\rho \mathbf{x}_{2}\right)=(1+\lambda) \mathbf{x}_{1}-\lambda \rho \mathbf{x}_{2}
$$

with

$$
\begin{align*}
\left\|\tilde{\mathbf{x}}_{1}\right\|^{2} & =(1+\lambda)^{2}\left\|\mathbf{x}_{1}\right\|^{2}-2(1+\lambda) \lambda \rho \mathbf{x}_{1}^{\prime} \mathbf{x}_{2}+(\lambda \rho)^{2}\left\|\mathbf{x}_{2}\right\|^{2} \\
& =(1+\lambda)^{2}-\rho^{2}\left(\lambda^{2}+2 \lambda\right) \tag{6}
\end{align*}
$$

Equivalently with $\left\|\tilde{\mathbf{x}}_{1}\right\|^{2}=1+2 \lambda\left(1-\rho^{2}\right)+\lambda^{2}\left(1-\rho^{2}\right)$ with inner product $\tilde{\mathbf{x}}_{1}^{\prime} \mathbf{x}_{2}=$ $\left(\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right)^{\prime} \mathbf{x}_{2}=\rho+\lambda \mathbf{e}_{1}^{\prime} \mathbf{x}_{2}=\rho$.

Denote $\widetilde{\mathbf{X}}=\left[\widetilde{\mathbf{x}}_{1}^{\prime}, \mathbf{x}_{2}^{\prime}\right]^{\prime}$. With $\lambda>0,\left\|\tilde{\mathbf{x}}_{1}\right\|^{2}>\left\|\mathbf{x}_{1}\right\|^{2}$, and we say $\mathbf{x}_{1}$ has been raised from an angle $\theta_{1}$ to an angle $\widetilde{\theta}_{1}$ with $\cos \left(\theta_{1}\right)=\left\|\mathbf{P}_{[1]} \mathbf{X}_{(1)}\right\| /\left\|\mathbf{x}_{1}\right\|$ and $\cos \left(\widetilde{\theta}_{1}\right)=\left\|\mathbf{P}_{[1]} \mathbf{X}_{(1)}\right\| /\left\|\tilde{\mathbf{x}}_{1}\right\|$. The Variance Inflation Factor $V I F_{1}$ for $\widetilde{\mathbf{X}}$ is functionally related to the angle $\widetilde{\theta}_{1}$ by the rule $\widetilde{\theta}_{1}=\arccos \left(\sqrt{1-1 / V I F_{1}}\right)$, for example Jensen and Ramirez (2013). Thus as $\lambda \rightarrow \infty, \widetilde{\theta}_{1} \rightarrow 90^{\circ}$ and the variance inflation factor $V I F_{1}$ converges to one indicating that collinearity is being eliminated, as in García et al. (2011, Theorem 4.2).

If we replace $\mathbf{x}_{1}$ by $\tilde{\mathbf{x}}_{1}$ in the original linear regression from Eq. (1), the raised model will be $\mathbf{y}=\beta_{1}(\lambda) \tilde{\mathbf{x}}_{1}+\beta_{2}(\lambda) \mathbf{x}_{2}+\mathbf{u}$ where the estimated parameters depend on $\lambda$ and will be denoted as $\beta_{1}(\lambda)$ and $\beta_{2}(\lambda)$.

Summarizing, in the case of two variables, $\mathbf{X}=\left[\mathbf{x}_{1}, \mathbf{x}_{2}\right]$, we regressed the first column using the remaining second column from $\mathbf{X}_{(1)}=\mathbf{X}_{[1]} \boldsymbol{\alpha}+\mathbf{v}$ with $\mathbf{x}_{1}=\rho \mathbf{x}_{2}+\mathbf{e}_{1}$ and $\tilde{\mathbf{x}}_{1}=\mathbf{x}_{1}+\lambda \mathbf{e}_{1}$. With the first column raised, the raised transformation $\mathbf{X} \rightarrow \mathbf{X}_{<1>}$ has the raised design matrix given by

$$
\begin{equation*}
\mathbf{X}_{<1>}=\left(\widetilde{\mathbf{x}}_{1}, \mathbf{x}_{2}\right)=\left(\left(1+\lambda_{1}\right) \mathbf{x}_{1}-\lambda_{1} \rho \mathbf{x}_{2}, \mathbf{x}_{2}\right) \tag{7}
\end{equation*}
$$

using $<1, \ldots, k>$ to indicate the column(s) that have been raised, notation that will be helpful in the general case. We note that

$$
\mathbf{X}_{<1>}^{\prime} \mathbf{X}_{<1>}=\left(\begin{array}{cc}
1+2 \lambda\left(1-\rho^{2}\right)+\lambda^{2}\left(1-\rho^{2}\right) & \rho  \tag{8}\\
\rho & 1
\end{array}\right)
$$

with only the $(1,1)$ element of $\mathbf{X}^{\prime} \mathbf{X}$ being effected by raising the first column of $\mathbf{X}$. We consider additional columns to be raised in Section 4.

With ridge regression, the coefficient of determination $R^{2}(k)$ monotonically decreases as $k \rightarrow \infty$, MacDonald (2010). We give the parallel result for surrogate regression for $R_{S}^{2}(k)$ in Appendix A. The importance of these results is that it allows the user to set a lower bound for changes to the original model, such as $R^{2}(k) \geq 0.95$ or $R_{S}^{2}(k) \geq 0.95$. One very desirable property of the raised regression method is that the coefficient of determination does not change with $\widetilde{R}^{2}(\lambda)=R^{2}(0)$ for all $\lambda \in(0, \infty)$. Additionally, the predicted values with the $O L S$ regression $\mathbf{X} \widehat{\boldsymbol{\beta}}$ and the predicted values with the raised regression $\widetilde{\mathbf{X}} \widehat{\boldsymbol{\beta}}(\lambda)$ are the same:

$$
\begin{equation*}
\mathbf{X} \widehat{\boldsymbol{\beta}}=\widetilde{\mathbf{X}} \widehat{\boldsymbol{\beta}}(\lambda) \tag{9}
\end{equation*}
$$

Thus raising a column vector in $\mathbf{X}$ is not effecting the basic $O L S$ regression model. These results follow from noting that the raised vector remains in the original $S p(\mathbf{X}), \mathbf{e}_{j}=\mathbf{X}_{(j)}-\mathbf{P}_{[j]} \mathbf{X}_{(j)} \in S p(\mathbf{X})$ so $S p(\widetilde{\mathbf{X}})=S p(\mathbf{X})$, as shown in García et al. (2011).

## 3 The stretched and rotated estimators

The raised column $\tilde{\mathbf{x}}_{j}=\mathbf{x}_{j}+\lambda \mathbf{e}_{j}$ with $\lambda \geq 0$ has two effects; namely, stretching the column and rotating the column. We study the effects separately with a case study. Set $\mathbf{X}_{0}=\left[\mathbf{1}, \mathbf{x}_{1}, \mathbf{x}_{2}, \mathbf{x}_{3}\right]$ the $20 \times 4$ matrix from Rawlings (1998, Ex. 11.11) which has been designed to be highly collinear. The pattern has $\mathbf{x}_{1}$ as the sequence from 20 to 29 and repeated; $\mathbf{x}_{2}$ is $\mathbf{x}_{1}-25$ with the $1^{\text {st }}$ and $11^{\text {th }}$ term changed from -5 to $-4 ; \mathbf{x}_{3}$ is the repeated sequence $5,4,3,2,1,2,3,4,5,6$.

The eigenvalues for $\mathbf{X}_{0}^{\prime} \mathbf{X}_{0}$ are [12439, 152.4, 40.11, 0.001515] with condition number for $\mathbf{X}_{0}^{\prime} \mathbf{X}_{0}$ as $\lambda_{\max } / \lambda_{\min }=0.821210^{7}$. Scaling $\mathbf{X}_{0} \rightarrow \mathbf{X}_{u}$ to have unit length reduces the eigenvalues for $\mathbf{X}_{u}^{\prime} \mathbf{X}_{u}$ to $[2.898,1.007,0.09505$, 0.00003872 ] with condition number for $\mathbf{X}_{u}^{\prime} \mathbf{X}_{u}$ as $0.748610^{5}$. Centering and rescaling $\mathbf{X}_{u} \rightarrow \mathbf{X}_{c}$ into correlation form without the constant term reduces the eigenvalues for $\mathbf{X}_{c}^{\prime} \mathbf{X}_{c}$ to $[2.167,0.8301,0.002898]$ with condition number for $\mathbf{X}_{c}^{\prime} \mathbf{X}_{c}$ as $0.747710^{3}$. The correlation between column 1 of $\mathbf{X}_{c(1)}$ and column 2 of $\mathbf{X}_{c(2)}$ is $\rho\left(\mathbf{X}_{c(1)}, \mathbf{X}_{c(2)}\right)=0.9956$. The variance inflation factors $V I F_{j}$ for $\mathbf{X}_{c}$ are the diagonal elements of $\left(\mathbf{X}_{c}^{\prime} \mathbf{X}_{c}\right)^{-1}$ and are [169.4, 175.7, 1.688] from which the angles in radians between the column vector $\mathbf{X}_{c(j)}$ and the span of the remaining columns $S p\left(\mathbf{X}_{c[j]}\right)$ can be computed by $\theta_{j}=\arccos \left(\sqrt{1-1 / V I F_{j}}\right)$, for example Jensen and Ramirez (2013). Converting to degree measure, the angles are $\left[4.407^{\circ}, 4.327^{\circ}, 50.34^{\circ}\right]$ supporting that one should first try to improve collinearity by modifying the second column.

Following the case study in García et al. (2011), the responses are generated for the linear model

$$
\mathbf{y}=\mathbf{X}_{0}\left(\begin{array}{c}
10.0 \\
0.4 \\
-0.2 \\
0.4
\end{array}\right)+\mathbf{u}
$$

with $u_{i}$ generated from a normal $N(0,1)$ distribution. Using this data set, the centered and standardize model is

$$
\mathbf{y}-\bar{y} \mathbf{1}_{n}=\mathbf{X}_{c} \boldsymbol{\beta}_{c}+\mathbf{u}=\mathbf{X}_{c}\left(\begin{array}{c}
5.138 \\
-2.440 \\
2.683
\end{array}\right)+\mathbf{u}
$$

with $O L S$ estimates $\widehat{\boldsymbol{\beta}}=(14.48,-13.18,4.493)^{\prime}$. The raised column will be the second column of the centered and standardized $\mathbf{X}_{c}$ with $\tilde{\mathbf{x}}_{c 2}=\mathbf{x}_{c 2}+\lambda \mathbf{e}_{2}$ with $\lambda \geq 0$. The empirical risk for the $O L S$ estimator given the true value $\boldsymbol{\beta}_{c}$ is $(\widehat{\boldsymbol{\beta}}-$ $\left.\boldsymbol{\beta}_{c}\right)^{\prime}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{c}\right)=\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_{c}\right\|^{2}=205.79$. The goal is to improve, that is decrease, the empirical risk by using the raised estimators and its two components separately; namely, the stretched estimators and the rotated estimators.

