International Journal of Statistics and Applications 2019, 9(5): 160-169
DOI: 10.5923/j.statistics.20190905.05

An Account of Principal Components Analysis and Some Cautions on Using the Correct Formulas and the Correct Procedures in SPSS

Dimitris Hatzinikolaou[1,*], Katerina Katsarou[2]

[1]University of Ioannina, Department of Economics, Ioannina, Greece, and Hellenic Open University, 18 Aristotelous, Patras, Greece
[2]Technische Universität Berlin, Service-centric Networking, Telecom Innovation Laboratories, Berlin, Germany

Abstract The paper provides an account of the principal components regression (PCR) and uses some examples from the literature to illustrate the following: (1) the importance of PCR in the presence of multicollinearity; (2) some cautions on its correct implementation in SPSS, as some researchers use it improperly; (3) the use of the correct formulas, in accordance with the choice of scaling the variables; (4) the choice of principal components to be dropped; (5) the conditions for the PCR to outperform ordinary least squares, in the minimum mean-square-error sense; and (6) the robustness of the estimates to substantial changes in the sample.

Keywords Multicollinearity, Principal Components, MSE, SPSS

1. Introduction

A problem that is frequently encountered in applied regression analysis is multicollinearity, i.e., high correlation among the explanatory variables (regressors), which causes the estimates to be imprecise, thus leading to erroneous inferences and imprecise forecasts. As Jackson (2003, p. 276) notes, a "salvation in some thorny regression problem" of this type may be achieved by using principal components (PCs) analysis. Unfortunately, however, some researchers often fail to implement it properly, despite the strong warnings that exist in the literature; see, e.g., Jolliffe (1982) and Hadi and Ling (1998).

For example, in a well cited paper, Liu, et al. (2003) drop the regressors that are not statistically significant at the 5-percent level *before* applying the principal components regression (PCR). This can lead to a wrong model, however, thus causing an omitted-variable bias, when in fact multicollinearity is to blame for the low values of the *t*-statistics, which, therefore, should not be taken to mean that the corresponding regressors are irrelevant (Chatterjee and Hadi, 2006, p. 299).

Also, Liu, et al. (2003) erroneously interpret the "component matrix" (produced by the SPSS Factor Analysis procedure) as the matrix of eigenvectors, which can be obtained by writing a program, or by modifying the "component matrix" (see section 3). Apparently, as Sharma (1996, p. 58) notes, this confusion often arises in packages where principal component analysis is embedded in the factor analysis procedure.

Finally, instead of using only a subset of the PCs in the model, Liu, et al. (2003) use *all* of them. Unless other errors are made, however, this procedure will return the original regression, thus nullifying the whole effort of implementing the PCR. Despite these errors, researchers still use Liu, et al. (2003) as a basic reference for the PCR, however; see, e.g., Ding, Ma, and Wang (2018) and Tran, et al. (2018).

The present paper provides an account of the PCR (Section 2) and shows (in Section 3) step-by-step how to implement it correctly in SPSS by replicating an example from Chatterjee and Hadi (2006). We choose to replicate an example from a standard textbook, rather than providing our own, in order to convince the reader that the steps taken here are the correct ones. The example demonstrates the importance of the PCR, as it produces estimates that have the expected signs and are statistically significant at the 1-percent level, whereas the ordinary least squares (OLS) regression fails in that respect. This result becomes stronger when we update the sample substantially. In addition, in Sections 2 and 4, we use two other data sets, one from Chatterjee and Hadi (2006) and another from Myers (1990), to illustrate other important aspects of the PCR, namely, the use of the correct formulas, in accordance with the choice of scaling the variables; the choice of PCs to be dropped from the PCR; and the conditions for the PCR to outperform OLS

* Corresponding author:
dhatzini@uoi.gr (Dimitris Hatzinikolaou)
Published online at http://journal.sapub.org/statistics

in the sense of the minimum mean square error (MSE) criterion. Section 5 provides a summary.

2. An Account of the PCR and Some Measures of Multicollinearity

2.1. Estimation of the Coefficients of Interest via the PCR

Consider the standard linear regression model with k regressors, X_1, ..., X_k,

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\mathbf{y}$ is a $n\times 1$ response vector; $\mathbf{X}$ is a $n\times(k+1)$ regressor matrix, whose first column is a vector of 1's; $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ is a $(k+1)\times 1$ vector of coefficients, where β_0 is the constant term (or intercept) and β_1, ..., β_k are the slopes (usually the only coefficients of interest) collected in the slope vector $\boldsymbol{\beta}^s = (\beta_1, ..., \beta_k)'$; $\boldsymbol{\varepsilon}$ is a $n\times 1$ vector of errors; and n is the number of observations. For the i-th observation, Equation (1) is

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i. \quad (1a)$$

Under the classical assumptions, the ordinary least squares (OLS) estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$, is the best linear unbiased estimator (BLUE).

To implement the PCR, we first write (1) in terms of standardized variables,

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\theta} + \tilde{\boldsymbol{\varepsilon}}, \quad (2)$$

where $\tilde{\mathbf{y}}$ is the $n\times 1$ vector of the standardized response variable, whose i-th element is defined as $\tilde{y}_i = (y_i - \bar{y})/s_y$, where $\bar{y}$ is the sample mean of y and s_y is its standard deviation, so that $\tilde{y}$ has zero mean and unit standard deviation; $\tilde{\mathbf{X}}$ is a $n\times k$ matrix (without a column of 1's) whose ij-th element is defined as $\tilde{X}_{ij} = (X_{ij} - \bar{X}_j)/s_j$, where $\bar{X}_j$ is the sample mean of X_j and s_j is its sample standard deviation ($s_j > 0$), $j = 1, ..., k$, $i = 1, ..., n$; $\boldsymbol{\theta}$ is $k\times 1$; and $\tilde{\boldsymbol{\varepsilon}}$ is a $n\times 1$ error vector whose i-th (unobserved) value is $\tilde{\varepsilon}_i = (\varepsilon_i - \bar{\varepsilon})/s_y$, where $\bar{\varepsilon}$ is the sample mean of ε. We assume that Equation (2) is correctly specified; the X's are stochastic, but strictly exogenous, implying that $E(\tilde{\varepsilon}_i \mid \mathbf{X}) = 0$, $i = 1, ..., n$; and $Cov(\tilde{\boldsymbol{\varepsilon}} \mid \mathbf{X}) = E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' \mid \mathbf{X}) = \sigma^2\mathbf{I}_n$, where $\sigma^2 > 0$ and $\mathbf{I}_n$ is the identity matrix of order n. Note that the literature on the PCR almost invariably assumes non-stochastic regressors, but here we adopt the assumption of stochastic regressors, because: (1) it is more realistic; (2) it renders the results more naturally interpretable, in that they are viewed as conditional on the observed values of the regressors; and (3) it has been adopted by famous modern econometrics textbooks, such as Hayashi (2000), Stock and Watson (2003), and Wooldridge (2006).

