Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v10n1p160

Biased power regression: a new biased estimation procedure in linear regression
By Qannari, El Ghaziri

Published: 26 April 2017

Electronic Journal of Applied Statistical Analysis
Vol. 10, Issue 01, April 2017, 160-179
DOI: 10.1285/i20705948v10n1p160

Biased power regression: a new biased estimation procedure in linear regression

El Mostafa Qannari*and Angélina El Ghaziri

StatSC, ONIRIS, INRA, 44320, Nantes, France

Published: 26 April 2017

In order to circumvent the effects of multicollinearity on the quality of a multiple linear regression, a new strategy of analysis is proposed. It is based on a biased estimation of the vector of coefficients. Properties of this approach of analysis are shown. Moreover, the link between this new strategy of analysis and existing strategies are discussed, particularly Ridge and Generalized Ridge regression. Illustrations on the basis of two datasets are also outlined and the outcomes are compared to those of Ridge regression.

keywords: Linear regression, Biased regression, Ridge regression, Generalized Ridge regression, Cross-Validation

1. Introduction

Consider the problem of estimating the vector of parameters in a multiple linear regression model. It is well know that the presence of multicollinearity among the predictors has a harmful impact on the quality and the stability of the fitted model. In such a situation, the parameter estimates are likely to have a poor numerical accuracy and large standard errors. This problem results from the ill conditioning of the variance-covariance matrix associated with the predictors. Biased estimation provides a way to circumvent this problem. The rationale behind this strategy of analysis is to trade the high variability of the parameter estimates for some (hopefully) negligible bias (Draper and Smith, 1998). In this context, a popular technique is Ridge regression (Hoerl and Kennard, 1970). It consists in improving the conditioning of the variance-covariance matrix associated with the predictors by augmenting its eigenvalues by a small quantity.

*Corresponding author: elmostafa.qannari@oniris-nantes.fr

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

We propose a new biased estimation procedure called biased power regression (or BP-regression, for short) which consists in setting the eigenvalues of the variance-covariance matrix of the predictors to the power $1-\alpha$, where α is a scalar which ranges between 0 and 1. We will show that this transformation improves the conditioning of the variance-covariance matrix therefore leading to an improvement of the prediction model. The paper is organized as follows. In section 2, we introduce BP-regression, and, in section 3, we investigate some of its properties. In section 4, we discuss a strategy for choosing the tuning parameter α. In section 5, BP-regression is illustrated and compared to Ridge regression on the basis of two datasets. We end the paper by sketching some concluding remarks in section 6.

2. BP-regression

We assume the multiple linear regression model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is an (n x *1*) vector (dependant variable), $\mathbf{X}$ is an (n x p) matrix (predictors), $\boldsymbol{\beta}$ is an (p x *1*) vector of unknown regression coefficients and $\boldsymbol{\epsilon}$ is an (n x *1*) vector of random errors. The ordinary least squares estimator is given by :

$$\mathbf{b}_0 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

It is well known that this estimator is likely to lead to an unstable model and poor predictions in presence of quasi-collinearity among variables or in the case of a small sample and high dimensional setting (*i.e.* large p, small n). Ridge regression was proposed as a regularization procedure to cope with this problem (Hoerl and Kennard, 1970). The Ridge estimators are given by:

$$\mathbf{b}_\kappa = (\mathbf{X}^T\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \tag{1}$$

where κ is a positive constant and $\boldsymbol{I}$, the identity matrix. Several procedures have been proposed to select an appropriate parameter κ (Hoerl et al., 1975; Golub et al., 1979). One of the simplest and more efficient strategies is to perform a cross-validation procedure (Stone and Brooks, 1990; Hastie et al., 2009) as sketched in section 4.

As an alternative to Ridge regression, we propose to estimate $\boldsymbol{\beta}$ as follows. Let $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{V}^T$ be the singular value decomposition of $\boldsymbol{X}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices and $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal entries $\lambda_1, \lambda_2, ..., \lambda_p$ are non-negative (Golub and Reinsch, 1970). It is well known that these quantities are the eigenvalues of $\mathbf{X}^T\mathbf{X}$. From now on, we shall assume that they are arranged in a decreasing order of magnitude and, for reasons that will be clear in subsequent sections, we shall assume that $0 < \lambda_j \leq 1$ ($j = 1, ..., p$). This can be obtained by applying a pre-treatment on the original dataset consisting in dividing it by its largest singular value (*i.e.* $\sqrt{\lambda_1}$). We consider a scalar α between 0 and 1. The BP-regression estimator of $\boldsymbol{\beta}$ is given by:

$$\mathbf{b}_\alpha = (\mathbf{X}^T\mathbf{X})^{\alpha-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

where $(\mathbf{X}^T\mathbf{X})^{\alpha-1} = \mathbf{V}\mathbf{\Lambda}^{\alpha-1}\mathbf{V}^T$. It is clear that for $\alpha = 0$, $\mathbf{b_0}$ is the ordinary least squares estimator. For $\alpha = 1$, $\mathbf{b_1}$ is proportional to the PLS1 estimator based on the first latent variable (de Jong, 1993).

3. Properties of BP-regression

3.1. Condition indices and an overall collinearity index

Belsley (1991) proposed to use the condition indices as collinearity diagnostics. These indices are given by:

$$\eta_j = \left(\frac{\lambda_1}{\lambda_j}\right)^{\frac{1}{2}} \qquad j = 1, ..., p$$

Obviously, $\eta_j \geq 1$ for all j. Large η_j indicate the presence of near collinearity among the predictors. By using Ridge regression, the condition indices become

$$\eta_j(\kappa) = \left(\frac{\lambda_1+\kappa}{\lambda_j+\kappa}\right)^{\frac{1}{2}} \qquad j = 1, ..., p$$

by deriving $(\eta_j(\kappa))^2$ with respect to κ it is easy to show that $\eta_j(\kappa)$ is a decreasing function of κ. The implication of this property is that Ridge regression improves the conditioning of matrix $\mathbf{X}^T\mathbf{X}$, which is likely to improve the stability of the regression model.

Likewise, by using BP-regression, matrix $\mathbf{X}^T\mathbf{X}$ is transformed into $(\mathbf{X}^T\mathbf{X})^{1-\alpha}$. Therefore, the associated condition indices are given by:

$$\eta_j(\alpha) = \left(\frac{\lambda_1}{\lambda_j}\right)^{\frac{1-\alpha}{2}} \qquad j = 1, ..., p$$

The derivative of this quantity with respect to α is:

$$\eta_j^{'}(\alpha) = \frac{1}{2} \times ln(\frac{\lambda_j}{\lambda_1}) \times (\frac{\lambda_1}{\lambda_j})^{\frac{1-\alpha}{2}}$$

which is negative since $\lambda_j \leq \lambda_1$. This indicates that $\eta_j(\alpha)$ is a decreasing function of α. This entails that by increasing α, we achieve a better conditioning of the regression problem.

It is clear that the purpose of the condition indices is to assess the relative importance of the eigenvalues λ_j $(j = 1, ..., p)$. Another way to assess this phenomenon is to compute the coefficient of variation (*i.e.* ratio of the standard deviation to the mean) of these quantities and examine their evolution as a function of the regularization parameter.

