Principal Covariates Clusterwise Regression (PCCR): Accounting for multicollinearity and population heterogeneity in hierarchically organized data

Wilderjans, T. F.[ab,*], Vande Gaer, E.[b], Kiers, H. A. L.[c], Van Mechelen, I.[b], & Ceulemans, E.[b]

[a] Methodology and Statistics Unit, Institute of Psychology, Faculty of Social and Behavioral Sciences, Leiden University, The Netherlands

[b] Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Belgium

[c] Faculty of Behavioural and Social Sciences, University of Groningen, The Netherlands

[*] corresponding author

Contact information

Tom F. Wilderjans

Methodology and Statistics Unit

Institute of Psychology

Faculty of Social and Behavioral Sciences

Leiden University

Wassenaarseweg 52 (Pieter de la Court Building)

2333 AK Leiden, The Netherlands

E-mail: t.f.wilderjans@fsw.leidenuniv.nl

Telephone: +31. 71. 527.60.58

Research Group of Quantitative Psychology and Individual Differences

Faculty of Psychology and Educational Sciences

KU Leuven

Tiensestraat 102, box 3713

3000 Leuven, Belgium

E-mail: tom.wilderjans@ppw.kuleuven.be

Telephone: +32. 16. 32.61.23 or Fax: +32. 16. 32.62.00

Acknowledgements

The first author is a post-doctoral Fellow of the Fund of Scientific Research (FWO) Flanders (Belgium). The research leading to the results reported in this paper was sponsored in part by the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles programme financed by the Belgian Government (IAP/P7/06).

Abstract

In the behavioral sciences, many research questions pertain to a regression problem in that one wants to predict a criterion on the basis of a number of predictors. Although in many cases ordinary least squares regression will suffice, sometimes the prediction problem is more challenging, for three reasons: First, multiple highly collinear predictors can be available, making it difficult to grasp their mutual relations as well as their relations to the criterion. In that case, it may be very useful to reduce the predictors to a few summary variables, on which one regresses the criterion and which at the same time yields insight into the predictor structure. Second, the population under study may consist of a few unknown subgroups that are characterized by different regression models. Third, the obtained data are often hierarchically structured, with for instance, observations being nested into persons or participants within groups or countries. Although some methods have been developed that partially meet these challenges (i.e., Principal Covariates Regression –PCovR–, clusterwise regression –*CR*–, and structural equation models), none of these methods adequately deals with all of them simultaneously. To fill this gap, we propose the Principal Covariates Clusterwise Regression (*PCCR*) method, which combines the key idea's behind *PCovR* (de Jong & Kiers, 1992) and *CR* (Späth, 1979). The *PCCR* method is validated by means of a simulation study and by applying it to cross-cultural data regarding satisfaction with life.

Keywords: clusterwise regression; component analysis; multicollinearity; population heterogeneity; hierarchically organized data

1 Introduction

In the behavioral sciences, many research questions pertain to the effect of one or several quantitative predictor variables on a quantitative criterion. These questions are traditionally addressed by ordinary linear least squares regression (*OLS*). Although *OLS* will often be adequate, at times it will fall short, because of three kind of complications that may simultaneously occur: First of all, sometimes multiple (e.g., in the range 10 to 20) predictors are involved, some of which might be highly correlated. In that case, without further insight into the underlying structure of the predictors, it becomes very hard to interpret each individual regression weight separately, all the more because multicollinearity problems might render instable regression weights. Second, often the population under study is not homogeneous with respect to the underlying regression model, but rather consists of a few unknown subgroups that are characterized by different underlying regression models. Finally, in many studies, the obtained data have a hierarchical structure in that the observations are nested within higher level units (e.g., measurement occasions nested within persons, individuals nested in cultural groups, or students nested in classes), where one is interested in the regression models of these higher level units.

To illustrate that these three challenges regularly occur simultaneously, let us take a look at three examples. As a first example, consider the paper of Stormshak et al. (1999) on the relationship between aggression and peer acceptance. These authors measured the extent to which first-grade students, that were nested into classes, exhibited five different kinds of aggressive behavior (predictors) and the extent to which they were socially accepted by their class mates (criterion). Stormshak et al. (1999) solved the interpretational problems of having multiple strongly collinear predictors by simply summing them. As such, all predictors received equal weight in the analysis, but no insight was given into their underlying structure. Regarding the regression part, it was found that classes differed with respect to the relation between aggression and peer acceptance: If aggressive behavior was accepted in the class, increasingly aggressive behavior led to increased acceptance. In contrast, in classes where aggressive behavior was frowned upon, more aggressive students were less liked by their class mates. Note that in this study the subgroups were not induced from the data, because one expected the regression relationships to be moderated by the class norm about aggressive behavior (acceptable or not). This allowed to model the class differences in a straightforward manner. However, this does not rule out that the data might contain other interesting but unknown subgroups.

As a second example, consider the study of DeSarbo and Edwards (1996) regarding the effect of several personality characteristics (e.g., perfectionism, self-esteem and avoidance coping) and psychiatric symptoms and disorders (e.g., depression, compulsiveness and anxiety) on self-reported compulsive buying behavior. They found that two groups of persons could be distinguished. In the first group, compulsive buying was mainly associated with psychological reasons while in the latter group, compulsive buying was more related to environmental factors. Because multiple correlated predictors were included, DeSarbo and Edwards (1996) were faced with multicollinearity problems, which they dealt with by imposing a positivity constraint on the regression weights. However, in this way no further insight was obtained into the predictor structure because all predictors were simply retained without modeling their interrelations.

As a final example, Ceulemans, Kuppens, and Van Mechelen (2012) examined the role of appraisals (predictors) in the elicitation of anger (criterion) and individual differences therein.

To this end, participants rated 24 situations with respect to 11 appraisals and with respect to anger. In this example, the observations are the situations, which are nested within persons. Using a Boolean method for binary data (*CLASSI*; Ceulemans & Van Mechelen, 2008), Ceulemans et al. (2012) simultaneously clustered the appraisals and the persons into three appraisal types and two person types. These two person types differed with respect to which appraisal types are necessary to become angry. Note that in order to apply *CLASSI*, Ceulemans and Van Mechelen (2008) had to dichotomize their data, implying a loss of information.

In the past, a few methods have been developed that tackle one or two of the three identified challenges (i.e., strongly collinear predictors, categorical differences in underlying regression models and nested data) by combining principal component analysis (*PCA*) or cluster analysis with regression analysis (see Figure 1). To obtain insight into the predictor structure and ease the interpretation of the regression weights, de Jong and Kiers (1992) developed Principal Covariates Regression (*PCovR*), which simultaneously reduces the predictors to a few components and regresses the criterion on these components. *PCovR* allows to attach different weights to the reconstruction of the predictors and the reconstruction of the criterion, and therefore encompasses Principal Component Regression (Coxe, 1986) as well as Reduced Rank Regression (Rao, 1964) as special cases. Another related method is Partial Least Squares (*PLS*) Regression (Wold, 1966), which emphasizes prediction of the criterion, at the expense of reconstructing the predictors. Specifically, *PLS* reduces the predictors to components such that these components are as much related (in terms of squared covariance) as possible to the criterion. Thus, a difference between both methods is that, unlike *PLS*, *PCCR* allows the user to flexibly weigh the reconstruction of the predictors and the prediction of the criterion. To detect categorical differences in underlying regression models on the other hand, Späth (1979, 1981) developed Clusterwise Regression (*CR*) that clusters observations on the basis of the underlying regression model. Related methods exist within the mixture and latent class framework (DeSarbo & Cron, 1988; Leisch, 2004; Wedel & DeSarbo, 1995). Note that extensions of *CR* exist that allow for nested data (DeSarbo, Oliver, & Rangaswamy, 1989).

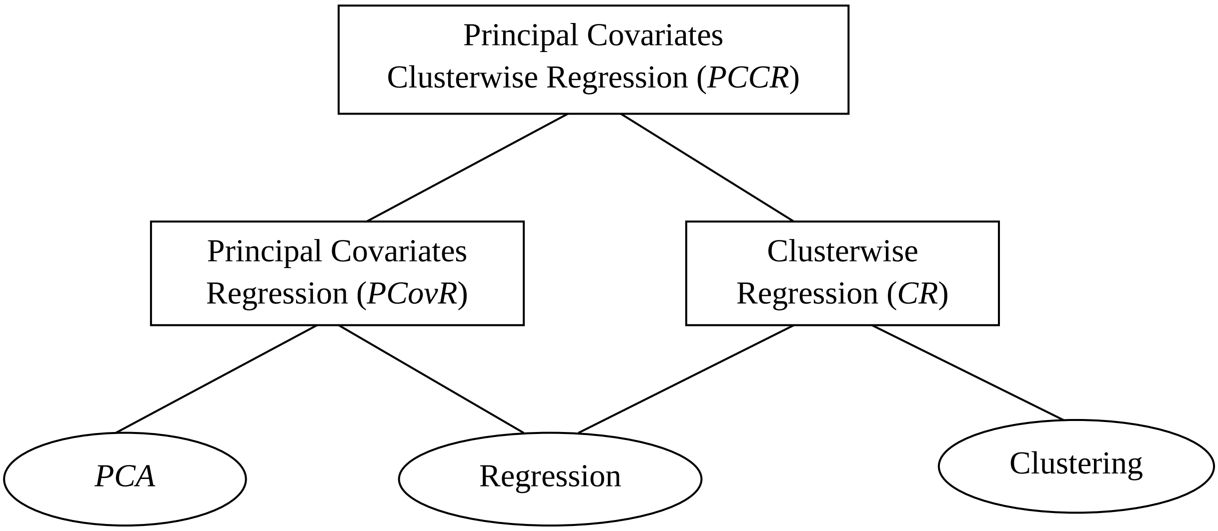

Figure 1. Graphical representation of the modeling features (ellipses) that are combined in different regression methods (rectangles)

Although all methods mentioned above can provide some useful insight into the data, none of them addresses all of the discussed challenges at the same time. In fact, the only methods that do combine dimension reduction of the predictors with detecting subgroups in the data on the basis of regression weights belong to the family of structural equation models and path models and are confirmatory in nature in that hypotheses are required about the predictor structure (Arminger & Stein, 1997; Hahn, Johnson, Herrmann, & Hubert, 2002; Sarstedt & Ringle, 2010). As in practice often no hypotheses are available about which variables are measuring what construct or such hypotheses prove to be incorrect, an exploratory counterpart of the latter methods can be a very useful tool. Therefore, we propose such a method, called Principal Covariates Clusterwise Regression (*PCCR*), which combines the key idea's behind *PCovR* (de Jong & Kiers, 1992) and *CR* (Späth, 1979) into one model (see Figure 1).

The remainder of the paper will be organized as follows: In Section 2, we discuss the data structure and preprocessing, the *PCCR* model, and the link to related models. In the third section, the aim of a *PCCR* analysis is described and an algorithm is presented for estimating *PCCR* models; we also briefly consider model selection. Section 4 reports on a simulation study that we conducted to evaluate the performance of the *PCCR* method. In Section 5 we apply the *PCCR* method to cross-cultural data. Finally, in Section 6 we discuss when to apply *PCCR* and reflect on some of the more technical choices that were made.

2 The Principal Covariates Clusterwise Regression (*PCCR*) model

2.1 Data structure and preprocessing

PCCR requires I predictor data blocks $\boldsymbol{X}_i$ $(N_i \times J)$ and criterion data vectors $\boldsymbol{y}_i$ $(N_i \times 1)$ that respectively contain scores on J predictors and on a criterion, where the number of observations N_i $(i = 1, \dots , I)$ in the data blocks may differ. These data blocks can be concatenated into an N (observations) $\times J$ (variables) predictor data matrix $\boldsymbol{X}$ and an $N \times 1$ criterion data vector $\boldsymbol{y}$, where $N = \sum_{i=1}^{I} N_i$. A graphical presentation of the data structure is given in Figure 2. It should be noted that *PCCR* can be used to analyze many types of nested data (e.g., students nested within classes and situations nested within persons). To indicate this, in the remainder of the manuscript, we will refer to the data blocks as the level-2 units and to the rows within each data block as the level-1 units.

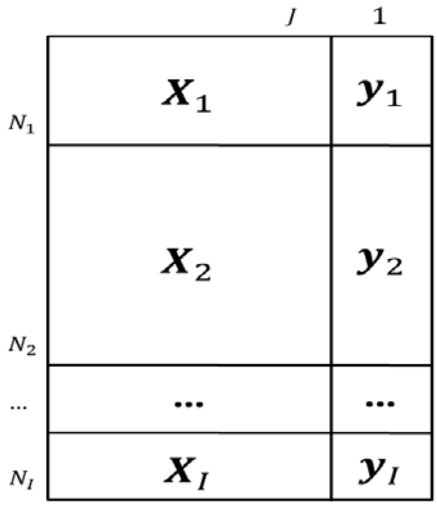

Figure 2. Graphical representation of the data needed for *PCCR* analysis

Because (1) *PCCR* aims at revealing the linear relationships within the predictor data set as well as between predictors and criterion rather than modeling individual differences in means and (2) Brusco, Cradit, Steinley, and Fox (2008) have conjectured that the results of *CR* methods tend to be dominated by differences in means, we center the variables per level-2 unit. Furthermore, to remove arbitrary scale differences, one may consider to standardize (i.e., a variance of one) the variables across level-2 units, as is often done in *PCA* analysis (see also Section 6 for a discussion on different scaling options).