Under these assumptions, the OLS estimator of $\boldsymbol{\theta}$, denoted as $\hat{\boldsymbol{\theta}}_{OLS}$, is BLUE. The vector $\boldsymbol{\theta}$ is related to the slope vector $\boldsymbol{\beta}^s$ as follows: $\theta_j = (s_j/s_y)\beta_j$, $j = 1, ..., k$, or

$$\boldsymbol{\theta} = s_y^{-1}\mathbf{S}\boldsymbol{\beta}^s, \quad (3a)$$

where $\mathbf{S} = \text{diag}(s_1, ..., s_k)$ is a $k\times k$ diagonal matrix, with s_1, ..., s_k in its main diagonal, so it is positive definite (Hadley, 1961, p. 260). Thus, $\beta_j = (s_y/s_j)\theta_j$, $j = 1, ..., k$, so we can estimate the βs through the θs (Chatterjee and Hadi, 2006, pp. 242 and 260):

$$\hat{\beta}_j = (s_y/s_j)\hat{\theta}_j, \quad \hat{\beta}_0 = \bar{y} - \Sigma_{j=1}^{k}\hat{\beta}_j\bar{x}_j, \quad j = 1, ..., k. \quad (3b)$$

Principal components are k orthogonal variables, C_1, ..., C_k, defined as the following linear combinations of the standardized regressors:

$$\mathbf{C} = \tilde{\mathbf{X}}\mathbf{V}. \quad (4)$$

Here, $\mathbf{V}$ is a $k\times k$ matrix of the eigenvectors of the correlation matrix of the regressors with the property $\mathbf{VV'} = \mathbf{I}$, hence $\mathbf{V'} = \mathbf{V}^{-1}$ and $(\mathbf{V'})^{-1} = \mathbf{V}$, where $\mathbf{V'}$ is the transpose of $\mathbf{V}$. Thus, inserting $\mathbf{VV'}$ into (2) and using (4), Equation (2) can be restated in terms of the PCs, since $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\mathbf{VV'}\boldsymbol{\theta} + \tilde{\boldsymbol{\varepsilon}}$ can be written as

$$\tilde{\mathbf{y}} = \mathbf{C}\boldsymbol{\alpha} + \tilde{\boldsymbol{\varepsilon}}, \quad (5)$$

where $\boldsymbol{\alpha}$ is a $k\times 1$ vector of new coefficients, defined as $\boldsymbol{\alpha} = \mathbf{V'}\boldsymbol{\theta}$, hence $\boldsymbol{\theta} = \mathbf{V}\boldsymbol{\alpha}$. Thus, the OLS estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are related as follows:

$$\hat{\boldsymbol{\theta}}_{OLS} = \mathbf{V}\hat{\boldsymbol{\alpha}}. \quad (6)$$

To prove (6), post-multiply (4) by $\mathbf{V'}$, use $\mathbf{VV'} = \mathbf{I}$, to get $\tilde{\mathbf{X}} = \mathbf{CV'}$, and note that $\hat{\boldsymbol{\theta}}_{OLS} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} = (\mathbf{VC'CV'})^{-1}\mathbf{VC'}\tilde{\mathbf{y}} = (\mathbf{V'})^{-1}(\mathbf{C'C})^{-1}\mathbf{V}^{-1}\mathbf{VC'}\tilde{\mathbf{y}} = \mathbf{V}(\mathbf{C'C})^{-1}\mathbf{C'}\tilde{\mathbf{y}} = \mathbf{V}\hat{\boldsymbol{\alpha}}$, since it is obvious from (5) that, when all the (k) PCs are retained, the OLS estimator of $\boldsymbol{\alpha}$ is $\hat{\boldsymbol{\alpha}} = (\mathbf{C'C})^{-1}\mathbf{C'}\tilde{\mathbf{y}}$, which, under the classical assumptions, is BLUE. Pre-multiplying (6) by $\mathbf{V'}$ and using $\mathbf{V'} = \mathbf{V}^{-1}$ yields

$$\hat{\boldsymbol{\alpha}} = \mathbf{V'}\hat{\boldsymbol{\theta}}_{OLS}. \quad (7)$$

So far, we have included all the (k) PCs, so the PCR results in the same BLUE $\hat{\boldsymbol{\theta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{OLS}$ that would be obtained by applying OLS to Equations (1)-(2), so it is of no practical interest, as the idea of the PCR is to escape from these imprecise estimators in the presence of multicollinearity. In practice, we always want to drop d PCs whose variances are close to zero or have no predictive power for $\tilde{y}$ in (5), and hence also drop the corresponding d columns of $\mathbf{V}$ and the corresponding d elements of $\hat{\boldsymbol{\alpha}}$. Let $\hat{\boldsymbol{\theta}}_{PC}$ and $\hat{\boldsymbol{\beta}}_{PC}$ denote the resulting PCR estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, which are biased (Myers, 1990, p. 415). Thus, instead of (6), we want to have

a relation between $\hat{\boldsymbol{\theta}}_{PC}$ and the $k - d$ retained elements of $\hat{\boldsymbol{\alpha}}$, collected in the $(k\text{-}d)\times 1$ sub-vector $\hat{\boldsymbol{\alpha}}_{k-d}$. Let $\hat{\boldsymbol{\alpha}}_d$ denote the $d\times 1$ sub-vector of the d dropped elements of $\hat{\boldsymbol{\alpha}}$, and partition $\mathbf{V}$ as $\mathbf{V} = [\mathbf{V}_{k-d} \vdots \mathbf{V}_d]$, where $\mathbf{V}_{k-d}$ is the $k\times(k\text{-}d)$ sub-matrix of the k-d retained eigenvectors, and $\mathbf{V}_d$ is the $k\times d$ sub-matrix of the d dropped eigenvectors. Thus, (6) is replaced by

$$\hat{\boldsymbol{\theta}}_{PC} = \mathbf{V}_{k-d}\hat{\boldsymbol{\alpha}}_{k-d}. \tag{8}$$

Clearly, if we retain all the (k) PCs, i.e., if $d = 0$, then Equation (8) reduces to (6), i.e., $\hat{\boldsymbol{\theta}}_{PC} = \hat{\boldsymbol{\theta}}_{OLS}$, and hence, using Equation (3b), $\hat{\boldsymbol{\beta}}_{PC} = \hat{\boldsymbol{\beta}}_{OLS}$; see Chatterjee and Hadi (2006, p. 264, Table 10.3, and p. 231, Table 9.7) and Rawlings (1988, p. 360).

Finally, since $(\mathbf{C}'\mathbf{C})/(n\text{–}1) = \mathbf{V}'[(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})/(n\text{–}1)]\mathbf{V} = \boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_k)$ is the covariance matrix of the k PCs, where $\lambda_1, ..., \lambda_k$ are the eigenvalues of the correlation matrix of the regressors, $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})/(n-1)$ (Hadley, 1961, p. 248), it is useful to partition $\boldsymbol{\Lambda}$ as follows:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{k-d} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_d \end{bmatrix}, \tag{9}$$

where $\boldsymbol{\Lambda}_{k-d}$ and $\boldsymbol{\Lambda}_d$ are diagonal matrices of order k-d and d, respectively, whose elements in the main diagonal are the eigenvalues associated with the retained and the dropped PCs, respectively; the upper-right zero sub-matrix is $(k\text{-}d)\times d$, and the lower-left one is $d\times(k\text{-}d)$. Note that, since $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})/(n-1)$ is positive definite, it follows that $\lambda_1 > 0$, ..., $\lambda_k > 0$ (Goldberger, 1964, p. 34), so $\boldsymbol{\Lambda}$ is also positive definite, and so are $\boldsymbol{\Lambda}_{k-d}$ and $\boldsymbol{\Lambda}_d$. Applying the result of inverting a partitioned nonsingular matrix to (9) (Hadley, 1961, pp. 107-109), we can now write the OLS estimator of $\boldsymbol{\alpha}$ as follows:

