Empower European Universities

European higher education policies

and the problem of estimating a complex model with a small cross-section

Gabriele Marconi

EEU Publications, Maastricht, 23/11/2014

European higher education policies and the problem of estimating a complex model with a small cross-section

Gabriele Marconi

Maastricht University – g.marconi@maastrichtuniversity.nl

Keywords: principal components regression – OLS – small sample – explorative research – higher education policies – Montecarlo simulation

Abstract

This paper discusses the components on components regression, a statistical technique suitable for explorative analyses of small datasets containing multiple independent, mediating and dependent variables. This method is compared to ordinary least squares and principal component regression by means of discussion of their properties and the assumptions underlying these estimators, a simulation and an empirical application to European higher education policy, and economic innovativeness in 32 countries. In the datasets used in this paper, the components on components regression yields more precise estimates of the coefficients of association between independent, mediating and dependent variables, compared to ordinary least squares. Compared to the principal components regression, it leads to a more parsimonious empirical model. The simulation also shows that the standard errors of the coefficients estimated with the components on components regression can be obtained by bootstrapping.

1. Introduction

This paper presents a method for analysing statistical relationships between multiple independent, mediating and independent variables, which will be called components on components (CoC) regression hereafter. This method starts by extracting components from the sets of independent, mediating and dependent variables via principal component analysis. Then, the relationships between the extracted components are estimated by least squares. This is very similar to principal components regression (PCR), in which a dependent variable is regressed on a set of components extracted from a set of independent variables. The difference is that CoC involves regressing components on other components.

The contribution of this paper is that it highlights the differences between CoC, PCR and Ordinary Least Squares (OLS), and compares the relative performance of the three methods through a simulation and an empirical application. Furthermore, the simulation shows that it is possible to derive unbiased estimates of the standard errors of CoC for the parameters estimated by CoC. As the discussion focuses on small datasets, the asymptotic properties of the estimators are not derived in this paper.

CoC has been specifically developed for a research project conducted by Hoareau and colleagues, whose results are published in Hoareau et al. (2012, 2013). In these studies, the authors explore the relationships between variables on higher education policy, higher education performance and economic innovativeness at the country level, using 32 countries. In the conceptual model of the authors, higher education policies are related to economic innovativeness through the performance of the higher education sector. The small sample size and the explorative nature of their analysis motivated the search for a suitable empirical method. The aim of this paper is to study some properties of this method, and to compare to alternative statistical methods.

CoC can be considered an adaptation of PCR to the context of multiple dependent and independent variables, related through mediators. PCR (see Jolliffe 2002, Chapter 8) is suitable for situations in which some variables are multi-collinear or in which the sample size is small compared to the number of variables. Although fewer results have been obtained for small sample size, simulation studies have confirmed that, when the problem of multi-collinearity is severe, PCR yields more precise estimates than OLS (e.g. Mittelhammer & Baritelle, 1977). However, differently than CoC, PCR has not been designed for dealing with multiple dependent variables. CoC can also be considered as a special case of structural modelling with latent variables estimated by principal components. This makes it similar to methods such as structural equation modelling (see Kaplan, 2000) and partial least squares-path modelling[1] (see Vinzi et al., 2009), which have been designed for multiple dependent variables and are able to accommodate relatively complex relationships such as mediation effects. However, these techniques require to identify the number of components to be extracted *a priori*. Hence, differently than CoC, these methods are more suitable for confirmatory than explorative analysis.

Besides the methods that have just been described, a wide variety of research methods have been suggested for dealing with small sample size (e.g. Hoyle, 1999), multi-collinearity (e.g. Belsley, Kuh, & Welsch, 2004), and for carrying on explorative research (e.g. Jambu, 1991; Stebbins, 2001). Often,

[1] Partial-least squares-path modelling is a technique related to partial least squares regression. The latter is similar to PCR, and it allows to work with multiple dependent variables. However, differently than partial least squares-path modelling, it is not suited for modelling relationships on multiple levels, in which the variables are linked through mediators.

however, the specific nature of the research problems encountered in the social sciences requires to combine or adapt some of these methods. CoC provides a tool for researchers who are estimating parameters of a linear model for which the following is true:

- A number of variables mediate the relationship between independent and dependent variables;
- There are multiple dependent, mediating and independent variables;
- There is multi-collinearity among some variables and/or sample size is small relative to the number of variables included in the estimation;
- The investigation is of an explorative nature: although the researcher may have a model in mind, the theoretical specification is not rich enough to allow to specify in advance the model in terms of the latent variables.

As similar problems are likely to be encountered in applied research, CoC can be useful for future research work. The increase in available data at the country level often allows researchers to work with a large number of variables, but only a small number of countries or points in time. This context differs from the "data-rich environment" described by Bernanke and Boivin (2003), which encourages the use of PCR and related methods in finance. The expression "data-rich environment" refers to the availability of datasets with a large number of variables, observations, and points in time. However, it can be argued that the available data are often "rich" relative to the number of observations, in the sense that a substantial amount of information is available for a small number of observations. If this is the case, researchers might be interested in empirical techniques allowing to describe the structure of a dataset through a few key parameters or statistics.

It is not the aim of this paper to give a systematic attempt of the conditions under which CoC outperforms other empirical methods. It is rather to show that, at least under certain conditions, CoC can be a useful empirical approach. After all, the utilization of this empirical approach must be motivated not only by data problems such as small sample size and multi-collinearity, but also by the types of relationships that are believed to better fit the data, and by the objectives of the research project. The research problems for which CoC can be a useful statistical method, and potential alternative methods, are discussed in Section 2. Section 3 describes the three empirical approaches compared in this paper (CoC, PCR and OLS) in the context of a given data generation process.

The simulations presented in this paper, of which the design is presented in Section 4, confirm the conclusions of the discussion of Section 3. Given the very small sample size of the generated dataset, PCR and CoC generally outperform OLS, and relative performance improves further if the problem of multi-collinearity is more severe. CoC is the method that allows to describe the relationships in the data with the lowest number of parameters, which is useful in explorative analysis.[2] Furthermore, boostrapping yields accurate estimates of the standard errors of the coefficients estimated by CoC. One more surprising result is that the ideal number of components to be retained in PCR and CoC can

[2] Note that here I use the term "describe", as opposed to "determine". The reason is that the focus of this paper is on explorative approaches, in which the task of the researcher is to describe the relationships between the observed variables (or, to put it with Jambu (1991, pp. 3–4), "synthesise the content of data") in a parsimonious way, rather than determining the size of some causal relationships. Similarly to other statistical methodologies based on factor or principal components analysis, CoC (despite requiring the estimation of many parameters when estimating the loadings of the principal components), provides the opportunity to describe these relationships in a parsimonious way.

be different depending on the type of relationship that the researcher investigates. For example, when performing PCR, Jolliffe (2002, Chapter 8) suggests excluding components with an eigenvalue lower than a threshold lying between 0.01 and 0.1. The simulations presented in this paper confirm the validity of this rule of thumb for estimating direct relationships between variables, but not for estimating mediated relationships. These results are reported in Section 5.

Section 6 presents an application of CoC to the dataset of Hoareau et al. (2012, 2013). The application is related to the empirical analysis carried out in that paper. However, some additional statistics and estimators are computed. Section 7 draws some conclusions, which must be interpreted with the design of the simulation in mind. In particular, the simulation is based on samples of small size and on the assumption of normality of the generated variables.

2. Definition of the problem

The research problem in Hoareau et al. (2012, 2013)

It is useful to start by describing the studies by Hoareau et al. (2012, 2013), since this paper is motivated by the empirical problems encountered there. Hoareau et al. (2012, 2013) investigate the relationships between variables on higher education policy, higher education performance and economic innovativeness in 32 European countries. They collect 18 indicators at the country level. Six indicators represent higher education policy, for example, organizational autonomy of universities or expenditures per student. The six indicators represent different dimensions of what the authors call "empowerment" of universities. Ten variables represent higher education performance: for example, grants won from the European Research Council per million inhabitants, or proportion of international students in tertiary education. Finally, two variables represent economic innovativeness: the proportion of the labour force employed in knowledge intensive sectors and labour productivity. In the conceptual model of the authors, higher education policies are related to the performance of the higher education sector, which in turn is related to economic innovativeness. The authors perform three principal component factor analyses for the three groups of variables separately.

Principal component factor analysis is a family of techniques for performing factor analysis starting from the extracted principal components (see e.g. Basilevsky, 1994, Chapter 6; Jolliffe, 2002, Chapter 7). In Hoareau et al. (2012, 2013), the factors extracted are in fact standardised principal components. The authors retain and rotate components with an associated eigenvalue greater than one, resulting in one component representing innovativeness, three components representing higher education performance, and three components representing policies. They regress the innovativeness components on the performance components, and the performance components on the policy components. Finally, the authors interpret and analyse the resulting coefficients. In doing so, they comment on the association among different components or variables, without claiming to be uncovering causal relationships. Nonetheless, they interpret the predicted value of the innovativeness component for a certain country (given the value of the policy components) as a measure of the contribution of university policy to economic innovativeness in that country. In line with their conceptual model, they find that higher education policies are associated with higher education performance, which in turn is associated with economic innovativeness in a given country.

Furthermore, they conclude that university policy is best tuned to the innovativeness of the economy in Norway, followed by Cyprus, the UK and a number of Northern and Central European countries.

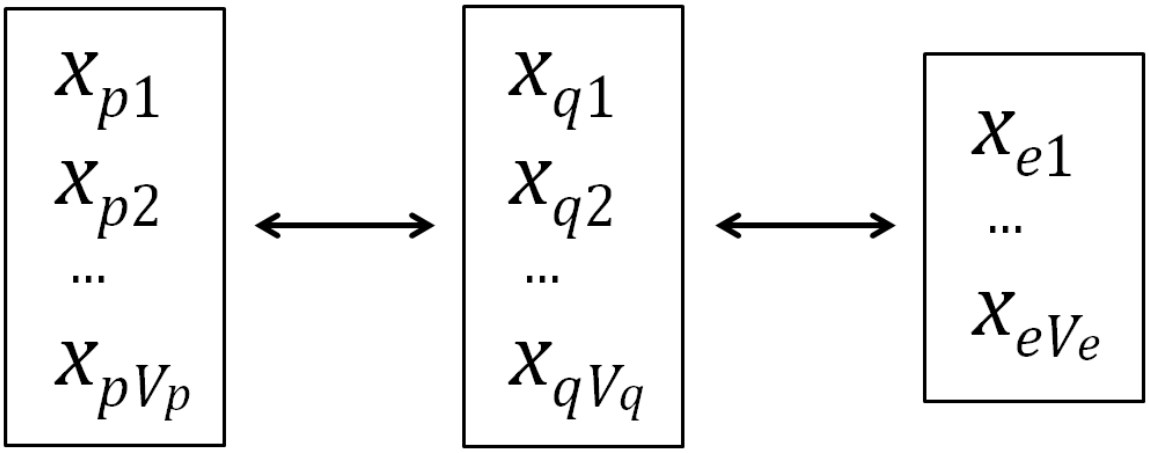
Notice that regressing components on components results in running five regressions, with three independent variables each. This allows the authors not only to find some significant relationships between the dimensions that they investigate (despite the very small sample size), but also to describe these relationships in a parsimonious way. These aspects are discussed in the remainder of this paper, after generalizing the research problem of Hoareau et al. (2012, 2013).

The research problem generalized

Suppose that there are three types of variables: p, q, and e. The variable of the type q are perfect mediators between p and e. This means that the variables of the type p are related to the variables of type e only through the variables of the type q. These relationships do not necessarily have to be causal relationships. For example, it could be that a variable of the type p influences some variables of the type q, but that it is itself affected by some variables of type q.[3] In this paper, the focus is on a closed model, in which each relevant variable can be classified in one of the three types p, q, or e. However, it would not be difficult to extend the model so that it includes other types of variables (for example, exogenous variables that influence some of the variables belonging to the group of the mediating or of the dependent variables).

This is illustrated by the path diagram in Figure 1, where V_p, V_q, and V_e are used to denote the total number of variables in the respective categories, and x_{cj} indicates the jth variable in type c (c=p,q,e).

Figure 1 Path diagram of the relationship between variables



Suppose that the researcher is interested in estimating the expected value of the variables of type e conditional on the variables of type p, and all relevant coefficients of association between couples of variables that allow to construct this estimate. By making appropriate assumptions about the relationships (linearity of the relationship being just one of these), this estimation could be performed by running a set of OLS regressions (see Greene, 2003, Chapters 2–4 for a review of OLS). However,

[3] The fact that the relationships are possibly bi-directional means that the model described in this section is a non-recursive model. Non-recursive models potentially generate complex dynamics for the effects among different variables (Kaplan, 2000). However, in this paper these complications are avoided, as we look at a static model, where the associations are evaluated at a given point in time.

using OLS can be problematic in the context of multi-collinearity and / or small sample size (relative to the number of variables). A broad and intuitive definition of multi-collinearity is given by Belsley et al. (2004, Chapter 86): "[multi-]collinearity exists if there is a high multiple correlation when one of the variates is regressed on the others". The consequences of multi-collinearity for OLS results are well known: in the presence of multi-collinearity the estimates are imprecise, that is, they have high variance. If the sample size is small relative to the variables used in the analysis, the problem is different but the effect is similar. Adding too many variables to the estimation decreases the degrees of freedom. Hence, the precision of the estimates decreases and standard errors increase.