2.2 Model

PCCR has a twofold aim: First, it captures the underlying structure of the predictor data block. Assuming that the underlying constructs or mechanisms are the same for all level-2 units, we reduce the predictors to *R* components. Given that *PCCR* belongs to the family of component analysis methods, the components are weighted sums of the predictors. Second, regressing the criterion on these components, it detects K subgroups or clusters of level-2 units that are characterized by different regression models.

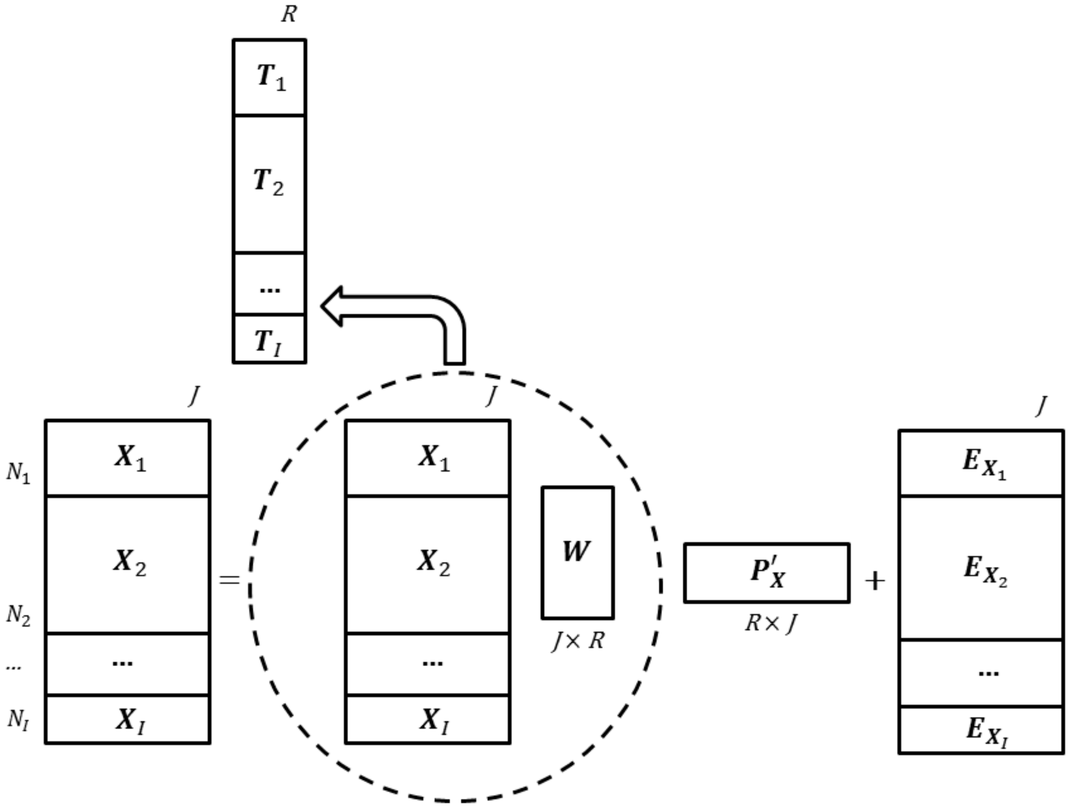

Figure 3. Decomposition of the predictor data blocks $\boldsymbol{X}_i$ in the *PCCR* model

Regarding the first goal, the predictor data block $\boldsymbol{X}_i$ of each level-2 unit i is decomposed as follows (see Figure 3):

$$\boldsymbol{X}_i = \boldsymbol{X}_i\boldsymbol{W}\boldsymbol{P}'_{\boldsymbol{X}} + \boldsymbol{E}_{\boldsymbol{X}_i} = \boldsymbol{T}_i\boldsymbol{P}'_{\boldsymbol{X}} + \boldsymbol{E}_{\boldsymbol{X}_i} \qquad (1)$$

In (1) $\boldsymbol{T}_i$ indicates the N_i by R component score matrix for level-2 unit i. These component scores are a weighted combination of the predictor variables, with the weights being represented in the J by R weighting matrix $\boldsymbol{W}$. Moreover, to avoid multicollinearity problems, the component scores are restricted to be orthogonal across level-2 units: $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}N$, with $\boldsymbol{T}$

$(N \times R)$ comprising the concatenated $\boldsymbol{T}_i$ matrices $(i = 1, \dots, I)$. Given that the variables are centered per level-2 unit, this orthogonality constraint implies that the components are uncorrelated across level-2 units. $\boldsymbol{P}_{\boldsymbol{X}}$ is a J by R loading matrix, which contains the loadings of the J variables on the R components. Note that these loadings amount to the correlations between the predictor variables in $\boldsymbol{X}$ and the component scores in $\boldsymbol{T}$ when the predictors are standardized across level-2 units. As $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{W}$ are restricted to be identical for all level-2 units, the interpretation of the components is the same for all level-2 units. Finally, $\boldsymbol{E}_{\boldsymbol{X}_i}$ $(N_i \times J)$ represents the predictor residuals for level-2 unit i.

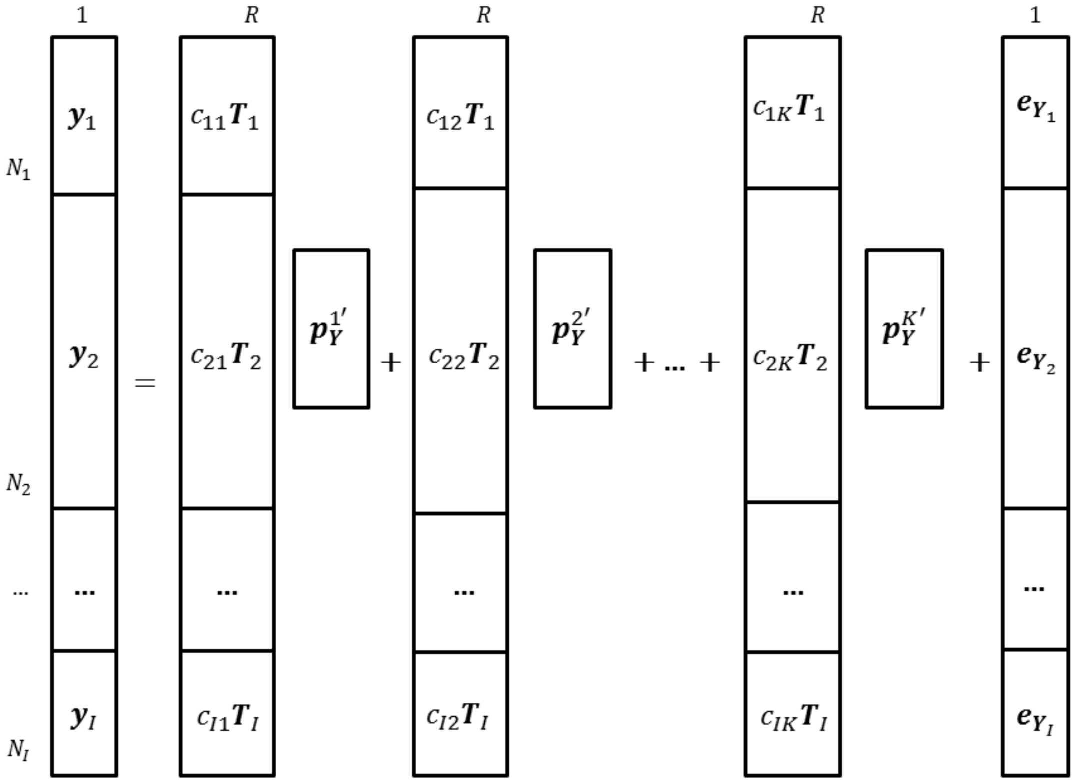

Figure 4. Decomposition of the criterion data vectors $\boldsymbol{y}_i$ in the *PCCR* model

Regarding the second goal, the decomposition rule for the criterion data vector $\boldsymbol{y}_i$ of each level-2 unit i is given by (see Figure 4):

$$\boldsymbol{y}_i = \sum_{k=1}^{K} \boldsymbol{c}_{ik} \boldsymbol{T}_i \boldsymbol{p}_{\boldsymbol{Y}}^{k\prime} + \boldsymbol{e}_{\boldsymbol{Y}_i} \qquad (2)$$

In (2), $\boldsymbol{c}_{ik}$ denotes the entries of the I by K partition matrix $\boldsymbol{C}$, which equal one if level-2 unit i belongs to cluster k, and zero otherwise. Furthermore, $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ indicates the 1 by R regression weight vector of cluster k $(k = 1, \dots, K)$, which specifies the cluster specific regression model of cluster k, regressing the criterion on the components. Because the data are centered per level-2 unit, no intercepts are included in these regression models. Finally, $\boldsymbol{e}_{\boldsymbol{Y}_i}$ $(N_i \times 1)$ indicates the criterion residuals of level-2 unit i.

As is the case for most component analysis models, *PCCR* solutions have rotational freedom. More specifically, one can rotate either the loading matrix $\boldsymbol{P}_{\boldsymbol{X}}$, the component score matrix $\boldsymbol{T}$ or the regression weight vectors $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ $(k = 1, \dots, K)$, provided that this rotation is compensated for in the other two matrices. One could, for instance, apply the Varimax rotation (Kaiser, 1958) on the loadings, which rotates them towards simple structure and makes them easier to interpret. Another option would be to rotate the component scores in such a way that they become as orthogonal as possible for each level-2 unit, rather than across level-2 units only, making the interpretation of the regression weights more clear. Indeed, if the components are orthogonal per level-2 unit, the level-2 unit specific regression weights on which the clustering of the level-2 units is based, only depend on the level-2 unit specific correlations between the components and the criterion and not on the level-2 unit specific intercorrelations among the components. To find the rotation matrix that makes all level-2 unit specific component scores as orthogonal as possible, one can use the *INDORT* algorithm for fitting constrained *INDSCAL* solutions (Kroonenberg, 1983, 2008; Kiers, 1989), as the rotation criterion that has to be optimized can be rewritten into the *INDORT* criterion (see Appendix I).

2.3 Relations to other models

As stated in the introduction, the *PCCR* model encompasses both *PCovR* (de Jong & Kiers, 1992) and *CR* for nested data (DeSarbo et al., 1989) as special cases. More specifically, the *PCCR* model boils down to *PCovR* when the number of clusters K equals one and is equivalent to *CR* when the number of components R equals the number of variables J.

3 Data analysis

3.1 Aim

Given a pre-specified number of components R, number of clusters K and weight α $(0 \leq \alpha \leq 1)$, the aim of *PCCR* is to find a weighting matrix $\boldsymbol{W}$, a partition matrix $\boldsymbol{C}$, a loading matrix $\boldsymbol{P_X}$ and K regression weight vectors $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ $(k = 1, \ldots, K)$ such that the following least squares loss function is minimized, subject to the restriction that $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}N$:

$$L = \alpha \frac{\|\boldsymbol{E_X}\|^2}{\|\boldsymbol{X}\|^2} + (1-\alpha)\frac{\|\boldsymbol{e_Y}\|^2}{\|\boldsymbol{y}\|^2} =$$

$$\frac{\alpha}{\|\boldsymbol{X}\|^2}\sum_{i=1}^{I}\|\boldsymbol{X}_i - \boldsymbol{X}_i\boldsymbol{W}\boldsymbol{P}_{\boldsymbol{X}}'\|^2 + \frac{(1-\alpha)}{\|\boldsymbol{y}\|^2}\sum_{i=1}^{I}\left\|\boldsymbol{y}_i - \sum_{k=1}^{K}\boldsymbol{c}_{ik}\boldsymbol{X}_i\boldsymbol{W}\boldsymbol{p}_{\boldsymbol{Y}}^{k'}\right\|^2 \qquad (3)$$

with $\|\ldots\|^2$ denoting the norm of a matrix/vector (i.e., sum of its squared elements). The loss function in (3) implies that the dimension reduction step and the clusterwise regression step are conducted simultaneously instead of sequentially. The weight α indicates the importance of reconstructing $\boldsymbol{X}$ when optimizing the loss function, whereas $(1-\alpha)$ denotes the importance of predicting $\boldsymbol{y}$. This loss function (3) can be rewritten (up to a constant) as:

$$L_2 = \beta \sum_{i=1}^{I} \|\boldsymbol{X}_i - \boldsymbol{X}_i \boldsymbol{W} \boldsymbol{P}_{\boldsymbol{X}}'\|^2 + (1-\beta) \sum_{i=1}^{I} \left\|\boldsymbol{y}_i - \sum_{k=1}^{K} \boldsymbol{c}_{ik} \boldsymbol{X}_i \boldsymbol{W} \boldsymbol{p}_{\boldsymbol{Y}}^{k'}\right\|^2 \qquad (4)$$

with $\beta = \frac{\alpha\|\boldsymbol{y}\|^2}{\alpha\|\boldsymbol{y}\|^2 + (1-\alpha)\|\boldsymbol{X}\|^2}$ (Vervloet, Van Deun, Van den Noortgate, & Ceulemans, 2013).