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\alpha}}_{k-d} \\ \hat{\boldsymbol{\alpha}}_d \end{pmatrix} &= (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\tilde{\mathbf{y}} = \frac{1}{n-1}\boldsymbol{\Lambda}^{-1}\mathbf{V}'\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ &= \frac{1}{n-1}\begin{pmatrix} \boldsymbol{\Lambda}_{k-d}^{-1}\mathbf{V}_{k-d}'\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ \boldsymbol{\Lambda}_d^{-1}\mathbf{V}_d'\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \end{pmatrix}. \end{aligned} \tag{10}$$

2.2. Variance of $\hat{\boldsymbol{\theta}}_{PC}$ and $\hat{\boldsymbol{\beta}}_{PC}$, t-ratios, Bias, and Mean Square Error

Consider the variance-covariance matrix of $\hat{\boldsymbol{\theta}}_{PC}$, obtained from (8),

$$Cov(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) = \mathbf{V}_{k-d}\,Cov(\hat{\boldsymbol{\alpha}}_{k-d} \mid \tilde{\mathbf{X}})\mathbf{V}_{k-d}'. \tag{11}$$

But

$$\begin{aligned} Cov\begin{pmatrix} \hat{\boldsymbol{\alpha}}_{k-d} \\ \hat{\boldsymbol{\alpha}}_d \mid \tilde{\mathbf{X}} \end{pmatrix} &= \sigma^2(\mathbf{C}'\mathbf{C})^{-1} = \frac{\sigma^2}{n-1}\boldsymbol{\Lambda}^{-1} \\ &= \frac{\sigma^2}{n-1}\begin{bmatrix} \boldsymbol{\Lambda}_{k-d}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_d^{-1} \end{bmatrix} \end{aligned}, \tag{12}$$

so

$$Cov(\hat{\boldsymbol{\alpha}}_{k-d} \mid \tilde{\mathbf{X}}) = \frac{\sigma^2}{n-1}\boldsymbol{\Lambda}_{k-d}^{-1}, \tag{13a}$$

i.e.,

$$\begin{aligned} &Var(\hat{\alpha}_j \mid \tilde{\mathbf{X}}) = \sigma^2/[(n-1)\lambda_j], \\ &Cov(\hat{\alpha}_j, \hat{\alpha}_l \mid \tilde{\mathbf{X}}) = 0, \text{ for } j \neq l. \end{aligned} \tag{13b}$$

Substituting (13a) into (11) yields

$$Cov(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) = \frac{\sigma^2}{n-1}\mathbf{V}_{k-d}\boldsymbol{\Lambda}_{k-d}^{-1}\mathbf{V}_{k-d}'. \tag{14a}$$

Thus,

$$Var(\hat{\theta}_{j,PC} \mid \tilde{\mathbf{X}}) = \frac{\sigma^2}{n-1}\Sigma_{i=1}^{k-d}\left(\frac{v_{ji}^2}{\lambda_i}\right),\ j = 1, ..., k, \tag{14b}$$

which shows clearly that if any of the eigenvalues is close to zero, the variance of any or all of the elements of $\hat{\boldsymbol{\theta}}_{PC}$ (and hence of $\hat{\boldsymbol{\beta}}_{PC}$) may be inflated. Note that, since we use Chatterjee and Hadi's (2006, p. 240) second type of scaling the variables in (2), which are standardized with zero mean and unit standard deviation (*not* unit length); and since the λ's are the eigenvalues of the correlation matrix $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})/(n-1)$, the division by $n - 1$ in (14b) is correct; see McCallum (1970), Cheng and Iglarsh (1976), and Gunst and Mason (1980, pp. 114-115). We stress this point, as the various types of scaling used in the literature seem to be a source of error. For example, in Chatterjee and Hadi's (2006, pp. 249-251) application of the PCR to the advertising data, where the variables and the λ's are defined as above, the authors fail to divide their Equations (9.34) and (9.35) by $n - 1$. Their calculation of the standard error of $\hat{\theta}_1$ is correct, however, as it is based on their Equation (9.33), which *is* correct.[1]

Now, using (3a), we can write $\hat{\boldsymbol{\beta}}_{PC}^{s} = s_y\mathbf{S}^{-1}\hat{\boldsymbol{\theta}}_{PC}$, hence

[1] The data for this example are given in Table 9.9 of Chatterjee and Hadi (2006, p. 236), where $n = 22$. Using the exact figures, and not the three-digit approximations used by the authors, we confirmed that the eigenvalues of the correlation matrix are indeed those given on p. 251, and that if their Equations (9.34) and (9.35) are used, then the standard error of $\hat{\theta}_1$ is incorrectly calculated as 1.947. On the other hand, their Equation (9.33) and our Equation (14b) both give the correct estimate of this standard error, which is 0.425 (= $1.947/\sqrt{21}$). Note that Chatterjee and Hadi's estimate of this standard error is given on p. 253 and is slightly different, 0.438, because of rounding errors.

$$Cov(\hat{\boldsymbol{\beta}}^{s}_{PC} \mid \tilde{\mathbf{X}}) = s_y^2 \mathbf{S}^{-1} Cov(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) \mathbf{S}^{-1}. \quad (15a)$$

Substituting (14a) into (15a) and replacing σ^2 with its estimator (S^2) based on Equation (5) that retains all the (k) PCs yields the following estimator of (15a):

$$\hat{Cov}(\hat{\boldsymbol{\beta}}^{s}_{PC} \mid \tilde{\mathbf{X}}) = s_y^2 \frac{S^2}{n-1} \mathbf{S}^{-1} \mathbf{V}_{k-d} \boldsymbol{\Lambda}_{k-d}^{-1} \mathbf{V}_{k-d}' \mathbf{S}^{-1}. \quad (15b)$$

In addition, using (3b), we have that

$$\begin{aligned} Var(\hat{\beta}_{0,PC} \mid \tilde{\mathbf{X}}) &= \Sigma_{j=1}^{k} \bar{x}_j^2 Var(\hat{\beta}_{j,PC} \mid \tilde{\mathbf{X}}) \\ &+ 2 \sum_{j<l} \bar{x}_j \bar{x}_l Cov(\hat{\beta}_{j,PC}, \hat{\beta}_{l,PC} \mid \tilde{\mathbf{X}}). \end{aligned} \quad (15c)$$

The t-ratio of $\hat{\beta}_{j,PC}$ is the same as that of $\hat{\theta}_{j,PC}$, j = 1, ..., k, since

$$\begin{aligned} t_j &= \hat{\beta}_{j,PC} / s_{\hat{\beta}_{j,PC}} = [(s_y / s_j) \hat{\theta}_{j,PC}] / [(s_y / s_j) s_{\hat{\theta}_{j,PC}}] \\ &= \hat{\theta}_{j,PC} / s_{\hat{\theta}_{j,PC}}, \ j = 1, ..., k. \end{aligned} \quad (16)$$

As we noted earlier, $\hat{\boldsymbol{\theta}}_{PC}$ (and hence $\hat{\boldsymbol{\beta}}_{PC}$) is biased. To calculate its bias, we follow Myers (1990, p. 415) and begin by using (7), to obtain

$$\hat{\boldsymbol{\alpha}}_{k-d} = \mathbf{V}_{k-d}' \hat{\boldsymbol{\theta}}_{OLS} \quad (17a)$$

and

$$\hat{\boldsymbol{\alpha}}_{d} = \mathbf{V}_{d}' \hat{\boldsymbol{\theta}}_{OLS}. \quad (17b)$$