When facing one of these problems, one alternative to OLS is reducing the variables in the *p* and *q* categories to a smaller number of latent factors or of principal components. Strucutural equation modelling (Kaplan, 2000; Williams et al., 2003) and partial least squares-path modelling (see Tenenhaus et al., 2005; Vinzi et al., 2009) are examples of techniques that reduce the variables to factors or components with the purpose of estimating complex relationships (accommodating for moderation and mediation effects). However, these require identifying the latent factors *a priori*. The researcher may not be willing to do so. For example, Horeau et al. (2012, 2013) do not always have theoretical reasons to group the variables together. In fact, each of their higher education policy variables is collected precisely because it represents a different dimension of what they define as "empowerment" of universities (hence, each variable represents a theoretically distinct construct). Therefore, the authors are not interested in estimating relationships between pre-specified factors or components. Instead, they are interested in an explorative analysis of the structure of the variances and covariances in the dataset.[4]

Two more flexible alternatives that allow handling multi-collinearity or small sample size without imposing an *a priori* specification of the latent structure onto the data are principal component regression (Basilevsky, 1994, Chapter 10; Jolliffe, 2002, Chapter 8) and factor analysis regression (Basilevsky, 1994, Chapter 10; Kosfeld and Lauridsen, 2008). The former has received more attention in the statistical literature. Both techniques consist of extracting a number of principal components (or factors) from the set of independent variables, and using them as independent variables in an OLS estimation. The coefficients of the original linear relationship between the dependent and the independent variables can subsequently be recovered, generally with a bias. The estimates will, however, usually be more stable, and the standard errors reduced (Jolliffe, 2002, Chapter 8). Often PCR and factor analysis regression are used as explorative tools, so that the coefficients of the original linear relationship between the variables are not recovered. In that case, the interest lies in the relationship between the dependent variables and the factors (or principal components) derived by the explorative analysis. Principal components or factors can be used in combination with OLS (e.g. Corazzini, Grazzi, & Nicolini, 2011) or other empirical techniques, such as logit regression (e.g. Braun et al., 2013; Jakobsen et al., 2013) or spatial analysis (e.g. Perobelli and Oliveira, 2013).

[4] An alternative would be to split the explorative part of the analysis and the estimation of the parameters. For example, the analysis could be conducted in two steps. In the first step, the relevant components would be identified through an explorative principal components analysis, rotated and interpreted; in the second step, the researcher would use the information gained through the explorative analysis to identify the latent variables of a partial least squares-path modelling analysis. This alternative is not analysed in this paper, but it would be interesting to study its properties in further research.

One reason why the classic PCR model (as well as factor analysis regression and the applications that have been just mentioned) may not be suitable for the model in Figure 1 is that, in the model in Figure 1, there are multiple dependent variables. Another reason is that, in the model in Figure 1, the variables in the category *q* are both independent and dependent variables. These two differences have implications in terms of how to estimate the model once the factors (or principal components) have been extracted. These implications will be explored in the next section. In that section, I will first show how the model in Figure 1 could be estimated by running a large number of OLS or principal component regressions. Then, I will show how CoC can be applied to the problem.

Before closing this section, it is useful to mention that many statistical methods have been developed for dealing with multi-collinearity. Dormann et al. (2013) provide the potentially most comprehensive review to date, presenting 23 different methods for handling multi-collinearity and comparing them by a simulation. Yet, they do not review a number of other methods described in the literature. To give a few examples of methods that have not been included in their review, see those proposed by Chang and Yang (2012), Kiers and Smilde (2007), and Kosfeld and Lauridsen (2008). Most of these methods are related to PCR or structural equation modelling.[5] These two techniques have either been designed for the case of a single dependent variable, or they require the pre-specification of a latent factor structure.

3. OLS, PCR, and CoC

A model of linear relationships with perfect mediation

Suppose that the relationships depicted in Figure 1 are linear. In that case, they can be written as:

$$(1)\begin{cases} X_e = X_q\beta_e + \nu_e \\ X_q = X_p\beta_q + \nu_q \end{cases}$$

Where each X_c is a $N \times V_c$ full column rank matrix of N observations on V_c variables; β_e and β_q are, respectively, a $V_q \times V_e$ and a $V_p \times V_q$ matrices of unknown coefficients; ν_e and ν_q are, respectively, a $N \times V_e$ and a $N \times V_q$ matrices of independently and identically distributed disturbances. These disturbances are orthogonal to each other and are drawn from multivariate normal distributions with mean 0 and covariance matrices Σ_{ve} and Σ_{vq}, respectively, with all non-diagonal elements equal to 0.

If every row of X_p is independently drawn from a multivariate normal distribution, then the variables in the stack matrix X=[X_p,X_q,X_e] (which is a juxtaposition of the three matrices X_p, X_q and X_e) are multivariate normal. This is illustrated in Equation (2), representing the distribution of the variables for the *i*-th row of the data matrix X.

[5] A different family of methods to deal with multi-collinearity is ridge regressions (A. E. Hoerl and Kennard, 1970; R. W. Hoerl et al., 1986). However, one characteristic that ridge regression has in common with PCR-related methods that it is designed for a case with a single dependent variable. Furthermore, ridge regression does not reduce the number of independent variables in the regression. As a result, it is useful for dealing with the problem of multi-collinearity, but it is less useful for the problem of small sample size relative to the number of variables. In general, although multi-collinearity and small sample size have similar effects and can be sometimes dealt with using the same techniques, estimators have more often been designed and tested for dealing with multi-collinearity, rather than small sample size (Dormann et al., 2013).

$$(2)\ x_i = [x_i^p, x_i^q, x_i^e] \sim \mathcal{N}\left([0,0,0], \begin{bmatrix} \Sigma_{pp} & \Sigma_{pq} & \Sigma_{pe} \\ \Sigma_{qp} & \Sigma_{qq} & \Sigma_{qe} \\ \Sigma_{ep} & \Sigma_{eq} & \Sigma_{ee} \end{bmatrix}\right)$$

Where x_i is the stack vector of the variables characterising observation i, i=1,…,N, and it is derived as the juxtaposition of the three vectors of variables of type c=p,q,e observed for the unit i, x_i^p, x_i^q, and x_i^e; Σ_{cc}=E[x_i^c'·x_i^c] is the full-rank variance-covariance matrix of the vector x_i^c for every variable of type c=p,q,e; Σ_{cd}=E[x_i^c'·x_i^d] is the full-rank matrix of covariances between the variables in the vectors x_i^c and x_i^d, d=p,q,e.

Notice that, as it was previously mentioned, System of Equations (1) should not necessarily be read in terms of causality. Instead, System of Equations (1) depicts a linear relationship among the variables of different types which may be only correlational. Nonetheless, it implies that, conditional on the matrix X_q, the expected value of the matrix X_e does not depend on the matrix X_p:

$$(3)\ E[X_e|X_q] = E[X_e|X_q, X_p]$$

Given the linearity of the relationships among the variables, Equation (3) can be expressed as a restriction on the covariance matrices by using linear partitioned projection formulas (Greene, 2003, sec. 3.3; Wooldridge, 2002, App. 2A). The resulting restriction is:

$$(4)\ \Sigma_{pe} - \Sigma_{pq}\Sigma_{qq}^{-1}\Sigma_{qe} = 0$$

By using the two equations presented in System of Equations (1), the equation relating X_e and X_p can be derived:

$$(5)\ X_e = X_p\beta_q\beta_e + \nu_q\beta_e + \nu_e$$

Estimation by Ordinary Least Squares (OLS)

Estimates of the parameters of the system of Equations (1) can be obtained by OLS under appropriate assumptions (see e.g. Greene, 2003, Chapters 2–4). The results can be presented as three matrices containing the estimated coefficients: one for the relationship between X_e and X_q ($\hat{\beta}_{q\text{OLS}}$), one for the relationship between X_q and X_p ($\hat{\beta}_{e\text{OLS}}$), and one for the relationship between X_e and X_p ($\hat{\delta}_{\text{OLS}}$). The vector of the mediated effects, $\hat{\delta}_{\text{OLS}}$, is then obtained by multiplying $\hat{\beta}_{q\text{OLS}}$ by $\hat{\beta}_{e\text{OLS}}$ (Hicks and Tingley, 2011; MacKinnon, 2008).

Estimating the model by OLS can lead to two problems. Firstly, the number of parameters within the matrices $\hat{\beta}_{e\text{OLS}}$ and $\hat{\beta}_{q\text{OLS}}$ is potentially very large, being equal to V_e·V_q+V_q·V_p. If the researcher is interested in describing the relationships in Figure 1 by using a few key parameters, this is inconvenient. Secondly, although all the parameters can be estimated by OLS, the estimates may be very imprecise in case of multi-collinearity or small sample size (relative to the number of variables). Indeed, multi-collinearity and small-sample size are the two reasons indicated by Stone (1947) in his pioneering study for justifying the use of what would have later been known as principal components regression.

Estimation by Principal Components Regression (PCR)

System of Equations (1) can also be estimated by PCR. Let us start the discussion of PCR with the problem of extracting the principal components (by principal component analysis, hereafter PCA) or

the factors (by factor analysis, hereafter FA). PCA and FA are based on different assumptions about the underlying structure of the data, although in practice they often lead to similar results (Jolliffe, 2002, Chapter 7). The underlying idea of PCA on one side is to "reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset" (Jolliffe, 2002, p. 1).[6] PCA leads to the following decomposition of the three groups of variables p, q, and e:

$$(6)\begin{cases} X_p = Z_p A_p \\ X_q = Z_q A_q \\ X_e = Z_e A_e \end{cases}$$

Where each Z_c is a $N \times V_c$ vector of principal components; each A_c is the transpose of the $V_c \times V_c$ orthogonal, full-rank matrix of loadings. The loadings are determined by PCA, and they allow to uniquely determine the components of each full-rank matrix Z_c. On the other side, the basic idea underlying factor analysis is that a number of observed random variables can be expressed, with the exception of an error term, as linear functions of a smaller number of common factors (Jolliffe, 2002, p. 151).

The remainder of this section will focus on PCA rather FA, for a number of reasons. First of all, the fact that the matrix A_c is invertible ensures that Z_c is linear in X_c, which simplifies the discussion in this paper. Furthermore, the model in Figure 1 requires fewer assumptions than those required by a latent factor model, and this could be appropriate in a number of applied research settings. For example, Hoareau et al. (2012, 2013) collect their policy variables because they represent distinct theoretical dimensions. It may not be appropriate to postulate that these distinct dimensions are explained by a number of latent common factors. Another reason is that in applied research very often a factor analysis model is invoked, but the estimations of the factors are obtained by extracting the first principal components and by standardising them (examples are Corazzini et al., 2011; Hoareau et al., 2012, 2013; however, details of the exact estimation of the factors are often omitted from applied research papers). This practice can find a statistical justification in Tipping and Bishop (1999) who show that, under a particular structure of errors in the latent factor model, the factors could be estimated equivalently by PCA or maximum-likelihood FA.

Once the components have been chosen, they can be rotated by using one of a number of algorithms developed for this purpose (see e.g. Jolliffe, 2002, Chapter 11). This helps the interpretation of the components and of the coefficients estimated by principal component regressions. However, since the rotation merely generates linear combinations of the previously extracted components, the number of components involved, as well as the estimators $\hat{\beta}_e$, $\hat{\beta}_q$, and $\hat{\delta}$, remain the same. Hence the discussion and the conclusions of this paper are unaffected by the fact that the components may be rotated.

[6] This definition is more related to what the empirical researcher can do with PCA, than to the underlying assumptions imposed by PCA on the underlying structure of the data. This is consistent with the approach taken by Jolliffe (2002, Chapter 7), who maintains that PCA requires basically no such assumption, different from factor analysis. Indeed, in the DGP used for the simulation of this paper, no restriction on the rank of the data matrix is imposed (i.e., the variables in the dataset are not generated by a smaller number of common factors or components).

By using the systems of Equations (1) and (6), it is possible to re-write the relationships between the variables and principal components of the different groups as:

$$(7)\begin{cases} X_e = Z_q A_q \beta_e + \nu_e \equiv Z_q \gamma_e + \nu_e \\ X_q = Z_p A_p \beta_q + \nu_q \equiv Z_p \gamma_q + \nu_q \end{cases}$$

$$(8)\ X_e = Z_p A_p \beta_q A_q' A_q \beta_e + \nu_q A_q' A_q \beta_e + \nu_e \equiv Z_p \gamma_q A_q' \gamma_e + \nu_q A_q' \gamma_e + \nu_e$$

Notice that systems of Equations (7) and Equation (8) are equivalent to system of Equations (1) and Equation (5), respectively. In fact, the relationships implied by the system of Equations (7) can be estimated by OLS, obtaining estimates of the elements of the two matrices $A_q \cdot \beta_e \equiv \gamma_e$ and $A_p \cdot \beta_q \equiv \gamma_q$. Given that A_q and A_p are known and invertible[7], it is straightforward to derive estimates for β_e and β_q. If all the principal components are included as independent variables, then the results will be identical (in terms of the estimated coefficients and the standard errors of the estimates) to those obtained by running OLS regressions with the original variables.