3.2 Algorithm

To minimize the loss function L_2 in (4), subject to the constraint that $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}N$, the following alternating least squares algorithm was developed (MATLAB code that implements this algorithm is available from the authors):

1. *Initialize the partition matrix* $\boldsymbol{C}^{initial}$: An initial partition is obtained by randomly drawing the cluster membership of each level-2 unit from a multinomial distribution with K categories and probabilities equal to $\frac{1}{K}$. Notice that this procedure does not necessarily yield a partition in which all clusters are non-empty. When one (or more) cluster(s) is empty, the whole procedure is repeated until a partition is obtained in which each of the K clusters contains at least one level-2 unit. It should further be noted that this procedure favors an initial partition with more or less equally sized clusters. As a consequence, when the true clusters differ considerably in size, the *PCCR* algorithm may perform poorer. To solve this problem one can increase the number of runs of the algorithm and use other types of starts, like rationally obtained (i.e., based on some previous analysis of the data) and pseudo-random (i.e., slightly perturbing a rationally obtained start) initializations (for more information on different types of starting procedures, see Ceulemans, Van Mechelen, & Leenen, 2007). However, a small simulation study with unequally sized clusters shows that our

initialization procedure works reasonably well in that case, except for some specific situations (i.e., very noisy data with a large number of underlying clusters).[1] These results are in line with simulation studies regarding other methods that also combine clustering and component analysis (e.g., Wilderjans & Ceulemans, 2013; Wilderjans, Ceulemans, & Kuppens, 2012; De Roover, Ceulemans, Timmerman, Vansteelandt, Stouten, & Onghena, 2012).

2. *Initialize $\boldsymbol{P_X}$, $\boldsymbol{T}$ $(\boldsymbol{W})$ and $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ conditional on the initial partition matrix $\boldsymbol{C}^{initial}$*: A singular value decomposition is conducted on $\boldsymbol{X}$: $\boldsymbol{X} = \boldsymbol{KML'}$. Next, $\boldsymbol{P_X}$ is obtained by multiplying the first R columns of $\boldsymbol{L}$ with the R largest singular values, which can be found on the diagonal of $\boldsymbol{M}$, and $\boldsymbol{T}$ is set to the first R columns of $\boldsymbol{K}$. Next, in order to ensure $\boldsymbol{T'T} = \boldsymbol{I}N$, $\boldsymbol{T}$ is multiplied by $\sqrt{N}$ and $\boldsymbol{P_X}$ by $\sqrt{1/_{N}}$. $\boldsymbol{W}$ is computed as $(\boldsymbol{X'X})^{-1}\boldsymbol{X'T}$, with $(\dots)^{-1}$ denoting matrix inversion. Initial estimates for the cluster specific regression weights $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ are obtained by regressing $\boldsymbol{y}_k$ on $\boldsymbol{T}_k$, where $\boldsymbol{y}_k$ and $\boldsymbol{T}_k$ respectively correspond to the concatenated criterion and component scores of the level-2 units that belong to cluster k $(k = 1, \dots, K)$.

3. *Consecutively update the cluster memberships of the level-2 units, conditional upon the clustering of the other level-2 units, until convergence*. In each iteration, all level-2 unit memberships are updated and the resulting loss function value is computed. This

[1] We generated 360 data sets by manipulating the number of clusters (at three levels: two, three and four clusters), the cluster sizes (at two levels: one small cluster with 20% of the level-2 units and one large cluster with 80% of the level-2 units; the other level-2 units were equally spread across the other clusters) and the α-level (at six levels: $.001, .01, .05, .15, .25$ and $.35$); all other factors were held constant (i.e., 50 level 2-units with 25 level-1 units each, 20 predictors, two relevant and six irrelevant components, the same cluster specific regression weights $\boldsymbol{p}_{\boldsymbol{Y}}^{k^{true}}$ as in Table 2 and 20% noise in $\boldsymbol{X}$ and in $\boldsymbol{y}$); we used ten replications for each cell of the design. The recovery of the underlying clustering, quantified by the Adjusted Rand Index (*ARI*; see section 4.3.2), can be considered reasonable as the mean *ARI* value equals $.93$, whereas the mean *ARI* for the comparable conditions in the simulation study with equal sized underlying clusters amounts to $.96$. Recovery, however, drops when the number of clusters increases (i.e., mean *ARI* of $.94$, $.95$ and $.89$ for two, three and four clusters, respectively) and/or when there is one large cluster next to multiple small ones (i.e., mean *ARI* of $.99$ and $.87$ for the first and second level of the cluster size factor, respectively).

procedure is repeated until the improvement in the loss function value is smaller than a tolerance value (e.g., $.0000001$)[2]. To update the cluster membership of level-2 unit i, the level-2 unit in question is assigned to each of the K clusters and the associated conditionally optimal $\boldsymbol{W}$, $\boldsymbol{P_X}$ and $\boldsymbol{p_Y^k}$ are computed by repeating the following three steps until convergence[3]:

3.1. Update $\boldsymbol{W}$ conditionally upon $\boldsymbol{C}$, $\boldsymbol{P_X}$ and $\boldsymbol{p_Y^k}$ under the constraint that $\boldsymbol{T'T} = \boldsymbol{I} = \sum_{k=1}^{K} \boldsymbol{W'X'_kX_kW}$ (with $\boldsymbol{X}_k$ being obtained by vertically concatenating all data matrices $\boldsymbol{X}_i$ of level-2 units i belonging to cluster k). Because of the orthogonality constraint on $\boldsymbol{W}$, no closed form solution for this constrained problem is available. A way out consists of first solving the unconstrained problem[4], yielding $\boldsymbol{W}_{uncon}$, and next enforcing the constraint by means of a nonsingular transformation of $\boldsymbol{W}_{uncon}$. The solution to the unconstrained minimization problem equals (see Appendix II):

$$\boldsymbol{W}_{uncon} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'y}$$

where

$$\boldsymbol{Z'Z} = \sum_{k=1}^{K} \left[\left(\beta \boldsymbol{P'_X P_X} + (1-\beta) \boldsymbol{p_Y^{k'} p_Y^k} \right) \otimes \boldsymbol{X'_k X_k} \right]$$

[2] Note that taking $.0000001$ as convergence criterion might be too strict, resulting in a considerable lengthening of the computation time. A possible solution is to use $.0000001\ ||\boldsymbol{X}||^2$ instead (i.e., to look at the proportional decrease instead of the absolute decrease in fit).

[3] Note that step three of the *PCCR* algorithm consists of a double (nested) iterative procedure of which the outer iterations pertain to updating the level-2 unit memberships and the inner iterations to updating $\boldsymbol{W}$, $\boldsymbol{P_X}$ and $\boldsymbol{p_Y^k}$, conditional on $\boldsymbol{C}$.

[4] Another option is to directly solve the constrained problem (instead of the unconstrained one) by means of an iterative majorization approach (Kiers & ten Berge, 1992). The technical details of this approach are available upon request from the authors. The results of a pilot simulation study showed that both algorithms in most cases lead to the same final solution, but that the majorization approach consumes more time. Therefore, the majorization approach is not further discussed in this manuscript.

$$\boldsymbol{Z}'\boldsymbol{y} = vec\left[\beta\left(\sum_{k=1}^{K}\boldsymbol{X}_k'\boldsymbol{X}_k\right)\boldsymbol{P_X} + (1-\beta)\sum_{k=1}^{K}\boldsymbol{X}_k'\boldsymbol{y}_k\,\boldsymbol{p}_{\boldsymbol{Y}}^k\right]$$

and $\otimes$ denotes the Kronecker product of two matrices and vec the vectorization-operator for matrices (Schott, 2005). To enforce the orthogonality constraint, an eigenvalue decomposition is performed on $\boldsymbol{Z_W} = \boldsymbol{W}_{uncon}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{W}_{uncon}$, resulting in a diagonal matrix $\boldsymbol{P}$ that contains the eigenvalues and a matrix $\boldsymbol{A}$ that holds the associated normalized eigenvectors. Next, $\boldsymbol{W}$ can be computed as $\boldsymbol{W} = \boldsymbol{W}_{uncon}\boldsymbol{A}\left(\sqrt{\boldsymbol{P}}\right)^{-1}$.

3.2. Update $\boldsymbol{P_X}$ and the $\boldsymbol{p}_{\boldsymbol{Y}}^k$ $(k = 1, \ldots, K)$ vectors conditionally upon $\boldsymbol{C}$ and $\boldsymbol{W}$ as follows:

$$\boldsymbol{P_X} = (\boldsymbol{X}'\boldsymbol{X}\boldsymbol{W})(\boldsymbol{W}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{W})^{-1}$$

$$\boldsymbol{p}_{\boldsymbol{Y}}^k = \boldsymbol{y}_k'\boldsymbol{X}_k\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{X}_k'\boldsymbol{X}_k\boldsymbol{W})^{-1}$$

3.3. Check the loss function value for improvement. When the resulting decrease in loss function value is larger than a pre-specified tolerance value, continue by returning to step 3.1; otherwise, stop.

Next, level-2 unit i is assigned to the cluster k for which the resulting loss function L_2 is lowest and the corresponding $\boldsymbol{W}$, $\boldsymbol{P_X}$ and $\boldsymbol{p}_{\boldsymbol{Y}}^k$ are retained. If the $\boldsymbol{C}$ that is obtained at the end of an iteration contains empty clusters, the level-2 unit that fits its cluster the worst is assigned to (one of) the empty cluster(s) and the resulting $\boldsymbol{W}$, $\boldsymbol{P_X}$ and $\boldsymbol{p}_{\boldsymbol{Y}}^k$ are recomputed (as in steps 3.1 and 3.2)[5]; this procedure is repeated until all clusters are non-empty.

4. *Rescale the obtained solution such that* $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}N$. To this end, $\boldsymbol{T}$ and $\boldsymbol{W}$ are multiplied by $\sqrt{N}$ and $\boldsymbol{P_X}$ and $\boldsymbol{p}_{\boldsymbol{Y}}^k$ $(k = 1, \ldots, K)$ are divided by $\sqrt{N}$.

[5] When the worst fitting level 2-unit is the only level-2 unit in his/her cluster, we move on to the level-2 unit with the second worst fit, etcetera.

In order to minimize the risk that the $PCCR$ algorithm ends in a suboptimal solution (i.e., a local minimum), a multi-start procedure is performed. This procedure consists of running the algorithm many times (e.g., 50 or 100), each time using a different initial partition matrix $\boldsymbol{C}^{initial}$. The solution that has the lowest loss function value L_2 is retained as the final solution.

3.3 Model selection

As described above, the $PCCR$ algorithm requires the number of components R, the number of clusters K and the value of α to be specified. Sometimes, R, K and α can be chosen on the basis of substantive arguments. However, often no strong a priori information is available. In such cases, one may adopt the following strategy: First, estimate $PCCR$ solutions using several values for R, K and α. Next, select a model that fits well, is not too complex and that generalizes well to other data from the same population. To this end, one may use k-fold cross-validation with k, for example, being equal to ten. In this procedure, ten new cross-validation samples are created from the original data set by removing each time ten percent of the level-1 units within each level-2 unit (such that each level-1 unit is removed exactly once). On each of these cross-validation samples, $PCCR$ analyses are performed with all the different values of R, K and α and the criterion values of the deleted level-1 units are predicted based on the estimated parameters. Finally, the cross-validation error for each (R, K, α)-combination is computed by summing the squared prediction errors for the deleted level-1 units across the ten samples. The (R, K, α)-combinations with the lowest cross-validation errors are to be preferred. The use of this model selection strategy will be illustrated in the application section (see Section 5).

4 Simulation study

4.1 Aim

The aim of this simulation study is twofold. First, we want to evaluate the performance of the *PCCR* algorithm with regard to sensitivity to local minima, on the one hand, and the recovery of the underlying clustering, components and regression models, on the other hand. Second, we will compare the performance of *PCCR,* which simultaneously extracts components and predicts the criterion, and a sequential strategy (i.e., first extract components from $\boldsymbol{X}$; next, predict $\boldsymbol{y}$ based on these components).