Substituting (17a) into (8) yields $\hat{\boldsymbol{\theta}}_{PC} = \mathbf{V}_{k-d} \mathbf{V}_{k-d}' \hat{\boldsymbol{\theta}}_{OLS}$, so

$$E(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) = \mathbf{V}_{k-d} \mathbf{V}_{k-d}' E(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}}) = \mathbf{V}_{k-d} \mathbf{V}_{k-d}' \boldsymbol{\theta}, \quad (18)$$

since $\hat{\boldsymbol{\theta}}_{OLS}$ is unbiased. Now, since $\mathbf{VV'} = \mathbf{I}$, we have that $\mathbf{V}_{k-d} \mathbf{V}_{k-d}' + \mathbf{V}_d \mathbf{V}_d' = \mathbf{I}$, hence $\mathbf{V}_{k-d} \mathbf{V}_{k-d}' = \mathbf{I} - \mathbf{V}_d \mathbf{V}_d'$, so (18) becomes $E(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) = (\mathbf{I} - \mathbf{V}_d \mathbf{V}_d') \boldsymbol{\theta} = \boldsymbol{\theta} - \mathbf{V}_d \mathbf{V}_d' \boldsymbol{\theta}$. From (17b) we have $E(\hat{\boldsymbol{\alpha}}_d \mid \tilde{\mathbf{X}}) = \mathbf{V}_d' E(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}}) = \mathbf{V}_d' \boldsymbol{\theta} = \boldsymbol{\alpha}_d$. Substituting in the previous equation yields $E(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) = \boldsymbol{\theta} - \mathbf{V}_d \boldsymbol{\alpha}_d$, so

$$bias(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) = E(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) - \boldsymbol{\theta} = -\mathbf{V}_d \boldsymbol{\alpha}_d. \quad (19)$$

Goldberger (1964, p. 127) notes, however, that "unbiasedness is not sacred" and reminds us of the intuitively appealing minimum mean square error (MSE) criterion, "which selects a biased estimator if its variance is small enough to compensate for its bias." Of course, the minimum MSE and other criteria for choosing among competing estimators are widely discussed in the literature (see, e.g., McCallum, 1970, Gunst and Mason, 1977, and Wu, 2017). Using the MSE to choose between $\hat{\boldsymbol{\theta}}_{PC}$ and $\hat{\boldsymbol{\theta}}_{OLS}$, we must determine whether the $k \times k$ matrix $MSE(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}}) - MSE(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}})$ is positive semi-definite, in which case $\hat{\boldsymbol{\theta}}_{PC}$ will be preferable. Since $\hat{\boldsymbol{\theta}}_{OLS}$ is unbiased, we have $MSE(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}}) = Cov(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}})$. But, using (12), we obtain from (6) the following expression: $Cov(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}}) = \mathbf{V} Cov(\hat{\boldsymbol{\alpha}} \mid \tilde{\mathbf{X}}) \mathbf{V}' = [\sigma^2/(n-1)] \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}'$. Thus, we have

$$MSE(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}}) = \frac{\sigma^2}{n-1} \mathbf{V}_{k-d} \boldsymbol{\Lambda}_{k-d}^{-1} \mathbf{V}_{k-d}' + \frac{\sigma^2}{n-1} \mathbf{V}_d \boldsymbol{\Lambda}_d^{-1} \mathbf{V}_d'. \quad (20a)$$

From (14a) and (20a) we obtain the $k \times k$ matrix $Cov(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}}) - Cov(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) = [\sigma^2/(n-1)] \mathbf{V}_d \boldsymbol{\Lambda}_d^{-1} \mathbf{V}_d'$, which is positive semi-definite, since $\boldsymbol{\Lambda}_d$ is positive definite and the $k \times d$ matrix $\mathbf{V}_d$ has rank $d < k$ (Rencher and Schaalje, 2008, p. 26, Corollary 2). Thus, the variance of $\hat{\boldsymbol{\theta}}_{PC}$ will never be greater than that of $\hat{\boldsymbol{\theta}}_{OLS}$, so the burden of the choice between the two estimators falls on the size of the $bias(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}})$. If the cost (bias) of falsely omitting a PC (underfitting) outweighs the gain (lower variance), the PCR will fail to be a minimum MSE estimator.

By definition (see Goldberger, 1964, p. 129), we have

$$\begin{aligned} MSE(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) &= E[(\hat{\boldsymbol{\theta}}_{PC} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_{PC} - \boldsymbol{\theta})' \mid \tilde{\mathbf{X}}] \\ &= Cov(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) + [bias(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}})][bias(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}})]'. \end{aligned} \quad (20b)$$

Substituting (14a) and (19) in (20b) gives

$$MSE(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) = \frac{\sigma^2}{n-1} \mathbf{V}_{k-d} \boldsymbol{\Lambda}_{k-d}^{-1} \mathbf{V}_{k-d}' + \mathbf{V}_d \boldsymbol{\alpha}_d \boldsymbol{\alpha}_d' \mathbf{V}_d'. \quad (20c)$$

Subtracting (20c) from (20a) yields

$$\begin{aligned} &MSE(\hat{\boldsymbol{\theta}}_{OLS} \mid \tilde{\mathbf{X}}) - MSE(\hat{\boldsymbol{\theta}}_{PC} \mid \tilde{\mathbf{X}}) \\ &= \mathbf{V}_d \left(\frac{\sigma^2}{n-1} \boldsymbol{\Lambda}_d^{-1} - \boldsymbol{\alpha}_d \boldsymbol{\alpha}_d' \right) \mathbf{V}_d'. \end{aligned} \quad (21)$$

Following the same steps, one can show that[2]

$$\begin{aligned} &MSE(\hat{\boldsymbol{\beta}}^{s}_{OLS} \mid \tilde{\mathbf{X}}) - MSE(\hat{\boldsymbol{\beta}}^{s}_{PC} \mid \tilde{\mathbf{X}}) \\ &= s_y^2 \mathbf{S}^{-1} \mathbf{V}_d \left(\frac{\sigma^2}{n-1} \boldsymbol{\Lambda}_d^{-1} - \boldsymbol{\alpha}_d \boldsymbol{\alpha}_d' \right) \mathbf{V}_d' \mathbf{S}^{-1}, \end{aligned} \quad (22)$$

where $\hat{\boldsymbol{\beta}}^{s}_{OLS} = (\hat{\beta}_{1,OLS}, ..., \hat{\beta}_{k,OLS})'$, $\hat{\boldsymbol{\beta}}^{s}_{PC} = (\hat{\beta}_{1,PC}, ..., \hat{\beta}_{k,PC})'$, and $\mathbf{S}$ is defined in (3a).