Regarding BP-regression, the coefficient of variation associated with $(\lambda_j^{1-\alpha})$ $(j = 1, ..., p)$ is given by:

$$CV(\alpha) = \frac{\sqrt{p\sum_{j=1}^{p}\lambda_j^{2(1-\alpha)} - (\sum_{j=1}^{p}\lambda_j^{1-\alpha})^2}}{\sum_{j=1}^{p}\lambda_j^{1-\alpha}} \tag{3}$$

We can show that this index decreases as α increases (see appendix A). This means that the regularization operated by BP-regression reduces the discrepancy among the

eigenvalues λ_j $(j = 1, ..., p)$ and thus reduces the effect of multicollinearity. It is also worth noting that:

$$CV(\alpha) = \sqrt{\frac{p}{\psi(\alpha)} - 1}$$

where $\psi(\alpha) = \frac{(\sum_{j=1}^{p} \lambda_j^{1-\alpha})^2}{\sum_{j=1}^{p} \lambda_j^{2(1-\alpha)}}$ is known as the sphericity index and used to determine the degree of freedom for multivariate tests (Worsley and Friston, 1995; Abdi, 2010). It was also advocated using this index as a measure of the dimensionality or the complexity of the dataset at hand (Kazi-Aoual et al., 1995). We have $\psi(\alpha) = 1 \Leftrightarrow CV(\alpha) = \sqrt{p-1} \Leftrightarrow$ only λ_1 is different from 0 (extreme situation of collinearity). $\psi(\alpha) = p \Leftrightarrow CV(\alpha) = 0 \Leftrightarrow \lambda_1 = \lambda_2 = ... = \lambda_p$ (orthogonal design). $\psi(\alpha)$ increases $\Leftrightarrow CV(\alpha)$ decreases.

By comparison, for Ridge regression, we have:

$$CV(\kappa) = \frac{\sqrt{p \sum_{j=1}^{p} \lambda_j^2 - (\sum_{j=1}^{p} \lambda_j)^2}}{\sum_{j=1}^{p} (\lambda_j + \kappa)}.$$

Obviously, this is a decreasing function of κ. It is also easy to check that

$$CV(\kappa) = \sqrt{\frac{p}{\psi(\kappa)} - 1}$$

Where $\psi(\kappa) = \frac{(\sum_{j=1}^{p} (\lambda_j + \kappa))^2}{\sum_{j=1}^{p} (\lambda_j + \kappa)^2}$.

3.2. BP-regression shrinks

An interesting property shared by several biased estimators is that they shrink the vector of regression coefficients. This is the case for Ridge regression which is defined as a shrinking procedure since the Ridge estimator is defined as a least squared estimator under the contraint that the length of the vector of regression coefficient is smaller than a prespecified quantity. PLS regression also shrinks (de Jong, 1995). In order to prove that BP-regression shrinks the vector of regression coefficients, let us consider the squared length of $\mathbf{b}_\alpha$. From equation (2) and recalling that $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{V}^T$, with $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ (identity matrix), it follows:

$$\begin{aligned} l(\alpha) = ||\mathbf{b}_\alpha||^2 &= ||(\mathbf{X}^T\mathbf{X})^{\alpha-1}\mathbf{X}^T\mathbf{y}||^2 \\ &= ||\mathbf{V}\mathbf{\Lambda}^{\alpha-\frac{1}{2}}\mathbf{U}^T\mathbf{y}||^2 \\ &= (\mathbf{y}^T\mathbf{U}\mathbf{\Lambda}^{\alpha-\frac{1}{2}}\mathbf{V}^T)(\mathbf{V}\mathbf{\Lambda}^{\alpha-\frac{1}{2}}\mathbf{U}^T\mathbf{y}) \\ &= \mathbf{y}^T\mathbf{U}\mathbf{\Lambda}^{2\alpha-1}\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p} (\mathbf{u}_j^T\mathbf{y})^2 \lambda_j^{2\alpha-1} \end{aligned}$$

Where $\mathbf{u}_j$ is the j^{th} column of matrix $\mathbf{U}$. By deriving $l(\alpha)$ with respect to α, it follows:

$$l^{'}(\alpha) = 2\sum_{j=1}^{p}(\mathbf{u}_j^T\mathbf{y})^2 ln(\lambda_j)\lambda_j^{2\alpha-1}$$

Since we have assumed that for $j = 1$ to p, $\lambda_j \leq 1$, we have $l^{'}(\alpha) \leq 0$. This implies that the length of $\mathbf{b}_\alpha$ decreases with α.

3.3. Biased estimator and Mean Squared Error

From equation (2) which introduces $\mathbf{b}_\alpha$, it readily follows that $\mathbf{b}_\alpha = (\mathbf{X}^T\mathbf{X})^\alpha\mathbf{b}_0$. Thus, $\mathbf{b}_\alpha$ is a linear transform of the ordinary least squares estimator, $\mathbf{b}_0$. It also follows that for $\alpha \neq 0$, $\mathbf{b}_\alpha$ is a biased estimator since $E(\mathbf{b}_\alpha) = (\mathbf{X}^T\mathbf{X})^\alpha E(\mathbf{b}_0) = (\mathbf{X}^T\mathbf{X})^\alpha\boldsymbol{\beta}$. The mean squared error associated with $\mathbf{b}_\alpha$ reflects how, on average, $\mathbf{b}_\alpha$ is far removed from the true parameter, $\boldsymbol{\beta}$. We have:

$$\begin{aligned}
MSE(\alpha) &= E(\| \mathbf{b}_\alpha - \boldsymbol{\beta} \|^2) \\
&= E(||(\mathbf{X}^T\mathbf{X})^\alpha\mathbf{b}_0 - \boldsymbol{\beta}||^2) \\
&= E(||\mathbf{V}\boldsymbol{\Lambda}^\alpha\mathbf{V}^T\mathbf{b}_0 - \boldsymbol{\beta}||^2) \\
&= E((\mathbf{V}\boldsymbol{\Lambda}^\alpha\mathbf{V}^T\mathbf{b}_0 - \boldsymbol{\beta})^T(\mathbf{V}\boldsymbol{\Lambda}^\alpha\mathbf{V}^T\mathbf{b}_0 - \boldsymbol{\beta})) \\
&= E(\mathbf{b}_0^T\mathbf{V}\boldsymbol{\Lambda}^{2\alpha}\mathbf{V}^T\mathbf{b}_0) - 2E(\mathbf{b}_0^T\mathbf{V}\boldsymbol{\Lambda}^\alpha\mathbf{V}^T\boldsymbol{\beta}) + E(\boldsymbol{\beta}^T\boldsymbol{\beta}) \\
&= E(\mathbf{b}_0^T\mathbf{V}\boldsymbol{\Lambda}^{2\alpha}\mathbf{V}^T\mathbf{b}_0) - 2\boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\Lambda}^\alpha\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta} \qquad (4)
\end{aligned}$$

Let us recall that for a quadratic form, $\mathbf{z}^T\mathbf{A}\mathbf{z}$, associated with a random variable $\mathbf{z}$, we have $E(\mathbf{z}^T\mathbf{A}\mathbf{z}) = trace(\mathbf{A}\boldsymbol{\Sigma}) + \mu^T\mathbf{A}\mu$ where μ and $\boldsymbol{\Sigma}$ are respectively the mean and variance-covariance matrix of $\mathbf{z}$ (Draper and Smith, 1998).