On the contrary, if some of the components are dropped (or, equivalently, if some of the coefficients in β_q and β_e are constrained to be equal to 0), then the literature on PCR suggests that the coefficients will be estimated with a bias, but (in case of multi-collinearity or small sample size relative to the number of variables) the standard error of the estimate will decrease (Jolliffe, 2002, Chapter 7). Only if the constrained coefficients are truly equal to zero, like in the case of measurement error (e.g. Basilevsky, 1994, Chapter 10), then the coefficients estimated by PCR are unbiased.

Hence, the researcher must decide which coefficients to set equal to zero or, in other words, which components to leave out of the estimation. Usually, in applied research, the choice is to leave out the components with the smallest eigenvalue, although alternative criteria exist and may be more efficient (e.g. Jolliffe, 2002, Chapter 8). In the simulations and the application carried out in the paper, this criterion is applied, leaving out the components with an associated eigenvalue below a pre-specified threshold. This also suits the explorative nature of many applied research papers, whose interest lies in exploring the correlation structure and computing relationships among a few components which carry most of the variance in the data.

After it has been decided which components to exclude, estimators for γ_e and γ_q (hereafter, $\hat{\gamma}_{e\mathrm{PCR}}$ and $\hat{\gamma}_{q\mathrm{PCR}}$) are obtained by a set of PCR regressions. From these estimators, it is possible to obtain estimators for β_q and β_e ($\hat{\beta}_{q\mathrm{PCR}}$ and $\hat{\beta}_{e\mathrm{PCR}}$). The multiplication of these two estimators can, in turn, be used as an estimator for δ ($\hat{\delta}_{\mathrm{PCR}}$).

Notice that the number of parameters in $\hat{\gamma}_{e\mathrm{PCR}}$ and $\hat{\gamma}_{q\mathrm{PCR}}$ (estimated by means of a number of PCRs) is lower than the number of parameters in $\hat{\beta}_{e\mathrm{OLS}}$ and $\hat{\beta}_{q\mathrm{OLS}}$ (that have to be estimated if OLS is used). If a number $r_c < V_c$ of principal components is retained for any category of variables c, then the total number of parameters estimated by PCR is $V_e \cdot r_q + V_q \cdot r_p$. This number is lower than with OLS, which can help in the exposition and interpretations of the results (provided that a useful interpretation of the components, before or after rotation, exists).

Estimation by Components-on-components regression (CoC)

[7] Indeed, since the matrices A_c are orthogonal matrices, their inverse is equivalent to their transpose. This fact is used in the derivations of this section and of Appendix A.

The assumption of a linear relationship among the variables is equivalent to the assumption of a linear relationship among the extracted principal components. Indeed, by combining the systems of Equations (1) and (6), it is possible to write the relationships between principal components as:

$$(9)\ \begin{cases} Z_e = Z_q A_q \beta_e A'_e + \nu_e A'_e \equiv Z_q \tau_e + \nu_e A'_e \\ Z_q = Z_p A_p \beta_q A'_q + \nu_q A'_q \equiv Z_p \tau_q + \nu_q A'_q \end{cases}$$

$$(10)\qquad Z_e = Z_p A_p \beta_q A'_q A_q \beta_e A'_e + \nu_q A'_q A_q \beta_e A'_e + \nu_e A'_e \equiv Z_p \tau_q \tau_e + \nu_q A'_q \tau_e + \nu_e A'_e$$

where system of Equations (9) and Equation (10) are equivalent to system of Equations (1) and Equation (5), respectively, as $\tau_e \equiv A_q \beta_e A'_e$ and $\tau_q \equiv A_p \beta_q A'_q$. Once the components have been extracted, estimates of the parameters of the system of Equations (9) can be obtained by OLS. The difference is that principal components are used not only as independent variables, but also as dependent variables. Estimates for the parameters in β_e, β_q and δ can be recovered; these will coincide with the OLS estimates if all components are used. Alternatively, some components can be excluded from the analysis, as is the case for PCR.

PCR and CoC are thus very similar, as the only difference is that the two sets of dependent variables in system of Equations (7) are replaced with two different and related sets of dependent variables in Equation (9). Hence, it is not unreasonable to expect that some of the characteristics of the two methodologies will be similar. In particular, it is not unreasonable to expect that dropping some of the components among the independent variables will possibly introduce bias in the estimates, but that it may also increase their stability under specific circumstances (typically, the presence of multi-collinearity and / or small sample size). However, excluding a number of components from the set of the dependent variables has the advantage of reducing the number of the estimated parameters in the vectors $\hat{\tau}_e$ and $\hat{\tau}_q$. If a number $r_c < V_c$ of principal components is retained for any category of variables c, then the total number of parameters estimated by CoC is $r_e \cdot r_q + r_q \cdot r_p$. This number is lower than for OLS and PCR, which may help to summarise and interpret the results concisely (provided that a useful interpretation of the components, before or after rotation, can be found).

It is interesting to notice that Hoareau et al. (2012) use CoC in their analysis, but impose a restriction on the estimate of Z_e, by replacing all negative elements of $\hat{\beta}_e$ and $\hat{\beta}_q$ with 0. This introduces further bias into the estimation of Z_e, in addition to the bias introduced by discarding the components with the smallest eigenvalue. Unlike the latter source of bias (excluding some of the components), imposing a minimum value of 0 for the components of $\hat{\beta}_e$ and $\hat{\beta}_q$ does not necessarily benefit the analysis by increasing the precision of the estimates. Hence these restrictions are not imposed in the application shown in this paper, and the estimates are obtained using the procedure that has just been described.

Relationship between OLS, PCR, and CoC

OLS and PCR yield identical estimates of the parameters β_e, β_q and δ if all components are included (see Jolliffe, 2002, for a discussion of the relationship between OLS and PCR). The same holds true for CoC. Furthermore, as can be seen in Appendix A, it is sufficient that all the components in Z_e are included in the estimation procedure to obtain the same estimator $\hat{\delta}$ with PCR and CoC.

4. Simulation design

Dataset design

In light of the previous discussion, PCR and CoC can be expected to be more efficient than OLS under specific circumstances, such as small sample size and multi-collinearity. Furthermore, CoC is expected to estimate the relationships between the different groups of variables with the smallest number of parameters. In this section, two datasets (Dataset 1 and Dataset 2) are generated to evaluate the relative performance of OLS, PCR and CoC in a statistical environment similar to that of Horeau et al. (2012, 2013), and to see if it is possible to estimate correctly the standard errors of CoC estimates by bootstrapping in this environment.

Each dataset contains data on 18 variables (six of the type p, ten of the type q, and two of the type e – the same numbers of variables as in Hoareau et al., 2012, 2013) for a hypothetical population of 10000 observations. The simulation consists of repeatedly extracting samples of 32 observations from this population, estimating the relationships between the variables according to each of the three statistical methods, and recording the statistics of interest. This subsection describes the procedure used for generating the datasets. The next subsection explains which indicators were used for the comparison of the three methods.

Dataset 1 was generated to mimic some characteristics of the model described by Equations (1) – (5) and some characteristics of the dataset used by Hoareau et al. (2012, 2013). The generated variables are distributed according to a multivariate normal distribution, where for each group of variables the correlation matrix is the same as in the dataset used by Hoareau et al. (2012, 2013). This full-rank correlation matrix is reported in Appendix B. Each variable has an expected value equal to 0 and unit standard deviation. The data were generated as follows:

$$(11) \qquad \begin{cases} W_p = \varepsilon_p \\ X_p = W_p B_p \\ W_q = W_p T_q + \varepsilon_q \\ X_q = W_q B_q \\ W_e = W_q T_e + \varepsilon_e \\ X_e = W_e B_e \end{cases}$$

Where the notation is explained in the following paragraphs.

Each X_c is a full-rank matrix containing the values for the variables of type c=p,q,e. Hence, X_p is a 10000 X 6 data matrix, X_q is a 10000 X 10 matrix, and X_e is a 10000 X 2 matrix.

W_p, W_q, and W_e are sets of vectors only used to generate the data matrices X_p, X_q, and X_e, respectively. Each of them is characterised by the same dimensionality as its respective data matrix.

T_q and T_e are, respectively, a $V_p \times V_q$ and a $V_q \times V_e$ matrix of parameters. Each of the parameters is independently drawn from a standard normal distribution.

Each B_c is a $V_c \times V_c$ full rank matrix of parameters generated in a manner so that the covariance matrix for X_c is the same as the respective correlation matrix in Hoareau et al. (2012) (which is reported in

Appendix B), and that the variance of every variable in the matrix X_c is equal to one.[8] Note that the fact that B_c is full rank implies that there is no perfect collinearity in X_c, so that there will be some bias in the PCR and CoC estimates of the parameters in the simulation. Hence, the simulated environment is favourable to OLS in this respect. It would be interesting, in further research, to generate data matrices of reduced rank, so that the PCR and CoC estimates could be unbiased.

Each ε_c is a set of vectors $\varepsilon_{c1},\ldots,\varepsilon_{cV_c}$ of disturbances independently drawn from a multivariate normal distribution with mean 0 and the following covariance matrix $\Sigma_{\varepsilon\varepsilon}$:

$$
(12) \qquad \Sigma_{\varepsilon\varepsilon} = \begin{array}{c|cccccccc}
 & \varepsilon_{p1} & \cdots & \varepsilon_{p6} & \varepsilon_{q1} & \cdots & \varepsilon_{q10} & \varepsilon_{e1} & \varepsilon_{e2} \\
\varepsilon_{p1} & 1 & & & & & & & \\
\cdots & \cdots & \cdots & & & & & & \\
\varepsilon_{p6} & 0 & 0 & 1 & & & & & \\
\varepsilon_{q1} & 0 & 0 & 0 & \sigma^2_{\varepsilon q} & & & & \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & & & \\
\varepsilon_{q10} & 0 & 0 & 0 & 0 & 0 & \sigma^2_{\varepsilon q} & & \\
\varepsilon_{\varepsilon 1} & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{\varepsilon e} & \\
\varepsilon_{\varepsilon 2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2_{\varepsilon e}
\end{array}
$$

$\sigma^2_{\varepsilon q}$ is chosen in a manner that the expected fraction of unexplained variance over explained variance in the equation $W_q = W_p T_q + \varepsilon_q$ (i.e., the equivalent of the statistic $1/(1\text{-}R^2)$ in an OLS regression) is equal to 1.79. This is the average value for the statistic $1/(1-R^2)$ that is obtained after regressing all variables of type q on the X_p matrix, using the dataset of Hoareau et al. (2012, 2013). The parameter $\sigma^2_{\varepsilon p}$ is similarly chosen so that the expected fraction of unexplained variance over explained variance is equal to 1.97.

An important feature of this dataset is that, given the procedure with which it has been constructed, X_p and X_e are correlated, but they are not correlated conditional on X_q. Hence, the restrictions provided by Equations (3) and (4) hold and X_q is a perfect mediator in the relationship between X_p and X_e.

Dataset 2 is generated using an identical procedure, but the correlation between any two variables belonging to a given category is the square root of the absolute value of the correlation between the same two variables in Dataset 1. Hence, the problem of multi-collinearity in Dataset 2 is more severe. Only two different datasets have been constructed because the purpose of the paper is not to extensively describe the relative performance of OLS, PCR, and CoC in a wide range of situations, but to discuss the CoC estimator and show that it has satisfying properties in a number of circumstances.

Indicators used for the comparison

In previous literature, the comparison between OLS and PCR is often found among comparisons of a larger number of techniques (e.g. Faber and Kowalski, 1997; R. W. Hoerl et al., 1986; Kiers and Smilde, 2007; Kosfeld and Lauridsen, 2008; Merola and Abraham, 2001; Mittelhammer and Baritelle, 1977). The results described in the aforementioned literature confirm the conclusions of the theoretical

[8] Each B_c is a trasformation of the square root of the estimated correlation matrix for the variables of the respective category *c* used in the application. To obtain B_c, each column of this matrix is multiplied by a scalar to ensure that the resulting variance of the variables of type *q* is equal to 1.

literature: if multi-collinearity is a serious problem, PCR outperforms OLS in terms of accuracy of the estimators and out-of-sample prediction. In these studies, the indicator which is most often used for comparing different estimators is the mean squared error (or a closely related measure, like its square root), both for accuracy of the parameters estimators, and for out-of-sample prediction. For this reason, the mean square error will be used as the main criterion for comparing different methods in this study.