The $k \times k$ symmetric matrix in (21) will be positive semi-definite if and only if all its eigenvalues are

[2] The proof of Equation (22) is given in an appendix that is available from the first author upon request.

nonnegative (Goldberger, 1964, p. 37). Note that if the matrix in (21) is positive semi-definite, then so is that in (22), since the latter comes from the former by multiplying it by the positive scalar s_y^2 and by pre- and post-multiplying it by the positive definite matrix $\mathbf{S}^{-1}$ (see Hadley, 1961, p. 255). In fact, (21) and (22) will be positive semi-definite if

$$\frac{\sigma^2}{n-1}\mathbf{\Lambda}_d^{-1} - \boldsymbol{\alpha}_d\boldsymbol{\alpha}_d' \text{ is a positive semi-definite matrix.} \quad (23)$$

Necessary, but not sufficient, conditions for (23) are (see Goldberger, 1964, p. 37)

$$\sigma^2/[(n-1)\lambda_j] - \alpha_j^2 \geq 0,\ j = 1, ..., d. \quad (24)$$

Thus, as a test of dropping the "optimal" number of PCs (in the sense of the minimum MSE), we can start from the PC with the smallest eigenvalue, or from the most insignificant one in the regression equation (5), and keep dropping such PCs until the conditions (24) are violated. Note that for $d = 1$, (24) is also a sufficient condition, since in this case $[\sigma^2/(n-1)]\mathbf{\Lambda}_d^{-1} - \boldsymbol{\alpha}_d\boldsymbol{\alpha}_d'$ is a scalar, so it can be factored out in (21), and the $k\times k$ matrix that emerges, $\mathbf{V}_d\mathbf{V}_d'$, is positive semi-definite [Goldberger, 1964, p. 37, Property (7.15) with $\mathbf{P} = \mathbf{V}_d'$]. Note also that versions of (23) already exist in the literature (see, e.g., McCallum, 1970, Farebrother, 1972, and Ö zkale, 2009). In particular, McCallum's (1970, p. 112) condition (12) can be shown to be a special case of (23) for $k = 2$ and considering the MSE of only one coefficient.

Consider the factors that enter (23) and favor the PCR over the OLS estimator. First, the larger the error variance (σ^2) is, the more crucial it becomes to reduce the coefficient variances via the PCR. Second, for the same reason, the smaller the size of the sample (n), the higher the level of uncertainty, and hence the larger the need for precision of the coefficient estimates gained by applying the PCR. Third, the smaller the eigenvalues associated with the dropped PCs are, the more severe the multicollinearity problem is, hence the more meaningful the application of the PCR becomes. Fourth, the smaller the (absolute) values of the coefficients of the dropped PCs ($\boldsymbol{\alpha}_d$) are, the weaker the effects of these PCs on the dependent variable, and hence the more justifiable their removal from the PCR becomes.

Note that a difficulty with condition (23) is that it involves unknown parameters. A way out is to use their unbiased estimators (see McCallum, 1970, p. 112, Farebrother, 1972, p. 335, and Ö zkale, 2009, p. 546). Since σ^2 is inherited from Equation (2) or (5), its estimate (s^2), as well as the estimate of $\boldsymbol{\alpha}_d$, should be obtained from the regression equation (5) that retains *all* the PCs. In sum, we have the following

Proposition 1: $\hat{\boldsymbol{\beta}}_{PC}$ outperforms $\hat{\boldsymbol{\beta}}_{OLS}$, in the minimum MSE sense, if and only if the eigenvalues of the symmetric d×d matrix $[s^2/(n-1)]\mathbf{\Lambda}_d^{-1} - \hat{\boldsymbol{\alpha}}_d\hat{\boldsymbol{\alpha}}_d'$ are all nonnegative. Necessary (but not sufficient) conditions are $s^2/[(n-1)\lambda_j] - \hat{\alpha}_j^2 \geq 0$, j = 1, ..., d, where s^2 and $\hat{\alpha}_j$ (an element of $\hat{\boldsymbol{\alpha}}_d$) are obtained from regression (5) that retains all the PCs. As a test of dropping the "optimal" number of PCs, one can start from the PC with the smallest eigenvalue or from the most insignificant one in (5), and keep dropping PCs until the condition is violated. For d = 1, the condition is also sufficient.

2.3. Some Measures of Multicollinearity

The simplest measures of multicollinearity that one could think of are the absolute values of Pearson's pairwise correlation coefficients (r_{ij}) among the regressors. If $k = 2$, this criterion is reliable, in that a "low" value of r_{12} means absence of multicollinearity and a "high" value of r_{12} means that multicolinearity is present. If $k > 2$, however, this criterion is *not* reliable, in that, although "high" values of r_{ij} (at least one of them) still imply that multicolinearity is present, nevertheless "low" values of r_{ij} do not necessarily imply absence of multicollinearity (Chatterjee and Hadi, 2006, pp. 233-237). Kmenta (1971, pp. 382-384) presents an example with $k = 3$, where there exists an exact linear relationship among the three regressors, i.e., there is *perfect* multicollinearity, and yet none of the three r_{ij}s exceeds 0.5 in absolute value.

According to another simple criterion, multicollinearity is considered harmful if, at a level of significance, say, 5-percent, the standard F statistic (for the hypothesis that the joint effect of all the regressors is zero) is significant, but all the t-statistics for the individual slope coefficients are insignificant. As Kmenta (1971, p. 390) points out, however, this criterion is too strong, since it considers multicollinearity harmful only when *all* the t-statistics for the slopes are insignificant, which makes it difficult to disentangle the individual effects of the regressors on the dependent variable.

Chatterjee and Hadi (2006, p. 233) suggest that researchers should pay attention to the following indications of multicollinearity: (i) large changes in the estimated coefficients if a regressor is added or dropped, or if a data point is altered or dropped; (ii) insignificant t-statistics for regressors that are important, according to the pertinent theory; and (iii) the signs of some of the estimated slope coefficients do not conform to those expected (based on theoretical grounds).

A well-known statistic that measures multicollinearity is the *variance inflation factor* (*VIF*), defined as $VIF_j = 1/Tolerance_j$, where $Tolerance_j = 1 - R_j^2$ and R_j^2 = the coefficient of determination in the (auxiliary) regression of X_j on the other regressors. Clearly, if the Xs are orthogonal among themselves, then $R_j^2 = 0$ and $VIF_j = Tolerance_j = 1$, $j = 1, ..., k$. According to Chatterjee and Hadi (2006, p. 238), "a VIF in excess of 10 is an indication that multicollinearity may be causing problems in estimation."

Another indication of the presence of multicollinearity is that some eigenvalues are close to zero. Thus, as another measure of multicollinearity, some authors suggest the *condition index* (κ), defined as $\kappa = \sqrt{\lambda_1/\lambda_p}$, where λ_1 and λ_p are, respectively, the largest and the smallest eigenvalue of

the matrix **X**'**X**. By definition, $\kappa > 1$. A large value of κ is evidence of strong multicollinearity, suggesting that the inversion of **X**'**X** will be sensitive to small changes in **X**. As an empirical rule, multicollinearity is considered to be harmful when $\kappa > 15$ (Chatterjee and Hadi, 2006, pp. 244-245).

The diagnostics of multicollinearity are often complemented by the "variance proportions" in assessing the effect of each linear dependency among the regressors on the coefficient variances. In the OLS regression Equation (1), if any of the eigenvalues of **X**'**X** is close to zero (indicating a serious linear dependency), the variance of any or all of the coefficients in $\hat{\boldsymbol{\beta}}_{OLS}$ may be inflated. The variance proportion p_{ji} is the proportion of the variance of the coefficient $\hat{\beta}_{i,OLS}$ attributed to the linear dependency characterized by the eigenvalue λ_j (see Myers, 1990, pp. 371-379).

3. Step-by-Step PCR in SPSS by Replicating and Updating an Example

Chatterjee and Hadi (2006, ch. 9) illustrate the PCR by estimating a linear imports function using French annual aggregate data, 1949-1959 ($n = 11$), on Imports (IMPORT, y), Gross Domestic Product (DOPROD, X_1), increase in Inventories (STOCK, X_2), and Consumption (CONSUM, X_3), all measured in billions of French francs at 1959 prices. Some useful descriptive statistics are $\bar{y}$ =21.891, $\bar{x}_1$ =194.591, $\bar{x}_2$ = 3.3, $\bar{x}_3$ = 139.736, s_y = 4.5437, s_1 = 30, s_2 = 1.6492, and s_3 = 20.6344.