Following the Stein procedure, we consider the stretched estimators based on the transformation $\mathbf{X}_{c(2)} \rightarrow(1+\lambda) \mathbf{X}_{c(2)}$; that is, the second column of $\mathbf{X}_{c}$ will be stretched by the factor $(1+\lambda)$ with $\lambda \geq 0$. Denote the stretched estimators by $\widehat{\boldsymbol{\beta}}_{s t}(\lambda)$. They are given by $\widehat{\boldsymbol{\beta}}_{s t}(\lambda)=\left(\widehat{\beta}_{1}, \widehat{\beta}_{2} /(1+\lambda), \widehat{\beta}_{3}\right)^{\prime}$ with $\left\|\widehat{\boldsymbol{\beta}}_{s t}(\lambda)\right\|^{2}$ monotonically decreasing to $\left\|\left(\widehat{\beta}_{1}, 0, \widehat{\beta}_{3}\right)^{\prime}\right\|^{2}$. For this case study, the
empirical risk for the stretched estimators $\left\|\widehat{\boldsymbol{\beta}}_{s t}(\lambda)-\boldsymbol{\beta}_{c}\right\|^{2}$ can be reduced from $\left\|\widehat{\boldsymbol{\beta}}_{s t}(0)-\boldsymbol{\beta}_{c}\right\|^{2}=205.79$ to $\min _{\lambda \geq 0}\left\|\widehat{\boldsymbol{\beta}}_{s t}(\lambda)-\boldsymbol{\beta}_{c}\right\|^{2}=90.64$ for $1+\lambda=5.402$, noting that $5.402=(-13.18) /(-2.440)$. Using the Jackknife procedure, the minimum for the average empirical risk for the stretched estimators over the $n=20$ replications is $\min _{\lambda \geq 0} \operatorname{avg}_{n=20}\left(\|\left(\widehat{\boldsymbol{\beta}}_{s t}(\lambda)-\boldsymbol{\beta}_{c} \|^{2}\right)=96.87\right.$ for $1+\lambda=5.58$. As this is a scaling procedure, the angle between $(1+\lambda) \mathbf{X}_{c(2)}$ and $S p\left(\mathbf{X}_{c[2]}\right)$ has not changed, and thus the VIFs have not changed.

The rotated estimators are similar to the raised estimators with the added constraint that the columns length remains at unit length. This allows one to track the improvement in ill-conditioning without confounding with a change in scale. The rotated estimators are based on the transformation $\mathbf{X}_{c(2)} \rightarrow$ $\left[\mathbf{X}_{c(2)}+\lambda \mathbf{e}_{2}\right] /\left\|\mathbf{X}_{c(2)}+\lambda \mathbf{e}_{2}\right\|$ with $\lambda \geq 0$; that is, the second column of $\mathbf{X}_{c}$ is raised, but then scaled back to unit length. For the same $\lambda$, both the rotated model and the raised model will have the same correlation matrix, the same angle between the second column of the design matrix and the span of the remaining columns, and the same variance inflation factors. Denote the rotated estimators by $\widehat{\boldsymbol{\beta}}_{r o t}(\lambda)$. As the second column is perturbed by a multiple of $\mathbf{e}_{2}$, and since $\mathbf{e}_{2}$ is orthogonal to the other columns, both the rotated estimators $\widehat{\boldsymbol{\beta}}_{\text {rot }}(\lambda)$ and the raised estimators $\boldsymbol{\beta}(\lambda)$ have the same values except in the row corresponding to the raised column. For this case study, the empirical risk for the rotated estimators $\left\|\widehat{\boldsymbol{\beta}}_{\text {rot }}(\lambda)-\boldsymbol{\beta}_{c}\right\|^{2}$ can be reduced from $\left\|\widehat{\boldsymbol{\beta}}_{\text {rot }}(0)-\boldsymbol{\beta}_{c}\right\|^{2}=$ 205.79 to $\min _{\lambda \geq 0}\left\|\widehat{\boldsymbol{\beta}}_{r o t}(\lambda)-\boldsymbol{\beta}_{c}\right\|^{2}=2.365$ for $\lambda=3.502$. For $\lambda=3.502$, the VIFs are $[9.394,9.618,1.121]$ with associated angle $\theta_{2}=18.81^{\circ}$. Using the Jackknife procedure, the minimum for the average empirical risk for the rotated estimators over the $n=20$ replications is $\min _{\lambda \geq 0} \operatorname{avg}_{n=20}\left(\left\|\widehat{\boldsymbol{\beta}}_{\text {rot }}(\lambda)-\boldsymbol{\beta}_{c}\right\|^{2}\right)$ $=2.950$ for $\lambda=3.501$.

For this case study, the empirical risk for the raised estimators $\| \boldsymbol{\beta}(\lambda)-$ $\boldsymbol{\beta}_{c} \|^{2}$ can be reduced from $\left\|\boldsymbol{\beta}(0)-\boldsymbol{\beta}_{c}\right\|^{2}=205.79$ to $\min _{\lambda \geq 0}\left\|\boldsymbol{\beta}(\lambda)-\boldsymbol{\beta}_{c}\right\|^{2}$ $=2.772$ for $\lambda=3.410$. For $\lambda=3.410$, VIFs are [9.744, 9.981, 1.122] with associated angle $\theta_{2}=18.45^{\circ}$. Using the Jackknife procedure, the minimum for the average empirical risk for the raised estimators over the $n=20$ replications is $\min _{\lambda \geq 0} a v g_{n=20}\left(\left\|\boldsymbol{\beta}(\lambda)-\boldsymbol{\beta}_{c}\right\|^{2}\right)=2.772$ for $\lambda=3.410$. These values are comparable to those using the rotated estimator.

Summarizing, the raised estimators are based on stretching and rotating the column to be improved. Both effects should be positive as shown with the stretching estimators and the rotating estimators. Computationally, the raised estimators are easier to compute lacking the square root in the denominator of the transformed columns. For this case study, the rotated and raised methods yielded comparable results and both performed better than the stretching method. The Jackknife procedure showed that the average over the replications for the raised estimators was smaller (better) than the average over the replications for the stretched estimators.

## 4 The MSE risk for two variables

We consider the linear model $\mathbf{y}=\mathbf{X} \boldsymbol{\beta}+\mathbf{u}$ where $E(\mathbf{u})=\mathbf{0}, E\left(\mathbf{u u}^{\prime}\right)=\sigma^{2} \mathbf{I}_{n}$ and $\mathbf{X}$ is a full rank matrix $n \times p$. In this section $p=2$. For convenience, we often take $\sigma^{2}=1$, however, for these calculations we need to track the value of $\sigma^{2}$. We assume that the variables are centered and standardized, that is, $\mathbf{X}^{\prime} \mathbf{X}$ is in correlation form. With respect to the MSE risk, Risk $(\widehat{\boldsymbol{\theta}})=$ $(E(\widehat{\boldsymbol{\theta}})-\boldsymbol{\theta})^{\prime}(E(\widehat{\boldsymbol{\theta}})-\boldsymbol{\theta})+\operatorname{tr}(\operatorname{var}(\widehat{\boldsymbol{\theta}}))$, we compare the three estimators: the raised $\boldsymbol{\beta}(\lambda)$, the stretched $\widehat{\boldsymbol{\beta}}_{s t}(\lambda)$, and the rotated $\widehat{\boldsymbol{\beta}}_{r o t}(\lambda)$.

For the raised model with column 1 raised with $\mathbf{X} \rightarrow \mathbf{X}_{<1>}$, the design matrix and moment matrix are given in Eqs. (7) and (8). The calculations follow from standard results for $O L S$ estimators and where we denote $E(\widehat{\boldsymbol{\beta}})=\boldsymbol{\delta}$. From Eq. (9) the predicted values using $O L S$ estimators and raised estimators are the same with $\widehat{\beta}_{1} \mathbf{x}_{1}+\widehat{\beta}_{2} \mathbf{x}_{2}=\mathbf{X} \widehat{\boldsymbol{\beta}}=\widetilde{\mathbf{X}} \boldsymbol{\beta}(\lambda)=\beta_{1}(\lambda) \widetilde{\mathbf{x}}_{1}+\beta_{2}(\lambda) \mathbf{x}_{2}=\beta_{1}(\lambda)[(1+$ $\left.\lambda) \mathbf{x}_{1}-\lambda \rho \mathbf{x}_{2}\right]+\beta_{2}(\lambda) \mathbf{x}_{2}=(1+\lambda) \beta_{1}(\lambda) \mathbf{x}_{1}+\left[\beta_{2}(\lambda)-\lambda \rho \beta_{1}(\lambda)\right] \mathbf{x}_{2}$ and thus

$$
\begin{aligned}
\beta_{1}(\lambda) & =\frac{1}{1+\lambda} \widehat{\beta}_{1} \\
\beta_{2}(\lambda) & =\widehat{\beta}_{2}+\frac{\lambda \rho}{1+\lambda} \widehat{\beta}_{1} .
\end{aligned}
$$

A more general result is given in García et al. (2011). It follows that $E\left(\beta_{1}(\lambda)\right)=$ $\frac{1}{1+\lambda} \delta_{1}, E\left(\beta_{2}(\lambda)\right)=\delta_{2}+\frac{\lambda \rho}{1+\lambda} \delta_{1}$, with squared bias $(E(\boldsymbol{\beta}(\lambda))-\boldsymbol{\delta})^{\prime}(E(\boldsymbol{\beta}(\lambda))-$ $\boldsymbol{\delta}))=\frac{\lambda^{2}\left(1+\rho^{2}\right)}{(1+\lambda)^{2}} \delta_{1}^{2}$. The variances are given by

$$
\begin{aligned}
\frac{\operatorname{var}\left(\widehat{\beta}_{1}(\lambda)\right)}{\sigma^{2}} & =\left(\frac{1}{1+\lambda}\right)^{2}\left(\frac{1}{1-\rho^{2}}\right) \\
\frac{\operatorname{var}\left(\widehat{\beta}_{2}(\lambda)\right)}{\sigma^{2}} & =\left(\frac{1}{1-\rho^{2}}\right)+2\left(\frac{\lambda \rho}{1+\lambda}\right)\left(\frac{-\rho}{1-\rho^{2}}\right)+\left(\frac{\lambda \rho}{1+\lambda}\right)^{2}\left(\frac{1}{1-\rho^{2}}\right)
\end{aligned}
$$

The minimum value for $\operatorname{Risk}(\boldsymbol{\beta}(\lambda))$ can be shown to be

$$
\widetilde{\lambda}=\frac{\sigma^{2}}{\left(1-\rho^{2}\right) \delta_{1}^{2}}
$$