Once the dataset has been constructed, 500 random samples of 32 observations were drawn for each of the three techniques described in Section 3 (OLS, PCR and CoC). Thirty-two was chosen as the number of observations because it is the same as in Hoareau et al. (2012, 2013). For each sample and method I computed the estimators $\hat{\beta}_q$, $\hat{\beta}_e$, and $\hat{\delta}$.[9]

In order to compare the three different methods with each other, I computed the mean of the square difference between estimated and true parameters for the 500 estimates, distinctly for the parameters belonging to $\hat{\beta}_q$ or $\hat{\beta}_e$ and for those belonging to $\hat{\delta}$:[10]

$$(13) \qquad MSE_\beta = \frac{1}{500}\sum_{k=1}^{500}\frac{1}{80}\left[\sum_{i=1}^{10}\sum_{j=1}^{6}\left(\hat{\beta}_{qijk}-\beta_{qij}\right)^2+\sum_{i=1}^{10}\sum_{h=1}^{2}\left(\hat{\beta}_{eihk}-\beta_{eih}\right)^2\right]$$

$$(14) \qquad MSE_\delta = \frac{1}{500}\sum_{k=1}^{500}\frac{1}{12}\left[\sum_{j=1}^{6}\sum_{h=1}^{2}\left(\hat{\delta}_{jhk}-\delta_{jh}\right)^2\right]$$

Where: β_{qij} is the parameter describing the relationship between the jth variable of the type p and the ith variable of the type q; β_{eih} is the parameter describing the relationship between the hth variable of the type e and the ith variable of the type q; δ_{ih} is the parameter describing the relationship between the hth variable of the type e and the jth variable of the type p; $\hat{\beta}_{qijk}$, $\hat{\beta}_{eihk}$, and $\hat{\delta}_{ihk}$ are the estimators for the respective parameter for the sample k, k=1,…,500. Notice that PCR and CoC are two-steps estimation procedures. In the first step, an estimate of the components is produced for the sets of variables of each type c=p,q,e, based on the sample estimate of the components' loadings. In the second step, the estimated components are used to estimate the relationships between the different sets of variables. As a result, the errors in the parameter estimates of β_e, β_q, and δ (which are aggregated and summarised by the indicators presented in Equations (13) and (14)) can be generated

[9] It is important to notice that when estimating the parameters by PCR and CoC, the loadings of the components are estimated for each sample by PCA, as described in Section 3. An alternative, which could be considered in future research, is using the 'true' loadings as defined in the data generating process. This could shed light on whether the so-called 'generated regressor problem' (see Pagan, 1984; or Westerlund and Urbain, 2011 for a more recent discussion with a focus on principal components) plays an important role in the context of the simulation presented in this study.

[10] These measures represent the arithmetic average of the mean squared errors across different parameter estimates, following the approach used in the aforementioned literature. However, it could be the case that the researcher is particularly concerned with the possibility that, although most of the estimators are close to the true parameters, a few estimators are very far from the true values. In this case, it could be interesting to use, instead of the arithmetic average, a weighted average of the squared errors across different parameter estimates, with weights depending on the precision with which the parameters are estimated. Additionally, some of the estimated parameters (belonging to the vectors $\hat{\beta}_q$, $\hat{\beta}_e$, or $\hat{\delta}$) may be more interesting than others, for example, in light of policy reasons. Also in this case, the comparison between OLS, PCR and CoC could be based on a weighted average of the squared errors across different parameter estimates. The weights, in this case, would depend on the relative importance, in light of policy considerations, of different parameter estimates. In the simulation carried out in this paper, there is no particular reason to believe that some of the generated parameters would be more or less important than others.

by each of the two steps: the estimation of the loadings for each sample and the choice of which components to retain; and the estimation of the parameters of the regressions for each sample.

For comparing the predictive ability between samples, the following indicator is used:

$$(15) \qquad MSE_e = \frac{1}{500}\sum_{k=1}^{500}\sum_{l=1}^{10000}\frac{1}{10000}\left\{\frac{1}{2}\left[(\hat{x}_{e1lk} - x_{e1l})^2 + (\hat{x}_{e2lk} - x_{e2l})^2\right]\right\}$$

where x_{e1l} is the true value of the variable x_{e1} (of the e-type variables) for the lth observation in the generated dataset; x_{e2l} is the respective value for the variable x_{e2l}; and $\hat{x}_{e1l}$ and $\hat{x}_{e2l}$ are their predicted values given the parameters estimated based on the kth sample.

An additional indicator used for comparing OLS, PCR and CoC is the average number of regression coefficients that have to be estimated to describe the relationships between the three groups of variables. As discussed, this is equal to $V_e \cdot V_q + V_q \cdot V_p$ in the case of OLS, which is 80 in the datasets used for this simulation. For PCR, it is $V_e \cdot r_q + V_q \cdot r_p$, and for CoC it is $r_e \cdot r_q + r_q \cdot r_p$. The average number of regression coefficients to be estimated is labelled *PARS*. The choice to use this particular indicator in the comparison of the three estimation techniques is not due to statistical reasons, but to more general reasons that have to do with the possibility of concisely interpreting the results. As it was mentioned in Section 3, if the researcher can suggest a good interpretation for the components after the exploratory data analysis, and if the number of regression coefficients is not large, then the results can be commented upon in a concise and effective way. Hence, in the context of the type of problem analysed in this paper, a lower number of estimated regression coefficients is potentially a desirable characteristic of an estimator.[11]

Each of these four indicators has been computed for OLS, PCR and CoC. Furthermore, for PCR and CoC, different rules were applied to select the principal components to be retained.

As discussed in Section 3, when applying PCR and CoC in this paper, the estimated principal components are retained and used in the analysis only if their estimated eigenvalue is greater than a specified threshold. Hence, it is of great interest to compare OLS, PCR and CoC when using different thresholds to select the principal components. The threshold, representing the minimal eigenvalue above which components are retained and used in the analysis, is the same for the variables of each group p, q, and e and it varies in the simulation from 0 to 1 with intervals of 0.1. Notice that choosing zero as the minimum eigenvalue makes the PCR and CoC estimators identical to the OLS estimator, since all principal components are retained. Conversely, if the minimum eigenvalue is set equal to one, then only those principal components with an eigenvalue equal to or greater than one will be retained, following the so-called Kaiser rule (Guttman, 1954; Kaiser, 1960). The choice which components to retain is guided by two partially conflicting objectives (reducing the variance of the estimate and avoiding the introduction of excessive bias – Jolliffe, 2002, Chapter 8). Thus, it seems likely that a trade-off will emerge. Increasing the minimum eigenvalue will likely increase the precision of the estimate at the cost of increasing the bias. The combined effect on the *MSE*s is therefore not clear *a priori*.

Before closing this paragraph, note that much wider range of possible eigenvalues is explored than what is usually suggested in the literature. For example, Jolliffe (2002, Chapter 8) suggests minimal

[11] Notice that *PARS* refers to the number of estimated regression coefficients, and not to the number of parameters in $\hat{\beta}_q$, $\hat{\beta}_e$, and $\hat{\delta}$, which is always equal to $V_e \cdot V_q$, $V_q \cdot V_p$ and $V_e \cdot V_p$, respectively (which is equal, in the simulation presented in this paper, to 20, 60 and 12 parameters, respectively).

eigenvalues between 0.01 and 0.1. However, in explorative analysis the need to reduce the dataset to a smaller number of components can often render greater minimal eigenvalues more desirable.

Besides comparing these four indicators for PCR, OLS and CoC, I also compute standard errors by bootstrapping for the CoC estimates. Bootstrapping is a suitable method for estimating standard errors for CoC estimates, because it allows deriving standard errors without knowing the theoretical distribution of the parameters (Davidson and MacKinnon, 2006; Efron and Tibshirani, 1986), and it performs reasonably well in small samples (Yung and Chan, 1999).

For each of the 500 samples, the parameters were estimated by CoC, and the standard errors of the parameters were computed by bootstrapping. This allowed to compute two key statistics, that can be compared to each other. One of these statistics is the average of the standard errors estimated by bootstrapping across the 500 estimates. The other one is the "true" standard error of the estimated parameters or, in other words, the standard deviation of the estimated parameter across the 500 sample estimates. If the values of the two statistics are close to each other, then bootstrapping can be regarded as an appropriate tool for computing standard errors, at least in situations resembling the simulated datasets (with homoscedastic error terms in the data-generation process). In order to generate the bootstrap statistics, I involved what Efron and Tibshirani (1986, p. 57) called "two levels of Monte Carlo". For 500 times, first a sample was drawn from the generated data and then, with this sample fixed, 500 bootstrap samples were drawn. For each bootstrap sample, the 12 parameters in δ were estimated, along with the respective standard errors (remember that in the simulated dataset, V_e=2 and V_p=6, so that the number of parameters in δ is 12). As a result, for each of these parameters, it has been possible to compute and compare the average estimated standard error, the variance of the estimated standard error, and the real standard deviation of the parameter estimate. The bootstrap samples were generated with what Davidson and MacKinnon (2006, p. 820) define as the "non-parametric procedure", which is probably the most commonly used. This procedure "amounts to drawing each observation of a bootstrap sample randomly, with replacement, from the original sample".

5. Simulation results

Comparison between OLS, PCR, and CoC

Figure 2 shows the comparison between the estimated MSE_β and MSE_δ for Dataset 1. On the horizontal axis, it is shown the eigenvalue representing the threshold above which components are retained in PCR and CoC. For example, on the left side of the figure (eigenvalue=0) the comparison between OLS, PCR and CoC has been carried out for the case in which all components have been retained. Conversely, on the right side of the figure (eigenvalue=1) the so-called Kaiser rule has been used to choose the components to use in the PCR and CoC estimation. The first indicator (MSE_β) measures the accuracy of the estimators $\hat{\beta}_e$ and $\hat{\beta}_q$ for each of the three different methods. In other words, it measures the ability to yield an accurate estimate for the direct relationships between, on one side, X_p and X_q; and, on the other side, X_q and X_e. The second indicator (MSE_δ) measures the accuracy of the estimator matrix $\hat{\delta}$. In other words, it measures the ability to yield an accurate estimate for the overall, mediated relationship between X_p and X_e. The *MSE*s are on the vertical axis. The dotted lines represent MSE_β and MSE_δ for OLS. These are horizontal lines, because changes in the minimum eigenvalues are relevant only for the

PCR and CoC estimators, but not for OLS. The solid and dashed lines represent MSE_β and MSE_δ for CoC and PCR, respectively. They are close to each other, which means that PCR and CoC yield similar estimates for Dataset 1. Note that OLS, PCR and CoC are identical if the minimal eigenvalue is equal to zero, so that the values for MSE_β and MSE_δ are the same if the minimal eigenvalue is equal to 0 (except for a small sampling error).

In general, MSE_β is much lower than MSE_δ for all three techniques (often MSE_β is less than half as large as MSE_δ). Furthermore, both MSE_δ and MSE_β decrease for small minimal eigenvalues, indicating that there is a gain in the accuracy of the estimates when using PCR or CoC. However, if the minimal eigenvalue increases above 0.1, MSE_β increases, and PCR and CoC perform worse than OLS for eigenvalues greater than 0.7. Conversely, MSE_δ decreases from approximately 3% if the minimal eigenvalue is set equal to zero, to approximately 2.2% if the eigenvalue is 0.5, and it stabilises for larger eigenvalues. This is a very interesting result, which indicates that the optimal retention rule may differ according to which parameters the researcher wants to estimate. In other words, the trade-off between the stability of the estimates and the amount of bias apparently affects the estimator differently for different parameters. As mentioned in the previous section this is particularly interesting because an often-used rule of thumb in PCR suggests to use a minimal eigenvalue between 0.01 and 0.1. In light of Figure 2, this is justified in the classic framework of PCR, which estimates the parameters of a direct, "X→Y" relationship. These parameters can be compared to the parameters in the matrices β_e and β_q. However, Figure 2 also shows that a different choice of the minimal eigenvalue may be optimal when estimating parameters of a mediated relationship such as the one between X_p and X_e.

Figure 3 shows the same indicators for Dataset 2, where the problem of multi-collinearity is more severe. The overall pattern is very similar to that in Figure 2, but PCR and CoC perform better than OLS. For example, if the minimum eigenvalue equals 0.5, the ratio between the value of MSE_β for CoC and for OLS is equal to 0.95 for the simulation using Dataset 1, and to 0.61 using Dataset 2. Similarly, the ratio between the value of MSE_δ for CoC and for OLS is equal to 0.77 for the simulation using Dataset 1, and to 0.42 using Dataset 2.