Note that in this example there are economic as well as econometric reasons to believe that the classical assumptions fail. For example, from the point of view of correct specification, instead of including STOCK and CONSUM as regressors, we would include the real exchange rate; and from the point of view of time-series econometrics, we would consider the problems of nonstationarity, endogeneity of the regressors, and serial correlation. We refrain from these issues here, however, and focus on the correct application of the PCR.

In step 1, we apply OLS to Equation (1) and use the above criteria to decide whether multicollinearity is harmful. After entering the data in SPSS and selecting

Analyze > Regression > Linear > Statistics > Collinearity Diagnostics

we get R^2 = 0.992 (coefficient of determination) and Tables 1-2. The second and the fourth column of Table 1 report the elements of $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\theta}}_{OLS}$. Note that the coefficient $\hat{\beta}_{1,OLS}$ = -0.051 has the wrong sign and is insignificant at all conventional levels (p-value = 0.488). Economic theory suggests that domestic income exerts a *positive* influence on imports, so we blame multicollinearity for these unexpected results, and *keep* DOPROD for the PCR.

Table 1. OLS estimates of **β** and **θ**[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-10.128	1.212		-8.355	.000		
	DOPROD	-.051	.070	-.339	-.731	.488	.005	185.997
	STOCK	.587	.095	.213	6.203	.000	.981	1.019
	CONSUM	.287	.102	1.303	2.807	.026	.005	186.110

a. Dependent Variable: IMPORT

Table 2. Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	DOPROD	STOCK	CONSUM
1	1	3.838	1.000	.00	.00	.01	.00
	2	.148	5.086	.01	.00	.94	.00
	3	.013	17.073	.77	.00	.03	.00
	4	5.447E-5	265.461	.22	1.00	.02	1.00

a. Dependent Variable: IMPORT

We obtain a large value of $VIF_1 \approx 186$ for DOPROD (see Table 1), suggesting that multicollinearity is present indeed. The high correlation coefficient between DOPROD and CONSUM (r_{13} = 0.997, see Table 3) confirms this conclusion. The condition index is $\kappa = 265.46 > 15$ (Table 2), so this criterion, too, suggests that multicollinearity may be harmful.[3] The linear dependency between DOPROD and CONSUM is also revealed by the small value of the last eigenvalue, λ_3 = 0.00005447, accompanied by the extremely

[3] For an excellent theoretical discussion of the collinearity indices reported in Tables 2 and 3, including their marginal values, see Myers (1990, pp. 123-133, 369-371) and Rawlings (1988, pp. 273-281).

high variance proportions of $Var(\hat{\beta}_{1,OLS})$ and $Var(\hat{\beta}_{3,OLS})$, namely, p_{31} = 0.9984 and p_{33} = 0.9989 (see Table 2, where both of these values are rounded to 1). That is to say, 99.84% of $Var(\hat{\beta}_{1,OLS})$ and 99.89% of $Var(\hat{\beta}_{3,OLS})$ can be attributed to the above linear dependency.

Table 3. Simple correlations

	DOPROD	STOCK	CONSUM
DOPROD	1	.026	.997**
STOCK	.026	1	.036
CONSUM	.997**	.036	1

**. Correlation is significant at the 0.01 level (2-tailed).

Thus, in step 2, we estimate Equation (5). First, we need to standardize the original variables by selecting

Analyze > Descriptive Statistics > Descriptives > (bring into the dialog box all four variables) IMPORT, DOPROD, STOCK, CONSUM > Save Standardized Values as Variables (denoted as ZIMPORT, ZDOPROD, ZSTOCK, ZCONSUM).

We can now obtain the PCs, the matrices **V** and **C**, and the vector $\hat{\boldsymbol{\alpha}}$ by selecting

File > New > Syntax

and by writing the following program in the command syntax window that appears:

```
matrix.        /* Comment: The dot at the end of each command is necessary.
get x /variables ZDOPROD, ZSTOCK, ZCONSUM.
compute xtx=t(x)*x/10.    /*We divide by n–1=10, since we use the correlation matrix.
call eigen(xtx,eigvec,eigval).
print eigval.    /* eigval is the vector of the eigenvalues of the correlation matrix.
print eigvec.    /* eigvec is the correct matrix V.
compute c=x*eigvec.    /* see Equation (4).
print c.
compute thetaols={-0.339; 0.213; 1.303}.    /* see Table 1, 4th column.
compute alpha=t(eigvec)*thetaols.    /* see Equation (7).
print alpha.
end matrix. /* To execute this program, do a right click and select Run All.
```

Although this program produces the correct matrix **V** directly, we will also construct it manually, in order to see the error made by Liu, et al. (2003). First, select

Analyze > Dimension Reduction > Factor > (insert into the dialog box the variables) DOPROD, STOCK, CONSUM > Extraction > Fixed Number of Factors > Factors to Extract > (enter into the dialog box) 3 (the number of the original regressors) > Continue > OK.

Tables 4-5 report the results. The 3×3 "component matrix" (Table 5) differs from that in Chatterjee and Hadi (2006, p. 243) and does not satisfy the property **V**′**V** = **I**, so it is *not* the correct matrix of eigenvectors, as Liu, et al. (2003) erroneously assume. Its elements need to be "normalized," i.e., its columns must be divided by the square root of the corresponding eigenvalue, given by the second column of Table 4. That is, the first column of this matrix must be divided by $\sqrt{1.999}$, the second by $\sqrt{0.998}$, and the third by $\sqrt{0.002691}$. We thus obtain the correct matrix of eigenvectors:

$$\mathbf{V} = \begin{bmatrix} 0.707 & 0.036 & -0.707 \\ 0.044 & -0.999 & -0.007 \\ 0.707 & 0.026 & 0.707 \end{bmatrix}. \qquad (25)$$

Table 4. Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	1.999	66.638	66.638
2	.998	33.272	99.910
3	.002691[a]	.090	100.000

Extraction Method: Principal Component Analysis.
[a]In the Table printed by SPSS this figure was rounded to 0.003.

Table 5. Component Matrix[a]

	Component		
	1	2	3
DOPROD	.999	-.036	.037
STOCK	.062	.998	0.000362[b]
CONSUM	.999	-.026	-.037

Extraction Method: Principal Component Analysis.
a. 3 components extracted.
b. In the Table printed by SPSS this figure was rounded to 0.000.

As we noted earlier, λ_i is the variance of the *i*-th PC. Here, λ_3 = 0.002691, which is close to zero, suggesting that C_3 is almost a constant, and can be omitted, whereas keeping it would inflate the s.e. of $\hat{\theta}_{j,PC}$ and hence that of $\hat{\beta}_{j,PC}$; see Equations (14b), (15a)-(15c). If C_3 is included in the regression equation (5), along with C_1 and C_2 (and no intercept), its coefficient is $\hat{\alpha}_3$ = 1.16, which is insignificant at the 5-percent level (*p*-value = 0.095), whereas the other estimates are the same as those of Table 6. Thus, we estimate (5), using only C_1 and C_2 as regressors (and no intercept), implying that the third column of **V** in (25) is dropped.[4] Table 6 reports the results.

Equation (5) does not suffer from multicollinearity, since the PCs are orthogonal. The estimates $\hat{\alpha}_1 = 0.690$ and $\hat{\alpha}_2 = -0.191$ (Table 6) have no natural interpretation, however, since the PCs are linear *combinations* of the

[4] The regressor sets {*C*1, *C*3}, {*C*2, *C*3}, {*C*2}, {*C*3} (and no intercept) all produce insignificant coefficients.

original variables, so they are used only as an intermediate step to estimate the βs.