For the stretched estimator $\widehat{\boldsymbol{\beta}}_{s t}(\lambda)$,

$$
\begin{aligned}
\widehat{\beta}_{s t, 1}(\lambda) & =\frac{1}{1+\lambda} \widehat{\beta}_{1} \\
\widehat{\beta}_{s t, 2}(\lambda) & =\widehat{\beta}_{2}
\end{aligned}
$$

with $E\left(\widehat{\beta}_{s t, 1}(\lambda)\right)=\frac{1}{1+\lambda} \delta_{1}, E\left(\widehat{\beta}_{s t, 2}(\lambda)\right)=\delta_{2}$, with squared bias $\left(E\left(\widehat{\boldsymbol{\beta}}_{s t}(\lambda)\right)-\right.$
$\left.\boldsymbol{\delta})^{\prime}\left(E\left(\widehat{\boldsymbol{\beta}}_{s t}(\lambda)\right)-\boldsymbol{\delta}\right)\right)=\frac{\lambda^{2}}{(1+\lambda)^{2}} \delta_{1}^{2}$, and with variances

$$
\begin{aligned}
& \frac{\operatorname{var}\left(\widehat{\beta}_{s t, 1}(\lambda)\right)}{\sigma^{2}}=\left(\frac{1}{1+\lambda}\right)^{2}\left(\frac{1}{1-\rho^{2}}\right) \\
& \frac{\operatorname{var}\left(\widehat{\beta}_{s t, 2}(\lambda)\right)}{\sigma^{2}}=\left(\frac{1}{1-\rho^{2}}\right)
\end{aligned}
$$

with the minimum value for $\operatorname{Risk}\left(\widehat{\boldsymbol{\beta}}_{s t}(\lambda)\right)$ when

$$
\lambda_{s t}=\frac{\sigma^{2}}{\left(1-\rho^{2}\right) \delta_{1}^{2}}
$$

We note that this is the same $\lambda$ value as with the raised estimator.
For the rotated estimator $\widehat{\boldsymbol{\beta}}_{r o t}(\lambda)$ the calculations are more tedious. The design matrix is $\mathbf{X}_{<1>\text { rot }}=\left(\left(\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right) /\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|, \mathbf{x}_{2}\right)=\left(\left((1+\lambda) \mathbf{x}_{1}-\lambda \rho \mathbf{x}_{2}\right) / \| \mathbf{x}_{1}+\right.$ $\left.\lambda \mathbf{e}_{1} \|, \mathbf{x}_{2}\right)$ with moment matrix

$$
\mathbf{X}_{<1>r o t}^{\prime} \mathbf{X}_{<1>r o t}=\left(\begin{array}{cc}
1 & \frac{\rho}{\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|} \\
\frac{\rho}{\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|} & 1
\end{array}\right)
$$

recalling that $\mathbf{e}_{1} \perp \mathbf{x}_{2}$. As the predicted responses are the same for both the OLS model and the rotated model, $\widehat{\beta}_{1} \mathbf{x}_{1}+\widehat{\beta}_{2} \mathbf{x}_{2}=\mathbf{X} \widehat{\boldsymbol{\beta}}=\mathbf{X}_{<1>r o t} \widehat{\boldsymbol{\beta}}_{r o t}(\lambda)=$ $\widehat{\beta}_{r o t, 1}(\lambda) \widetilde{\mathbf{x}}_{1} /\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|+\widehat{\beta}_{r o t, 2}(\lambda) \mathbf{x}_{2}=\widehat{\beta}_{r o t, 1}(\lambda)\left[\left((1+\lambda) \mathbf{x}_{1}-\lambda \rho \mathbf{x}_{2}\right) /\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|\right]+$ $\widehat{\beta}_{r o t, 2}(\lambda) \mathbf{x}_{2}=\frac{1+\lambda}{\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|} \widehat{\beta}_{r o t, 1}(\lambda) \mathbf{x}_{1}+\left[\widehat{\beta}_{r o t, 2}(\lambda)-\frac{\lambda \rho}{\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|} \widehat{\beta}_{r o t, 1}(\lambda)\right] \mathbf{x}_{2}$. And thus

$$
\begin{aligned}
& \widehat{\beta}_{r o t, 1}(\lambda)=\frac{\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|}{1+\lambda} \widehat{\beta}_{1} \\
& \widehat{\beta}_{r o t, 2}(\lambda)=\widehat{\beta}_{2}+\frac{\lambda \rho}{(1+\lambda)} \widehat{\beta}_{1},
\end{aligned}
$$

with $E\left(\left(\widehat{\beta}_{r o t, 1}\right)\right)=\frac{\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|}{1+\lambda} \delta_{1}, E\left(\widehat{\beta}_{r o t, 2}(\lambda)\right)=\delta_{2}+\frac{\lambda \rho}{1+\lambda} \delta_{1}$, with squared bias $\left(E\left(\widehat{\boldsymbol{\beta}}_{r o t}(\lambda)\right)-\boldsymbol{\delta}\right)^{\prime}\left(E\left(\widehat{\boldsymbol{\beta}}_{r o t}(\lambda)\right)-\boldsymbol{\delta}\right)=\left(1-\frac{\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|}{1+\lambda}\right)^{2} \delta_{1}^{2}+\left(\frac{\lambda \rho}{1+\lambda}\right)^{2} \delta_{1}^{2}$, and with variances

$$
\begin{aligned}
& \frac{\operatorname{var}\left(\widehat{\beta}_{r o t, 1}(\lambda)\right)}{\sigma^{2}}=\left(\frac{\left\|\mathbf{x}_{1}+\lambda \mathbf{e}_{1}\right\|}{1+\lambda}\right)^{2}\left(\frac{1}{1-\rho^{2}}\right) \\
& \frac{\operatorname{var}\left(\widehat{\beta}_{r o t, 2}(\lambda)\right)}{\sigma^{2}}=\left(\frac{1}{1-\rho^{2}}\right)+2\left(\frac{\lambda \rho}{1+\lambda}\right)\left(\frac{-\rho}{1-\rho^{2}}\right)+\left(\frac{\lambda \rho}{1+\lambda}\right)^{2}\left(\frac{1}{1-\rho^{2}}\right) .
\end{aligned}
$$

The value of $\lambda_{\text {rot }}$ for the minimum of $\operatorname{Risk}\left(\widehat{\boldsymbol{\beta}}_{\text {rot }}(\lambda)\right)$ was found using Maple. First one solves for the positive root $t_{0}$ of the quartic polynomial $p_{4}(t)=a_{0} t^{4}+$ $a_{1} t^{3}+a_{2} t^{2}+a_{3} t+a_{4}$ with $a_{0}=1-\rho^{2}, a_{1}=2\left(1-\rho^{2}\right), a_{2}=\left(1-\rho^{2}\right)^{2}-4(1-$ $\left.\rho^{2}+\left(\frac{\sigma}{\delta_{1}}\right)^{2}\right)^{2}, a_{3}=-2 \rho^{2}\left(1-\rho^{2}\right), a_{4}=-\rho^{2}\left(1-\rho^{2}\right)$. Appendix B establishes that
real roots exist for $p_{4}(t)$. The value of $\lambda_{\text {rot }}$ is then given by

$$
\lambda_{\text {rot }}=\frac{\left(t_{0}-1\right)\left(1-\rho^{2}\right)+2 t_{0}\left(\frac{\sigma}{\delta_{1}}\right)^{2}}{\left(t_{0}+1\right)\left(1-\rho^{2}\right)}
$$

As an example, with $\left\{\rho=0.95, \delta_{1}=2, \sigma=1\right\}$ :
(1) for the raised estimator the $\operatorname{Risk}(\boldsymbol{\beta}(0))=20.51$ and decreases to $\operatorname{Risk}(\boldsymbol{\beta}(\widetilde{\lambda}))=6.475$ with $\widetilde{\lambda}=2.564$
(2) for the stretched estimator, the risk decreases to $\operatorname{Risk}\left(\widehat{\boldsymbol{\beta}}_{s t}(\lambda)\right)=13.13$ with $\lambda_{s t}=2.564$ and
(3) for the rotated estimator $p_{4}(t)=0.0975 t^{4}+0.1950 t^{3}-0.4735 t^{2}-0.1760 t-$ 0.08799 with positive root $t_{0}=1.691, \lambda_{\text {rot }}=3.479$, and $\operatorname{Risk}\left(\widehat{\boldsymbol{\beta}}_{s t}\left(\lambda_{r o t}\right)\right)=$ 6.651 .

In summary, the raised and stretched estimators have their minimum risk at the same $\lambda$ value; the raised and rotated estimators have comparable results; and the rotated estimators are more complex to compute.

## 5 Variance inflation factors and the metric number

We consider the linear model $\mathbf{y}=\mathbf{X} \boldsymbol{\beta}+\mathbf{u}$ where $E(\mathbf{u})=\mathbf{0}, E\left(\mathbf{u u}^{\prime}\right)=\sigma^{2} \mathbf{I}_{n}$ and $\mathbf{X}$ is a full rank matrix $n \times p$. We assume that the variables are centered. Reorder $\mathbf{X}=\left[\mathbf{X}_{[p]}, \mathbf{X}_{(p)}\right]$ with $\mathbf{X}_{(p)}=\mathbf{x}_{p}$ the $p^{t h}$ column and $\mathbf{X}_{[p]}$, the design matrix $\mathbf{X}$ without the $p^{t h}$ column, the resting columns. The problem in this section is to measure the effect of adding the last column $\mathbf{X}_{(p)}$ to $\mathbf{X}_{[p]}$. An ideal column would be orthogonal to the previous columns with the entries in the off diagonal elements of the $p^{t h}$ row and $p^{t h}$ column of $\mathbf{X}^{\prime} \mathbf{X}$ all zeros. Denote by $\mathbf{M}_{p}$ the idealized moment matrix

$$
\mathbf{M}_{p}=\left[\begin{array}{cc}
\mathbf{X}_{[p]}^{\prime} \mathbf{X}_{[p]} & \mathbf{0} \mathbf{1}_{p-1} \\
\mathbf{0} \mathbf{1}_{p-1}^{\prime} & \mathbf{x}_{p}^{\prime} \mathbf{x}_{p}
\end{array}\right]
$$

The metric number associated to $\mathbf{x}_{p}$ is defined by

$$
M N\left(\mathbf{x}_{p}\right)=\sqrt{\frac{\operatorname{det}\left(\mathbf{X}^{\prime} \mathbf{X}\right)}{\operatorname{det}\left(\mathbf{M}_{p}\right)}}
$$

The metric number is easy to compute and has been used in García et al. (2011) as a measure of collinearity. A similar measure of collinearity is mentioned in Footnote 2 in Wichers (1975) and Theorem 1 of Berk (1977). The geometry for the metric number has been shown in García et al. (1999). The case study in García et al. (2011) suggests the functional relationship between $M N\left(\mathbf{x}_{p}\right)$ and the variance inflation factor for $\widehat{\beta}_{p}$ as

$$
\begin{equation*}
\operatorname{VIF}\left(\widehat{\beta}_{p}\right)=\frac{1}{M N\left(\mathbf{x}_{p}\right)^{2}} \tag{10}
\end{equation*}
$$

We note that this relationship holds, and so the metric number $M N\left(\mathbf{x}_{j}\right)$ is also functionally equivalent to the collinearity indices $\kappa_{j}$ of Stewart (1987), the variance inflation factors $V I F_{j}$, and the angle between $\left(\mathbf{X}_{(j)}, S p\left(\mathbf{X}_{[j]}\right)\right)$.