The performance of OLS, PCR and CoC in the two different datasets with respect to their predictive ability (as measured by MSE_e) is shown in Figure 4. The value of the mean squared error is approximately 0.0001, but it is higher when using Dataset 1 than when using Dataset 2. Again, the indicators for OLS are on the horizontal dotted lines, whereas the indicators for PCR and CoC (which are consistently very close to each other) are on the dashed and the solid line, respectively. MSE_e is (as expected) approximately equal for all the three methods if the eigenvalue equals zero, but it is lower for PCR and CoC if the eigenvalue is greater than zero, indicating that the latter methods perform better. Again, the difference between OLS and the other methods is larger in Dataset 2 (the ratio between the MSE_e using CoC and using OLS is 0.98 for Dataset 1 and 0.94 for Dataset 2). This confirms expectations, because the problem of multi-collinearity is more severe in Dataset 2.

Figure 2 Comparison between the *MSE*s obtained by using different estimation techniques and retention rules (Dataset 1)

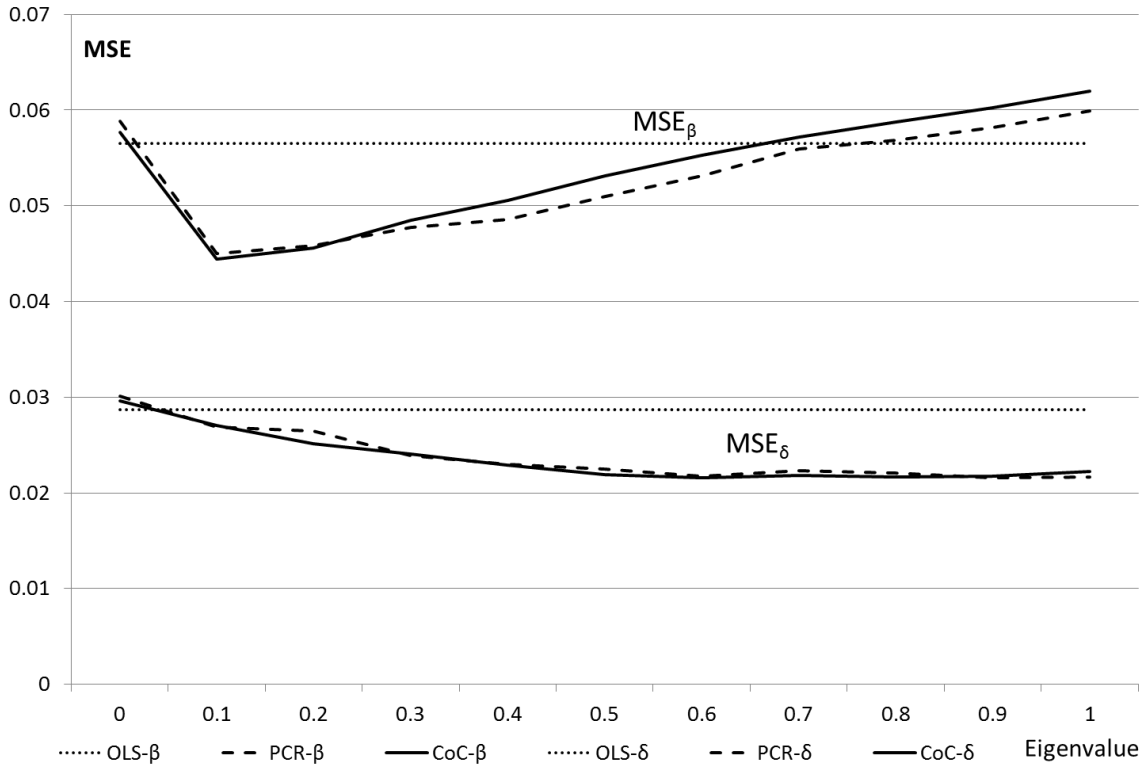


Figure 3 Comparison between the *MSEs* obtained by using different estimation techniques and retention rules (Dataset 2)

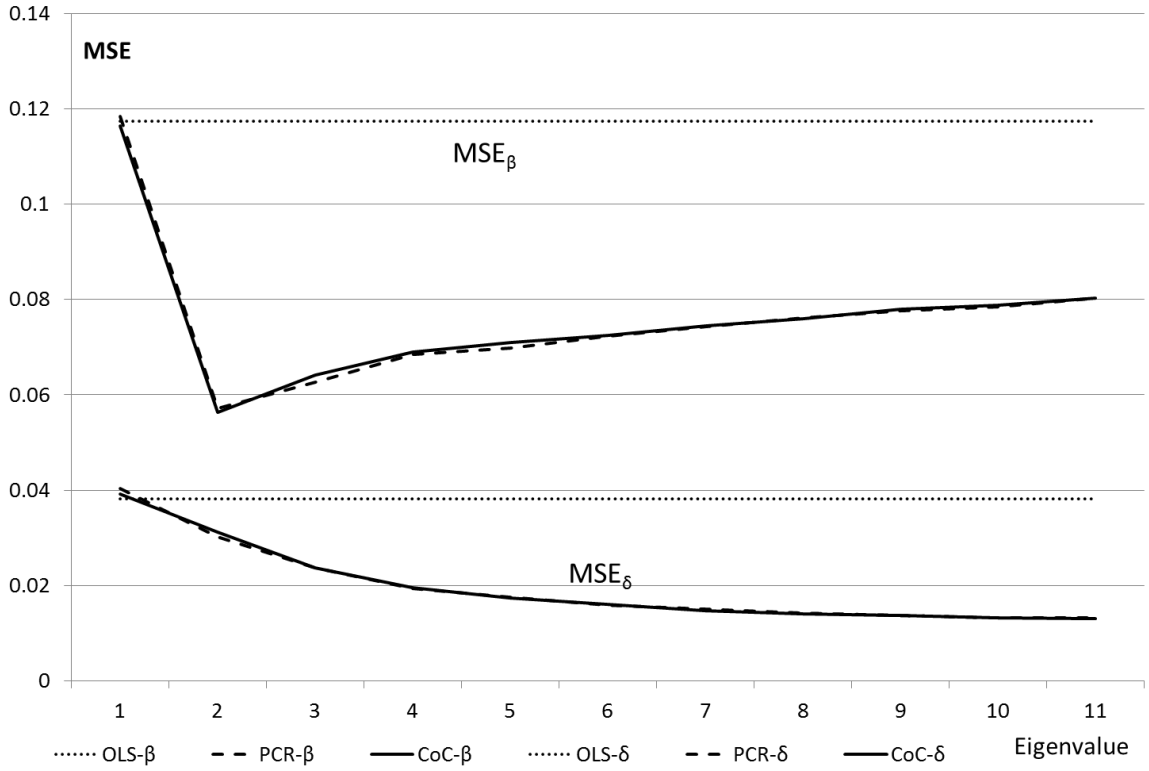


Figure 4 Comparison of MSE_e for OLS, PCR, and CoC, using the two different datasets

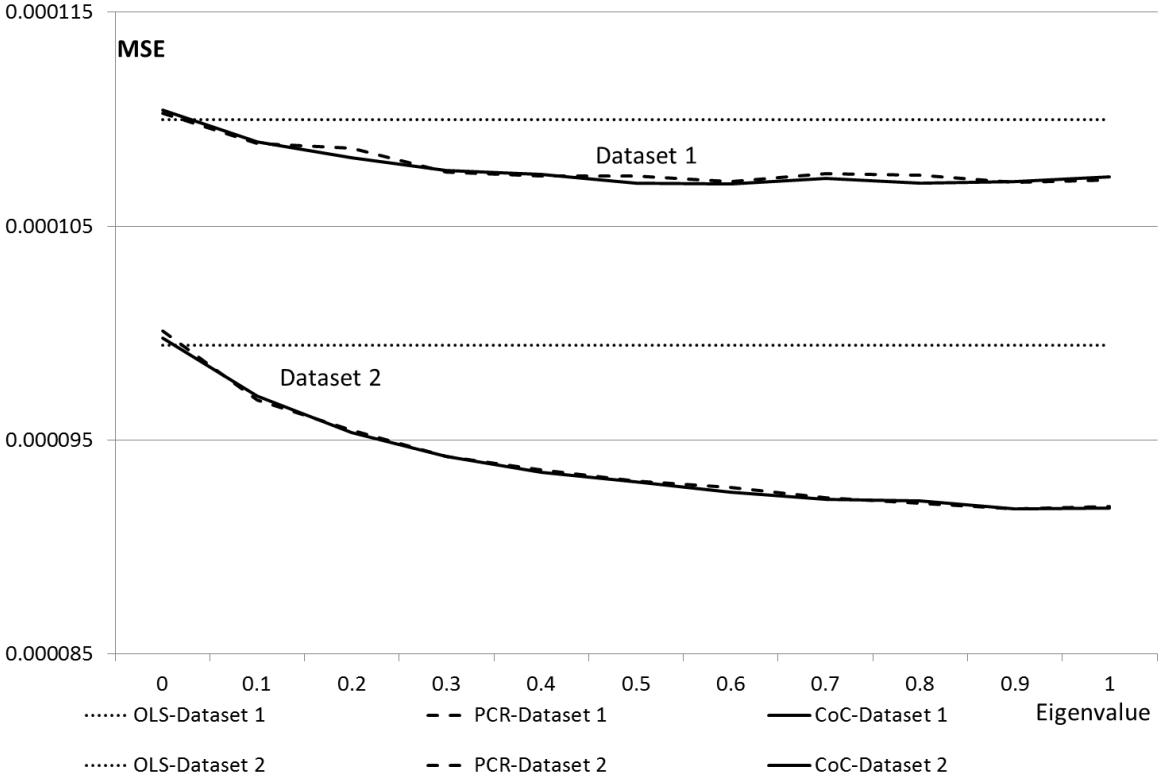
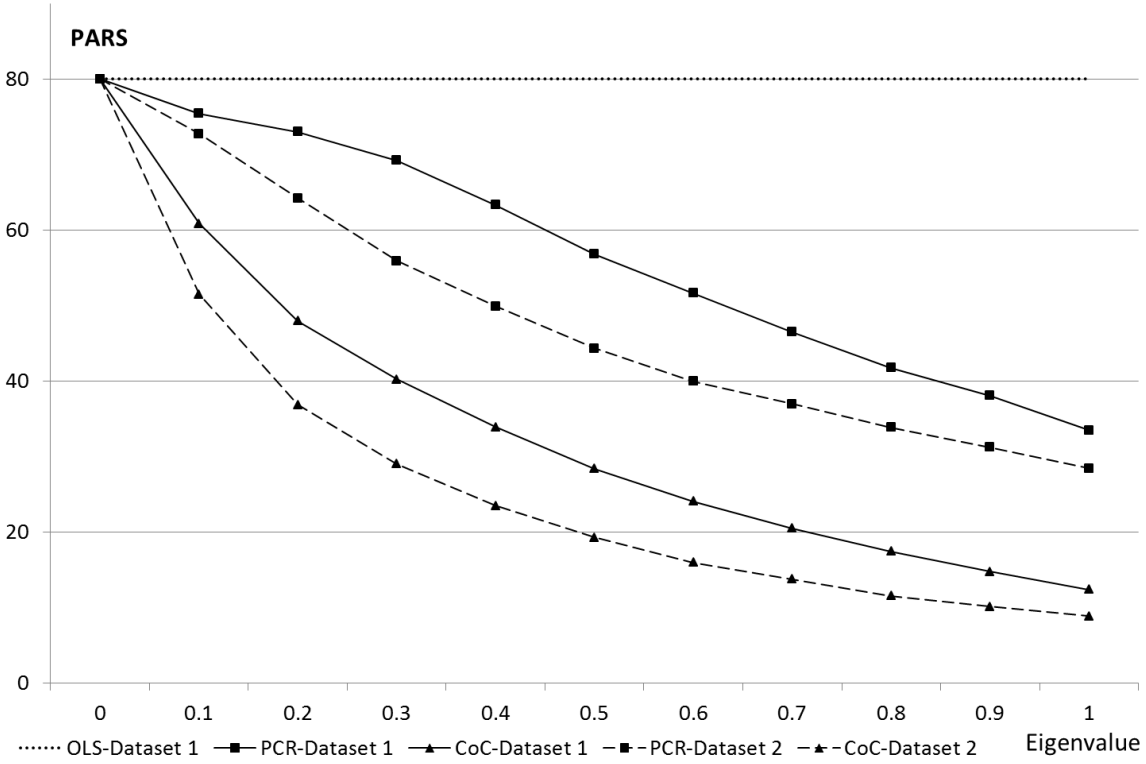


Figure 5 Comparison of *PARS* for OLS, PCR, and CoC, using the two different datasets



The indicator $PARS$ indicates how many parameters are used, on average, to describe the relationships in Figure 1. This indicator is reported in

Figure 5. If its value is low, then the researcher can describe the relationships in the data in a more parsimoniously (provided that an interesting interpretation exists for the components associated to the parameters). As discussed, this value must be lowest for CoC and highest for OLS. For OLS, $PARS=V_e \cdot V_q + V_q \cdot V_p$, which is equal to 80 (horizontal dotted line in

Figure 5). For CoC and PCR, $PARS$ is equal to 80 if the minimal eigenvalue is equal to zero, and it decreases steadily for increasing eigenvalues. Using Dataset 1 for example, $PARS$ decreases at more or less a constant rate of 14-17% per decimal point of the minimal eigenvalue for CoC (solid line marked by triangles), and it is equal to 12.4 if the minimal eigenvalue is equal to one. In general, $PARS$ for CoC is lower than for PCR, and it is lower using Dataset 2 than using Dataset 1.