Regarding the second aim, it is useful to make a distinction between relevant and irrelevant components in $\boldsymbol{X},$ with irrelevant components having zero regression weights. We hypothesize that when the strongest components in $\boldsymbol{X}$ (in terms of explained variance) are all relevant, *PCCR* and the sequential strategy will perform equally well[6]. However, clear performance differences are expected when $\boldsymbol{X}$ contains a weak but relevant component and one or more irrelevant components, of which some are strong (Vervloet et al., 2013). In these situations, we predict that the sequential strategy will extract a strong but irrelevant component rather than the weak but relevant one, because the criterion variable is not taken into account and because most model selection procedures focus on strong components only (Ceulemans & Kiers, 2009). In contrast, we hypothesize that *PCCR* will ignore the irrelevant components and will retain the relevant ones, because it takes both the information in $\boldsymbol{X}$ and $\boldsymbol{y}$ into account. Herewith, we expect that *PCCR* will perform best when α is small (Vervloet et

[6] The results of a pilot study in which the number of components and clusters (i.e., two and four), the amount of noise in $\boldsymbol{X}$ and $\boldsymbol{y}$ (i.e., 10% and 40%) and the cluster sizes (see Brusco & Cradit, 2001; Steinley, 2003) have been manipulated, indeed reveal that *PCCR* and a sequential approach perform equally well, when all components are strong and relevant.

al., 2013). To test this hypothesis, we conducted a simulation study in which we added a number of irrelevant components and varied their strength.

4.2 Design and procedure

We kept the following data characteristics fixed: (1) 50 level-2 units and 25 level-1 units per level-2 unit, (2) 20 predictors, (3) two relevant components, (4) the cluster specific regression weights $\boldsymbol{p}_{\boldsymbol{Y}}^{k^{true}}$ of these relevant components (see Table 2), (5) the amount of noise on $\boldsymbol{y}$ is set to 20%, and (6) equal cluster sizes. We manipulated the following three characteristics:

1) *Number of irrelevant components*, at two levels: four and six irrelevant components. The amount of variance explained by each relevant and irrelevant component is displayed in Table 1. For all data sets, the first relevant component explains 38% of the variance and the weak but relevant second component only 2%. Half of the irrelevant components (i.e., two or three in case of four and six irrelevant components, respectively) are weak too and explain about the same amount of variance as the weak relevant component, whereas the other half of the irrelevant components are strong and thus explain a sizeable amount of variance in $\boldsymbol{X}$.
2) *The amount of noise in* $\boldsymbol{X}$, at three levels: 10%, 20% and 40%.
3) *The number of clusters*, at three levels: two, three and four clusters.

Table 1

Percentage of explained variance per relevant and irrelevant component

	relevant components		irrelevant components					
N_{comp}	1	2	3	4	5	6	7	8
4	38%	2%	30%	25.7%	2.2%	2.1%	---	---
6	38%	2%	25%	15%	13.4%	2.3%	2.2%	2.1%

N_{comp}: number of irrelevant components

Each data set was created as follows: The true partition matrix $\boldsymbol{C^{true}}$ was obtained by randomly assigning level-2 units to a cluster taking into account the desired number of clusters and cluster size. The true component scores on both the relevant and irrelevant components, stored in the matrix $\boldsymbol{T^{true}}$, were drawn from a multivariate normal distribution with the $\mathbf{0}$ vector as mean vector and the identity matrix as variance-covariance matrix. Next, these component scores were centered per level-2 unit and orthogonalized (i.e., $\boldsymbol{T^{true}}'\boldsymbol{T^{true}} = \boldsymbol{I}$) and standardized across level-2 units[7]. True relevant and irrelevant component loadings $\boldsymbol{P_X^{true}}$ were obtained by first randomly generating numbers between -1 and 1 from a uniform distribution and next rescaling these numbers in order to get the desired percentages of explained variance (see Table 1). The true regression weights $\boldsymbol{p_Y^{k^{true}}}$ that have been used are presented in Table 2[8]. Finally, data matrices $\boldsymbol{X}$ and $\boldsymbol{y}$ were obtained by adding error matrices

[7] $\boldsymbol{W^{true}}$ can be computed from $\boldsymbol{T^{true}}$ as follows: $\boldsymbol{W^{true}} = \left(\boldsymbol{X^{true}}'\boldsymbol{X^{true}}\right)^{-1}\boldsymbol{X^{true}}'\boldsymbol{T^{true}}$, with $\boldsymbol{X^{true}} = \boldsymbol{T^{true}}\boldsymbol{P_X^{true}}'$.

[8] This choice of cluster-specific regression weights implies that the clusters are clearly separated in terms of their underlying cluster specific regression models. To quantify the degree of cluster separation, we computed the ratio $z = \frac{\sum_{r=1}^{R}\left|p_{Y_r}^{k^{true}} - p_{Y_r}^{k'^{true}}\right|}{\sum_{r=1}^{R}\left|p_{Y_r}^{k^{true}}\right|}$ for each pair of clusters, with $z > .50$ implying that clusters are well separated

$\boldsymbol{E_X}$ and $\boldsymbol{e_Y}$ to $\boldsymbol{X^{true}}$ and $\boldsymbol{y^{true}}$, respectively, with $\boldsymbol{X^{true}}$ and $\boldsymbol{y^{true}}$ resulting from combining the true partitioning, component scores, loadings, and regression weights by the decomposition rules (1) and (2). The elements of $\boldsymbol{E_X}$ and $\boldsymbol{e_Y}$ were independently sampled from a standard normal distribution and rescaled so as to obtain the desired amount of noise. At the end, the data per level-2 unit (both $\boldsymbol{X}$ and $\boldsymbol{y}$) were centered.

Table 2

Cluster-specific regression weight $\boldsymbol{p}_{\boldsymbol{Y}}^{k^{true}}$ for each manipulated number of clusters

	Regression weights $\boldsymbol{p}_{\boldsymbol{Y}}^{k^{true}}$ for cluster k			
	$k = 1$	$k = 2$	$k = 3$	$k = 4$
2 clusters	$\sqrt{.01}$ and $\sqrt{.99}$	$\sqrt{.80}$ and $\sqrt{.20}$		
3 clusters	$\sqrt{.01}$ and $\sqrt{.99}$	$\sqrt{.80}$ and $\sqrt{.20}$	$\sqrt{.01}$ and $-\sqrt{.99}$	
4 clusters	$\sqrt{.01}$ and $\sqrt{.99}$	$\sqrt{.80}$ and $\sqrt{.20}$	$\sqrt{.01}$ and $-\sqrt{.99}$	$-\sqrt{.80}$ and $\sqrt{.20}$

Note: all cluster-specific regression weights $\boldsymbol{p}_{\boldsymbol{Y}}^{k^{true}}$ are chosen such that $\|\boldsymbol{p}_{\boldsymbol{Y}}^{k^{true}}\| = 1$

The design has 2 (number of irrelevant components) × 3 (amount of noise in $\boldsymbol{X}$) × 3 (number of clusters) $=$ 18 conditions and 10 data sets were generated per condition, resulting in 180 data sets. To investigate how α influences the performance of *PCCR*, each of these data sets will be analyzed with the following six α values: .35, .25, .15, .05, .01 and .001. We selected relatively low α values, because based on results of the pilot simulation study, both *PCCR* and the sequential strategy are predicted to perform equally well when α increases (i.e., when

(Kiers & Smilde, 2007) and $p_{Y_r}^{k^{true}}$ $\left(p_{Y_r}^{k'^{true}}\right)$ indicating the r^{th} true regression weight for cluster k (k'). For the generated datasets with two clusters the z-ratio equals 1.23, whereas the mean z-ratio (computed over all possible cluster pairs) for the three- and four-cluster datasets equals 1.57 (with separate values of 1.23, 1.82 and 1.67) and 4.47 (with separate values of 1.23, 2.64, 8.76, 1.15, 7.33 and 5.71), respectively. Note that for generated datasets in the three- and four-cluster conditions, the clusters are more clearly separated than in the two-cluster conditions. Notice further that the larger mean z-ratio for the solutions with four clusters is mainly due to the fourth cluster being clearly separated from the other three clusters (with the z-ratios for the other three clusters being in the range of the z-ratios for the generated datasets with two and three clusters).

α increases, *PCCR* approaches the sequential strategy, with both strategies being equivalent when $\alpha = 1$). During the analysis, the true number of clusters and two components were extracted. We used a tolerance value of $.0000001$ and a multi-start procedure with 100 runs (see Section 3.2). Each run started from a different randomly generated initial clustering $\boldsymbol{C}^{initial}$ in which each level-2 unit had a probability of $\frac{1}{K}$ to be assigned to each cluster and in which empty clusters were avoided. From these 100 runs, the solution with the lowest loss function value (4) was retained, yielding the reconstructed data matrix $\widehat{\boldsymbol{X}}$ and data vector $\widehat{\boldsymbol{y}}$. To obtain results for the sequential strategy, standard *PCA* was applied to each data set $\boldsymbol{X}$ retaining R components, which yields $\boldsymbol{W}$, $\boldsymbol{T}$ and $\boldsymbol{P}_{\boldsymbol{X}}$. Next a clusterwise regression analysis with K clusters was conducted in which $\boldsymbol{y}$ was regressed on $\boldsymbol{T}$, subject to the constraint that all participants of specific level-2 units are assigned to the same cluster; this step yields the $\boldsymbol{C}$ and $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ matrices. The computations were run in *MATLAB* (version 2013b and 2014a) on a cluster of computers from the Flemish Supercomputer Center (*VSC*, https://vscentrum.be/nl/en). Averaged over all simulated data sets, the computation time of a single *PCCR* run equals 38.7 seconds (1 hour and a few minutes for 100 runs), whereas a run of the sequential strategy takes 1.4 seconds on average (about 2.5 minutes for 100 runs). The computation time of a *PCCR* run is mostly influenced by the number of clusters (i.e., mean computation time of 20.3, 36.6 and 59.1 seconds for two, three, and four clusters, respectively) and the α–level (i.e., mean computation time of 23.6, 51.0 and 73.5 seconds for α equaling $.001$, $.25$ and $.35$, respectively).

4.3 Results

4.3.1 Local minima

To assess how often *PCCR* yields a local minimum, we computed the following $L_{2_{diff}}$ measure:

$$L_{2_{diff}} = \left(\beta\left\|\boldsymbol{X}-\widehat{\boldsymbol{X}}\right\|^2 + (1-\beta)\|\boldsymbol{y}-\widehat{\boldsymbol{y}}\|^2\right) - \left(\beta\left\|\boldsymbol{X}-\widehat{\boldsymbol{X}}^{\boldsymbol{proxy}}\right\|^2 + (1-\beta)\|\boldsymbol{y}-\widehat{\boldsymbol{y}}^{\boldsymbol{proxy}}\|^2\right),$$

which compares the loss function value (4) of the estimated *PCCR* solution with the loss function value of the *PCCR* solution with $\widehat{\boldsymbol{X}}^{\boldsymbol{proxy}}$ and $\widehat{\boldsymbol{y}}^{\boldsymbol{proxy}}$, which is obtained by seeding the *PCCR* algorithm with the true clustering $\boldsymbol{C}^{\boldsymbol{true}}$. When the computed $L_{2_{diff}}$-value is positive, the algorithm ended in a local minimum for sure, in that at least one other solution exists (i.e., $\widehat{\boldsymbol{X}}^{\boldsymbol{proxy}}$ and $\widehat{\boldsymbol{y}}^{\boldsymbol{proxy}}$) that fits the data better. Note that a negative $L_{2_{diff}}$-value does not necessarily mean that the global optimum of the loss function is reached, as $\left(\beta\left\|\boldsymbol{X}-\widehat{\boldsymbol{X}}^{\boldsymbol{proxy}}\right\|^2 + (1-\beta)\|\boldsymbol{y}-\widehat{\boldsymbol{y}}^{\boldsymbol{proxy}}\|^2\right)$ is only a proxy (i.e., estimate) of the global minimum, which is always unknown when the data contain noise. In general, *PCCR* appears not to have any problems with local minima (i.e., all $L_{2_{diff}}$-values are negative).

A second way to get some idea whether or not the *PCCR* algorithm has a large problem with locally optimal solutions, is to check the number of runs of the algorithm that yielded the optimal loss function value. Indeed, when most of the runs end in the retained solution, it can be assumed that the algorithm does not suffer too much from local optima. On average, 96.8 (72.9 for the sequential strategy) of the 100 *PCCR* runs per simulated data set resulted in the same optimal loss function value.