Table 6. Equation (5) with C_1 and C_2 as regressors[a,b,c]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	C1	.690	.026	.976	27.032	.000
	C2	-.191	.036	-.191	-5.296	.000

a. Dependent Variable: Zscore(IMPORT)
b. Linear Regression through the Origin
c. $R^2 = 0.988$

Thus, in step 3, we get $\hat{\boldsymbol{\theta}}_{PC}$ and $\hat{\boldsymbol{\beta}}_{PC}$ and their standard errors (s.e.). Using (8), where $\mathbf{V}_{k\text{-}d}$ is 3×2, and the above estimates of α_1 and α_2, we get

$$(\hat{\theta}_{1,PC}, \hat{\theta}_{2,PC}, \hat{\theta}_{3,PC}) = (0.481, 0.221, 0.483). \quad (26)$$

Using Equation (14a), after replacing σ^2 with the estimate $s^2 = 0.010129$ [obtained from the regression equation (5) when all the three PCs are retained], we calculate

$$\hat{Cov}(\hat{\boldsymbol{\theta}}_{PC}) = \begin{bmatrix} 0.000259 & -0.000021 & 0.000258 \\ & 0.001030 & -0.000011 \\ & & 0.000258 \end{bmatrix} \quad (27)$$

(symmetric terms are omitted). In SPSS, (26) and (27) can be obtained by running the following program:

```
matrix.
compute alpha2={0.690; -0.191}.
compute v2={0.707, 0.036; 0.044, -0.999; 0.707, 0.026}.
compute thetapc=v2*alpha2.
print thetapc.
        /* This gives the estimates in (26).
compute Lamda2={1.999, 0; 0, 0.998}.
compute  covtheta=(0.01029/10)*v2*inv(Lamda2)*t(v2).
   /* This is Equation (14a).
print covtheta.
end matrix. /* To execute this program, do a right click and select Run All.
```

Next, using (3b), we calculate the values of the $\hat{\beta}_{PC}$s. These estimates are reported in Table 7 (the PCR) and are the same as those obtained by Chatterjee and Hadi (2006, p. 263).[5] Table 7 also reports the estimated s.e.s of the $\hat{\beta}_{PC}$s and their *t*-ratios. The s.e.s are obtained from the matrix $\hat{Cov}(\hat{\boldsymbol{\beta}}_{PC})$, which is calculated in accordance with (15a)-(15c) and (27) and is reported below in (28) (the covariances between $\hat{\beta}_{0,PC}$ and the slope coefficients are not reported, as they are almost never useful):

$$\hat{Cov}(\hat{\boldsymbol{\beta}}_{PC}) = \begin{bmatrix} 1.005 & - & - & - \\ - & 0.0000059 & -0.0000088 & 0.0000086 \\ - & & 0.0078182 & -0.0000065 \\ - & & & 0.0000125 \end{bmatrix}. \quad (28)$$

Comparing the results of Table 7 with those of Table 1 (OLS), we observe that the major difference is that the coefficient of DOPROD has now the expected sign and is highly statistically significant. We conclude that the original OLS estimate of this coefficient (-0.051) involves a large sampling error, whereas the PCR yields a precise estimate with the expected sign (0.0728) and s.e. = 0.0024, which is about 30 times smaller than that of Table 1 (0.0703). The other coefficients also have the expected signs and are highly significant. Thus, the PCR is a substantial improvement over the OLS estimator, and our decision not to drop DOPROD turned out to be correct. Unfortunately, however, the minimum MSE criterion does not support this conclusion, as Proposition 1 fails, since $0.010129/(10\times0.002691) - 1.16^2 = -0.97 < 0$, apparently because the coefficient $\hat{\alpha}_3 = 1.16$ is relatively large.

Table 7. Estimation of Equation (1) via the PCR, French annual data, 1949-1959[a]

Variable	Coefficient	Standard error	*t*-statistic[b]
Constant	-9.1407	1.0024	-9.12***
DOPROD	0.0728	0.0024	29.90***
STOCK	0.6093	0.0884	6.89***
CONSUM	0.1063	0.0035	30.07***

[a] Dependent Variable: IMPORT; [b] the asterisks *** denote significance at the 1-percent level.

To check the robustness of these findings to substantial changes in the sample, we now re-estimate Equation (1) with French annual aggregate data, 1960-2018 (n = 59). The variables are defined as before, but they are now measured in billions of euros at 2010 prices. The source of the data is the European Commission (AMECO Online). Again, we refrain from the theoretical and the econometric issues referred to earlier.

In the updated sample, multicollinearity is again strong, as $r_{13} = 0.999$, $VIF_1 = 570$ for DOPROD, $VIF_3 = 576$ for CONSUM, $\kappa = 172$, the value of the smallest eigenvalue is $\lambda_3 = 0.0001$, and the variance proportions of $Var(\hat{\beta}_{1,OLS})$ and $Var(\hat{\beta}_{3,OLS})$ are $p_{31} \approx p_{33} \approx 0.9994$. Thus, following exactly the same steps as before, we obtain and report the OLS and the PCR estimates side by side in Table 8.

Again, the PCR is a substantial improvement over the OLS estimator in that it produces estimates that have the expected sign and are highly statistically significant. In particular, the OLS estimate of the coefficient of DOPROD (-0.848) is negative *and statistically significant* at the 1-percent level, an unacceptable result from the point of view of economic theory, whereas the PCR eliminates this obvious sampling error. Recall that in the case of the

[5] Chatterjee and Hadi (2006, p. 263) report a slightly different estimate of the intercept, namely -9.106, apparently because of rounding errors.

1949-1959 data, this coefficient was wrongly signed, but at least it was *in*significant at any conventional level. Thus, in the updated sample, the PCR proves to be even more important.

Table 8. Estimation of Equation (1), French annual aggregate data, 1960-2018[a]

	OLS			PCR		
Variable	**Coeff.**	**s.e.**	***t*-ratio[b]**	**Coeff.**	**s.e.**	***t*-ratio[b]**
Constant	-216.39	16.54	-13.1***	-217.76	15.4212	-14.12***
DOPROD	-0.848	0.255	-3.33***	0.175	0.0063	27.93***
STOCK	3.157	0.550	5.74***	3.592	0.5346	6.72***
CONSUM	2.199	0.469	4.69***	0.319	0.0113	28.35***

[a] Dependent Variable: IMPORT; [b] the asterisks *** denote significance at the 1-percent level.

Proposition 1 fails again, however. Here, $s^2 = 0.038767$, $\lambda_3 = 0.000873$, $n - 1 = 58$, and $\hat{\alpha}_3 = -3.543$, so $0.038767/(58\times0.000873) - 3.543^2 = -11.78 < 0$. The failure of the minimum MSE criterion to support a theoretically and empirically sound result suggests that other criteria for comparing the PCR with the OLS estimator should also be used. This is beyond the purpose of this paper, however; see, e.g., Wu (2017).