To summarise, the simulations confirm what was expected from the discussion of the three methods. Given the very small sample size, PCR and CoC generally outperform OLS, and relative performance improves further the more severe the problem of multi-collinearity is. CoC is the method that allows to describe the relationships in the data with the lowest number of parameters, which can be useful in an explorative analysis. Although it is not the goal of the simulations to systematically show the relative performance of the three methods under different conditions, this exercise shows that, under certain conditions, CoC is a more attractive statistical tool OLS and PCR.

Bootstrapped standard errors for CoC

Bootstrapped standard errors were computed for the parameters in the matrix δ, estimated by CoC, and for three different minimal eigenvalues: 0, 0.5, and 1. I focused on δ and the three eigenvalues mentioned above in order to keep the exposition brief, but the results are qualitatively unchanged for the parameters in β_e and β_q, and for different eigenvalues.

Table 1 and Table 2 report (for Dataset 1 and Dataset 2, respectively) three statistics related to the standard error of each $\hat{\delta}_{ij}$, where $\hat{\delta}_{ij}$ is the estimator of the effect of variable x_j (belonging to X_p) on variable x_i (belonging to X_e), i=1,…,6, j=1,2. The first statistic is the average of the estimated standard error of $\hat{\delta}_{ij}$, labelled as $\hat{\sigma}$. In other words, the standard error of each parameter $\hat{\delta}_{ij}$ is estimated by bootstrapping for each of the 500 samples, and then $\hat{\sigma}$ is defined as its average. The second statistic is the standard deviation of the estimated standard error of $\hat{\delta}_{ij}$, $\mathrm{SD}(\hat{\sigma})$. This means that the standard error of each parameter $\hat{\delta}_{ij}$ is estimated by bootstrapping for each of the 500 samples, and then $\mathrm{SD}(\hat{\sigma})$ is defined as its standard deviation. The third statistic is the "true" standard error of $\hat{\delta}_{ij}$, labelled as σ. This is obtained in the following way: first, each parameter $\hat{\delta}_{ij}$ is estimated for each of the 500 samples; second, σ is computed as the standard deviation of the 500 parameter estimates. Table 1 and Table 2 report these three statistics for the three cases of minimal eigenvalues equal to 0, 0.5 or 1. The goal of the comparison of these three statistics is to see how close $\hat{\sigma}$ and σ are to each other, especially relative to $\mathrm{SD}(\hat{\sigma})$.

Table 1 shows that the standard deviation of the standard errors estimates, $\mathrm{SD}(\hat{\sigma})$, is equal to approximately 15 to 20% of the average estimate $\hat{\sigma}$ or of the "true" parameter σ (depending on the exact $\hat{\delta}_{ij}$). In most cases, the average estimated standard error $\hat{\sigma}$ is close to the "true" standard error (for example, in 78% of the cases, σ lies within the interval $\hat{\sigma}\pm\mathrm{SD}(\hat{\sigma})$, and it always lies within the interval $\hat{\sigma}\pm2\cdot\mathrm{SD}(\hat{\sigma})$). Furthermore, there is no clear evidence of an upward or downward bias (in 25 of 36 cases the relationship is $\hat{\sigma}>\sigma$, while in the remaining 11 cases the contrary is true). All three statistics are generally lower for higher minimal eigenvalues, reflecting an increase in the stability of the estimate if more components are excluded. The results are very similar in Table 2.

As expected (given the extensive literature on the properties of bootstrapping) it can be concluded that bootstrapped standard errors are good estimators for the real standard errors of the estimators in $\hat{\delta}$. This means that a measure of the precision of the CoC estimators can be obtained by bootstrapping. However, in the simulation reported in this paper, the errors in the data-generating process are homoscedastic, and all variables are approximately normally distributed. Researchers applying CoC in contexts in which these assumptions do not hold should evaluate the appropriateness of the bootstrap procedure case by case.

Table 1 Estimated standard errors (with respective standard deviation) and "true" standard errors, Dataset 1

Eigenv.	Stat.	$\hat{\delta}_{11}$	$\hat{\delta}_{12}$	$\hat{\delta}_{13}$	$\hat{\delta}_{14}$	$\hat{\delta}_{15}$	$\hat{\delta}_{16}$	$\hat{\delta}_{21}$	$\hat{\delta}_{22}$	$\hat{\delta}_{23}$	$\hat{\delta}_{24}$	$\hat{\delta}_{25}$	$\hat{\delta}_{26}$
0	$\hat{\sigma}$	0.186	0.187	0.186	0.196	0.196	0.182	0.200	0.195	0.195	0.204	0.208	0.189
	σ	0.191	0.190	0.180	0.190	0.163	0.182	0.203	0.211	0.205	0.157	0.196	0.157
	$\mathrm{SD}(\hat{\sigma})$	0.033	0.033	0.033	0.034	0.037	0.033	0.033	0.032	0.031	0.034	0.035	0.030
0.5	$\hat{\sigma}$	0.108	0.106	0.101	0.102	0.111	0.105	0.109	0.106	0.101	0.102	0.114	0.105
	σ	0.144	0.102	0.130	0.090	0.085	0.094	0.139	0.103	0.130	0.091	0.085	0.094
	$\mathrm{SD}(\hat{\sigma})$	0.020	0.019	0.017	0.019	0.022	0.017	0.021	0.019	0.017	0.018	0.023	0.018
1	$\hat{\sigma}$	0.067	0.066	0.064	0.065	0.071	0.066	0.067	0.066	0.064	0.065	0.071	0.066
	σ	0.065	0.061	0.065	0.056	0.063	0.061	0.065	0.061	0.065	0.056	0.063	0.061
	$\mathrm{SD}(\hat{\sigma})$	0.013	0.014	0.013	0.012	0.014	0.014	0.013	0.014	0.013	0.012	0.014	0.014

Table 2 Estimated standard errors (with respective standard deviation) and "true" standard errors, Dataset 2

Eigenv.	Stat.	$\hat{\delta}_{11}$	$\hat{\delta}_{12}$	$\hat{\delta}_{13}$	$\hat{\delta}_{14}$	$\hat{\delta}_{15}$	$\hat{\delta}_{16}$	$\hat{\delta}_{21}$	$\hat{\delta}_{22}$	$\hat{\delta}_{23}$	$\hat{\delta}_{24}$	$\hat{\delta}_{25}$	$\hat{\delta}_{26}$
0	$\hat{\sigma}$	0.22	0.237	0.256	0.255	0.244	0.227	0.229	0.246	0.266	0.263	0.252	0.235
	σ	0.173	0.238	0.235	0.334	0.238	0.196	0.185	0.222	0.214	0.324	0.27	0.225
	$\mathrm{SD}(\hat{\sigma})$	0.042	0.044	0.05	0.052	0.049	0.043	0.044	0.047	0.054	0.053	0.05	0.045
0.5	$\hat{\sigma}$	0.097	0.098	0.104	0.099	0.095	0.103	0.097	0.098	0.104	0.099	0.095	0.103
	σ	0.111	0.088	0.09	0.076	0.1	0.111	0.111	0.088	0.09	0.076	0.1	0.111
	$\mathrm{SD}(\hat{\sigma})$	0.019	0.02	0.021	0.021	0.018	0.02	0.019	0.02	0.021	0.021	0.018	0.02
1	$\hat{\sigma}$	0.056	0.065	0.06	0.055	0.06	0.055	0.056	0.065	0.06	0.055	0.06	0.055
	σ	0.045	0.054	0.057	0.052	0.062	0.058	0.045	0.054	0.057	0.052	0.062	0.058
	$\mathrm{SD}(\hat{\sigma})$	0.014	0.014	0.015	0.013	0.014	0.014	0.014	0.014	0.015	0.013	0.014	0.014

6. Application

In this section, the dataset analysed by Hoareau et al. (2012, 2013) is used to illustrate how CoC can be applied in practice. The units of observation are 32 European countries, and the variables are

standardised before the analysis. However, the coefficients of association between different variables that are estimated in this section are reported in terms of the original units of measurement.

Table 3 List of variables

Category	Indicator	Year	Source
Policy	Organisational autonomy (1 to 6 scale)	2008	CHEPS
	Policy autonomy (1 to 6 scale)	2008	CHEPS
	Financial autonomy (1 to 6 scale)	2008	CHEPS
	Expenditures per student as a % of GDP per capita (tertiary education)	2008	OECD and World Bank
	Expenditure on financial aid as a % of total public expenditures on education (tertiary education)	2006-2008	Eurostat
	Role of formulas and contract in funding mechanisms (1 to 6 scale)	2008	CHEPS
Performance	Scientific publications within the 10% most cited scientific publications worldwide as a % of total scientific publications per country	2007	European Commission Innovation Unit
	Number of universities in the top 500 Academic Ranking of World Universities (ARWU) per million inhabitants	2011	ARWU
	Number of incoming Marie Curie fellows per million inhabitants	2008-2009	European Commission Innovation Unit
	European Research Council Starting grant wins per million inhabitants	2011	European Research Council
	Public-private co-publications per million inhabitants	2008	Pro Innovation Union Scoreboard
	Employment rates of 18-34 years old, 3 years or less after leaving tertiary education	2010	Eurostat
	Graduates to enrolments ratio (tertiary education)	2010	Eurostat
	Transition: students entering tertiary education through an alternative route (vocational training, accreditation of prior learning, etc.)	2008-2011	Eurostudent
	Students in tertiary education aged 20 as a % of total population in the corresponding age	2010	Eurostat
	International students: inward mobile students as a % of the student population	2009	Eurostat
Economic innovativeness	Employment in knowledge intensive activities as a % of total employment	2009	European Commission Innovation Unit
	GDP per hour worked in Purchasing Power Standard units	2009-2010	European Commission Innovation Unit

Table 3 is adapted from Hoareau et al. (2012, pp. 11–13). It shows the variables used in that study, as well as in this section. The variables are grouped into three categories: policy, performance and innovativeness. In the theoretical framework of the authors, the university policies of a country impact the performance of the university system, which in turn affects economic innovativeness. Performance perfectly mediates the relationship between policy and innovativeness.

The policy variables are intended to capture the policy tools that can be used by governments to influence the performance of the national higher education system. The policy variables are: three variables representing different aspects of autonomy (organisational, policy, and financial autonomy); one variable intended to capture the efficiency of the incentive system that the government creates for universities (the extent to which public, higher education funding is allocated by means of formulas and contracts – see CHEPS, 2008); one variable measuring the expenditures per student in higher education, relative to GDP per capita; and one variable indicating how much of these expenditures is devoted to student aid.

The performance category includes ten indicators, representing several dimensions of higher education performance at country level such as research (e.g. the number of grants from the European Research Council won in a given country, relative to population), internationalisation (e.g. international students) or student retention (graduates to enrolment ratio).

Finally, there are two variables linked to economic innovativeness: the proportion of workers employed in knowledge-intensive activities and labour productivity (measured as GDP per hour worked). For more details on the variables chosen and on the coding, as well as on the rationale behind the choice of the variables, the reader is referred to Hoareau et al. (2012, 2013).

Following the authors of these studies, no control variable is included such as the overall level of employment, or indicators for the level of the physical capital stock in a country. Notice that, while for the sake of tractability the assumption of multivariate normality of the variables was maintained throughout Sections 3 to 5, several of the variables in this dataset are discrete variables. This is a limitation of this paper. Since in applied work non-normally distributed variables are often encountered, it will be interesting to relax the normality assumption in future studies on CoC. Finally, it must be mentioned that there is no claim of uncovering causal relationships in this section. The interest is rather in the relative ability of CoC, OLS and PCR of yielding precisely estimated association coefficients, conditional on the other variables included in the estimation.

Table 4, Table 5 and Table 6 report estimated coefficients and standard errors (computed by bootstrapping) for the estimated association between the policy variables and the innovativeness variables for CoC, PCR and OLS, respectively. The association between policy and performance indicators or between performance and innovativeness indicators can also be estimated according to the procedure described in the previous sections, but it is not reported here, as it does not add substantially to the comparison between the three methods analysed in this paper.

The eigenvalue under which components are discarded for CoC and PCR is one. The components can be rotated in any way, yielding identical results.

The estimates are derived under the assumption that this association is perfectly mediated by the performance variables. Note that CoC allows to test this assumption based on a regression of the innovativeness component on the performance components and the policy components. The coefficients for the performance components in the regression are jointly significant at a 1% confidence level. In contrast, the coefficients for the policy components are jointly not significant (p-value = 0.66), indicating that the policy components are not associated with the innovativeness component conditional on the performance components. Interestingly, if the performance components are not included in the regression, then the coefficients for the policy components are jointly significant at a 10% confidence level, indicating that they are associated with the innovativeness component.