Table 3

Mean recovery results for PCCR (with $\alpha = .001$, $\alpha = .35$ *and across all six* α *levels) and the sequential approach*

Recovery	***PCCR*** **with** $\boldsymbol{\alpha = .001}$	***PCCR*** **with** $\boldsymbol{\alpha = .35}$	***PCCR*** **across** **all six** $\boldsymbol{\alpha}$ **levels**	**Sequential** **approach**
Clustering (*ARI*)	.97	.97	.97	.62
Perfect clustering (*ARI* = *1*)	132 / 180	126 / 180	785 / 1080	8 / 180
	(73.33%)	(70%)	(72.69%)	(4.44%)
Loadings ($\Phi_{\boldsymbol{P_X}}$)	.99	.97	.98	.59
Regression weights ($MAD_{\boldsymbol{p_Y}}$)	.20	.20	.20	.64

4.3.2 Goodness-of-recovery

Recovery of the clustering. To assess the recovery of the clustering of the level-2 units, we made use of the Adjusted Rand Index (*ARI*; Hubert & Arabie, 1985; Steinley, 2004). An *ARI* value of one indicates perfect recovery and an *ARI* value of zero indicates that the obtained clustering does not resemble the true clustering more than can be expected by chance. One can see in Table 3 that, on average (i.e., across all levels of α), *PCCR* clearly outperforms the sequential strategy in terms of recovering the true clustering (i.e., a difference in ARI of $.35$ and in percentage perfect recovery of 68.25%). In Figure 5, one can see that, in general, the *PCCR* recovery of the clustering somewhat deteriorates when the amount of noise in the data increases. However, in the worst case recovery stays acceptable with mean $ARI > .88$. The effects of the other factors (and their interactions) are even smaller and not univocal.

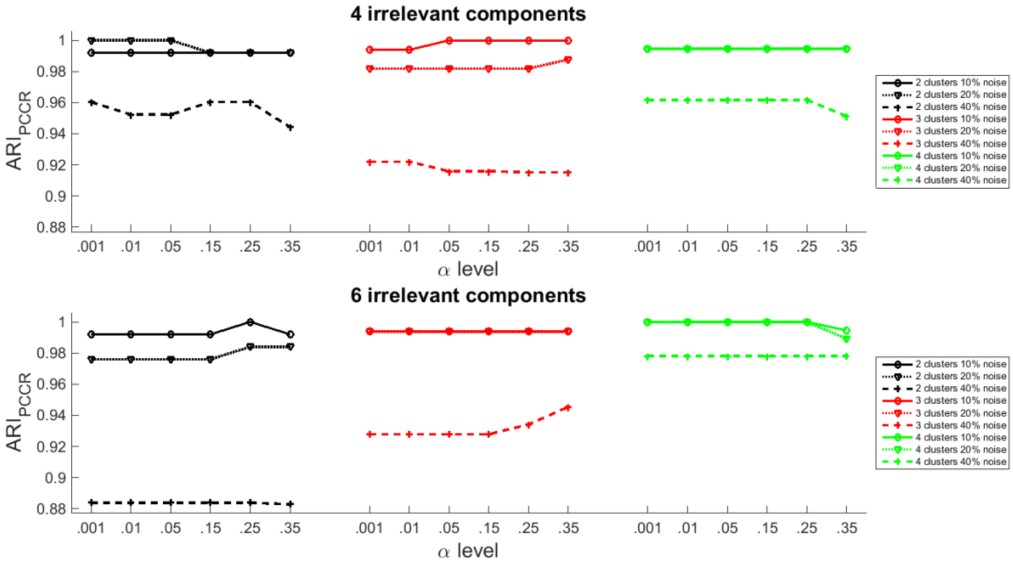

Figure 5. Mean *ARI* for *PCCR* solutions as a function of the number of irrelevant components (top part: four irrelevant components; bottom part: six irrelevant components), the number of clusters (left panels: two clusters; middle panels: three clusters; right panels: four clusters), the amount of noise in the data and α.

Recovery of the loadings. To check to what extent *PCCR* is able to recover the relevant components, the recovery of the corresponding loadings in $\boldsymbol{P_X}$ was evaluated by calculating the Tucker Phi coefficient (Φ; Tucker, 1951; Korth & Tucker, 1975) between the true and estimated loadings per relevant component and averaging (the absolute value of) this coefficient over the components. Before calculating $\Phi_{\boldsymbol{P_X}}$, the estimated loadings $\boldsymbol{\hat{P}_X}$ were orthogonally rotated (ten Berge, 1977) towards the true loadings $\boldsymbol{P_X^{true}}$. A $\Phi_{\boldsymbol{P_X}}$ value of one indicates perfect recovery, whereas a $\Phi_{\boldsymbol{P_X}}$ value of zero implies no recovery at all. Across all 360 analyses (i.e., 60 data sets analyzed with six levels of α), the average difference in $\Phi_{\boldsymbol{P_X}}$ between *PCCR* and the sequential strategy equals $.39$ (see Table 3). In Figure 6, one can see that *PCCR* recovers the loadings a bit better when α becomes smaller (i.e., less influence of $\boldsymbol{X}$

when extracting the components) and slightly better when the data contain less noise, whereas there is almost no effect of the number of irrelevant components.

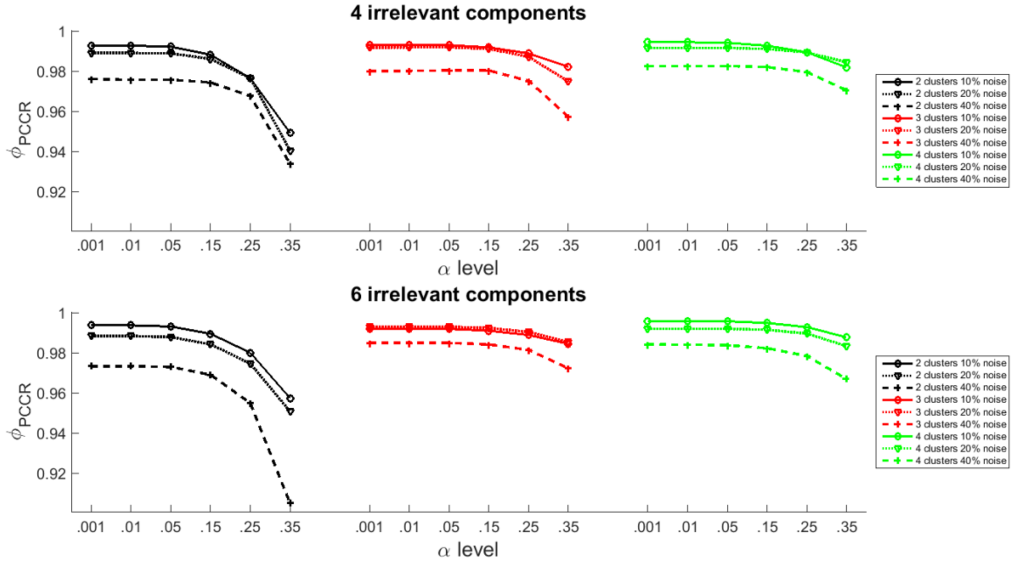

Figure 6. Mean $\Phi_{\boldsymbol{P_X}}$ for *PCCR* solutions as a function of the number of irrelevant components (top part: four irrelevant components; bottom part: six irrelevant components), the number of clusters (left panels: two clusters; middle panels: three clusters; right panels: four clusters), the amount of noise in the data and α.

Recovery of the regression weights. To investigate the recovery of the regression weights $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ of the relevant components, the $MAD_{\boldsymbol{p_Y}}$ measure was used (Kiers & Smilde, 2007):

$$MAD_{\boldsymbol{p_Y}} = \frac{\sum_{k=1}^{K} \sum_{r=1}^{R} \left| p_{\boldsymbol{Y}_r}^{k^{\boldsymbol{true}}} - \hat{p}_{\boldsymbol{Y}_r}^{k} \right|}{\sum_{k=1}^{K} \sum_{r=1}^{R} \left| p_{\boldsymbol{Y}_r}^{k^{\boldsymbol{true}}} \right|}.$$

$MAD_{\boldsymbol{p_Y}}$ is zero in case of perfect recovery (i.e., $\hat{\boldsymbol{p}}_{\boldsymbol{Y}}^{k}$ equaling $\boldsymbol{p}_{\boldsymbol{Y}}^{k^{\boldsymbol{true}}}$ for all $k = 1, \dots, K$) and higher values indicate worse recovery. Note that the cluster specific regression weights were rotated adopting the same rotation matrix that was used in the orthogonal target rotation of the loadings $\boldsymbol{P_X}$ (see earlier). Moreover, we matched the true and estimated clusters so as to obtain the lowest MAD_{p_Y}-value. On average, $PCCR$ recovers the true regression weights clearly better than the sequential strategy. In particular, as can be seen in Table 3, $PCCR$ has a good MAD_{p_Y} of $.20$, whereas the MAD_{p_Y} of $.64$ for the sequential strategy is terrible. As one can see in Figure 7, $PCCR$ better recovers the cluster specific regression weights when there is less noise in the data. There is hardly any effect of the number of irrelevant components and recovery seems to deteriorate a little when α becomes larger than $.25$.

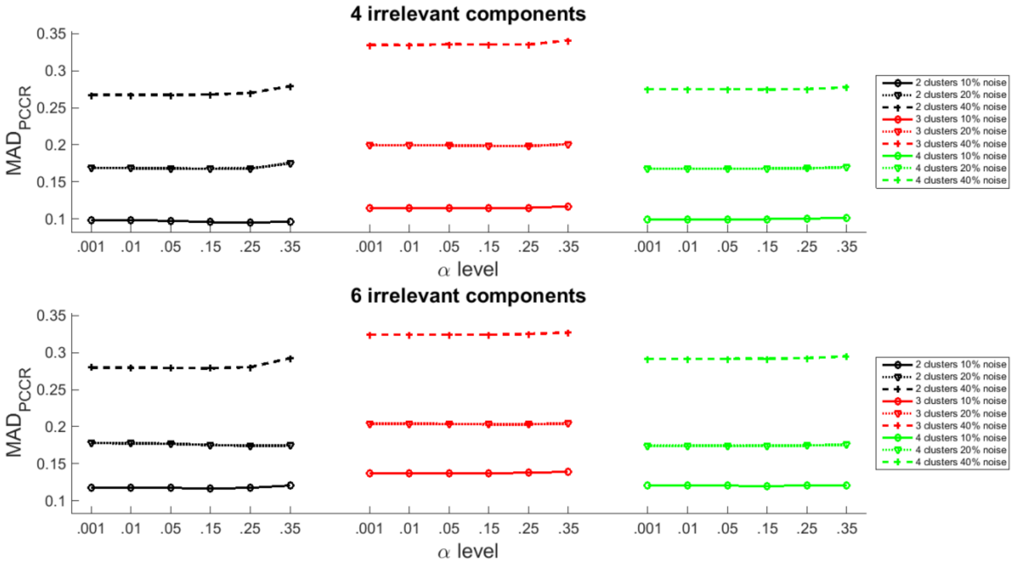

Figure 7. Mean MAD_{p_Y} for $PCCR$ solutions as a function of the number of irrelevant components (top part: four irrelevant components; bottom part: six irrelevant components), number of clusters (left panels: two clusters; middle panels: three clusters; right panels: four clusters), the amount of noise in the data and α. Note that larger values on the Y-axis imply a worse recovery.

Analyses with more components. When performing analyses with three/four components (i.e., the number of strong components in the four and six irrelevant components conditions) and the true number of clusters, *PCCR* still clearly outperforms the sequential strategy (i.e., a mean *ARI* value of .98 versus .56).

Conclusion. It can be concluded that, for our settings, *PCCR* with 100 runs recovers the underlying clustering, loadings and cluster specific regression weights excellently. Moreover, when the data contain one or more strong but irrelevant components as well as a weak but relevant one, *PCCR* clearly outperforms the sequential approach in terms of recovering the underlying clustering, component loadings and cluster specific regression weights. Extracting as many components as there are strong ones is not a solution as *PCCR* still clearly outperforms the sequential strategy in that case.

5 Application

To illustrate the usefulness of *PCCR*, we will analyze cross-cultural data on satisfaction with life. Specifically, we will look for linear combinations of values, personality characteristics and beliefs about happiness that have a differential predictive value for satisfaction with life. The differences are assumed to be categorical in that we will look for clusters of countries (i.e., the level-2 units in this application refer to countries) that are characterized by different regression models.

The data that will be analyzed are part of the International College Survey (ICS) 2001 (Kuppens, Ceulemans, Timmerman, Diener, & Kim-Prieto, 2006), a large-scale paper-and-pencil questionnaire-based study regarding subjective wellbeing and its determinants, in which 10,018 inhabitants from 48 different countries (see Table 6) took part. We used the satisfaction with life score of each participant as the criterion variable and a set of 34 variables (see Table 4 for more details) regarding which characteristics people value, people’s personality and their beliefs on happiness as the predictors. The satisfaction with life scores were obtained by summing participants’ answers (rated on a seven-point Likert scale) on five items (i.e., “In most ways my life is close to my ideal”, “the conditions of my life are excellent”, “I am satisfied with my life”, “so far I have gotten the important things I want in life” and “if I could live my life over, I would change almost nothing”). We list-wise deleted the participants with missing data, resulting in a final sample of 9054 respondents.