Applying this property to $\mathbf{b}_0^T\mathbf{V}\boldsymbol{\Lambda}^{2\alpha}\mathbf{V}^T\mathbf{b}_0$ and recalling that the mean and the variance-covariance matrix associated with $\mathbf{b}_0$ are respectively $\boldsymbol{\beta}$ and $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, it follows:

$$E(\mathbf{b}_0^T\mathbf{V}\boldsymbol{\Lambda}^{2\alpha}\mathbf{V}^T\mathbf{b}_0) = \sigma^2 trace(\boldsymbol{\Lambda}^{2\alpha-1}) + \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\Lambda}^{2\alpha}\mathbf{V}^T\boldsymbol{\beta}$$

Replacing this expression in equation (4), we are led to:

$$\begin{aligned}
MSE(\alpha) &= \sigma^2 trace(\boldsymbol{\Lambda}^{2\alpha-1}) + \boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\Lambda}^{2\alpha}\mathbf{V}^T\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\mathbf{V}\boldsymbol{\Lambda}^\alpha\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta} \\
&= \sigma^2 trace(\boldsymbol{\Lambda}^{2\alpha-1}) + ||(\boldsymbol{\Lambda}^\alpha - \mathbf{I})\mathbf{V}^T\boldsymbol{\beta}||^2 \\
&= \sigma^2\sum_{j=1}^{p}\lambda_j^{2\alpha-1} + \sum_{j=1}^{p}(\lambda_j^\alpha - 1)^2\gamma_j^2 \qquad (5)
\end{aligned}$$

Where γ_j is the j^{th} element of vector $\mathbf{V}^T\boldsymbol{\beta}$. The first term in the expression of $MSE(\alpha)$ (equation (5)) represents the variance associated with $\mathbf{b}_\alpha$ whereas the second term corresponds to the bias.

Deriving $MSE(\alpha)$ with respect to α leads us to:

$$MSE'(\alpha) = 2\sigma^2 \sum_{j=1}^{p} ln(\lambda_j)\lambda_j^{2\alpha-1} - 2\sum_{j=1}^{p}(\lambda_j^{\alpha}-1)ln(\lambda_j)(\lambda_j^{\alpha})\gamma_j^2 \tag{6}$$

Since we have assumed that $\lambda_j \leq 1$, it follows that $ln(\lambda_j) \leq 0$ and $(\lambda_j^{\alpha}-1) \leq 0$. Therefore, the first term which corresponds to the derivative of the variance of b_α is negative. This means that the variance of $\mathbf{b}_\alpha$ decreases with α. Contrariwise, the second term in equation (6), which corresponds to the derivative of the bias of $\mathbf{b}_\alpha$ is positive. This indicates that the bias increases. All in all, we expect that the decrease of the variance associated with $\mathbf{b}_\alpha$ more than compensates the increase of the bias, resulting in a decrease of the mean squared error.

We can show that the BP-regression yields a family of regressors, $\mathbf{b}_\alpha$, which is admissible in that sense that there always exists a scalar α such that $MSE(\alpha) \leq MSE(0)$; $MSE(0) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$ being the mean squared error associated with the ordinary least squares estimator, $\mathbf{b}_0$. We have:

$$MSE^{'}(\alpha) = \sum_{j=1}^{p} f_j(\alpha)$$

Where $f_j(\alpha) = 2ln(\lambda_j)\lambda_j^{\alpha}(\sigma^2\lambda_j^{\alpha-1} + \gamma_j^2\lambda_j^{\alpha} - \gamma_j^2)$.

Because of the pre-treatment that we have applied, we have $\lambda_1 = 1$ and $\lambda_j < 1$ $(j = 2, ..., p)$. Thus :

$$MSE^{'}(\alpha) = \sum_{j=2}^{p} f_j(\alpha)$$

It is easy to check that for $j = 2, ..., p$, $f_j(\alpha) \leq 0$ is equivalent to:

$$\alpha \leq \frac{ln\left(\gamma_j^2/(\frac{\sigma^2}{\lambda_j}+\gamma_j^2)\right)}{ln(\lambda_j)}$$

The quantity in the right side of this inequality is positive since both the numerator and the denominator are negative. It follows that if we choose $\alpha_0 = \min_{j=2,...,p}\left(\frac{ln\left(\gamma_j^2/(\frac{\sigma^2}{\lambda_j}+\gamma_j^2)\right)}{ln(\lambda_j)}\right)$, $MSE^{'}(\alpha)$ will be negative, and therefore, $MSE(\alpha)$ will decrease. This indicates that there exists a parameter α between 0 and α_0 such that $MSE(\alpha) \leq MSE(0)$.

3.4. Correlation between the observed and the predicted y

For each tuning parameter α, the predicted variable $\hat{\boldsymbol{y}}_\alpha$ is given by $\hat{\boldsymbol{y}}_\alpha = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{\alpha-1}\mathbf{X}^T\boldsymbol{y}$. The squared coefficient of correlation, $R^2(\alpha)$, between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\alpha$ reflects the quality of the adjustmet as a function of α. We have:

$$\begin{aligned} R^2(\alpha) &= \frac{cov^2(\boldsymbol{y}, \hat{\boldsymbol{y}}_\alpha)}{var(\boldsymbol{y})var(\hat{\boldsymbol{y}}_\alpha)} \\ &= \frac{(\boldsymbol{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{\alpha-1}\mathbf{X}^T\boldsymbol{y})^2}{\boldsymbol{y}^T\boldsymbol{y}\boldsymbol{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{\alpha-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{\alpha-1}\mathbf{X}^T\boldsymbol{y}} \end{aligned}$$

By expressing $\mathbf{X}$ in terms of its singular value decomposition: $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^T$, it is easy to show that:

$$R^2(\alpha) = \frac{\left(\sum_{j=1}^{p} \lambda_j^{\alpha} (\mathbf{u}_j^T \boldsymbol{y})^2\right)^2}{\boldsymbol{y}^T\boldsymbol{y} \sum_{j=1}^{p} \lambda_j^{2\alpha} (\mathbf{u}_j^T \boldsymbol{y})^2}$$

Where $\mathbf{u}_j$ is the j^{th} column of matrix $\mathbf{U}$. We show that $R^2(\alpha)$ is a decreasing function of α (appendix B.1). By depicting the curve of $R^2(\alpha)$ as a function of α, it is possible to get some insight into how to customize the parameter α for the data at hand. The general idea is that we may accept a loss in terms of $R^2(\alpha)$ providing that this loss does not exceed a reasonable threshold. By way of comparison, it is worth noting that the squared coefficient of correlation associated to Ridge regression, $R^2(\kappa)$, can be written as:

$$R^2(\kappa) = \frac{\left(\sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j+\kappa} (\boldsymbol{u}_j^T \boldsymbol{y})^2\right)^2}{\boldsymbol{y}^T\boldsymbol{y} \sum_{j=1}^{p} \left(\frac{\lambda_j}{\lambda_j+\kappa}\right)^2 (\boldsymbol{u}_j^T \boldsymbol{y})^2}$$

We can show that $R^2(\kappa)$ is also a decreasing function of κ (appendix B.2)

3.5. Comparison of methods

The concept of continuum regression was firstly formalized by Stone and Brooks (1990) and had since gained popularity. The rationale behind this strategy of analysis is to consider a family of regression estimators depending on a parmater that needs to be customized to the data at hand in order to improve the performance of the model in terms of stability, prediction ability... Generally, the proposed continuum strategies encompass ordinary least squares, PLS and Principal Components regression.