As Table 4 shows, for most of the policy variables the association estimated by CoC is lower than the standard deviation in absolute values. The exceptions are expenditures per student and financial aid, whose coefficient is more than twice as large as the standard error (the coefficient-to-standard error ratio equals 2.47 and 2.94, respectively). An increase of 1% in the expenditure per student relative to GDP per capita is associated with a 0.25% increase in the share of workforce employed in knowledge-intensive activities, and with a 0.40€ increase in GDP per hour worked. An increase of 1% in the

expenditure related to student aid relative to total public expenditures in higher education is associated with a 0.16% increase in the share of the workforce employed in knowledge-intensive activities, and with an increase in GDP per hour worked approximately equal to 0.26€.

Notice that since there are only two variables in the innovativeness category, only one component with an eigenvalue greater than one is extracted. As a result, the relative association coefficient of different policy variables with the two innovativeness variables is always the same. For example: the ratio between the effect of expenditures on knowledge-intensive employment and on labour productivity is equal to 0.64; and the ratio between the effect of student aid on knowledge-intensive employment and on labour productivity is equal to 0.64 as well. This is not a general characteristic of CoC, but it is a restriction which is necessarily imposed if there are only two dependent variables and the minimum eigenvalue is chosen to be one.

Table 4 Estimated association between policy and innovativeness variables – CoC

	Employment in knowledge intensive activities		GDP per hour worked	
	Coefficient	Std. Error	Coefficient	Std. Error
Organisational autonomy	0.657	0.752	1.02	1.169
Policy autonomy	0.391	1.042	0.607	1.619
Financial autonomy	-0.898	1.162	-1.396	1.805
Expenditure per student	0.254	0.103	0.395	0.16
Student financial aid	0.165	0.058	0.256	0.09
Role of formulas and contract	0.112	0.413	0.174	0.641

Table 5 shows that PCR leads to very similar estimates as CoC. Again, the only two variables with a coefficient to standard error ratio greater than two are expenditure and student aid. Also the size of the estimated coefficients is similar. Therefore, on the basis of the CoC and PCR estimation, it is possible to conclude that expenditure per student and student aid are robustly associated with the share of the workforce employed in knowledge-intensive activities and with labour productivity.

Two differences between CoC and PCR are worth mentioning. The first difference lies in the number of regression coefficient that have to be estimated to compute the coefficients reported in Table 4 and Table 5. This is equal to 12 for CoC (only four regressions need to be estimated), and to 36 for PCR (with 12 regressions to be estimated). Thus, if there is a plausible interpretation for the components extracted from the dataset, CoC allows a more parsimonious interpretation of the results. A good example of this can be found in Hoareau et al. (2012, 2013). The second difference is that the ratio between the two association coefficients of each policy variable with the two innovativeness variables is not constrained to be the same in PCR.

Table 5 Estimated association between policy and innovativeness variables – PCR

	Employment in knowledge intensive activities		GDP per hour worked	
	Coefficient	Std. Error	Coefficient	Std. Error
Organisational autonomy	0.527	0.718	1.222	1.265
Policy autonomy	0.315	1.096	0.725	1.754
Financial autonomy	-0.623	1.054	-1.805	1.861
Expenditure per student	0.233	0.112	0.428	0.168
Student financial aid	0.147	0.062	0.284	0.104
Role of formulas and contract	0.031	0.413	0.301	0.744

OLS yields less precise estimates, as Table 6 shows. The coefficient-to-standard error ratio never exceeds two, and its maximum is 1.68 (for the association between expenditures per student and knowledge-intensive employment). As a result, it is not possible to conclude that any of the policy variables are robustly associated with any of the innovativeness variables on the basis of the OLS estimation.

Table 6 Estimated association between policy and innovativeness variables – OLS

	Employment in knowledge intensive activities		GDP per hour worked	
	Coefficient	Std. Error	Coefficient	Std. Error
Organisational autonomy	-0.144	1.642	1.807	2.402
Policy autonomy	0.771	1.823	1.72	2.749
Financial autonomy	-0.299	1.545	-2.736	2.494
Expenditure per student	0.455	0.27	0.719	0.414
Student financial aid	0.131	0.115	0.174	0.17
Role of formulas and contract	0.713	0.831	1.948	1.227

Hoareau et al. (2012, 2013) interpret the predicted value of the innovativeness component on the basis of the policy components as a measure of the suitability of national higher education policies to economic innovativeness. They rank European countries accordingly. Given the importance that the authors attribute to this constructed variable, it is interesting to see how using the three estimation methods analysed in this paper affects it. Instead of the predicted value of the innovativeness component as in Hoareau et al. (2012, 2013), the predicted value for labour productivity is reported here. The two measures are perfectly correlated for CoC, because of the implicitly imposed restriction that the ratio between the association coefficient between each policy variable and the two innovativeness variables must be the same. Hence, these lead to the same country ranking in the case of CoC. The predicted values based on CoC, PCR and OLS are computed by multiplying each of the policy variables by the respective coefficient in the third column of Table 4, Table 5 and Table 6, respectively.

Table 7 shows that the scores are almost perfectly correlated (correlation coefficient = 0.998) for CoC and PCR. This implies that constructing the ranking based on the PCR and CoC coefficients leads to very similar results. For example, the first ten countries remain the same for CoC and PCR: Cyprus, Norway, Sweden, Germany, Denmark, the UK, Iceland, Austria and Belgium. However, there are small changes within the top ten: For example, in the ranking computed on the basis of CoC Cyprus comes first, whereas if PCR is used, Norway scores higher. OLS leads to different conclusions. The correlation coefficient between the score for OLS and CoC is equal to 0.843. Germany, which came fourth and third in the CoC and PCR ranking (respectively), has the highest predicted labour productivity. Cyprus, which was in the top two positions, comes fifteenth if the ranking is based on the OLS results. The position of Italy, just below the median position in the rankings based on CoC and PCR, worsens by eight ranks.

The application shown in this section illustrates how CoC can be applied to a dataset with many variables compared to the number of observations, yielding more precise estimates than OLS. PCR yields very similar estimates to CoC, but the latter allows a more parsimonious interpretation of the results.

Table 7 Predicted value for labour productivity in Euro (and respective rank) for the sample countries, according to CoC, PCR and OLS

country	Pred. Lab. Prod. – CoC	Rank – CoC	Pred. Lab. Prod. – PCR	Rank – PCR	Pred. Lab. Prod. – OLS	Rank – OLS
Cyprus	41	1	41.5	2	32.9	15
Norway	40.9	2	41.9	1	40.6	2
Sweden	36.4	3	36.6	5	37.4	3
Germany	36.4	4	37.7	3	44.3	1
Denmark	36.1	5	36.6	6	36.1	8
Netherlands	36.1	6	36.8	4	36.6	6
UK	36	7	36.5	7	37.2	4
Iceland	32.5	8	32.9	8	33.6	12
Austria	32.4	9	32.8	9	34.7	10
Belgium	32.3	10	32.5	10	34.2	11
Spain	31.8	11	32.3	11	36.2	7
Portugal	31.6	12	31.7	13	33.5	14
Hungary	31.4	13	31.9	12	33.5	13
Finland	31.2	14	31.3	16	35.9	9
Ireland	31.2	15	31.3	17	32.3	16
Slovenia	31.2	16	31.5	15	30.9	17
France	31.1	17	31.5	14	36.6	5
Italy	30	18	30	18	24.7	26
Switzerland	28.8	19	28.5	19	30.4	18
Croatia	28.7	20	28.1	21	25.4	25
Turkey	28.5	21	28.3	20	28.8	19
Lithuania	28.2	22	27.9	22	23.7	27
Estonia	27.4	23	27	23	28.4	20
Romania	27	24	26.6	24	27.2	21
Bulgaria	26.8	25	26.3	25	26.2	24
Latvia	26.4	26	26.1	26	26.4	23
Malta	26.3	27	25	29	20.5	32
Slovakia	25.8	28	25.2	27	20.7	31
Poland	25.8	29	25.1	28	26.7	22
Luxembourg	25.5	30	24.9	30	23	28
Greece	24	31	23.5	31	21.8	30
Czech Rep.	23.5	32	22.6	32	22.2	29

7. Conclusions

In this paper, the estimator used by Hoareau et al. (2012, 2013) for their explorative analysis of university policies is described and compared to alternative estimators using a discussion, a simulation and an application of the dataset used by Hoareau et al. (2012, 2013). CoC is suitable for explorative analyses with multiple independent, mediating and dependent variables, and where there are problems of multi-collinearity or small sample size.

CoC is an adaptation of principal component regression (PCR) in the context of multiple dependent variables and a mediated relationship between variables. It can also be considered as a special case of structural modelling with latent variables estimated by principal components. Like all other methods designed for multi-collinearity or small sample size, however, CoC does not “solve” the problem. Multi-collinearity and small sample size relate to an insufficiency of information in the data that cannot be eliminated. However, even in the presence of these problems some methods might be superior to others in terms of robust model fitting, prediction, or required assumptions.

CoC has the potential to lead to a more parsimonious empirical model, and smaller standard errors for the estimated coefficients, than ordinary least squares or PCR. Furthermore, standard errors can be

satisfactorily estimated by bootstrapping, at least if the error terms are homoscedastic. However, these conclusions rest on a simulation based on samples of small size (about 30 observations) and on particular assumptions on the data generating process (in particular, normality of the generated errors and variables). To generalise the results beyond this particular setting, it is necessary to investigate the properties of CoC in different contexts (e.g. large samples or non-normally distributed variables).

References

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods*. New York: Wiley.

Belsley, D. A., Kuh, E., and Welsch, R. E. (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (2nd ed.). Hoboken (NJ): John Wiley & Sons.

Bernanke, B. S., and Boivin, J. (2003). Monetary Policy in a Data-rich Environment. *Journal of Monetary Economics*, *50*, 525–546.

Braun, A., Müller, K., and Schmeiser, H. (2013). What Drives Insurers' Demand for Cat Bond Investments? Evidence from a Pan-European Survey. *The Geneva Papers on Risk and Insurance - Issues and Practice*, *38*, 580–611.

Chang, X., and Yang, H. (2012). Combining Two-Parameter and Principal Component Regression Estimators. *Statistical Papers*, *53*, 549–562.

CHEPS. (2008). *Progress in Higher Education Reform across Europe - Governance and Funding Reform - Volume 2: Methodology, Performance Data, Literature Survey, National System Analyses and Case Studies* (Vol. 2). Brussels: European Commission.

Corazzini, L., Grazzi, M., and Nicolini, M. (2011). Social Capital and Growth in Brazilian Municipalities. In P. Nijkamp & I. Siedschlag (Eds.), *Innovation, Growth and Competitiveness - Dynamic Regions in the Knowledge-Based World Economy* (pp. 195–217). Berlin: Springer.

Davidson, R., and MacKinnon, J. G. (2006). Bootstrap Methods in Econometrics. In T. C. Mills & K. Patterson (Eds.), *Palgrave Handbook of Econometrics* (pp. 812–838). New York: Palgrave Macmillan.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: a Review of Methods to Deal With It and a Simulation Study Evaluating Their Performance. *Ecography*, *36*, 27–46.

Efron, B., and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, *1*, 54–75.

Faber, K., and Kowalski, B. R. (1997). Propagation of Measurement Errors for the Validation of Predictions Obtained by Principal Component Regression and Partial Least Squares. *Journal of Chemometrics*, *11*, 181–238.

Greene, W. H. . (2003). *Econometric Analysis*. *Prentice Hall* (Vol. 97). Upper Saddle River (NJ): Prentice Hall.

Guttman, L. (1954). Some Necessary Conditions for Common-factor Analysis. *Psychometrika*, *19*, 149–161.

Hicks, R., and Tingley, D. (2011). Causal Mediation Analysis. *The Stata Journal*, *11*, 1–15.

Hoareau, C., Ritzen, J., and Marconi, G. (2012). *The State of University Policy for Progress in Europe - Technical report*. Maastricht: EEU.

Hoareau, C., Ritzen, J., and Marconi, G. (2013). Higher Education and Economic Innovation, a Comparison of European Countries. *IZA Journal of European Labor Studies*, *2*, 24.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*, 55–67.

Hoerl, R. W., Schuenemeyer, J. H., and Hoerl, A. E. (1986). A Simulation of Biased Estimation and Subset Selection Regression Techniques. *Technometrics*, *28*, 369–380.

Hoyle, R. H. (1999). *Statistical Strategies for Small Sample Research*. (R. H. Hoyle, Ed.). Thousand Oaks (CA): Sage Publications.

Jakobsen, T. G., De Soysa, I., and Jakobsen, J. (2013). Why Do Poor Countries Suffer Costly Conflict? Unpacking per Capita Income and the Onset of Civil War. *Conflict Management and Peace Science*, *30*, 140–160.

Jambu, M. (1991). *Exploratory and Multivariate Data Analysis*. London: Academic Press Limited.