To evaluate $\operatorname{VIF}\left(\widehat{\beta}_{p}\right)=V I F_{j}$, transform $\mathbf{X}^{\prime} \mathbf{X}$ into correlation form $\mathbf{R}$ with $\mathbf{S}=\operatorname{diag}\left(\mathbf{x}_{1}^{\prime} \mathbf{x}_{1}, \ldots, \mathbf{x}_{p}^{\prime} \mathbf{x}_{p}\right)$ the diagonal matrix with entries from the diagonal of $\mathbf{X}^{\prime} \mathbf{X}$ with $\mathbf{R}=\mathbf{S}^{-1 / 2}\left(\mathbf{X}^{\prime} \mathbf{X}\right) \mathbf{S}^{-1 / 2}$. The VIFs are the diagonal entries of $\mathbf{R}^{-1}=\mathbf{S}^{1 / 2}\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1} \mathbf{S}^{1 / 2}$. It remains to note that the inverse $\mathbf{R}^{-1}$ can be computed using cofactors $\mathbf{C}_{i, j}$, and, in particular,

$$
\begin{aligned}
\operatorname{VIF}\left(\widehat{\beta}_{p}\right) & =\mathbf{R}_{p, p}^{-1}=\left[\mathbf{S}^{1 / 2}\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1} \mathbf{S}^{1 / 2}\right]_{p, p} \\
& =\left(\mathbf{x}_{p}^{\prime} \mathbf{x}_{p}\right)^{1 / 2} \frac{\operatorname{det}\left(\mathbf{C}_{p, p}\right)}{\operatorname{det}\left(\mathbf{X}^{\prime} \mathbf{X}\right)}\left(\mathbf{x}_{p}^{\prime} \mathbf{x}_{p}\right)^{1 / 2} \\
& =\frac{\operatorname{det}\left(\mathbf{M}_{p}\right)}{\operatorname{det}\left(\mathbf{X}^{\prime} \mathbf{X}\right)}=\frac{1}{M N\left(\mathbf{x}_{p}\right)^{2}}
\end{aligned}
$$

With the example in the previous section with $\{p=2, \rho=0.95\}, V I F_{2}$ is easy to compute from Eq. (10) with:
(1) for the raised estimator $\operatorname{VIF}\left(\beta_{2}(0)\right)=10.26$ and decreases to $\operatorname{VIF}\left(\beta_{2}(\widetilde{\lambda})\right)=1.729$ with $\widetilde{\lambda}=2.564$
(2) for the stretched estimator, $\operatorname{VIF}\left(\widehat{\beta}_{s t, 2}\left(\lambda_{s t}\right)\right)=10.26$ for any $\lambda_{s t} \in(0, \infty)$ since there is no change in the angle between $\left(\mathbf{X}_{(2)}, S p\left(\mathbf{X}_{[2]}\right)\right)$
(3) for the rotated estimator $\operatorname{VIF}\left(\widehat{\beta}_{\text {rot }, 2}\left(\lambda_{\text {rot }}\right)\right)=1.080$ for $\lambda_{\text {rot }}=3.479$, noting that VIF for the rotated estimator is smaller than the raised estimator since the rotated estimator has the larger angle between $\left(\mathbf{X}_{(2)}, S p\left(\mathbf{X}_{[2]}\right)\right)$ as $\lambda_{\text {rot }}>\widetilde{\lambda}$.

## 6 How to overcome the ill-conditioned of the matrix $X^{\prime} \mathbf{X}$ by successive raises

We will focus on finding a natural way to get the perturbation matrix $\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{I}_{p}$ from the raised estimators as this matrix is an essential component in ridge regression. We will show that the raised estimators allow an estimation by $O L S$ which retains the coefficient of determination. Firstly, we will present the estimation for two variables and, secondly, to the general case.

### 6.1 For two variables

In the case of two variables, $\mathbf{X}=\left[\mathbf{x}_{1}, \mathbf{x}_{2}\right]$ with the first column raised, $\mathbf{X} \rightarrow \mathbf{X}<1>$ with the raised design matrix given by $\mathbf{X}_{<1>}=\left(\widetilde{\mathbf{x}}_{1}, \mathbf{x}_{2}\right)=\left(\left(1+\lambda_{1}\right) \mathbf{x}_{1}-\right.$ $\left.\lambda_{1} \rho \mathbf{x}_{2}, \mathbf{x}_{2}\right)$ and

$$
\mathbf{X}_{<1>}^{\prime} \mathbf{X}_{<1>}=\left(\begin{array}{cc}
1+2 \lambda_{1}\left(1-\rho^{2}\right)+\lambda_{1}^{2}\left(1-\rho^{2}\right) & \rho \\
\rho & 1
\end{array}\right)
$$

After having raised column 1 with $\lambda_{1}>0$, we will raise column 2 with $\lambda_{2}>0$. First regress the second column of $\mathbf{X}_{<1>}$ using the remaining column(s) as $\mathbf{X}_{<1>(2)}=\alpha \mathbf{X}_{<1>[2]}+\mathbf{v}$; that is, $\mathbf{x}_{2}=\alpha \widetilde{\mathbf{x}}_{1}+\mathbf{v}$. Scaling $\widetilde{\mathbf{x}}_{1}$ to unit length gives $\mathbf{x}_{2}=\left(\alpha\left\|\widetilde{\mathbf{x}}_{1}\right\|\right)\left(\widetilde{\mathbf{x}}_{1} /\left\|\widetilde{\mathbf{x}}_{1}\right\|\right)+\mathbf{v}$ with $\left(\alpha\left\|\widetilde{\mathbf{x}}_{1}\right\|\right)=\rho$ as in Eq. (5); and with $\left\|\widetilde{\mathbf{x}}_{1}\right\|^{2}$ $=\left(1+\lambda_{1}\right)^{2}-\rho^{2}\left(\lambda_{1}^{2}+2 \lambda_{1}\right)$ from Eq. (6). Thus

$$
\widehat{\mathbf{x}}_{2}=\frac{\rho}{\left\|\widetilde{\mathbf{x}}_{1}\right\|} \widetilde{\mathbf{x}}_{1}=\frac{\rho}{\left\|\widetilde{\mathbf{x}}_{1}\right\|}\left(\left(1+\lambda_{1}\right) \mathbf{x}_{1}-\lambda_{1} \rho \mathbf{x}_{2}\right)
$$

with residual $\mathbf{e}_{2}$ from $\mathbf{x}_{2}=\hat{\mathbf{x}}_{2}+\mathbf{e}_{2}=\frac{\rho \tilde{\mathbf{x}}_{1}}{\left\|\widetilde{\mathbf{x}}_{1}\right\|}+\mathbf{e}_{2}$ so $\mathbf{e}_{2}=\mathbf{x}_{2}-\rho \tilde{\mathbf{x}}_{1} /\left\|\widetilde{\mathbf{x}}_{1}\right\|$ with $\mathbf{e}_{2} \perp \tilde{\mathbf{x}}_{1}$.

We now raise the vector $\mathbf{x}_{2}$ using the residual $\mathbf{e}_{2}$ by the rule

$$
\tilde{\mathbf{x}}_{2}=\mathbf{x}_{2}+\lambda_{2} \mathbf{e}_{2}=\mathbf{x}_{2}+\lambda_{2}\left(\mathbf{x}_{2}-\frac{\rho \tilde{\mathbf{x}}_{1}}{\left\|\widetilde{\mathbf{x}}_{1}\right\|}\right)=\left(1+\lambda_{2}\right) \mathbf{x}_{2}-\frac{\lambda_{2} \rho \tilde{\mathbf{x}}_{1}}{\left\|\widetilde{\mathbf{x}}_{1}\right\|}
$$

The doubly raised matrix is given by

$$
\mathbf{X}_{<1,2>}=\left(\begin{array}{cc}
\left(1+\lambda_{1}\right) \mathbf{x}_{1}-\lambda_{1} \rho \mathbf{x}_{2} \quad\left(1+\lambda_{2}\right) \mathbf{x}_{2}-\frac{\lambda_{2} \rho \tilde{\mathbf{x}}_{1}}{\left\|\tilde{\mathbf{x}}_{1}\right\|}
\end{array}\right) .
$$

For any values $k_{1}>0, k_{2}>0$, it is possible to show that there exist raising parameters $\lambda_{1}>0, \lambda_{2}>0$ such that

$$
\mathbf{X}_{<1,2>}^{\prime} \mathbf{X}_{<1,2>}=\left(\begin{array}{cc}
1+k_{1} & \rho  \tag{11}\\
\rho & 1+k_{2}
\end{array}\right)
$$

the matrix used in generalized ridge regression. This follows from noting that $\left\|\tilde{\mathbf{x}}_{1}\right\|^{2}=1+2\left(1-\rho^{2}\right) \lambda_{1}+\left(1-\rho^{2}\right) \lambda_{1}^{2}$ viewed as a polynomial in $\lambda_{1}$ has positive coefficients so it can achieve any value greater than one; and similarly for $\left\|\widetilde{\mathbf{x}}_{2}\right\|^{2}=$ $1+2\left(1-\frac{\rho^{2}}{\sqrt{1+k_{1}}}\right) \lambda_{2}+\left(1-\rho^{2}\left(\frac{2}{\sqrt{1+k_{1}}}-1\right)\right) \lambda_{2}^{2}$. The off-diagonal entries in Eq. (11) are not effected by raising column 2 since $\tilde{\mathbf{x}}_{1}^{\prime} \tilde{\mathbf{x}}_{2}=\tilde{\mathbf{x}}_{1}^{\prime}\left(\mathbf{x}_{2}+\lambda_{2} \mathbf{e}_{2}\right)=\tilde{\mathbf{x}}_{1}^{\prime} \mathbf{x}_{2}$ since $\mathbf{e}_{2} \perp \widetilde{\mathbf{x}}_{1}$; and earlier we had established that $\tilde{\mathbf{x}}_{1}^{\prime} \mathbf{x}_{2}=\left(\mathbf{x}_{1}+\lambda_{1} \mathbf{e}_{1}\right)^{\prime} \mathbf{x}_{2}=\mathbf{x}_{1}^{\prime} \mathbf{x}_{2}=\rho$ since $\mathbf{e}_{1} \perp \mathbf{x}_{2}$. Thus, in particular, given $k>0$ there are raising parameters $\lambda_{1}$ and $\lambda_{2}$ such that

$$
\mathbf{X}_{<1,2>}^{\prime} \mathbf{X}_{<1,2>}=\left(\begin{array}{cc}
1+k & \rho \\
\rho & 1+k
\end{array}\right)=\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{I}_{p}
$$

the perturbation matrix usually used in ridge regression.
Some advantages of the raise method for addressing the problem of collinearity are:

- The raise estimators are estimated by $O L S$ and thus confidence intervals can be computed.
- It retains the coefficient of determination of the initial regression.
- The VIFs associated with the raise estimators are monotone functions decreasing with $k$, see García et al. (2015).