BP-regression draws from a procedure of continuum regression called "Continuum Power Partial Least Squares Regression" (Wise and Ricker, 1993; de Jong et al., 2001). This consists in performing a PLS regression of $\boldsymbol{y}$ on $\mathbf{X}^{\alpha} = \mathbf{U}\mathbf{\Lambda}^{\alpha/2}\mathbf{V}^T$ with α varying between 0 and 1. By comparison, our approach pertains to the biased regression framework and does not involve the derivation of latent components as it is the case for PLS regression.

From another stand point, we stated above that Ridge regression amounts to augmenting the eigenvalues, λ_j $(j = 1, ..., p)$ of $\mathbf{X}^T\mathbf{X}$ by a positive constant, κ. This was readily generalized to a strategy of analysis called Generalized Ridge regression where each eigenvalue λ_j $(j = 1, ..., p)$ is augmented by a specific constant, κ_j (Hoerl and Kennard, 1970). This method of analysis seems to be more intuitively appealing than Ridge regression since some directions in the space spanned by the predictors are more in need of a cure from the effect of multicollinearity than others. However, the problem of selecting the appropriate parameters κ_j $(j = 1, ..., p)$ becomes even more acute than when dealing with a single constant (*i.e.* Ridge regression). We show that BP-regression stands at a mid-point between Ridge regression and Generalized Ridge regression since

each eigenvalue is somehow augmented differentially; the smaller eigenvalues being augmented by a quantity larger than those associated with larger eigenvalues. Yet, only one regularization parameter is actually involved.

As noted above, by using BP-regression, each eigenvalue λ_j $(j = 1, ..., p)$ is transformed into $\lambda_j^{1-\alpha}$ $(j = 1, ..., p)$. Because of the pretreatment that we have applied to $\mathbf{X}$, we have $0 < \lambda_j \leq 1$. Therefore, λ_j can be written as $\lambda_j = 1 - h_j$ where $h_j = 1 - \lambda_j$. We have $0 \leq h_j < 1$ and $h_1 \leq h_2 ... \leq h_p$. It follows that $\lambda_j^{1-\alpha} = (1 - h_j)^{1-\alpha}$ can be approximated by $1-(1-\alpha)h_j = \lambda_j + \alpha h_j$. This amounts to augmenting each eigenvalue, λ_j $(j = 1, ..., p)$, by $k_j = \alpha h_j$ $(j = 1, ..., p)$. The smaller an eigenvalue is, the larger is its associated regularization quantity.

4. Choice of an appropriate α

BP-regression yields a family of estimators indexed by $\alpha \in [0, 1]$. From a practical point of view, the model that may be eventually retained can be obtained by means of a validation technique such as cross-validation. This consists in scanning values of α between 0 and 1 and selecting the parameter, α^*, that corresponds to the minimum of predicted residual error sum of squares (PRESS) statistic or some connected statistic such as the root mean squared error (RMSE). In the illustrations discussed below, we performed a k fold cross-validation which consists in partitioning the dataset in k segments. Thereafter, each segment is left aside and the model parameters are estimated using the remaining segments. Eventually, the segment that was left aside is used for the validation of the model. This procedure is reiterated by setting aside each segment in turn. The particular case where there are as many segments as observations is referred to leave-one-out cross-validation. For more details regarding this strategy of selection of the tuning parameter, we refer to Stone and Brooks (1990).

The plot of the PRESS or RMSE statistics against α is of paramount interest since it may suggest a whole range of appropriate values, α. The general idea is that the global minimum corresponding to α^* often results in overfitting. A lower value of α than α^* may be more appropriate providing that its associated PRESS or RMSE statistics are not significantly larger than those of α^*. The common idea is to choose a regularization parameter where the PRESS curve 'flattens out' (Varmuza and Filzmoser, 2009).

5. Illustration

5.1. Orange juice

The first example relates to an ^{1}H NMR spectroscopy study on orange juice authentication (Vigneau and Thomas, 2012). The dataset $\mathbf{X}$ is composed of 480 variables and 150 observations. This is a situation where the number of variables exceeds the number of observations. Some of the observations are a mix of pure orange juice and Clementine juice. The response variable $\boldsymbol{y}$ to be predicted is the percentage of Clementine juice. The dataset can be found in Vigneau and Chen (2014). We performed both BP-regression

and Ridge regression on these data. Figure 1 shows the evolution of the coefficient of variation associated with the eigenvalues of the matrices $(\mathbf{X}^T\mathbf{X})^{1-\alpha}$ (left) and $\mathbf{X}^T\mathbf{X}+\kappa\mathbf{I}$ (right).

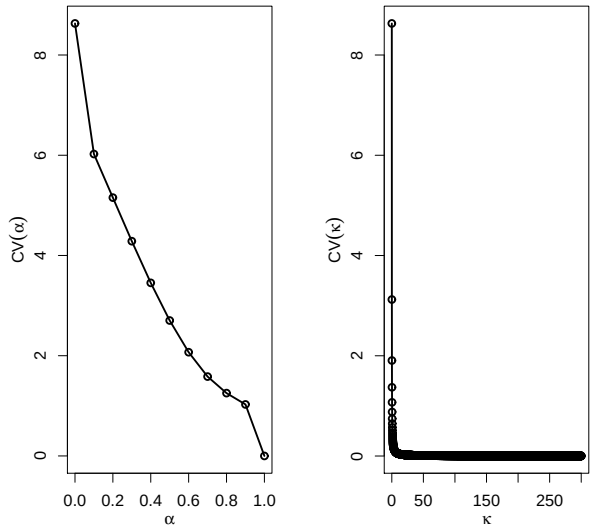

Figure 1: Orange juice: Evolution of the coefficient of variation associated with the eigenvalues of the matrix $(\mathbf{X}^T\mathbf{X})^{1-\alpha}$ (left) and $\mathbf{X}^T\mathbf{X}+\kappa I$ (right) with α between 0 and 1, and κ between 0 and 300.

For BP-regression, the coefficient of variation decreases steadily with α. It is equal to 1 for α around 0.9. With Ridge regression, the coefficient of variation decreases abruptly with κ: it starts with a value around 8.6 for κ equal to 0 and reaches the value 1 with κ around 0.4.

Figure 2 shows the decrease of the squared correlation coefficient between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\alpha$ in function of α. This curve decreases almost linearly with α, starting with a value of 1 with $\alpha = 0$ and becomes as small as 0.5 for $\alpha = 1$. By way of comparison, figure 2 also depicts the decrease of the correlation coefficient between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\kappa$ obtained by means of Ridge regression. We can see in this latter figure, that the coefficient of correlation decreases sharply for small values of κ.