Table 4

34 predictors and their keywords for the cross-cultural data regarding satisfaction with life

Keyword	**Item**
How much do you value (from 1 “do not value it at all” to 9 “value it extremely”)	
Happiness	Happiness
Intelligence and knowledge	Intelligence and knowledge
Material wealth	Material wealth
Physical attractiveness	Physical attractiveness
Physical comforts	Physical comforts
Excitement and arousal	Excitement and arousal
Competition	Competition
Getting to heaven	Getting to heaven, achieving a happy afterlife
Self-sacrifice	Self-sacrifice
Success	Success
Fun	Fun (personal enjoyment)
Describe yourself (from 1 “very inaccurate” to 5 “very accurate”)	
Partylife	Am the life of the party
No talk	Don’t talk a lot

Background	Keep in the background
Start conversation	Start conversations
Talk people	Talk to a lot of different people at parties
Quiet strangers	Am quiet around strangers
How much do you agree (from 1 “strongly disagree” to 9 “strongly agree”)	
Harmony	It is important for me to maintain harmony within my group
Outperforming	It is important for me that I do my job better than others
Aging parents	We should keep our aging parents with us at home
Unique	I am a unique individual
Competition	Without competition it is not possible to have a good society
Competition friends	I often feel like I am competing with my friends
Competition family	I often feel like I am competing with my family members
Family success	The success of my family is more important than my own pleasure
Enjoy present	I would rather enjoy the present and not worry about the future
Work hard	It is better to work hard now because happiness can be saved up and enjoyed later
Family pessimism	When I think about my friends and family, I think more about what might go wrong than I think about rewards
School pessimism	When I think about my school work, I think more about what might go wrong than I think about rewards
Cycle	Happiness and unhappiness are like night and day – one follows the other in a regular cycle
Misfortune	Talking about personal happiness will turn my fate to

	misfortune
Born happy	Happiness is something you are born with – you are either happy or unhappy and that can't be changed
Resent happy	If you talk about how happy you are, people will resent you
Create happiness	A person has to create happiness for himself or herself

Before the analysis, the data were preprocessed as follows: First, between-country differences in variable means were removed by centering the variables (criterion and predictors) per country. Next, in order to remove arbitrary scale differences and give each variable the same weight in the analysis, all variables were standardized (i.e., a variance of one) per country. As such, it is ensured that the *PCCR* clustering is influenced only by between-country differences in the correlations between the 34 predictors and the criterion (i.e., satisfaction with life). On the thus standardized data, *PCCR* analyses were performed with the number of components (R) and the number of clusters (K) going from two to four[9]. For each (R, K) combination, the analysis was performed using $.050$, $.010$ and $.001$ as α–values. Note that these low α–values were chosen as the simulation study demonstrated that low α–values yield the best results (see Section 4). For each (R, K, α) combination, the *PCCR* algorithm was run 100 times to avoid local minima and the solution with the best fit was retained.

[9] We believe that models with more than four clusters and/or four components are too complex for our data (i.e., 48 countries and 34 variables). In particular, the fit of these complex models is not substantially larger than the fit of less complex models. Furthermore, we did not consider solutions with one cluster and/or one component as such models are too simplistic.

Table 5

Cross-validation (CV) prediction error for the PCCR and the sequential solution for each considered model for the cross-cultural data regarding satisfaction with life

Number of components	**Number of clusters**	**α-value**	**CV error *PCCR***	**CV error sequential**
3	2	.050	7986.5	8265.8
3	3	.050	7992.0	8287.9
4	2	.050	7992.3	8135.6
2	2	.050	7992.3	8261.1
3	2	.001	8000.7	8265.8
2	2	.010	8001.1	8263.1
4	2	.010	8001.2	8139.2
2	2	.001	8001.7	8262.6
4	2	.001	8001.7	8135.6
3	4	.050	8006.0	8230.9
3	4	.010	8007.2	8230.6
3	2	.010	8007.3	8265.8
2	3	.050	8008.4	8246.3
3	4	.001	8008.6	8230.6
2	3	.010	8009.8	8246.3
2	3	.001	8010.0	8246.3
4	3	.050	8022.8	8147.6
3	3	.010	8041.7	8283.6
4	4	.050	8042.0	8114.7
4	3	.001	8044.3	8147.6

3	3	.001	8044.4	8287.3
4	3	.010	8044.7	8150.5
2	4	.010	8054.3	8276.4
2	4	.050	8057.7	8276.4
2	4	.001	8064.7	8276.4
4	4	.001	8075.5	8114.7
4	4	.010	8078.1	8114.7

Performing ten-fold cross-validation (see Section 3.3) revealed, as can be seen in Table 5, that for each considered (R, K, α) combination, the *PCCR* solution had a clearly smaller cross-validation error than the associated sequential solution and that both solutions yielded a very different clustering (as measured by means of the Adjusted Rand Index; Hubert & Arabie, 1985). We decided to retain the *PCCR* solution with three components, three clusters and an α-value of $.05$. Although this solution only has the second lowest cross-validation error (see Table 5), compared to the best solution, this solution yields cluster specific regression weights that differ in a more pronounced way and as such allows us to better illustrate the richness of the obtained results. For the selected number of clusters and components and α-value, the computation time for an average *PCCR* run equaled 209.5 seconds (and, hence, a little bit less than six hours for 100 runs[10]), whereas an average run of the sequential strategy took 1.5 seconds (and, hence, about two minutes and a half for 100 runs); 14 of the 100 *PCCR* runs resulted in the same optimal solution (with the same being true for the sequential strategy).

[10] These computations are based on the use of a single core at the same time. Modern computers, however, are able to use up to four cores simultaneously, reducing the computation time to two hours.

Table 6

Clustering of the countries in the PCCR solution ($\alpha = .05$) with three components and three clusters for the cross-cultural data regarding satisfaction with life

Cluster 1	**Cluster 2**	**Cluster 3**
Nigeria	Thailand	United States
Uganda	Philippines	Canada
Ghana	Indonesia	Australia
Cameroon	Hong Kong	Singapore
Malaysia	China	Japan
Bangladesh	Nepal	Korea Republic
Greece	India	Chile
Bulgaria	Switzerland	Brazil
	Turkey	Mexico
	Cyprus	Colombia
	Georgia	Venezuela
	Russia	Germany
	South Africa	Belgium
	Zimbabwe	Netherlands
	Egypt	Austria
		Portugal
		Spain
		Italy
		Slovenia
		Slovakia
		Poland

		Croatia
		Hungary
		Iran
		Kuwait

The obtained clustering of the countries is presented in Table 6. To validate and interpret this clustering, the clusters were related to a set of nation-level variables, including wealth and development measures (e.g., GDP, human development index, life expectancy), power distance, egalitarianism, etc. For each variable separately, we computed the ratio of the between cluster variance to the total variance (i.e., the amount of explained variance). In Table 7, the cluster specific means are presented for all variables that yielded a ratio larger than .13 (Cohen, 1992). It can be concluded that the countries in the first cluster have more power distance (i.e., a more hierarchical society), less individualism, a smaller IQ, smaller gender egalitarianism, lower health expenditure, lower life expectancy, lower GDP, lower human development and a lower health life expectancy than the countries in the third cluster. The countries in the second cluster take an intermediate position as their variable means are in between the means of the other two groups. In sum, the first cluster mainly groups less developed countries and the third cluster mainly well-developed countries, whereas the countries in the second cluster are situated somewhere in between.

Table 7

Cluster specific means on several nation-level variables for the three obtained clusters for the cross-cultural values data regarding satisfaction with life

Variable	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Ratio**
Power distance	78.00	71.18	58.45	.16
Individualism	25.16	33.70	49.37	.16
Gender egalitarianism	4.24	4.20	4.72	.29
IQ	79.87	87.33	95.80	.28
Health expenditure (% of GDP; 2002)	2.32	2.58	5.03	.36
Health expenditure per capita (US$; 2002)	372	561	1559	.23
GDP per capita (US$; 2003)	3067	6626	16723	.23
Life expectancy at birth (in years; 1998)	60.75	67.26	74.88	.37
Life Expectancy Index (1998)	.60	.72	.83	.38
Human Development Index (1998)	.60	.70	.86	.41
Length of life (in years)	62.88	60.91	74.50	.22
Health life expectancy	50.86	55.31	68.04	.30

In Table 8, the Varimax rotated loadings $\boldsymbol{P_X}$ and cluster specific regression weights $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ for the *PCCR* ($\alpha = .05$) and sequential solution with three components and three clusters for the cross-cultural data set are displayed. From this table one can see that the first *PCCR* component pertains to valuing external appearances (i.e., being rich and beautiful and outperforming others) along with the notion that one has to create happiness for him/herself. The second component is characterized by being extravert, living in the present and not focusing on what can go wrong but on possible rewards instead. The third component, finally, values intelligence and knowledge and not hiding yourself (i.e., extraversion) along with focusing on rewards in the future and believing that happiness is one's own responsibility.

Table 8

Loadings and cluster specific regression weights (after Varimax rotation) of the PCCR ($\alpha = .05$) and sequential solution with three components and three clusters for the cross-cultural values data regarding satisfaction with life[*]

	PCCR			**Sequential strategy**		
	Comp 1	*Comp 2*	*Comp 3*	*Comp 1*	*Comp 2*	*Comp 3*
Happiness	.24	.04	**.38**	**.43**	.17	-.02
Intelligence and knowledge	-.03	-.03	**.34**	.15	.19	.15
Material wealth	**.31**	-.05	-.04	**.69**	-.02	.00
Physical attractiveness	**.53**	.04	-.15	**.75**	-.02	-.02
Physical comforts	.28	.07	-.04	**.73**	-.05	-.02
Excitement and arousal	**.54**	-.02	.02	**.66**	.08	-.06
Competition	**.45**	.04	.11	**.69**	.05	.09
Getting to heaven	**.38**	.16	.03	**.57**	-.01	.01
Self-sacrifice	**.50**	.11	-.00	**.54**	.00	.09
Success	.23	.02	.25	**.45**	.21	.22
Fun	-.10	.09	**.38**	.07	.24	.14
Partylife	-.05	**.53**	.20	.03	**.62**	.01
No talk	.07	-.23	**-.34**	.01	**-.64**	.09
Background	.11	**-.43**	**-.32**	-.03	**-.64**	.15
Start conversation	.10	**.46**	.11	.04	**.67**	-.02
Talk people	.05	**.48**	.07	.01	**.69**	.01

Quiet strangers	-.06	**-.36**	-.07	.01	**-.57**	.15
Harmony	.10	.27	.09	.23	.21	.12
Outperforming	**.35**	-.19	.06	.25	.18	**.40**
Aging parents	.19	.11	.01	.11	.07	.19
Unique	.27	.16	.16	.15	.22	-.01
Competition	.10	-.02	.12	.16	.20	**.38**
Competition friends	.00	-.17	.00	.12	.12	**.55**
Competition family	.12	-.19	-.18	.01	.05	**.43**
Family success	.29	.01	.07	.09	.03	**.30**
Enjoy present	.07	**.54**	**-.37**	-.04	.05	-.05
Work hard	.29	-.17	.08	.12	.02	**.43**
Family pessimism	-.21	**-.33**	**-.44**	-.08	-.19	**.55**
School pessimism	-.13	-.26	**-.41**	-.03	-.19	**.51**
Cycle	-.06	-.04	-.28	.01	-.03	**.43**
Misfortune	-.12	-.12	-.28	-.10	-.09	**.49**
Born happy	-.06	-.00	**-.46**	-.12	-.10	**.39**
Resent happy	.10	-.15	-.22	-.06	-.08	**.44**
Create happiness	**.32**	.01	.12	.19	.15	.08
	Regression weights					
	PCCR			*Sequential strategy*		
Cluster 1	.28	.05	-.03	.09	.20	-.19
Cluster 2	.03	.38	.08	.08	.17	.04
Cluster 3	.15	.23	.31	.16	.35	-.14

* loadings (in absolute value) ≥ .30 are indicated in bold

Table 8 also shows the regression weights for the three country clusters and reveals that the relevance of the predictor components clearly differs across the clusters. In the cluster of least developed countries (cluster 1), satisfaction with life is related to keeping up appearances, doing it yourself and doing better than others (component 1). In the cluster of well-developed countries (i.e., third cluster) the third component (valuing intelligence, extraversion) has the strongest unique contribution. Finally, in the moderately developed countries (cluster 2), satisfaction with life is predicted by extraversion and by focusing more on immediate instead of anticipated rewards (component 2).