4. Another Example, Where More than One PCs are Dropped

To illustrate how Proposition 1 is implemented when more than one PCs (not necessarily consecutive) are to be dropped from the regression equation (5), we employ the "Hospital manpower data" given in Myer's (1990, pp. 132-133) Table 3.8, where $n = 17$ and $k = 5$. In this example, too, multicollinearity is strong, as the bivariate correlation coefficients are high and highly statistically significant, e.g., $r_{13} = 0.9999$, $r_{14} \approx r_{34} \approx 0.94$, $r_{12} \approx r_{23} \approx r_{24} \approx 0.91$, whose *p*-values for two-tailed tests are all 0.000; $VIF_1 = 9598$ and $VIF_3 = 8933$; the condition index is $\kappa = 427$; the last three eigenvalues of the $\mathbf{X'X}$ matrix are 0.0447, 0.0082, and 0.00002848; and the two highest variance proportions are $p_{51} \approx p_{53} \approx 0.999$. In this example, we have $s^2 = 0.012223$; the last three eigenvalues of the correlation matrix are $\lambda_3 = 0.0946332$, $\lambda_4 = 0.040712$, and $\lambda_5 = 0.00005397$; and the PCs C_3 and C_5 are statistically insignificant in (5), since the *p*-values of their estimated coefficients, $\hat{\alpha}_3 = 0.064$ and $\hat{\alpha}_5 = -1.301$, are 0.493 and 0.735 (whereas the *p*-values of the coefficients of C_1, C_2, and C_4 are 0.000000, 0.000976, and 0.001859). Thus, if we drop C_5 only, Proposition 1 gives

$$0.012223/(16\times0.00005397) - 1.301^2 = 12.46 > 0. \quad (29)$$

The choice of PCs to be deleted is a debatable issue in the literature. There are two strategies. The first deletes the PCs that are associated with the smallest eigenvalues of the correlation matrix, whereas the second deletes those that are not significant in (5). Gunst and Mason (1980, pp. 327-328) argue that "the first strategy often works better in practice than the second, although the individual *t* tests can be more effective if a very small significance level is used (say $\alpha = .001$). The rationale behind this suggestion is that the decrease in variance associated with the deletion of multicollinear components generally is much greater than the bias incurred by doing so." Myers (1990, p. 419) favors the second strategy based on the individual *t*-values, which "should be rank ordered and components be considered for elimination beginning with the *smallest t-value*, in magnitude" (Myers's emphasis). Jackson's (2003, p. 44) advice is: "do NOT include pc's in the model that do not belong there statistically" (Jackson's emphasis). With these suggestions in mind, we choose to drop C_3 and C_5, because, as we demonstrated earlier, they are highly insignificant.

Thus, in accordance with our Proposition 1, we must calculate the two eigenvalues of the following 2×2 symmetric matrix:

$$\frac{0.012223}{17-1}\begin{bmatrix}0.0946332 & 0\\ 0 & 0.00005397\end{bmatrix}^{-1} - \begin{bmatrix}0.064\\ -1.301\end{bmatrix}\begin{bmatrix}0.064 & -1.301\end{bmatrix} = \begin{bmatrix}0.003977 & 0.083264\\ 0.083264 & 12.462516\end{bmatrix}. \quad (30)$$

They are 12.46 and 0.0034. Since both are nonnegative, we conclude that the matrix is positive semi-definite, and hence $\hat{\boldsymbol{\beta}}_{PC}$ outperforms $\hat{\boldsymbol{\beta}}_{OLS}$ in the minimum MSE sense.

5. Summary

In this paper, we revisit the PCR and show step-by-step how to implement it properly in SPSS by replicating an example of Chatterjee and Hadi (2006), in which the regressors are highly collinear. The PCR proves to be important, as it produces estimates that have the expected signs and are statistically significant at the 1-percent level, whereas the OLS fails in that respect. This result becomes stronger when we update the sample substantially. Our main motivation has been the fact that some researchers still fail to implement this useful estimation method properly, despite the strong warnings that already exist in the literature. As an example of such a failure, we have briefly commented on the paper by Liu, et al. (2003).

In addition, we use two more data sets from the literature to illustrate other important aspects of the PCR, namely, the use of the correct formulas, in accordance with the choice of scaling the variables, the choice of PCs to drop, and the conditions for the PCR to outperform OLS in the sense of the

minimum MSE criterion.

REFERENCES

[1] Chatterjee, S., and Hadi, A.S., 2006, Regression Analysis by Example, 2nd Ed., John Wiley & Sons, NJ.

[2] Cheng, D.C., and Iglarsh, H.J., 1976, Principal component estimators in regression analysis, The Review of Economics and Statistics 58:2, 229-234.

[3] Ding, Y., Ma, X., and Wang, Y., 2018, Health status monitoring for ICU patients based on locally weighted principal component analysis, Computer Methods and Programs in Biomedicine 156, 61-71.

[4] Farebrother, R.W., 1972, Principal component estimators and minimum mean square error criteria in regression analysis, The Review of Economics and Statistics 54:3, 332-336.

[5] Goldberger, A.S., 1964, Econometric Theory, John Wiley & Sons, New York.

[6] Gunst, R.F., and Mason, R.L., 1977, Biased estimation in regression: an evaluation using mean square error, Journal of the American Statistical Association 72:359, 616-628.

[7] Gunst, R.F., and Mason, R.L., 1980, Regression analysis and its application: A data-oriented approach, Marcel Dekker, New York.

[8] Hadi, A.S., and Ling, R.F., 1998, Some cautionary notes on the use of principal components regression, The American Statistician 52:1, 15-19.

[9] Hadley, G., 1961, Linear Algebra, Addison-Wesley, Reading, MA.

[10] Hayashi, F., 2000, Econometrics, Princeton University Press, Princeton, NJ.

[11] Jackson, J.E., 2003, A User's Guide to Principal Components, John Wiley & Sons, Hoboken, NJ.

[12] Jolliffe, I.T., 1982, A note on the use of principal components in regression, Journal of the Royal Statistical Society, Series C (Applied Statistics) 31, No. 3, 300-303.

[13] Kmenta, J., 1971, Elements of Econometrics, Macmillan, New York.

[14] Liu, R.X., Kuang, J., Gong, Q., and Hou, X.L., 2003, Principal component regression analysis with SPSS, Computer Methods and Programs in Biomedicine 71:141-147.

[15] Massy, W.F., 1965, Principal components regression in exploratory statistical research, Journal of the American Statistical Association 60:309, 234-256.

[16] McCallum, B.T., 1970, Orthogonalization in regression analysis, The Review of Economics and Statistics 52:1, 110-113.

[17] Myers, R.H., 1990, Classical and Modern Regression with Applications, Duxbury Press, Belmont, CA.

[18] Özkale, M.R., 2009, Principal component regression estimator and a test for the restrictions, Statistics 43:6, 541-551.

[19] Rawlings, J.O., 1988, Applied Regression Analysis: A Research Tool, Wadsworth and Brooks, Pacific Grove, CA.

[20] Rencher, A.C., and Schaalje, G.B., 2008, Linear Models in Statistics, 2nd Ed., John Wiley & Sons, Hoboken, New Jersey.

[21] Sharma, S., 1996, Applied Multivariate Techniques, John Wiley & Sons, Hoboken, NJ.

[22] SPSS Statistics, version 25.

[23] Stock, J.H., and Watson, M.W., 2003, *Introduction to Econometrics*, Addison Wesley, Boston, MA.

[24] Tran, H., Kim, J., Kim, D., Choi, M., Choi, M., 2018, Impact of air pollution on cause-specific mortality in Korea: Results from Bayesian Model Averaging and Principle Component Regression approaches, Science of The Total Environment 636, 1020-1031.

[25] Wooldridge, J.M., 2006, Introductory Econometrics: A Modern Approach, 3rd Edition, Mason, OH: Thomson South-Western.

[26] Wu, J., 2017, The small sample properties of the restricted principal component regression estimator in linear regression model, Communications in Statistics – Theory and Methods 46:4, 1661-1667.