A three fold cross-validation (CV) was applied to compare the prediction ability of BP-regression and Ridge regression. Figure 3 shows the variation of the Root Mean Squared Error associated with the three fold CV (RMSECV) for BP-regression (left) and Ridge regression (right) in function of α and κ, respectively. Globally, the RMSECV associated with BP-regression reaches smaller values than that of Ridge regression. In BP-regression, the minimum value is equal to 7.66 and it is achieved for $\alpha = 0.03$ whereas, with Ridge regression, the minimum value of the RMSECV is around 7.94 for $\kappa = 0$. In order to better substantiate this finding, we run fifty times the three fold CV,

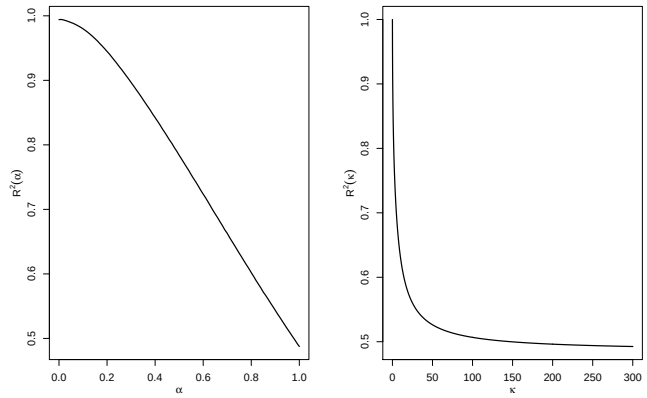

Figure 2: Orange juice: Evolution of the squared correlation coefficients between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\alpha$ in function of α (left) and between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\kappa$ in function of κ (right).

and at each time we selected the minimum value of the RMSECV for both BP-regression and Ridge regression. Figure 4 depicts the results associated to these fifty repetitions of the three fold CV. We can see that BP-regression outperforms Ridge regression.

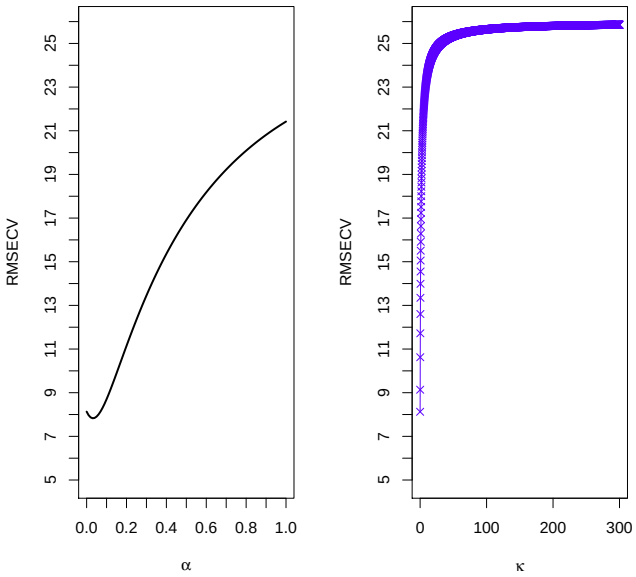

Figure 3: Orange juice: comparison of the RMSE associated with a three fold cross-validation applied on BP-regression and Ridge regression in function of α ($0 \leq \alpha \leq 1$) and κ (between 0 and 300).

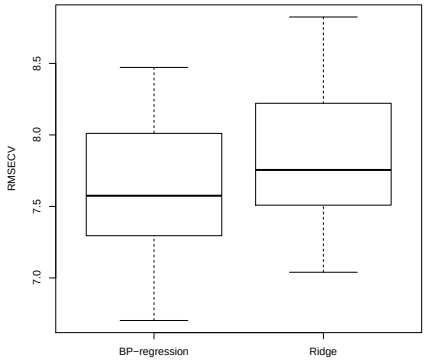

Figure 4: Orange juice: Box-plot of fifty minimum values of RMSE selected from fifty repetitions of a three fold cross-validation using BP-regression and Ridge regression models.

Figure 5 shows the regression coefficients estimated by means of BP-regression (α=0.03) and Ridge regression (κ=0). It is clear that the two vectors of coefficients are very similar to each other.

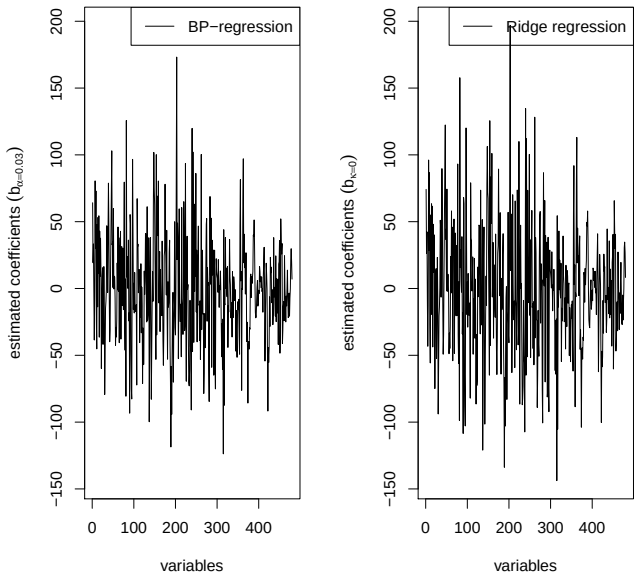

Figure 5: Orange juice: the regression coefficients associated to Ridge ($\kappa = 0$) and BP-regressions ($\alpha = 0.03$).

5.2. Economics dataset

The second example pertains to an economics study used in Gruber (1998). The data (table 1) consist of a dependent variable $\boldsymbol{y}$ which corresponds to the percentage of research and development expenditures in USA and four dependent variables representing the same percentages for France, Germany, Japan and the former Soviet Union, respectively for ten years extending from 1972 to 1986. The condition indices are $\eta_2 = 15.8$, $\eta_3 = 37.5$ and $\eta_4 = 152.2$ indicating the presence of near collinearity among the predictors.

YEAR	$\boldsymbol{y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$
1972	2.3	1.9	2.2	1.9	3.7
1975	2.2	1.8	2.2	2.0	3.8
1979	2.2	1.8	2.4	2.1	3.6
1980	2.3	1.8	2.4	2.2	3.8
1981	2.4	2.0	2.5	2.3	3.8
1982	2.5	2.1	2.6	2.4	3.7
1983	2.6	2.1	2.6	2.6	3.8
1984	2.6	2.2	2.6	2.6	4.0
1985	2.7	2.3	2.8	2.8	3.7
1986	2.7	2.3	2.7	2.8	3.8

Table 1: Economics dataset: total national Research and development expenditures between 1972 and 1986.

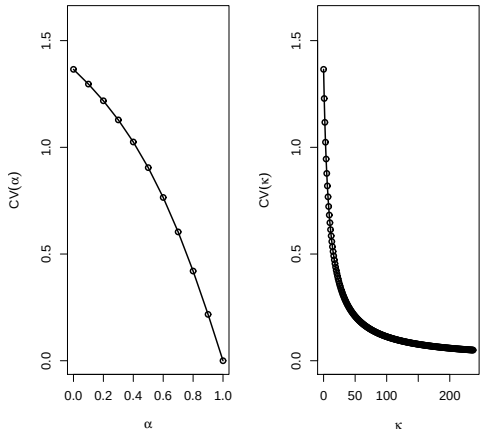

Figure 6: Economics data, evolution of the coefficient of variation associated with the eigenvalues of the matrices $(\mathbf{X}^T\mathbf{X})^{1-\alpha}$ (left) and $\mathbf{X}^T\mathbf{X} + \kappa\mathbf{I}$ (right) with α between 0 and 1, and κ between 0 and 250.

We start by investigating the effect of the parameters α and κ in reducing the collinearity among the $\mathbf{X}$-variables using the coefficient of variation. Figure 6 (left) shows the evolution of the coefficient of variation associated with the eigenvalues of $(\mathbf{X}^T\mathbf{X})^{1-\alpha}$ in function of α (α varies between 0 and 1). Figure 6 (right) shows the evolution of the coefficient of variation associated with the eigenvalues of $\mathbf{X}^T\mathbf{X} + \kappa\mathbf{I}$ (κ varies between 0 and 250). With BP-regression, the coefficient of variation decreases steadily with α, it reaches the value 1 for α equals to 0.4. For Ridge regression, the coefficient of variation decreases sharply for small values of κ then, starting from κ around 10, it decreases slowly. It is equal to 1 for κ around 3.