### 6.2 For the general case

In the general case we start with a standardized matrix $\mathbf{X}$ with the following column vectors $\mathbf{X}=\left(\mathbf{x}_{1}, \mathbf{x}_{2}, \ldots, \mathbf{x}_{p}\right)$, and the raising procedure will be given by the following steps:

- Step 1: Firstly, we raise the vector $\mathbf{x}_{1}$ from the regression of $\mathbf{x}_{1}$ with the resting vectors $\mathbf{X}_{[1]}=\left(\mathbf{x}_{2}, \mathbf{x}_{3}, \ldots \mathbf{x}_{p}\right)$. For this, we take the residual $\mathbf{e}_{1}$ from $\tilde{\mathbf{x}}_{1}=\mathbf{x}_{1}+\lambda_{1} \mathbf{e}_{1}$ such that $\mathbf{e}_{1} \perp S p\left(\mathbf{X}_{[1]}\right)$. The raised design matrix is denoted $\mathbf{X}_{<1>}=\left(\widetilde{\mathbf{x}}_{1}, \mathbf{x}_{2}, \ldots, \mathbf{x}_{p}\right)$.
- Step 2: Next, we raise the vector $\mathbf{x}_{2}$ from the regression of $\mathbf{x}_{2}$ with the resting vectors from $\mathbf{X}_{<1>}$, namely $\mathbf{X}_{<1>[2]}=\left(\tilde{\mathbf{x}}_{1}, \mathbf{x}_{3}, \ldots, \mathbf{x}_{p}\right)$. Then, we get the residual $\mathbf{e}_{2}$ from $\tilde{\mathbf{x}}_{2}=\mathbf{x}_{2}+\lambda_{2} \mathbf{e}_{2}$ such that $\mathbf{e}_{2} \perp S p\left(\mathbf{X}_{<1>[2]}\right)$. The raised design matrix is denoted $\mathbf{X}_{<1,2>}=\left(\widetilde{\mathbf{x}}_{1}, \tilde{\mathbf{x}}_{2}, \ldots, \mathbf{x}_{p}\right)$.
- Step $j$ : We raise the vector $\mathbf{x}_{j}$ from the regression of $\mathbf{x}_{j}$ with the resting vectors from $\mathbf{X}_{<1, \ldots, j-1>}$, namely $\mathbf{X}_{<1, \ldots, j-1>[j]}=\left(\tilde{\mathbf{x}}_{1}, \ldots, \tilde{\mathbf{x}}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_{p}\right)$. Then, we get the residual $\mathbf{e}_{j}$ from $\tilde{\mathbf{x}}_{j}=\mathbf{x}_{j}+\lambda_{j} \mathbf{e}_{j}$ such that $\mathbf{e}_{j} \perp S p\left(\mathbf{X}_{<1, \ldots, j-1>[j]}\right)$. The raised design matrix is denoted $\mathbf{X}_{<1, \ldots, j>}=$ $\left(\tilde{\mathbf{x}}_{1}, \ldots, \tilde{\mathbf{x}}_{j}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_{p}\right)$.
- Step $p$ : We raise the vector $\mathbf{x}_{p}$ from the regression of $\mathbf{x}_{p}$ with the resting vectors from $\mathbf{X}_{<1, \ldots, p-1>}$, namely $\mathbf{X}_{<1, \ldots, p-1>[p]}=\left(\tilde{\mathbf{x}}_{1}, \ldots, \tilde{\mathbf{x}}_{p-1}\right)$. Then, we get the residual $\mathbf{e}_{p}$ from $\tilde{\mathbf{x}}_{p}=\mathbf{x}_{p}+\lambda_{p} \mathbf{e}_{p}$ such that $\mathbf{e}_{p} \perp \operatorname{Sp}\left(\mathbf{X}_{<1, \ldots, p-1>[p]}\right)$. The raised design matrix is denoted $\mathbf{X}_{<1, \ldots, p>}=\left(\tilde{\mathbf{x}}_{1}, \ldots, \tilde{\mathbf{x}}_{p}\right)$.

Suppose columns $1, \ldots, j-1$ have been raised to form $\mathbf{X}_{<1, \ldots, j-1>}$. The $j^{\text {th }}$ column of $\mathbf{X}_{<1, \ldots, j-1>}$ is raised to define $\mathbf{X}_{<1, \ldots, j>}$, with $\mathbf{X}_{<1, \ldots, j-1>[j]}$ the resting columns, by the rule $\mathbf{X}_{<1, \ldots, j>(j)}=\mathbf{X}_{<1, \ldots, j-1>(j)}+\lambda_{j} \mathbf{e}_{j}$ with $\mathbf{e}_{j} \perp S p\left(\mathbf{X}_{<1, \ldots, j-1>[j]}\right)$. Noted that $\mathbf{X}_{<1, \ldots, j>}^{\prime} \mathbf{X}_{<1, \ldots, j>}$ will differ from $\mathbf{X}_{<1, \ldots, j-1>}^{\prime} \mathbf{X}_{<1, \ldots, j-1>}$ only in the $(j, j)$ entry as the $j^{t h}$ raised column is perturbed by the vector $\lambda_{j} \mathbf{e}_{j}$ orthogonal to the span of the remaining columns. The $[j, j]$ entry for the $j^{t h}$ raised moment matrix is

$$
\begin{aligned}
\mathbf{X}_{<1, \ldots, j>}^{\prime} \mathbf{X}_{<1, \ldots, j>}[j, j]= & \mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime} \mathbf{X}_{<1, \ldots, j-1>(j)}+ \\
& 2 \lambda_{j} \mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime} \mathbf{e}_{j}+\lambda_{j}^{2} \mathbf{e}_{j}^{\prime} \mathbf{e}_{j}
\end{aligned}
$$

with the surplus term denoted by $k_{j}=2 \lambda_{j} \mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime} \mathbf{e}_{j}+\lambda_{j}^{2} \mathbf{e}_{j}^{\prime} \mathbf{e}_{j}$. We view $\mathbf{X}_{<1, \ldots, j>}^{\prime} \mathbf{X}_{<1, \ldots, j\rangle}[j, j]$ as a quadratic polynomial in $\lambda_{j}$; and since the coefficients of $\lambda_{j}^{2}$ and $\lambda_{j}$ are both positive, the quadratic polynomial can achieve any value larger than the original value $\mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime} \mathbf{X}_{<1, \ldots, j-1>(j)}$. Let $\mathbf{P}_{[j]}$ be the projection matrix for $\operatorname{Sp}\left(\mathbf{X}_{<1, \ldots, j-1>[j]}\right)$. That the coefficient of $\lambda_{j}$ is positive follows from

$$
\begin{align*}
\mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime} \mathbf{e}_{j} & =\mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime}\left(\mathbf{I}_{p}-\mathbf{P}_{[j]}\right) \mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime}  \tag{12}\\
& =\mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime}\left(\mathbf{I}_{p}-\mathbf{P}_{[j]}\right)^{2} \mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime} \\
& =\left\|\left(\mathbf{I}_{p}-\mathbf{P}_{[j]}\right) \mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime}\right\|^{2}=\left\|\mathbf{e}_{j}\right\|^{2}>0
\end{align*}
$$

We started with $\mathbf{X}^{\prime} \mathbf{X}$ in correlation form and with final raising matrix having moment matrix

$$
\begin{aligned}
\mathbf{X}_{<1,2, \ldots, p>}^{\prime} \mathbf{X}_{<1,2, \ldots, p>} & =\left(\begin{array}{ccccc}
1+k_{1} & \rho_{12} & \rho_{12} & \ldots & \rho_{1 p} \\
\rho_{12} & 1+k_{2} & \rho_{23} & \ldots & \rho_{2 p} \\
\rho_{13} & \rho_{23} & 1+k_{3} & \ldots & \rho_{3 p} \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
\rho_{1 p} & \rho_{2 p} & \rho_{3 p} & \ldots & 1+k_{p}
\end{array}\right) \\
& =\mathbf{X}^{\prime} \mathbf{X}+\mathbf{K},
\end{aligned}
$$

with $\mathbf{K}$ the diagonal matrix $\mathbf{K}=\operatorname{diag}\left(k_{1}, \ldots, k_{p}\right)$ with $k_{j} \geq 0$ for $1 \leq j \leq p$. Thus the raised regression perturbation matrix is equivalent to a generalized ridge regression perturbation matrix. And conversely, any generalized ridge regression matrix has a corresponding raised regression matrix.

Given ridge parameters $\left(k_{1}, \ldots, k_{p}\right)$, the associated raised parameters are given from Eq. (12) by $k_{j}=2 \lambda_{j} \mathbf{X}_{<1, \ldots, j-1>(j)}^{\prime} \mathbf{e}_{j}+\lambda_{j}^{2} \mathbf{e}_{j}^{\prime} \mathbf{e}_{j}=2 \lambda_{j}\left\|\mathbf{e}_{j}\right\|^{2}+$ $\lambda_{j}^{2}\left\|\mathbf{e}_{j}\right\|^{2}=\left(2 \lambda_{j}+\lambda_{j}^{2}\right)\left\|\mathbf{e}_{j}\right\|^{2}$, and thus

$$
\begin{align*}
\lambda_{j} & =\sqrt{1+\frac{k_{j}}{\left\|\mathbf{e}_{j}\right\|^{2}}}-1  \tag{13}\\
k_{j} & =\left\|\mathbf{e}_{j}\right\|^{2}\left(\lambda_{j}^{2}+2 \lambda_{j}\right) \tag{14}
\end{align*}
$$

## 7 Empirical application

We have chosen the following data set for an empirical application because it is relatively recent and presents a high level of collinearity. In the course of the financial crisis in the United States and over the whole world there is a big discussion about the life of the American people on credit. The original data set is presented in Table 1 and it is taken from the Economic Report of the President (2008) where $\mathbf{y}$ is the outstanding Mortgage Debt (in trillions of dollars) and the three independent variables are $\mathbf{X}_{1}$ Personal Consumption (in trillions of dollars), $\mathbf{X}_{2}$ Personal Income (in trillions of dollars), and $\mathbf{X}_{3}$ Consumer credit outstanding (in trillions of dollars).