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Kaplan, D. (2000). *Structural Equations Modeling: Foundations and Extensions*. Thousand Oaks (CA): Sage Publications.

Kiers, H. A. L., and Smilde, A. K. (2007). A Comparison of Various Methods for Multivariate Regression with Highly Collinear Variables. *Statistical Methods and Applications*, *16*, 193–228.

Kosfeld, R., and Lauridsen, J. (2008). Factor Analysis Regression. *Statistical Papers*, *49*, 653–667.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. New York: Erlbaum.

Merola, G. M., and Abraham, B. (2001). Dimensionality Reduction Approach to Multivariate Prediction. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, *29*, 191–200.

Mittelhammer, R. C., and Baritelle, J. L. (1977). On Two Strategies for Choosing Principal Components in Regression Analysis. *American Journal of Agricultural Economics*, *59*, 336–343.

Pagan, A. (1984). Econometric Issues in the Analysis of Regressions with Generated Regressors. *International Economic Review*, *25*, 221–247.

Perobelli, F. S., and Oliveira, C. C. C. De. (2013). Energy Development Potential: An Analysis of Brazil. *Energy Policy*, *59*, 683–701.

Stebbins, R. A. (2001). *Exploratory Research in the Social Sciences*. Thousand Oaks (CA): Sage Publications.

Stone, R. (1947). On the Interdependence of Blocks of Transactions. *Journal of the Royal Statistical Society*, *9*, 1–45.

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., and Lauro, C. (2005). PLS Path Modeling. *Computational Statistics and Data Analysis*, *48*, 159–205.

Tipping, M. E., and Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*, 611–622.

Vinzi, V. E., Chin, W. W., Henseler, J., and Wang, H. (2009). *Handbook of Partial Least Squares: Concepts, Methods and Applications*. *Handbook of Partial Least Squares*. Berlin: Springer.

Westerlund, J., and Urbain, J. (2011). *Cross Sectional Averages or Principal Components?* (No. 11/053). Maastricht: METEOR.

Williams, L. J., Edwards, J. R., and Vandenberg, R. J. (2003). Recent Advances in Causal Modeling Methods for Organizational and Management Research. *Journal of Management*, *29*, 903–936.

Wooldridge, J. D. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge (MA): MIT Press.

Yung, Y.-F., and Chan, W. (1999). Statistical Analyses Using Bootstrapping: Concepts and Implementation. In R. H. Hoyle (Ed.), *Statistical Strategies for Small Sample Research* (pp. 81–105). Thousand Oaks (CA): Sage Publications.

Appendix A

(a)

Suppose that the system of Equations (7) is estimated through a set of PCRs as described in Section 2, so that V_e+V_q regressions are run. After extracting the components from a data matrix X_c, a number m_c of components are retained for being used as dependent variables in the PCRs. Also, suppose that these components are arranged in the matrix M_c.

The estimators $\hat{\gamma}_e$ and $\hat{\gamma}_q$ are obtained by least squares regression of the variables in X_e and X_q on (respectively) the retained components M_q and M_p. As a result:

$$\hat{\gamma}_e = (M_q'M_q)^{-1}M_q'X_e = (M_q'M_q)^{-1}M_q'Z_eA_e \equiv \hat{\tau}_eA_e$$
$$\hat{\gamma}_q = (M_p'M_p)^{-1}M_p'X_q = (M_p'M_p)^{-1}M_p'Z_qA_q \equiv \hat{\tau}_qA_q$$

Where $\hat{\tau}_e$ is defined as the matrix of coefficients derived from regressing every principal component in Z_e on the retained principal components M_q; and $\hat{\tau}_q$ is similarly defined as the matrix of coefficients derived from regressing the principal components in Z_q on the retained principal components M_{p}.

The PCR estimators for β_e, β_q and δ are:

$$\hat{\beta}_{ePCR} = A_{Mq}'\hat{\gamma}_e = A_{Mq}'\hat{\tau}_eA_e$$
$$\hat{\beta}_{qPCR} = A_{Mp}'\hat{\gamma}_q = A_{Mp}'\hat{\tau}_qA_q$$
$$\hat{\delta}_{PCR} = \hat{\beta}_{ePCR} \cdot \hat{\beta}_{qPCR}$$

Where A'_{Mc} is made of the first m_c columns of the matrix A'_c.

(b)

Now, suppose that the system of Equations (9) is estimated by a set of PCRs, where the same m_c components are used as independent variables in the regressions as in point (a), but all the components are used as dependent variables, so that V_e+V_q regressions are run. The estimated coefficients can be arranged in the two matrices $\hat{\tau}_e$ and $\hat{\tau}_q$. Given the relationships in the system of Equations (6), it can be easily verified that the estimators for β_e and β_q are the same as in point (a). Hence, the estimation of the system of Equations (9) by PCR gives the same parameters as the estimation of the System of Equations (7) by PCR.

(c)

If the system of Equations (9) is estimated using a set of PCRs, where the same m_c components are used as independent variables in the regressions as in point (a), and only the retained components are used as dependent variables (so that V_e+r_q PCRs are run), then the same estimator for δ as in point (b) is obtained.

To understand why, let us define O_{c} as the matrix containing the principal components that have not been retained, so that all the vectors contained in Z_{c} are contained in either M_{c} or O_{c}. The matrix A_c can be decomposed accordingly into two matrices A_{Mp} and A_{Op}, so that the System of Equations (6) can be re-written as:

$$(6b)\begin{cases} X_p = M_p A_{Mp} + O_p A_{Op} \\ X_q = M_q A_{Mq} + O_q A_{Oq} \\ X_e = Z_e A_e \end{cases}$$

Where for every c, the dimensionality is as follows: M_c is $N \times r_c$, O_c is $N \times (V_c - r_c)$, A_{Mc} is $r_c \times r_c$, and A_{Oc} is $(V_c - r_c) \times (V_c - r_c)$.

Is possible to re-write the system of Equations (9):

$$(9b)\begin{cases} Z_e = M_q \tau_e^s + O_q \cdot \tau_e^r + error_{Ze} \\ M_q = M_p \tau_q^{sM} + O_p \cdot \tau_q^{rM} + error_{Mq} \\ O_q = M_p \tau_q^{sO} + M_p \tau_q^{rO} + error_{Oq} \end{cases}$$

Where the dimensionality of the matrices of parameters is as follow: $\tau^s{}_q$ is r_q X V_e; $\tau^r{}_q$ is $(V_q - r_q)$ X V_e; $\tau^{sM}{}_q$ is r_p X r_q; $\tau^{rM}{}_q$ is $(V_p - r_p)$ X r_q; $\tau^{sO}{}_q$ is r_p X $(V_q - r_q)$; $\tau^{rO}{}_q$ is $(V_p - r_p)$ X $(V_q - r_q)$. Consequently, the vectors of parameters in system of Equations (9) are re-written as:

$$\tau_e = \begin{bmatrix} \tau_e^s \\ \tau_e^r \end{bmatrix}$$
$$\tau_q = \begin{bmatrix} \tau_q^{sM} & \tau_q^{rM} \\ \tau_q^{sO} & \tau_q^{rO} \end{bmatrix}$$

Estimating the system of Equations (9) by PCR implies setting the estimators $\hat{\tau}^r{}_e$, $\hat{\tau}^{rM}{}_q$, and $\hat{\tau}^{rO}{}_q$ equal to zero. Estimating the system of Equations (9) by CoC implies setting $\hat{\tau}^r{}_e$, $\hat{\tau}^{rM}{}_q$, $\hat{\tau}^{rO}{}_q$, and $\hat{\tau}^{sO}{}_q$ equal to zero. Hence, the equations in the system of Equation (9b) are estimated by OLS. Notice that the estimators $\hat{\tau}^s{}_e$, $\hat{\tau}^r{}_e$, $\hat{\tau}^{sM}{}_q$, $\hat{\tau}^{rM}{}_q$, $\hat{\tau}^{rO}{}_q$, and $\hat{\tau}^{sO}{}_q$ do not depend on each other, since they are either in different equations or they estimate the parameters of orthogonal variables (because M_c and O_c are orthogonal by construction for every c). As a result, the estimators $\hat{\tau}^s{}_e$ and $\hat{\tau}^{sM}{}_q$ are identical when using CoC and PCR, because they were estimated the same way.

The estimate of δ by CoC is equal to:

$$\hat{\delta}_{CoC} = \hat{\beta}_{qCoC}\hat{\beta}_{eCoC} = A_p'\hat{\tau}_{qCoC}A_q A_q'\hat{\tau}_{eCoC}A_e = [A_{Mp}' \quad A_{Op}'] \begin{bmatrix} \hat{\tau}_q^{sM} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A_{Mq} \\ A_{Oq} \end{bmatrix} [A_{Mq}' \quad A_{Oq}'] \begin{bmatrix} \hat{\tau}_e^s \\ 0 \end{bmatrix} A_e$$

Algebraic manipulation leads to the following identity:

$$\hat{\delta}_{CoC} = A_{Mp}'\hat{\tau}_q^{sM}A_{Mq}A_{Mq}'\hat{\tau}_e^s A_e$$

The estimate of δ by PCR is equal to:

$$\hat{\delta}_{PCR} = \hat{\beta}_{qPCR}\hat{\beta}_{ePCR} = A_p'\hat{\tau}_{qPCR}A_q A_q'\hat{\tau}_{ePCR}A_e = [A_{Mp}' \quad A_{Op}'] \begin{bmatrix} \hat{\tau}_q^{sM} & 0 \\ \hat{\tau}_q^{sO} & 0 \end{bmatrix} \begin{bmatrix} A_{Mq} \\ A_{Oq} \end{bmatrix} [A_{Mq}' \quad A_{Oq}'] \begin{bmatrix} \hat{\tau}_e^s \\ 0 \end{bmatrix} A_e$$

Algebraic manipulation leads to the following identity:

$$\hat{\delta}_{PCR} = A_{Mp}'\hat{\tau}_q^{sM}A_{Mq}A_{Mq}'\hat{\tau}_e^s A_e = \hat{\delta}_{CoC}$$

Note that $\hat{\beta}_{ePCR}$=$\hat{\beta}_{eCoC}$ because the two vectors are estimated identically. However, it can be seen that:

$$\hat{\beta}_{qCoC} = A_p'\hat{\tau}_{qCoC}A_q = A_{Mp}'\hat{\tau}_q^{sM}A_{Mq} \neq A_{Mp}'\hat{\tau}_q^{sM}A_{Mq} + A_{Op}'\hat{\tau}_q^{sO}A_{Oq} = \hat{\beta}_{qPCR}$$

Appendix B

Table of correlation for Dataset 1.

	x_{e2}	x_{e1}	x_{q10}	x_{q9}	x_{q8}	x_{q7}	x_{q6}	x_{q5}	x_{q4}	x_{q3}	x_{q2}	x_{q1}	x_{p6}	x_{p5}	x_{p4}	x_{p3}	x_{p2}	x_{p1}
x_{p1}	0	0.03	-0.14	-0.13	-0.15	-0.09	-0.18	-0.16	-0.17	-0.06	-0.1	-0.13	-0.04	0.09	-0.09	0.07	0.31	1
x_{p2}	-0.02	0.06	-0.02	0.22	-0.1	-0.04	-0.02	0.01	0.04	-0.32	0.06	-0.17	-0.02	0.01	-0.02	0.19	1	
x_{p3}	0.04	0.03	0.09	-0.07	0.04	0.07	-0.17	-0.04	0.34	0.05	-0.13	0.17	-0.27	0.03	0.14	1		
x_{p4}	0.06	0.05	-0.16	0.04	-0.14	0.09	0.05	-0.02	-0.15	-0.32	-0.05	0.12	-0.11	0.39	1			
x_{p5}	0.04	0.06	0.2	0.12	0.25	0.23	0.2	0.02	-0.07	-0.3	0.01	0.32	-0.18	1				
x_{p6}	-0.01	-0.02	0.1	0.06	0.06	0.23	0.12	-0.11	-0.14	-0.08	-0.03	0.11	1					
x_{q1}	0.09	0.09	0.37	0.14	0.49	0.46	0.24	0.01	0.01	-0.22	0	1						
x_{q2}	-0.19	-0.11	0.16	0.06	0.13	0.03	0.33	-0.01	-0.05	0.22	1							
x_{q3}	-0.09	-0.17	-0.15	-0.53	-0.17	-0.37	-0.26	-0.22	0.06	1								
x_{q4}	0.03	-0.01	0.03	-0.04	0.13	-0.2	-0.24	0.33	1									
x_{q5}	0.2	0.2	0.35	0.42	0.35	0.16	0.42	1										
x_{q6}	-0.1	-0.08	0.81	0.68	0.75	0.76	1											
x_{q7}	-0.15	-0.11	0.83	0.65	0.79	1												
x_{q8}	-0.14	-0.11	0.91	0.56	1													
x_{q9}	-0.17	-0.08	0.63	1														
x_{q10}	-0.1	-0.06	1															
x_{e1}	0.83	1																
x_{e2}	1																	