When comparing the *PCCR* results with those of the sequential strategy, the advantages of *PCCR* are nicely illustrated. First, whereas the components from the sequential solution correspond with the three sets of variables included in the study (i.e., values, personality and beliefs regarding happiness), the *PCCR* components better reveal the links between these variable sets and disclose the more subtle differences between them. Second, the *PCCR* solution shows larger differences in cluster specific regression weights than the sequential solution. Moreover, in the latter solution the same component has the largest regression weight in each cluster. Third, *PCCR* yields a more insightful clustering of the countries than the sequential strategy. In particular, when relating the sequential clustering to the nation-level variables, the clusters only differ on a subset of them, whereas the *PCCR* clusters show clear differences on all these variables. Finally, when performing ten-fold cross-validation, the *PCCR* solution better generalizes to other data from the same population as the *PCCR* solution has a clearly lower cross-validation error than the sequential one (i.e., 7990.9 versus 8283.1).

6 Discussion

In this discussion, we will reflect on technical choices that may influence the kind of solutions that are obtained and, therefore, the conclusions that can be drawn. These choices are situated on the level of the data (preprocessing), on the level of the model (which constraints are imposed) and on the level of the data analysis (choice of the α weight in the loss function).

On the level of the data, let us consider the preprocessing options available to the user. In the literature, preprocessing mostly refers to centering and scaling. Regarding centering, we opt to center the data per level-2 unit. This choice was made because studies on clusterwise regression have revealed that this method often yields a clustering that is dominated by differences in intercepts rather than slopes (Brusco et al., 2008). Centering per level-2 unit is a straightforward remedy, because it implies that the cluster specific intercepts equal zero and can thus be omitted from the model. Regarding scaling, we will discuss the pros and cons of three options: no scaling, standardizing across level-2 units, or standardizing per level-2 unit. Of course, other scaling options exist (see, e.g., van den Berg, Hoefsloot, Westerhuis, Smilde, & van der Werf, 2006), but they are almost never used in behavioral research. Which option one selects should depend among other reasons on the expected source of the (level-2 unit specific) differences in variance between variables. If differences in variance are real and important, one should not rescale, implying that variables with a larger variance will have more weight in the analysis. However, if differences in variance (partly) reflect arbitrary differences in measurement scale, it might be better to standardize, implying that each variable has the same weight in the analysis and with the additional advantage that the obtained loadings can be interpreted as correlations between the variables and the components.

In case one chooses to standardize, one must decide whether it is better to standardize per level-2 unit or across level-2 units. The advantage of standardizing across level-2 units is that differences in variance between level-2 units, which may often be of interest, are retained in the analysis. However, the disadvantage is that the clustering of the level-2 units is not only determined by the correlations between the components and the criterion, but also by the level-2 unit specific variance of the component scores and the criterion, as the regression weights depend on both the correlations and the variances. As large inter-level-2 unit differences in the variance of the predictors will imply large inter-level-2 unit differences in the variances of the component scores, standardizing across level-2 units may thus yield a different clustering than standardizing per level-2 unit. It can be concluded that researchers should carefully consider which aspects of the data should influence the obtained clustering: is only the correlation structure between the predictors and the criterion of interest or should variance differences also be taken into account.

On the model level, we imposed two constraints that influence the interpretation of *PCCR* solutions. First, the *PCCR* model requires the dimensional reduction of the predictors to be the same across clusters. In other words, the obtained clusters only differ with respect to the underlying regression models, which makes it relatively easy to compare the clusters. However, for some data sets one could argue that it would make sense to also allow the components to be different across clusters, which implies that one would fit a separate *PCovR* model to each cluster. Although such a model would be more general, the interpretation would be more difficult, as the regression models of different clusters are not comparable anymore, since they are based on different components.

Second, we imposed an orthogonality constraint on the component scores in order to avoid multicollinearity issues. However, for now, we only managed to impose it across level-2 units instead of for each level-2 unit separately, that is, we imposed $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}N$ rather than $\boldsymbol{T}_i'\boldsymbol{T}_i = \boldsymbol{I}N_i$. Note that the latter constraint would completely rule out multicollinearity problems, whereas the former constraint does not completely guarantee that none of the components are strongly correlated within a level-2 unit. By means of the *INDORT* rotation (Kiers, 1989), orthogonality per level-2 unit can be approximated but not imposed (see Section 2.2). Note, however, that rotating towards orthogonality per level-2 unit may imply that the loadings become harder to interpret, because one no longer aims for simple structure.

Regarding the data analysis, the *PCCR* loss function (L_2) is a weighted sum of two other loss functions, one for the predictor data and one for the criterion data. By means of the α weight, the user may manipulate which part of the data is emphasized more. Yet, although several methods have been proposed that imply such a weighted loss function in the context of clustering and dimension reduction (Brusco, Cradit, & Tashchian, 2003; Van Deun, Smilde, van der Werf, Kiers, & Van Mechelen, 2009), choosing an optimal weight is still an issue that has not been fully resolved, with different ways of weighting sometimes leading to very different results (Van Deun et al., 2009; Wilderjans, Ceulemans, & Van Mechelen, 2009; Wilderjans, Ceulemans, Van Mechelen, & van den Berg, 2011; van den Berg, Van Mechelen, Wilderjans, Van Deun, Kiers, & Smilde, 2009). We propose to use cross-validation to determine the optimal α weight for a data set at hand (see Section 3.3).

To conclude, we presented a new regression method that simultaneously handles three challenges: nested data, many predictors of which the underlying structure is unclear, and categorical differences in underlying regression models. As the regression data that are gathered by behavioral scientists become increasingly complex, we believe this tool might be of great value.

Appendix I *INDORT* rotation towards orthogonality per level-2 unit

The aim of the *INDORT* rotation is to find the orthonormal rotation matrix $\boldsymbol{U}$ $(\boldsymbol{U'U} = \boldsymbol{UU'} = \boldsymbol{I})$ that minimizes the following loss function:

$$f = \sum_{i=1}^{I} ||\boldsymbol{U'W'X'_iX_iWU} - \boldsymbol{D}_i||^2 = \sum_{i=1}^{I} ||\boldsymbol{U'T'_iT_iU} - \boldsymbol{D}_i||^2,$$

where $\boldsymbol{D}_i$ is a diagonal matrix containing the variances of the component scores for level-2 unit i. This loss function equals zero when the component scores are at the same time orthogonal across level-2 units (i.e., $\boldsymbol{T'T} = \boldsymbol{I}N$), which will always be the case since we imposed $\boldsymbol{T}$ to be orthogonal (see Section 2.2), and orthogonal within each level-2 unit i (i.e., $\boldsymbol{T'_iT_i} = \boldsymbol{I}N_i$). This function can be rewritten as:

$$\sum_{i=1}^{I} ||\boldsymbol{U'T'_iT_iU} - \boldsymbol{D}_i||^2 = \sum_{i=1}^{I} \boldsymbol{tr}[(\boldsymbol{U'T'_iT_iU} - \boldsymbol{D}_i)(\boldsymbol{U'T'_iT_iU} - \boldsymbol{D}_i)']$$

$$= \sum_{i=1}^{I} \boldsymbol{tr}[(\boldsymbol{U'T'_iT_iU} - \boldsymbol{D}_i)\boldsymbol{U'U}(\boldsymbol{U'T'_iT_iU} - \boldsymbol{D}_i)'\boldsymbol{U'U}]$$

$$= \sum_{i=1}^{I} \boldsymbol{tr}[\boldsymbol{U}(\boldsymbol{U'T'_iT_iU} - \boldsymbol{D}_i)\boldsymbol{U'U}(\boldsymbol{U'T'_iT_iU} - \boldsymbol{D}_i)'\boldsymbol{U'}]$$

$$= \sum_{i=1}^{I} \boldsymbol{tr}[(\boldsymbol{UU'T'_iT_iUU'} - \boldsymbol{UD}_i\boldsymbol{U'})(\boldsymbol{UU'T'_iT_iUU'} - \boldsymbol{UD'_iU'})]$$

$$= \sum_{i=1}^{I} \boldsymbol{tr}[(\boldsymbol{T'_iT_i} - \boldsymbol{UD}_i\boldsymbol{U'})(\boldsymbol{T'_iT_i} - \boldsymbol{UD'_iU'})]$$

yielding the *INDORT* loss function, for which an alternating least squares estimation approach exists (Kroonenberg, 1983; Kiers, 1989).

Appendix II Unconstrained minimization of $\boldsymbol{W}$

To update $\boldsymbol{W}$ (without any constraint), the loss function in (4) can be rewritten as follows:

$$L_2(\boldsymbol{W}) = \beta \sum_{i=1}^{I} \|\boldsymbol{X}_i - \boldsymbol{X}_i\boldsymbol{W}\boldsymbol{P}'_{\boldsymbol{X}}\|^2 + (1-\beta)\sum_{i=1}^{I} \left\|\boldsymbol{y}_i - \sum_{k=1}^{K} \boldsymbol{c}_{ik}\boldsymbol{X}_i\boldsymbol{W}\boldsymbol{p}_{\boldsymbol{Y}}^{k'}\right\|^2$$

$$= \sum_{k=1}^{K} \left\|\sqrt{\beta}\boldsymbol{X}_k - \sqrt{\beta}\boldsymbol{X}_k\boldsymbol{W}\boldsymbol{P}'_{\boldsymbol{X}}\right\|^2 + \sum_{k=1}^{K} \left\|\sqrt{1-\beta}\boldsymbol{y}_k - \sqrt{1-\beta}\boldsymbol{X}_k\boldsymbol{W}\boldsymbol{p}_{\boldsymbol{Y}}^{k'}\right\|^2$$

$$= \sum_{k=1}^{K} \left\|vec\big(\sqrt{\beta}\boldsymbol{X}_k\big) - \sqrt{\beta}(\boldsymbol{P}_{\boldsymbol{X}} \otimes \boldsymbol{X}_k)vec(\boldsymbol{W})\right\|^2$$

$$+ \sum_{k=1}^{K} \left\|vec\big(\sqrt{1-\beta}\boldsymbol{y}_k\big) - \sqrt{1-\beta}\big(\boldsymbol{p}_{\boldsymbol{Y}}^{k} \otimes \boldsymbol{X}_k\big)vec(\boldsymbol{W})\right\|^2$$

$$= \left\|\begin{bmatrix} vec\big(\sqrt{\beta}\boldsymbol{X}_1\big) \\ \dots \\ vec\big(\sqrt{\beta}\boldsymbol{X}_K\big) \end{bmatrix} - \begin{bmatrix} \sqrt{\beta}(\boldsymbol{P}_{\boldsymbol{X}} \otimes \boldsymbol{X}_1) \\ \dots \\ \sqrt{\beta}(\boldsymbol{P}_{\boldsymbol{X}} \otimes \boldsymbol{X}_K) \end{bmatrix} vec(\boldsymbol{W})\right\|^2$$

$$+ \left\|\begin{bmatrix} vec\big(\sqrt{1-\beta}\boldsymbol{y}_1\big) \\ \dots \\ vec\big(\sqrt{1-\beta}\boldsymbol{y}_K\big) \end{bmatrix} - \begin{bmatrix} \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{1} \otimes \boldsymbol{X}_1) \\ \dots \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{K} \otimes \boldsymbol{X}_K) \end{bmatrix} vec(\boldsymbol{W})\right\|^2$$

$$= \left\|\begin{bmatrix} vec\big(\sqrt{\beta}\boldsymbol{X}_1\big) \\ \dots \\ vec\big(\sqrt{\beta}\boldsymbol{X}_K\big) \\ vec\big(\sqrt{1-\beta}\boldsymbol{y}_1\big) \\ \dots \\ vec\big(\sqrt{1-\beta}\boldsymbol{y}_K\big) \end{bmatrix} - \begin{bmatrix} \sqrt{\beta}(\boldsymbol{P}_{\boldsymbol{X}} \otimes \boldsymbol{X}_1) \\ \dots \\ \sqrt{\beta}(\boldsymbol{P}_{\boldsymbol{X}} \otimes \boldsymbol{X}_K) \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{1} \otimes \boldsymbol{X}_1) \\ \dots \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{K} \otimes \boldsymbol{X}_K) \end{bmatrix} vec(\boldsymbol{W})\right\|^2$$

$$= \|\boldsymbol{y}^* - \boldsymbol{Z}^* vec(\boldsymbol{W})\|^2$$

It appears that updating this loss function over $\boldsymbol{W}$ conditional on $\boldsymbol{P}_{\boldsymbol{X}}$, $\boldsymbol{p}_{\boldsymbol{Y}}^{k}$ and $\boldsymbol{C}$ boils down to solving a multivariate multiple linear regression problem, for which the following closed-form solution exists:

$$vec(\boldsymbol{W}) = \big(\boldsymbol{Z}^{*'}\boldsymbol{Z}^*\big)^{-1}\boldsymbol{Z}^{*'}\boldsymbol{y}^*$$