Table 1. $\mathbf{y}$ Mortgage debt, $\mathbf{X}_{1}$ Personal Consumption, $\mathbf{X}_{2}$ Personal Income, and $\mathbf{X}_{3}$ Consumer Credit.

| year | $\mathbf{y}$ | $\mathbf{X}_{1}$ | $\mathbf{X}_{2}$ | $\mathbf{X}_{3}$ |
| :---: | :---: | :---: | :---: | :---: |
| 1990 | 3.8051 | 4.7703 | 4.8786 | 808.23 |
| 1991 | 3.9458 | 4.7784 | 5.0510 | 798.03 |
| 1992 | 4.0579 | 4.9348 | 5.3620 | 806.12 |
| 1993 | 4.1913 | 5.0998 | 5.5585 | 865.65 |
| 1994 | 4.3585 | 5.2907 | 5.8425 | 997.30 |
| 1995 | 4.5453 | 5.4335 | 6.1523 | 1140.7 |
| 1996 | 4.8149 | 5.6194 | 6.5206 | 1253.4 |
| 1997 | 5.1286 | 5.8318 | 6.9151 | 1324.8 |
| 1998 | 5.6151 | 6.1258 | 7.4230 | 1420.5 |
| 1999 | 6.2249 | 6.4386 | 7.8024 | 1532.1 |
| 2000 | 6.7864 | 6.7394 | 8.4297 | 1717.5 |
| 2001 | 7.4944 | 6.9104 | 8.7241 | 1867.2 |
| 2002 | 8.3993 | 7.0993 | 8.8819 | 1974.1 |
| 2003 | 9.3951 | 7.2953 | 9.1636 | 2078.0 |
| 2004 | 10.680 | 7.5614 | 9.7272 | 2191.3 |
| 2005 | 12.071 | 7.8036 | 10.301 | 2284.9 |
| 2006 | 13.482 | 8.0441 | 10.983 | 2387.5 |

The ridge procedure perturbs the moment matrix $\mathbf{X}^{\prime} \mathbf{X}$ while the raised and surrogate procedures perturbs the design matrix $\mathbf{X}$. This second approach has the distinct advantage of allowing the user to specify, for each of the variables, a precision that the data will retain during the perturbation stages by restricting the mean absolute perturbations

$$
\begin{equation*}
\lambda_{j} \frac{1}{n} \sum_{i=1}^{n}\left|\mathbf{e}_{j, i}\right|=\pi_{j} . \tag{15}
\end{equation*}
$$

For the case study, we set the precision $\pi_{j}=0.005$ in the first trial, and for the second trial we set the precision $\pi_{j}=0.05$. Thus given a specified precision $\pi_{j}>0$, we raise column $j$ in $\mathbf{X}_{<1, \ldots, j>}$ to $\widetilde{\mathbf{x}}_{j}=\mathbf{x}_{j}+\lambda_{j} \mathbf{e}_{j}$ where $\lambda_{j}$ is solved from Eq. (15). The precision values should be based on the researcher's belief in the accuracy of the data. The raised parameters $\lambda_{j}$ are constrained to assure that the original data has not been perturbed more than what the researcher has permitted.

The original data $\mathbf{X}_{0}$ is centered and standardized $\mathbf{X}_{0} \rightarrow \mathbf{X}$. For convenience, the columns are raised in sequence $1,2,3$. In Table 2 (with precision $\pi_{j}=0.005$ for all $j$ ) and in Table 3 (with precision $\pi_{j}=0.05$ for all $j$ ), we report the raised parameters $\lambda_{j}$, the corresponding generalized ridge parameters $k_{j}$ the $V I F s$, the associated angles $\theta$, the condition number $\kappa$ for the moment matrices as the ratio of the largest to smallest eigenvalue for the moment matrices, and
the squared length of the coefficient vector $\left\|\boldsymbol{\beta}\left(\lambda_{j}\right)\right\|^{2}$. Recall from Eq. (7) that the final raised matrix $\mathbf{X}_{<1, \ldots, p>}$ has moment matrix $\mathbf{X}^{\prime} \mathbf{X}+\operatorname{diag}\left(k_{1}, \ldots, k_{p}\right)$.

Note that the initial VIFs present high values, and this fact justifies the need to correct the collinearity. With precision $\pi=0.05$, after the successive raising process all the VIFs from $\mathbf{X}_{<1,2,3>}$ are less than 10, namely 7.063, 6.121, 5.244. But what is more important is that the matrix we need to invert, namely $\mathbf{X}^{\prime} \mathbf{X}+\operatorname{diag}\left(k_{1}, \ldots, k_{p}\right)$, is better conditioned than the initial matrix $\mathbf{X}^{\prime} \mathbf{X}$; and it has been obtained by adding a parameter $k_{i} \geq 0$ in the main diagonal of the matrix $\mathbf{X}^{\prime} \mathbf{X}$, in a way similar to ridge estimation, and where we can preserve the original data to a specified precision.

It is important to note that our methodology offers a concrete method to determine the perturbation parameter $\lambda_{j}$ for the raise estimators, and therefore of $\mathbf{k}$, while the choice of the ridge paramater remains an open problem. As noted by one of the referees, "the authors have obtained, by using an intuitive way, the matrix $\mathbf{X}^{\prime} \mathbf{X}+\mathbf{k} \mathbf{I}_{\mathbf{p}}$ with the sucessive raising estimator, while other proposed justications, are a posteriori."

Table 2. With precision $\pi_{j}=0.005$, raised parameters $\lambda_{j}$, ridge parameters $k_{j}, V I F_{j}$, associated angles $\theta_{j}$, condition numbers $\kappa=\lambda_{\max } / \lambda_{\text {min }}$ and $\left\|\boldsymbol{\beta}\left(\lambda_{j}\right)\right\|^{2}$.

|  | X | $\mathbf{X}_{<1>}$ | $\mathbf{X}_{<1,2>}$ | $\mathbf{X}_{<1,2,3>}$ |
| :---: | :---: | :---: | :---: | :---: |
| $\lambda_{j}$ | 0 | 0.6145 | 0.3433 | 0.2541 |
| $k_{j}$ | 0 | 0.002724 | 0.005021 | 0.004676 |
| VIF | $\left[\begin{array}{l}589.8 \\ 281.9 \\ 189.5\end{array}\right]$ | $\left[\begin{array}{l}226.9 \\ 160.2 \\ 124.7\end{array}\right]$ | $\left[\begin{array}{l}179.0 \\ 89.23 \\ 122.5\end{array}\right]$ | $\left[\begin{array}{l}145.6 \\ 88.47 \\ 78.23\end{array}\right]$ |
| $\theta$ | $\left[\begin{array}{l}2.360^{\circ} \\ 3.415^{\circ} \\ 4.166^{\circ}\end{array}\right]$ | $\left[\begin{array}{l}3.807^{\circ} \\ 4.532^{\circ} \\ 5.137^{\circ}\end{array}\right]$ | $\left[\begin{array}{l}4.286^{\circ} \\ 6.077^{\circ} \\ 5.185^{\circ}\end{array}\right]$ | $\left[\begin{array}{l}4.754^{\circ} \\ 6.103^{\circ} \\ 6.492^{\circ}\end{array}\right]$ |
| $\kappa$ | 2676 | 1033 | 819.2 | 654.3 |
| $\left\\|\boldsymbol{\beta}\left(\lambda_{j}\right)\right\\|^{2}$ | 998.1 | 556.6 | 301.1 | 287.0 |

Table 3. With precision $\pi_{j}=0.05$, raised parameters $\lambda_{j}$, ridge parameters $k_{j}, V I F_{j}$, associated angles $\theta_{j}$, condition numbers $\kappa=\lambda_{\text {max }} / \lambda_{\text {min }}$ and $\left\|\boldsymbol{\beta}\left(\lambda_{j}\right)\right\|^{2}$.

|  | $\mathbf{X}$ | $\mathbf{X}_{<1>}$ | $\mathbf{X}_{<1,2>}$ | $\mathbf{X}_{<1,2,3>}$ |
| :---: | :---: | :---: | :---: | :---: |
| $\lambda_{j}$ | 0 | 6.145 | 2.477 | 1.010 |
| $k_{j}$ | 0 | 0.08488 | 0.1258 | 0.1675 |
| $V I F$ | $\left[\begin{array}{c}589.8 \\ 281.9 \\ 189.5\end{array}\right]$ | $\left[\begin{array}{c}12.53 \\ 88.32 \\ 86.51\end{array}\right]$ | $\left[\begin{array}{c}12.03 \\ 8.222 \\ 18.15\end{array}\right]$ | $\left[\begin{array}{c}7.063 \\ 6.121 \\ 5.244\end{array}\right]$ |
|  | $\left[\begin{array}{c}2.360^{\circ} \\ 3.415^{\circ} \\ 4.166^{\circ}\end{array}\right]$ | $\left[\begin{array}{c}16.41^{\circ} \\ 6.108^{\circ} \\ 6.172^{\circ}\end{array}\right]$ | $\left[\begin{array}{c}16.76^{\circ} \\ 20.41^{\circ} \\ 13.58^{\circ}\end{array}\right]$ | $\left[\begin{array}{c}22.10^{\circ} \\ 23.84^{\circ} \\ 25.89^{\circ}\end{array}\right]$ |
| $\boldsymbol{2 6 7 6}$ | 998.1 | 227.8 | 106.4 | 25.87 |
| $\left\\|\boldsymbol{\beta}\left(\lambda_{j}\right)\right\\|^{2}$ | 26.4 | 46.49 |  |  |

## 8 Conclusions

The ridge procedure perturbs the moment matrix $\mathbf{X}^{\prime} \mathbf{X} \rightarrow \mathbf{X}^{\prime} \mathbf{X}+\lambda \mathbf{I}_{p}$ but does not allow the user to compute the perturbed design so the changes to $\mathbf{X}$ will be unknown to the user.

The surrogate procedure has the advantage of allowing the user to explicitly compute the surrogate design $\mathbf{X}_{S}$ in terms of the singular values of $\mathbf{X}$. Thus the Mean Absolute Deviation $\left\|\mathbf{X}-\mathbf{X}_{S}\right\|_{M A D}=\frac{1}{n k} \sum_{i, j}\left|\mathbf{X}(i, j)-\mathbf{X}_{S}(i, j)\right|$ can be computed showing the average change in the design $\mathbf{X} \rightarrow \mathbf{X}_{S}$. However, the surrogate procedure does not allow for varying perturbations for the varying explanatory variables. The surrogate procedure is based on perturbating the singular values of $\mathbf{X}$ and thus it is not intuitive nor geometric. On the other hand, the raise procedure it both intuitive and geometric.

It is important to note that the raise procedure does yield the explicit new design with $\mathbf{X} \rightarrow \widetilde{\mathbf{X}}$, and that the mean absolute perturbations given in Eq. (15) are permitted to be different for each explanatory variable. Thus the user can set the mean absolute deviation to be smaller for variables which are known to be accurate and allow larger deviations for variables which are known to be less accurate.

The raised estimator $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})=\left(\widetilde{\mathbf{X}}^{\prime} \widetilde{\mathbf{X}}\right)^{-1 /} \widetilde{\mathbf{X}}^{\prime} \mathbf{y}$ obtained from the successive raising process in Section 4 improves collinearity. The raised regression retains the value of the coefficient of determination, contrary to what happens with the ridge and surrogate estimators. In this way, we can affirm that the successive raising estimator has improving the collinearity similar to the improvement with ridge regression.

The estimator obtained from the successive raising process is given by $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})=$ $\left(\mathbf{X}^{\prime} \mathbf{X}+\mathbf{K}\right)^{-1} \widetilde{\mathbf{X}}^{\prime} \mathbf{y}$ with $\mathbf{K}$ a diagonal matrix. If we want $\mathbf{K}=k \mathbf{I}_{p}$ to be a
constant diagonal matrix then this can be achieved with the proper choice of $\lambda=\left(\lambda_{1}, \ldots, \lambda_{p}\right)$ using Eq. (13). To conclude, if the user wants a better conditioned matrix than $\mathbf{X}^{\prime} \mathbf{X}$, the proposed methodology offers a procedure to improve the collinearity in the spirit of Levenberg (1944) with a perturbation matrix similar to the perturbation matrix used in ridge regression. Unlike ridge regression, the user can visualize the perturbations of the underlying model and easily control the amount of perturbations to the original data retaining a specified precision in the data.