Figure 7 depicts the squared coefficient of correlation obtained by means of BP-regression, $R^2(\alpha)$, (Figure 7, left) and Ridge Regression, $R^2(\kappa)$ (Figure 7, right). The squared coefficient of correlation $R^2(\alpha)$, decreases very smoothly from 0.975 ($\alpha = 0$) to 0.945 ($\alpha = 1$). This entails that a value of α around 1 will not result in a substantial decrease of $R^2(\alpha)$. By way of comparison, the coefficient of correlation $R^2(\kappa)$ decreases sharply and reach the smallest value (0.945) with the first values of κ.

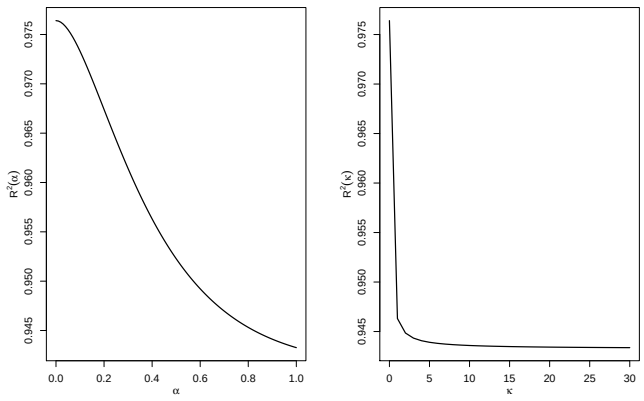

Figure 7: Economics data: Evolution of the squared correlation coefficient between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\alpha$ in function of α (left) and between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\kappa$ in function of κ (right).

Since the number of rows is small, a leave-one-out (LOO) cross-validation was performed and the RMSECV was computed. The evolution of RMSECV curve as a function of α (figure 8) indicates that this statistic decreases and flattens out starting from $\alpha = 0.4$. Any value of α above 0.4 could be chosen to ensure an optimal RMSECV. By comparison, the curve associated to Ridge regression (figure 8) indicates that an optimal value for RMSECV is reached for κ around 0.06. The optimal value for RMSECV for both strategies were practically identical (around 0.057).

Table 2 shows the correlation coeffients of the dependant and the indepant variables together with the regression coefficients estimated by means of BP-regression (α=0.4) and Ridge regression (κ=0.006). It can be seen that Ordinary Least Squares (OLS) regression, leads to an inconsistency since the correlation between $\boldsymbol{y}$ and $\boldsymbol{x}_2$ is very high

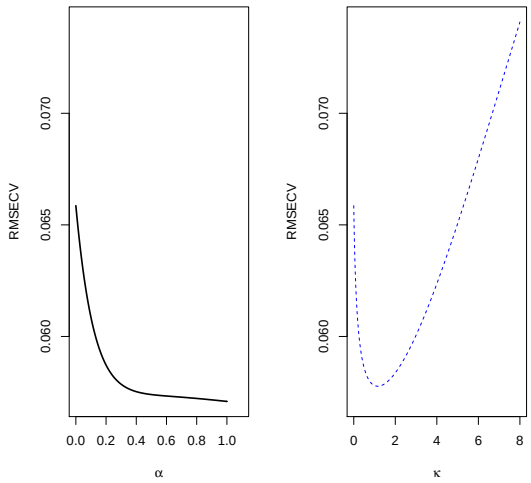

Figure 8: Economics data: comparison of the RMSE under LOO cross-validation between BP-regression (left) and Ridge regression (right).

and positive yet, the associated coefficient is negative. Both Ridge and BP-regressions yield consistant coefficients since these coefficients are all positive. Morover, the two vectors of coefficients bear a hight similarity to each other.

		Regression coefficients		
Dependant variables	Correlation with y	OLS	BP-regression	Ridge regression
x_1	0.9777	0.626	0.386	0.378
x_2	0.908	-0.115	0.179	0.094
x_3	0.956	0.287	0.405	0.281
x_4	0.348	0.025	0.064	0.071

Table 2: Correlation coefficients of the dependant and the indepant variables. Regression cofficients obtained by means of OLS, BP and Ridge regressions.

6. Conclusion

BP-regression is a biased regression strategy which shares several features with Ridge regression. Among these features, we single out: (i) it is a simple and straightforward regression method, (ii) it is a shrinkage procedure since the length of the vector of regression coefficients decreases as α increases, (iii) it depends on a single tuning parameter, (iv) it aims at achieving a profitable bias-variance trade off. Moreover, BP-regression seems to stand at a mid-point between Ridge regression and Generalized Ridge regression since, through a first order approximation, it turns out that BP-regression differentially augments the various eigenvalues of the predictors variance covariance matrix; yet it depends on a single regularization parameter, α.

The illustrations on the basis of two datasets seem to endorse the efficiency of the BP-regression since it seems that it yields a performance similar to that of Ridge regression.

Further investigations are needed to better clarify the merits of BP-regression over competing methods. For instance, it would be interesting to investigate whether this strategy of analysis could be supported by considerations pertaining to the Bayesian framework as it is the case for Ridge regression (Timothy, 2007).

Another line of investigation would be to extend BP-regression to a multidimensional setting where the aim is to predict a multivariate, $\mathbf{Y}$, from a set of predictors, $\mathbf{X}$.

Another extension of BP-regression which is worth investigating concerns the context of classification and discrimination. This would provide an alternative to the so-called regularized discriminant analysis (Friedman, 1989). The regularization proposed in this latter method draws from Ridge regression and, therefore, should easily be adapted to a regularization akin to BP-regression.

Ridge regression was not designed to select a subset of variables. However, as stated above, this method of analysis shrinks the regression coefficient estimates toward zero as κ increases. Some authors advocated using this property to discard those variables whose coefficients become very close to 0 (Draper and Smith, 1998). BP-regression shares the same property and thus can highlight those variables whose coefficients become very small. A better method than Ridge and BP-regressions to achieve a selection of variable is LASSO (Tibshirani, 1996). This method of analysis yields sparse models since the coefficients of the variables which are deemed to be unimportant are set to zeo. The Elastinet method (Zou and Hastie, 2005) stands as a combination of both Ridge and Lasso regression and is likely to lead to a better performance. This hints to a direction of research that combines BP-regression with a strategy of analysis for sparcity.
Throughout the paper, we have compared BP-regression with Ridge regression both from a conceptual point of view and in terms of their performance. Both methods are biased regression methods which are based on the principle of trading the high variability of the parameters for a (small) bias. Both methods seem to have the same performance in terms of prediction ability. Further research is indicated to better highlight the merits, if any, of BP-regression over Ridge regression.

7. Appendices

Prerequisite: Chebyshev's Weighted sum inequality (Cvetkovski, 2012)
We start by recalling Chebyshev's weighted sum inequality.