Unlike ridge regression where the choice of the ridge parameter remains an open problem, our methodology offers a concrete procedure for determining the perturbation parameter for the raise estimators.

## 9 Appendix A

In a full-rank model $\{\mathbf{Y}=\mathbf{X} \boldsymbol{\beta}+\varepsilon\}$ having zero-mean, uncorrelated, and homoscedastic errors, the Ordinary Least Squares $O L S$ estimator $\widehat{\boldsymbol{\beta}}$ solves the $p$ equations $\left\{\mathbf{X}^{\prime} \mathbf{X} \boldsymbol{\beta}=\mathbf{X}^{\prime} \mathbf{y}\right\}$. These solutions are unbiased with dispersion matrix $V(\widehat{\boldsymbol{\beta}})=\sigma^{2}\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}$. Near-dependency among the columns of $\mathbf{X}$, as illconditioning, often causes the $O L S$ solutions to have inflated size $\|\widehat{\boldsymbol{\beta}}\|^{2}$, and of questionable signs, and "very sensitive to small changes in X" (Belsley, 1986). Among standard remedies are the ridge system $\left\{\left(\mathbf{X}^{\prime} \mathbf{X}+k \mathbf{I}_{p}\right) \boldsymbol{\beta}=\mathbf{X}^{\prime} \mathbf{y} ; k \geq 0\right\}$ (Hoerl and Kennard, 1970) with solutions $\left\{\widehat{\boldsymbol{\beta}}_{R}(k) ; k \geq 0\right\}$. The $O L S$ solution at $k=0$ is known to be unstable with a slight movement away from $k=0$ giving completely different estimates of the coefficients.

McDonald $(2009,2010)$ showed that the square of the correlation coefficient $R^{2}(k)$ of the observed values $\mathbf{y}$ and the predicted values $\widehat{\mathbf{y}}_{R}(k)=\mathbf{X} \widehat{\boldsymbol{\beta}}_{R}(k)$ for the ridge regression estimators $\widehat{\boldsymbol{\beta}}_{R}(k)$ is a monotone decreasing function of the ridge parameter $k$. We show the corresponding result for the surrogate estimators, Jensen and Ramirez (2008, 2010b).

The singular decomposition $\boldsymbol{X}=\boldsymbol{P} \boldsymbol{D}_{\xi} \boldsymbol{Q}^{\prime}$, with $\boldsymbol{P}^{\prime} \boldsymbol{P}=\boldsymbol{I}_{p}$, together with $\boldsymbol{\theta}=\boldsymbol{Q}^{\prime} \boldsymbol{\beta}$ as an orthogonal reparametrization, give $\boldsymbol{y}=\boldsymbol{X} \boldsymbol{\beta}+\boldsymbol{\varepsilon} \rightarrow \boldsymbol{y}=$ $\boldsymbol{P} \boldsymbol{D}_{\xi} \boldsymbol{Q}^{\prime} \boldsymbol{\beta}+\boldsymbol{\varepsilon} \rightarrow \boldsymbol{U}=\boldsymbol{P}^{\prime} \boldsymbol{y}=\boldsymbol{D}_{\xi} \boldsymbol{\theta}+\boldsymbol{P}^{\prime} \boldsymbol{\varepsilon}$ as a canonical and equivalent model on $\mathbb{R}^{p}$. Gauss-Markov assumptions regarding $\boldsymbol{y}=\boldsymbol{X} \boldsymbol{\beta}+\boldsymbol{\varepsilon}$ stipulate that $E(\varepsilon)=\mathbf{0} \in \mathbb{R}^{n}$ and $V(\varepsilon)=\sigma^{2} \boldsymbol{I}_{n}$. From these assumptions, it follows that $E\left(\boldsymbol{P}^{\prime} \boldsymbol{\varepsilon}\right)=\mathbf{0} \in \mathbb{R}^{p}$ and $V\left(\boldsymbol{P}^{\prime} \boldsymbol{\varepsilon}\right)=\sigma^{2} \boldsymbol{P}^{\prime} \boldsymbol{I}_{n} \boldsymbol{P}=\sigma^{2} \boldsymbol{I}_{p}$, so that $E(\boldsymbol{U})=\boldsymbol{D}_{\xi} \boldsymbol{\theta}$ and $V(\boldsymbol{U})=\sigma^{2} \boldsymbol{I}_{p}$.

Specifically, the singular decomposition $\boldsymbol{X}=\boldsymbol{P} \boldsymbol{D}_{\xi} \boldsymbol{Q}^{\prime}$, and the surrogate $\boldsymbol{X}_{k}=\boldsymbol{P} \operatorname{Diag}\left(\left(\xi_{1}^{2}+k\right)^{\frac{1}{2}}, \ldots,\left(\xi_{p}^{2}+k\right)^{\frac{1}{2}}\right) \boldsymbol{Q}^{\prime}$ with singular values $\left\{\xi_{1}^{2} \geq \xi_{2}^{2} \geq \ldots \geq\right.$ $\left.\xi_{p}^{2}>0\right\}$, give $\boldsymbol{X}_{k}^{\prime} \boldsymbol{X}_{k}=\left(\boldsymbol{X}^{\prime} \boldsymbol{X}+k \boldsymbol{I}_{p}\right)$ with $\left\{\boldsymbol{X}_{k}^{\prime} \boldsymbol{X}_{k} \boldsymbol{\beta}=\boldsymbol{X}^{\prime} \boldsymbol{y}\right\}$ for the ridge estimators. For the surrogate model, $\left\{\boldsymbol{y}=\boldsymbol{X}_{k} \boldsymbol{\beta}+\boldsymbol{\varepsilon}\right\}$ is taken as an approximation, or surrogate, for the ill-conditioned model $\{\boldsymbol{y}=\boldsymbol{X} \boldsymbol{\beta}+\boldsymbol{\varepsilon}\}$ itself with $\left\{\boldsymbol{X}_{k}^{\prime} \boldsymbol{X}_{k} \boldsymbol{\beta}=\boldsymbol{X}_{k}^{\prime} \boldsymbol{y}\right\}$ for the ridge estimators.

The canonical estimators $\left\{\widehat{\theta}, \widehat{\theta}_{R}, \widehat{\theta}_{S}\right\}$ in terms of $\left\{k, \xi_{i}\right\}$ are given by

| Estimator | Definition |
| :---: | :---: |
| $\widehat{\theta}$ | $\mathbf{D}\left(\frac{1}{\xi_{i}}\right) \mathbf{U}$ |
| $\widehat{\theta}_{R}$ | $\mathbf{D}\left(\frac{\xi_{i}^{i}}{\xi_{i}^{2}+k}\right) \mathbf{U}$ |
| $\widehat{\theta}_{S}$ | $\mathbf{D}\left(\frac{1}{\sqrt{\xi_{i}^{2}+k}}\right) \mathbf{U}$ |

Following McDonald (2009, 2010), we assumed the model has been standardized with $\mathbf{1}^{\prime} \mathbf{y}=0$ and $\mathbf{y}^{\prime} \mathbf{y}=\mathbf{1}$ and with $\mathbf{X}^{\prime} \mathbf{X}$ having correlation form; in particular, $\mathbf{1}^{\prime} \mathbf{X}=\mathbf{0}$ with $\mathbf{1}^{\prime} \widehat{\mathbf{y}}_{R}=\mathbf{1}^{\prime} \mathbf{X} \widehat{\boldsymbol{\beta}}_{R}=\mathbf{0}=\mathbf{1}^{\prime} \mathbf{X} \widehat{\boldsymbol{\beta}}_{S}=\mathbf{1}^{\prime} \widehat{\mathbf{y}}_{S}$. It follows that the squared correlation coefficient $R_{S}^{2}(k)$ of the observed values $\mathbf{y}$ and the predicted values $\widehat{\mathbf{y}}_{S}(k)=\mathbf{X} \widehat{\boldsymbol{\beta}}_{S}$ in terms of the canonical variables is

$$
\begin{aligned}
& R_{S}^{2}(k)= {\left[(\mathbf{y}-\overline{\mathbf{y}})^{\prime}\left(\widehat{\mathbf{y}}_{S}(k)-\overline{\widehat{\mathbf{y}}}_{S}(k)\right)\right]^{2} } \\
& {\left[(\mathbf{y}-\overline{\mathbf{y}})^{\prime}(\mathbf{y}-\overline{\mathbf{y}})\right]\left[\left(\widehat{\mathbf{y}}_{S}(k)-\overline{\widehat{\mathbf{y}}}_{S}(k)\right)^{\prime}\left(\widehat{\mathbf{y}}_{S}(k)-\overline{\hat{\mathbf{y}}}_{S}(k)\right)\right] } \\
&=\frac{\left(\mathbf{y}^{\prime} \widehat{\mathbf{y}}_{S}(k)\right)^{2}}{\widehat{\mathbf{y}}_{S}^{\prime}(k) \widehat{\mathbf{y}}_{S}(k)} \cdot=\frac{\left(\mathbf{y}^{\prime} \mathbf{X} \widehat{\boldsymbol{\beta}}_{S}\right)^{2}}{\widehat{\boldsymbol{\beta}}_{S}^{\prime} \mathbf{X}^{\prime} \mathbf{X}_{\boldsymbol{\boldsymbol { \beta }}}^{S}} \mathrm{C}
\end{aligned}=\frac{\left(\mathbf{U}^{\prime} \mathbf{D}\left(\frac{\xi_{i}}{\sqrt{\xi_{i}^{2}+k}}\right) \mathbf{U}\right)^{2}}{\mathbf{U}^{\prime} \mathbf{D}\left(\frac{\xi_{i}^{2}}{\xi_{i}^{2}+k}\right) \mathbf{U}},
$$

with $R_{S}^{2}(0)=\left(\mathbf{U}^{\prime} \mathbf{U}\right)^{2} /\left(\mathbf{U}^{\prime} \mathbf{U}\right)=1$. For ridge estimators, McDonald (2010) gives the parallel result for ridge estimators

$$
R_{R}^{2}(k)=\frac{\left(\mathbf{U}^{\prime} \mathbf{D}\left(\frac{\xi_{i}^{2}}{\xi_{i}^{2}+k}\right) \mathbf{U}\right)^{2}}{\mathbf{U}^{\prime} \mathbf{D}\left(\frac{\xi_{i}^{4}}{\left(\xi_{i}^{2}+k\right)^{2}}\right) \mathbf{U}}
$$

with $d R_{R}^{2}(k) / d k<0$ for $k>0$.
For the surrogate estimators,

$$
\frac{d R_{S}^{2}(k)}{d k}=\frac{(2 G)\left(-\frac{1}{2} H\right)}{J}-\frac{G^{2}(-K)}{J^{2}}=\frac{G(-H J+G K)}{J^{2}}
$$

with

$$
\begin{array}{rr}
G=\sum_{i=1}^{p} u_{i}^{2} \frac{\xi_{i}}{\sqrt{\xi_{i}^{2}+k}} & H=\sum_{i=1}^{p} u_{i}^{2} \frac{\xi_{i}}{\left(\xi_{i}^{2}+k\right)^{3 / 2}} \\
J=\sum_{i=1}^{p} u_{i}^{2} \frac{\xi_{i}^{2}}{\xi_{i}^{2}+k} & K=\sum_{i=1}^{p} u_{i}^{2} \frac{\xi_{i}^{2}}{\left(\xi_{i}^{2}+k\right)^{2}}
\end{array}
$$