Let us consider p scalars (a_j), p scalars (b_j) and p positive scalars (q_j) $(j = 1, ..., p)$ such that

$$a_1 \geq a_2 ... \geq a_p$$
$$b_1 \geq b_2 ... \geq b_p$$
$$q_1 + q_2 + ... + q_p = 1$$

Then, we have:

$$(\sum_{k=1}^{p} a_k q_k)(\sum_{k=1}^{p} b_k q_k) \leq \sum_{k=1}^{p} a_k b_k q_k$$

A. Coefficient of variation

We aim at proving that $CV(\alpha)$ is a decreasing function of α. In order to study the evolution of $CV(\alpha)$ as a function of α, we consider $K(\alpha) = (CV(\alpha))^2$. Both $K(\alpha)$ and $CV(\alpha)$ have the same variation since $CV(\alpha) \geq 0$. We have:

$$\begin{aligned} K(\alpha) &= \frac{p\sum_{j=1}^{p} \lambda_j^{2(1-\alpha)} - (\sum_{j=1}^{p} \lambda_j^{1-\alpha})^2}{(\sum_{j=1}^{p} \lambda_j^{1-\alpha})^2} \\ &= p\frac{\sum_{j=1}^{p} \lambda_j^{2(1-\alpha)}}{(\sum_{j=1}^{p} \lambda_j^{1-\alpha})^2} - 1 \end{aligned}$$

It follows:

$$K'(\alpha) = \frac{A}{B} \times \frac{[-2(\sum_{j=1}^{p} ln(\lambda_j)\lambda_j^{2(1-\alpha)})(\sum_{j=1}^{p} \lambda_j^{1-\alpha}) + 2(\sum_{j=1}^{p} \lambda_j^{2(1-\alpha)})(\sum_{j=1}^{p} ln(\lambda_j)\lambda_j^{1-\alpha})]}{(\sum_{j=1}^{p} \lambda_j^{1-\alpha})^2}$$

with $A = p(\sum_{j=1}^{p} \lambda_j^{1-\alpha})$ and $B = (\sum_{j=1}^{p} \lambda_j^{1-\alpha})^2$ which are both positive. Thus, $K'(\alpha) \leq 0$ iff:

$$\frac{(\sum_{j=1}^{p} \lambda_j^{2(1-\alpha)})(\sum_{j=1}^{p} ln(\lambda_j)\lambda_j^{1-\alpha})}{(\sum_{j=1}^{p} \lambda_j^{1-\alpha})^2} \leq \frac{(\sum_{j=1}^{p} ln(\lambda_j)\lambda_j^{2(1-\alpha)})(\sum_{j=1}^{p} \lambda_j^{1-\alpha})}{(\sum_{j=1}^{p} \lambda_j^{1-\alpha})^2}$$

Or equivalently:

$$\left(\sum_{j=1}^{p} \lambda_j^{1-\alpha} \frac{\lambda_j^{1-\alpha}}{\sum_{j=1}^{p} \lambda_j^{1-\alpha}}\right)\left(\sum_{j=1}^{p} ln(\lambda_j) \frac{\lambda_j^{1-\alpha}}{\sum_{j=1}^{p} \lambda_j^{1-\alpha}}\right) \leq \left(\sum_{j=1}^{p} ln(\lambda_j)\lambda_j^{1-\alpha} \frac{\lambda_j^{1-\alpha}}{\sum_{j=1}^{p} \lambda_j^{1-\alpha}}\right)$$

Let us denote by $q_j = \frac{\lambda_j^{1-\alpha}}{\sum_{j=1}^{p} \lambda_j^{1-\alpha}}$. Obviously, we have $q_j \geq 0$ and $\sum_{j=1}^{p} q_j = 1$

The inequality above can be written as

$$\left(\sum_{j=1}^{p}\lambda_j^{1-\alpha}q_j\right)\left(\sum_{j=1}^{p}ln(\lambda_j)q_j\right)\leq\left(\sum_{j=1}^{p}ln(\lambda_j)\lambda_j^{1-\alpha}q_j\right)$$

Since $\lambda_1\geq\lambda_2...\geq\lambda_p$ and $0\leq\alpha\leq 1$, we have $(\lambda_1)^{1-\alpha}\geq(\lambda_2)^{1-\alpha}\geq...\geq(\lambda_p)^{1-\alpha}$ and $ln(\lambda_1)\geq ln(\lambda_2)\geq...\geq ln(\lambda_p)$.

It follows that the inequality above holds by a direct application of the Chebyshev's weighted sum inequality (see Prerequisite). Therefore $K^{'}(\alpha)\leq 0$. This shows that the coefficient of variation is a decreasing function of α.

B. Correlation coefficient

B.1. BP-regression correlation coefficient: $R^2(\alpha)$

We aim at proving that the squared coefficient of correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\alpha$, $R^2(\alpha)$, is a decreasing function of α. We recall that $R^2(\alpha)$ is given by:

$$R^2(\alpha)=\frac{\left(\sum_{j=1}^{p}\lambda_j^{\alpha}(\mathbf{u}_j^T\boldsymbol{y})^2\right)^2}{\boldsymbol{y}^T\boldsymbol{y}\sum_{j=1}^{p}\lambda_j^{2\alpha}(\mathbf{u}_j^T\boldsymbol{y})^2}$$

with $\mathbf{u}_j$ the j^{th} column of the matrix $\mathbf{U}$ associated with the singular value decomposition of $\mathbf{X}=\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{V}^T$. $R^2(\alpha)$ have the same variation as:

$$f(\alpha)=\frac{\left(\sum_{j=1}^{p}\lambda_j^{\alpha}a_j\right)^2}{\sum_{j=1}^{p}\lambda_j^{2\alpha}a_j}$$

Where $a_j=(\mathbf{u}_j^Ty)^2$. By deriving $f(\alpha)$ with respect to α, we have:

$$f^{'}(\alpha)=A\left[\left(\sum_{j=1}^{p}a_j(ln\lambda_j)\lambda_j^{\alpha}\right)\left(\sum_{j=1}^{p}\lambda_j^{2\alpha}a_j\right)-\left(\sum_{j=1}^{p}\lambda_j^{\alpha}a_j\right)\left(\sum_{j=1}^{p}(ln\lambda_j)\lambda_j^{2\alpha}a_j\right)\right]$$

with $A=\frac{2\left(\sum_{j=1}^{p}\lambda_j^{\alpha}a_j\right)}{\left(\sum_{j=1}^{p}\lambda_j^{2\alpha}a_j\right)^2}\geq 0$. Thus $f^{'}(\alpha)\leq 0$ iff

$$\left(\sum_{j=1}^{p}a_j(ln\lambda_j)\lambda_j^{\alpha}\right)\left(\sum_{j=1}^{p}\lambda_j^{2\alpha}a_j\right)-\left(\sum_{j=1}^{p}\lambda_j^{\alpha}a_j\right)\left(\sum_{j=1}^{p}(ln\lambda_j)\lambda_j^{2\alpha}a_j\right)\leq 0$$

Dividing the two members of this last inequality by $\left(\sum_{j=1}^{p}a_j\lambda_j^{\alpha}\right)^2$, it follows that $f^{'}(\alpha)\leq 0$ iff:

$$\left(\sum_{j=1}^{p}(ln\lambda_j)q_j\right)\left(\sum_{j=1}^{p}\lambda_j^{\alpha}q_j\right)-\left(\sum_{j=1}^{p}(ln\lambda_j)\lambda_j^{\alpha}q_j\right)\leq 0$$

where $q_j = \frac{a_j \lambda_j^\alpha}{\sum_{j=1}^p a_j \lambda_j^\alpha}$ ($q_j \geq 0$ and $\sum_{j=1}^p q_j = 1$). Since, $\lambda_1 \geq \lambda_2 ... \geq \lambda_p$ and $0 \leq \alpha \leq 1$, we have $(\lambda_1)^\alpha \geq (\lambda_2)^\alpha \geq ... \geq (\lambda_p)^\alpha$ and $ln(\lambda_1) \geq ln(\lambda_2) \geq ... \geq ln(\lambda_p)$. Therefore by applying the Chebyshev's weighted sum inequality (see Prerequisite) we have:

$$\left(\sum_{j=1}^p (ln\lambda_j) q_j\right) \left(\sum_{j=1}^p \lambda_j^\alpha q_j\right) - \left(\sum_{j=1}^p (ln\lambda_j) \lambda_j^\alpha q_j\right) \leq 0$$

which means that $R^2(\alpha)$ is a decreasing function of α.

B.2. Ridge regression correlation coefficient: $R^2(\kappa)$

We aim at proving that the squared coefficient of correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_\kappa$, $R^2(\kappa)$, is a decreasing function of α. We recall that $R^2(\kappa)$ is given by:

$$R^2(\kappa) = \frac{\left(\sum_{j=1}^p \frac{\lambda_j}{\lambda_j+\kappa} (\boldsymbol{u}_j^T \boldsymbol{y})^2\right)^2}{\boldsymbol{y}^T \boldsymbol{y} \sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j+\kappa}\right)^2 (\boldsymbol{u}_j^T \boldsymbol{y})^2}$$

with $\mathbf{u}_j$ the j^{th} column of the matrix $\mathbf{U}$ associated with the singular value decomposition of $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^T$. $R^2(\kappa)$ have the same variation as:

$$g(\kappa) = \frac{\left(\sum_{j=1}^p \frac{\lambda_j}{\lambda_j+\kappa} a_j\right)^2}{\sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j+\kappa}\right)^2 a_j}$$

Where $a_j = (\mathbf{u}_j^T y)^2$. By deriving $g(\kappa)$ with respect to κ, we have $g^{'}(\kappa)$ equal to:

$$A\left[\left(\sum_{j=1}^p a_j (\frac{\lambda_j}{\lambda_j+\kappa})\right)\left(\sum_{j=1}^p \frac{\lambda_j}{(\lambda_j+\kappa)} \frac{\lambda_j}{(\lambda_j+\kappa)^2} a_j\right) - \left(\sum_{j=1}^p \frac{\lambda_j}{(\lambda_j+\kappa)^2} a_j\right)\left(\sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j+\kappa}\right)^2 a_j\right)\right]$$

with $A = \frac{2\left(\sum_{j=1}^p \frac{\lambda_j}{\lambda_j+\kappa} a_j\right)}{\left(\sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j+\kappa}\right)^2 a_j\right)^2} \geq 0$. Thus $g^{'}(\kappa) \leq 0$ iff

$$\left(\sum_{j=1}^p a_j \frac{\lambda_j}{\lambda_j+\kappa}\right)\left(\sum_{j=1}^p \frac{\lambda_j}{(\lambda_j+\kappa)} \frac{\lambda_j}{(\lambda_j+\kappa)^2} a_j\right) - \left(\sum_{j=1}^p \frac{\lambda_j}{(\lambda_j+\kappa)^2} a_j\right)\left(\sum_{j=1}^p \left(\frac{\lambda_j}{\lambda_j+\kappa}\right)^2 a_j\right) \leq 0$$

Dividing the two members of this last inequality by $\left(\sum_{j=1}^p a_j \frac{\lambda_j}{(\lambda_j+\kappa)^2}\right)^2$, it follows that $g^{'}(\kappa) \leq 0$ iff:

$$\left(\sum_{j=1}^p (\lambda_j+\kappa) q_j\right)\left(\sum_{j=1}^p \frac{\lambda_j}{\lambda_j+\kappa} q_j\right) - \left(\sum_{j=1}^p \lambda_j q_j\right) \leq 0$$

where $q_j = \frac{a_j \frac{\lambda_j}{(\lambda_j+\kappa)^2}}{\sum_{j=1}^{p} a_j \frac{\lambda_j}{\lambda_j+\kappa}}$ ($q_j \geq 0$ and $\sum_{j=1}^{p} q_j = 1$). Since, $\lambda_1 \geq \lambda_2 ... \geq \lambda_p$ and $\kappa \geq 0$, we have $\lambda_1 + \kappa \geq \lambda_2 + \kappa \geq ... \geq \lambda_p + \kappa$ and $\frac{\lambda_1}{\lambda_1+\kappa} \geq \frac{\lambda_2}{\lambda_2+\kappa} \geq ... \geq \frac{\lambda_p}{\lambda_p+\kappa}$. Therefore by applying the Chebyshev's weighted sum inequality (see Prerequisite) we have $g^{'}(\kappa) \leq 0$ which means that $R^2(\kappa)$ is a decreasing function of κ.

References

Abdi, H. (2010). Congruence: Congruence coefficient, rv coefficient, and mantel coefficient. *Encyclopedia of Research Design*, pages 222–229.

Belsley, D. A. (1991). A guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, 4(1):33–50.

Cvetkovski, Z. (2012). *Inequalities*. Springer Berlin Heidelberg.

de Jong, S. (1993). Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251 – 263.

de Jong, S. (1995). PLS shrinks. *Journal of Chemometrics*, 9(4):323–326.

de Jong, S., Wise, B. M., and Ricker, N. L. (2001). Canonical partial least squares and continuum power regression. *Chemiometrics*, 15:85–100.

Draper, N. R. and Smith, H. (1998). Bias in regression estimates, and expected values of mean squares and sums of squares. In *Applied Regression Analysis*, pages 235–242. John Wiley & Sons, Inc.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numer. Math.*, 14(5):403–420.

Gruber, M. H. J. (1998). *Improving efficiency by shrinkage : the James-Stein and ridge regression estimators*. New York : Marcel Dekker. Includes bibliographical references (pages 591-617) and indexes.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, second edition.

Hoerl, A. E., Kannard, R. W., and Baldwin, K. F. (1975). Ridge regression:some simulations. *Communications in Statistics*, 4(2):105–123.

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

Kazi-Aoual, F., Hitier, S., Sabatier, R., and Lebreton, J.-D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis*, 20(6):643–656.

Stone, M. and Brooks, R. J. (1990). Continuum Regression: Cross-Validated sequentially

constructed prediction embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. *Journal of the Royal Statistical Society, Series B*, 52(2):237–269.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Timothy, D. (2007). A bayesian framework for multimodel regression. *Journal of Climate*, 20(12):2810–2826.

Varmuza, K. and Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Taylor & Francis, CRC Press: Boca Raton, FL, USA.

Vigneau, E. and Chen, M. (2014). *ClustVarLV: Clustering of variables around Latent Variables*. R package version 1.2.

Vigneau, E. and Thomas, F. (2012). Model calibration and feature selection for orange juice authentication by ^{1}H NMR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 117(0):22–30.

Wise, B. M. and Ricker, N. L. (1993). Identification of finite impulse response models with continuum regression. *Journal of Chemometrics*, 7(1):1–14.

Worsley, K. J. and Friston, K. J. (1995). Analysis of fMRI Time-Series Revisited–again. *NeuroImage*, 2(3):173–181.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.