We show that $-H J+G K<0$ for $k>0$ :

$$
-H J+G K=-\left(\sum_{i=1}^{p} H_{i}\right)\left(\sum_{j=1}^{p} J_{j}\right)+\left(\sum_{i=1}^{p} G_{i}\right)\left(\sum_{j=1}^{p} K_{j}\right)
$$

with $-H_{i} J_{i}+G_{i} K_{i}=0$. For $i \neq j$, we consider

$$
\begin{aligned}
& -H_{i} J_{j}-H_{j} J_{i}+G_{i} K_{j}+G_{j} K_{i} \\
= & -\left(u_{i}^{2} \frac{\xi_{i}}{\left(\xi_{i}^{2}+k\right)^{3 / 2}}\right)\left(u_{j}^{2} \frac{\xi_{j}^{2}}{\xi_{j}+k}\right)-\left(u_{j}^{2} \frac{\xi_{j}}{\left(\xi_{j}^{2}+k\right)^{3 / 2}}\right)\left(u_{i}^{2} \frac{\xi_{i}^{2}}{\xi_{i}+k}\right) \\
& +\left(u_{i}^{2} \frac{\xi_{i}}{\sqrt{\xi_{i}^{2}+k}}\right)\left(u_{j}^{2} \frac{\xi_{j}^{2}}{\left(\xi_{j}^{2}+k\right)^{2}}\right)+\left(u_{j}^{2} \frac{\xi_{j}}{\sqrt{\xi_{j}^{2}+k}}\right)\left(u_{i}^{2} \frac{\xi_{i}^{2}}{\left(\xi_{i}+k\right)^{2}}\right) \\
= & \frac{u_{i}^{2} u_{j}^{2} \xi_{i} \xi_{j} \sqrt{\xi_{i}^{2}+k} \sqrt{\xi_{j}^{2}+k}}{\left(\xi_{i}^{2}+k\right)^{5 / 2}\left(\xi_{j}^{2}+k\right)^{5 / 2}}\left[\left(\xi_{i}^{2}-\xi_{j}^{2}\right)\left(\xi_{j} \sqrt{\xi_{i}^{2}+k}-\xi_{i} \sqrt{\xi_{j}^{2}+k}\right)\right] \\
= & \frac{u_{i}^{2} u_{j}^{2} \xi_{i} \xi_{j} \sqrt{\xi_{i}^{2}+k} \sqrt{\xi_{j}^{2}+k}}{\left(\xi_{i}^{2}+k\right)^{5 / 2}\left(\xi_{j}^{2}+k\right)^{5 / 2}}\left[\left(\xi_{i}^{2}-\xi_{j}^{2}\right)\left(\sqrt{\xi_{i}^{2} \xi_{j}^{2}+k \xi_{j}^{2}}-\sqrt{\xi_{i}^{2} \xi_{j}^{2}+k \xi_{i}^{2}}\right)\right]
\end{aligned}
$$

with the term in the brackets negative as the two terms in parentheses will have opposite signs.

## 10 Appendix B

With $p=2$, to find the minimum $M S E$ risk for the rotated estimators required solving for the roots of the quartic polynomial $p_{4}(t)=a_{0} t^{4}+a_{1} t^{3}+a_{2} t^{2}+a_{3} t+a_{4}$ with $a_{0}=1-\rho^{2}, a_{1}=2\left(1-\rho^{2}\right), a_{2}=\left(1-\rho^{2}\right)^{2}-4\left(1-\rho^{2}+\left(\frac{\sigma}{\delta_{1}}\right)^{2}\right)^{2}$, $a_{3}=-2 \rho^{2}\left(1-\rho^{2}\right), a_{4}=-\rho^{2}\left(1-\rho^{2}\right)$. Yang (1999) has given the root classification for quartic polynomials with general coefficients. Following his notation, we compute

$$
\begin{aligned}
D_{4}= & 256 a_{0}^{3} a_{4}^{3}-27 a_{0}^{2} a_{3}^{4}-192 a_{0}^{2} a_{3} a_{4}^{2} a_{1}-27 a_{1}^{4} a_{4}^{2} \\
& -6 a_{0} a_{1}^{2} a_{4} a_{3}^{2}+a_{2}^{2} a_{3}^{2} a_{1}^{2}-4 a_{0} a_{2}^{3} a_{3}^{2}+18 a_{2} a_{4} a_{1}^{3} a_{3} \\
& +144 a_{0} a_{2} a_{4}^{2} a_{1}^{2}-80 a_{0} a_{2}^{2} a_{4} a_{1} a_{3}+18 a_{0} a_{2} a_{3}^{3} a_{1}-4 a_{2}^{3} a_{4} a_{1}^{2} \\
& -4 a_{1}^{3} a_{3}^{3}+16 a_{0} a_{2}^{4} a_{4}-128 a_{0}^{2} a_{2}^{2} a_{4}^{2}+144 a_{0}^{2} a_{2} a_{4} a_{3}^{2}
\end{aligned}
$$

A quartic polynomial has two distinct roots when $D_{4}<0$. For $p_{4}(t)$, this follows from factoring $D_{4}$ with $A, B, C, D, E, F, G, H, I>0$ as

$$
D_{4}=A B\left(-C-4 \sigma^{2} D-4 \sigma^{4} E-16 \sigma^{6} F-16 \sigma^{8} G-64 \sigma^{10} H-\sigma^{12} I\right)
$$

where

$$
\begin{aligned}
A & =\frac{64\left(1-\rho^{2}\right)^{2} \rho^{2}}{\delta_{1}^{16}}>0 \\
B & =\left(\delta_{1}^{2}-\delta_{1}^{2} \rho^{2}+\sigma^{2}\right)^{2}>0 \\
C & =27 \delta_{1}^{12}\left(1+\rho^{2}\right)^{2}(1-\rho)^{4}(1+\rho)^{4}>0 \\
D & =54 \delta_{1}^{10}\left(1-\rho^{2}+\rho^{4}\right)(1-\rho)^{3}(1+\rho)^{3}>0 \\
E & =9 \delta_{1}^{8}\left(19-35 \rho^{2}+19 \rho^{4}\right)(1-\rho)^{2}(1+\rho)^{2}>0 \\
F & =68 \delta_{1}^{6}(1-\rho)^{3}(1+\rho)^{3}>0 \\
G & =57 \delta_{1}^{4}(1-\rho)^{2}(1+\rho)^{2}>0 \\
H & =6 \delta_{1}^{2}(1-\rho)(1+\rho)>0 \\
I & =64>0
\end{aligned}
$$

## 11 References

1. Belsley, D.A. (1986). Centering, the constant, first-differencing, and assessing conditioning. In Model Reliability (E. Kuh and D.A. Belsley, D.A. eds.). Cambridge: MIT Press.
2. Berk, K. (1977). Tolerance and condition in regression computations. J. Amer. Stat. Assoc. 72:863-866.
3. Davidov, O. (2006). Constrained estimation and the theorem of KuhnTucker. J. Appl. Math. Decis. Sci. Article ID 92970:1-13.
4. García, J., Andújar, A., Soto, J. (1999). Acerca de la detección de la colinealidad en los modelos de regresión lineal. XIII Reunión AsepeltEspaña. Universidad de Burgos.
5. García, C.B., García, J. and Soto, J. (2011). The raise method: An alternative procedure to estimate the parameters in presence of collinearity. Quality and Quantity 45:403-423.
6. García, C.B., García, J., López Martin, M.M. and Salmeron, R. (2015). Collinearity: Revisiting the variance inflation factor in ridge regression. Journal of Applied Statistics 42:648-661.
7. Goldstein, M. and Smith, A. F. M. (1974). Ridge-type estimators for regression analysis. J. of R. Stat. Soc. B 36:84-91.
8. Hoerl, A.E. (1964). Ridge Analysis. Chemical Engineering Progress, Symposium Series 60:67-77.
9. Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55-67.
10. Jensen, D.R. and Ramirez, D.E. (2008). Anomalies in the foundations of ridge regression. Int. Stat. Rev. 76:89-105.
11. Jensen, D.R. and Ramirez, D.E. (2010a). Tracking MSE efficiencies in ridge regression. Advances and Applications in Statistical Sciences 1:381398.
12. Jensen, D.R. and Ramirez, D.E. (2010b). Surrogate models in ill-conditioned systems. Journal of Statistical Planning and Inference 140:2069-2077.
13. Jensen, D.R. and Ramirez, D.E. (2013). Revision: Variance inflation in regression. Adv. Decis. Sci. Article ID 671204:1-15.
14. Kapat, P. and Goel, P.K. (2010). Anomalies in the foundations of ridge Regression: Some Clarifications. Int. Stat. Rev. 78:209-215.
15. Levenberg, K. (1944). A method for the solution of certain non-linear problems in least-squares. Quart. of Appl. Math. 2:164-168.
16. Marshal, A.W. and Olkin, I. (1979). Inequalities : Theory of Majorization and Its Applications. New York: Academic Press.
17. McDonald, G.C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics 1:93-100.
18. McDonald, G.C. (2010). Tracing ridge regression coefficients. Wiley Interdisciplinary Review: Computational Statistics 2:695-793.
19. Marquardt, D.W. (1963). An algorithm for least-squares estimation for nonlinear parameters. J. Soc. Ind. Appl. Math. 11:431-441.
20. Mayer, L.S. and Willke, T.A.(1973). On Biased Estimation in Linear Models. Technometrics 15:497-508.
21. The Economic Report of the President (2008). http://www.gpo.gov/fdsys/browse/collection.action?collectionCode=CPD
22. Piegorsch, W. and Casella, G. (1989). The early use of matrix diagonal increments in statistical problems. SIAM Review 31:428-434.
23. Puntanen, S., Styan, G.P.H., and Isotalo, J. (2011). Matrix Tricks for Linear Statistical Models. Berlin: Springer-Verlag.
24. Rawlings, J.O., Pentula, S.G., and Dickey, D.A. (1998). Applied Regression Analysis: A Research Tool- $2^{\text {nd }} \mathrm{ed}$. New York: Springer.
25. Riley, J. (1955). Solving Systems of linear equations with a positive definite, symmetric but possibly ill-conditioned matrix. Math Tables Aids Comput. 9:96-101.
26. Seber, G.A.F. (1977). Linear Regression Analysis. New York: John Wiley and Sons.
27. Stein, C.M. (1960) Multiple regression, contribution to probability and Statistics. Chapter 37 in Essays in Honor of Harold Hotelling (I. Olkin, S. Ghurye, W. Hoeffding, W. Madow and H. Mann eds). Stanford: Stanford University Press.
28. Stewart, G.W. (1987). Collinearity and least squares regression. Stat. Sci. 2:68-84.
29. Wichers, R. (1975). The detection of multicollinearity: A comment. Review Econ. Stat. 57:366-368.
30. Yang, L. (1999). Recent advances on determining the number of real roots of parametric polynomials. J. Symbolic Computation 28:225-242.