The matrix $\boldsymbol{W}$ can be obtained by devectorizing $vec(\boldsymbol{W})$ into a matrix. Note that $\boldsymbol{Z}^{*'}\boldsymbol{Z}^{*}$ and $\boldsymbol{Z}^{*'}\boldsymbol{y}^{*}$ can be simplified as follows:

$$\boldsymbol{Z}^{*'}\boldsymbol{Z}^{*} = \begin{bmatrix} \sqrt{\beta}(\boldsymbol{P_X} \otimes \boldsymbol{X}_1) \\ \cdots \\ \sqrt{\beta}(\boldsymbol{P_X} \otimes \boldsymbol{X}_K) \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{1} \otimes \boldsymbol{X}_1) \\ \cdots \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{K} \otimes \boldsymbol{X}_K) \end{bmatrix}' \begin{bmatrix} \sqrt{\beta}(\boldsymbol{P_X} \otimes \boldsymbol{X}_1) \\ \cdots \\ \sqrt{\beta}(\boldsymbol{P_X} \otimes \boldsymbol{X}_K) \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{1} \otimes \boldsymbol{X}_1) \\ \cdots \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{K} \otimes \boldsymbol{X}_K) \end{bmatrix}$$

$$= \left(\sqrt{\beta}\boldsymbol{P}'_{\boldsymbol{X}}\sqrt{\beta}\boldsymbol{P_X} \otimes \boldsymbol{X}'_1\boldsymbol{X}_1\right) + \cdots + \left(\sqrt{\beta}\boldsymbol{P}'_{\boldsymbol{X}}\sqrt{\beta}\boldsymbol{P_X} \otimes \boldsymbol{X}'_K\boldsymbol{X}_K\right)$$

$$+ \left(\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{1'}\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{1} \otimes \boldsymbol{X}'_1\boldsymbol{X}_1\right) + \cdots + \left(\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{K'}\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{K} \otimes \boldsymbol{X}'_K\boldsymbol{X}_K\right)$$

$$= \sum_{k=1}^{K}\left[\left(\sqrt{\beta}\boldsymbol{P}'_{\boldsymbol{X}}\sqrt{\beta}\boldsymbol{P_X} \otimes \boldsymbol{X}'_k\boldsymbol{X}_k\right) + \left(\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{k'}\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{k} \otimes \boldsymbol{X}'_k\boldsymbol{X}_k\right)\right]$$

$$= \sum_{k=1}^{K}\left[\left(\sqrt{\beta}\boldsymbol{P}'_{\boldsymbol{X}}\sqrt{\beta}\boldsymbol{P_X} + \sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{k'}\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{k}\right) \otimes \boldsymbol{X}'_k\boldsymbol{X}_k\right]$$

$$= \sum_{k=1}^{K}\left[\left(\beta\boldsymbol{P}'_{\boldsymbol{X}}\boldsymbol{P_X} + (1-\beta)\boldsymbol{p}_{\boldsymbol{Y}}^{k'}\boldsymbol{p}_{\boldsymbol{Y}}^{k}\right) \otimes \boldsymbol{X}'_k\boldsymbol{X}_k\right]$$

$$\boldsymbol{Z}^{*'}\boldsymbol{y}^{*} = \begin{bmatrix} \sqrt{\beta}(\boldsymbol{P_X} \otimes \boldsymbol{X}_1) \\ \cdots \\ \sqrt{\beta}(\boldsymbol{P_X} \otimes \boldsymbol{X}_K) \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{1} \otimes \boldsymbol{X}_1) \\ \cdots \\ \sqrt{1-\beta}(\boldsymbol{p}_{\boldsymbol{Y}}^{K} \otimes \boldsymbol{X}_K) \end{bmatrix}' \begin{bmatrix} vec\left(\sqrt{\beta}\boldsymbol{X}_1\right) \\ \cdots \\ vec\left(\sqrt{\beta}\boldsymbol{X}_K\right) \\ vec\left(\sqrt{1-\beta}\boldsymbol{y}_1\right) \\ \cdots \\ vec\left(\sqrt{1-\beta}\boldsymbol{y}_K\right) \end{bmatrix} = \left(\sqrt{\beta}\boldsymbol{P}'_{\boldsymbol{X}} \otimes \boldsymbol{X}'_1\right)vec\left(\sqrt{\beta}\boldsymbol{X}_1\right) + \cdots +$$

$\left(\sqrt{\beta}\boldsymbol{P}'_{\boldsymbol{X}} \otimes \boldsymbol{X}'_K\right)vec\left(\sqrt{\beta}\boldsymbol{X}_K\right) + \left(\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{1'} \otimes \boldsymbol{X}'_1\right)vec\left(\sqrt{1-\beta}\boldsymbol{y}_1\right) + \cdots + \left(\sqrt{1-\beta}\boldsymbol{p}_{\boldsymbol{Y}}^{K'} \otimes \boldsymbol{X}'_K\right)vec\left(\sqrt{1-\beta}\boldsymbol{y}_K\right) = vec[\beta\boldsymbol{X}'_1\boldsymbol{X}_1\boldsymbol{P_X} + \cdots + \beta\boldsymbol{X}'_K\boldsymbol{X}_K\boldsymbol{P_X} + (1-\beta)\boldsymbol{X}'_1\boldsymbol{y}_1\boldsymbol{p}_{\boldsymbol{Y}}^{1} + \cdots + (1-\beta)\boldsymbol{X}'_K\boldsymbol{y}_K\boldsymbol{p}_{\boldsymbol{Y}}^{K}] = vec\left[\beta\left(\sum_{k=1}^{K}\boldsymbol{X}'_k\boldsymbol{X}_k\right)\boldsymbol{P_X} + (1-\beta)\sum_{k=1}^{K}\left(\boldsymbol{X}'_k\boldsymbol{y}_k\boldsymbol{p}_{\boldsymbol{Y}}^{k}\right)\right]$.

References

Arminger, G., & Stein, P. (1997). Finite mixtures of covariance structure models with regressors: Loglikelihood function, minimum distance estimation, fit indices, and a complex example. *Sociological Methods & Research, 26*(2), 148-182. doi: 10.1177/0049124197026002002

Brusco, M. J., & Cradit, J. D. (2001). A variable selection heuristic for K-means clustering. *Psychometrika, 66*, 249–270. doi: 10.1007/BF02294838

Brusco, M. J., Cradit, J. D., Steinley, D., & Fox, G. L. (2008). Cautionary remarks on the use of Clusterwise Regression. *Multivariate Behavioral Research, 43*(1), 29-49. doi: 10.1080/00273170701836653

Brusco, M. J., Cradit, J. D., & Tashchian, A. (2003). Multicriterion clusterwise regression for joint segmentation settings: An application to customer value. *Journal of Marketing Research, 40*(2), 225-234.

Ceulemans, E., & Kiers, H. A. L. (2009). Discriminating between strong and weak structures in three-mode principal component analysis. *British Journal of Mathematical & Statistical Psychology, 62*, 601-620. doi:10.1348/000711008X369474

Ceulemans, E., Kuppens, P., & Van Mechelen, I. (2012). Capturing the structure of distinct types of individual differences in the situation-specific experience of emotions: The case of anger. *European Journal of Personality, 26*, 484-495. doi:10.1002/per.847

Ceulemans, E., & Van Mechelen, I. (2008). CLASSI: A classification model for the study of sequential processes and individual differences therein. *Psychometrika, 73,* 107-124. doi: 10.1007/s11336-007-9024-1

Ceulemans, E., Van Mechelen, I., & Leenen, I. (2007). The local minima problem in hierarchical classes analysis: an evaluation of a simulated annealing algorithm and various multistart procedures. *Psychometrika, 72*, 377–391.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. doi:10.1037/0033-2909-112-1-155

Coxe, K. L. (1986). Principal components regression analysis. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.), *Encyclopedia of statistical sciences* (pp. 181-184). New York: Wiley.

de Jong, S., & Kiers, H. A. L. (1992). Principal covariates regression: Part I. Theory. *Chemometrics and Intelligent Laboratory Systems, 14*(1-3), 155-164. doi: 10.1016/0169-7439(92)80100-I

De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., & Onghena, P. (2012). Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. Psychological Methods, 17, 100-119. doi:10.1037/a0025385

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, *5*, 249-282.

DeSarbo, W. S., & Edwards, E. A. (1996). Typologies of compulsive buying behavior: A constrained clusterwise regression approach. *Journal of Consumer Psychology, 5*(3), 231-262.

DeSarbo, W. S., Oliver, R. L., & Rangaswamy, A. (1989). A simulated annealing methodology for clusterwise linear regression. *Psychometrika, 54*(4), 707-736.

Hahn, C., Johnson, M. D., Herrmann, A., & Huber, F. (2002). Capturing customer heterogeneity using a finite mixture PLS approach. *Schmalenbach Business Review, 54*(3), 243-269.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187-200.

Kiers, H. A. L. (1989). *Three-way methods for the analysis of qualitative and quantitative two-way data.* Leiden, The Netherlands: DSWO Press.

Kiers, H. A. L., & Smilde, A. (2007). A comparison of various methods for multivariate regression with highly collinear variables. *Statistical Methods & Applications, 16*(2), 193-228. doi: 10.1007/s10260-006-0025-5

Kiers, H. A. L., & ten Berge, J. M. F. (1992). Minimization of a class of matrix trace functions by means of refined majorization. *Psychometrika, 57*, 371-382.

Korth, B., & Tucker, L. R. (1975). The distribution of chance congruence coefficients from simulated data. *Psychometrika, 40*(3), 361-372.

Kroonenberg, P. M. (2008). *Applied multiway data analysis*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Kroonenberg, P. M. (1983). *Three-mode principal component analysis: Theory and applications*. Leiden: DSWO Press.

Kuppens, P., Ceulemans, E., Timmerman, M. E., Diener, E., & Kim-Prieto, C. (2006). Universal intracultural and intercultural dimensions of the recalled frequency of emotional experience. *Journal of Cross-Cultural Psychology, 37,* 491-515. doi:10.1177/0022022106290474

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software, 11*(8), 1–18.

Roa, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 26*(4), 329-358.

Sarstedt, M., & Ringle, C. M. (2010). Treating unobserved heterogeneity in PLS path modeling: a comparison of FIMIX-PLS with different data analysis strategies. *Journal of Applied Statistics, 37*(8), 1299-1318. doi: 10.1080/02664760903030213

Schott, J. R. (2005). *Matrix analysis for statistics (2nd ed.)*. Hoboken, New Jersey, USA: John Wiley & Sons, Inc.

Späth, H. (1979). Algorithm 39: Clusterwise linear regression. *Computing, 22*(4), 367-373. doi:10.1007/BF02265317

Späth, H. (1981). Correction to algorithm 39: Clusterwise linear regression. *Computing, 26*(3), 275-275. doi: 10.1007/BF02243486

Steinley, D. (2003). Local optima in K-means clustering: What you don't know may hurt you. *Psychological Methods, 8*, 294–304. doi: 10.1037/1082-989X.8.3.294

Steinley, D. (2004). Properties of the Hubert-Arabie Adjusted Rand Index. *Psychological Methods, 9*(3), 386–396. doi: 10.1037/1082-989X.9.3.386

Stormshak, E. A., Bierman, K. L., Bruschi, C., Dodge, K. A., Coie, J. D., & the Conduct Problems Prevention Research Group (1999). The relation between behavior problems and peer preference in different classroom contexts. *Child Development, 70*(1), 169-182.

ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika, 42*(2), 267-276.

Tucker, L. R. (1951). A method for synthesis of factor analysis studies. *Personnel Research Section Rapport # 984*. Washington, DC: Department of the Army.

van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling and transformations: improving the biological information content of metabolomics data. *BMC Genomics, 7*(1), 142-157. doi:10.1186/1471-2164-7-142

van den Berg, R. A., Van Mechelen, I., Wilderjans, T. F., Van Deun, K., Kiers, H. A. L., & Smilde, A. K. (2009). Integrating functional genomics data using maximum likelihood based simultaneous component analysis. *BMC Bioinformatics, 10*, 340. doi:10.1186/1471-2105-10-340

Van Deun, K., Smilde, A. K., van der Werf, M. J., Kiers, H. A. L., & Van Mechelen, I. (2009). A structured overview of simultaneous component based data integration. *BMC Bioinformatics, 10,* 246. doi:10.1186/1471-2105-10-246

Vervloet, M., Van Deun, K., Van den Noortgate, W., & Ceulemans, E. (2013). On the selection of the weighting parameter value in principal covariates regression. *Chemometrics and Intelligent Laboratory Systems, 123*, 36-43. doi:10.1016/j.chemolab.2013.02.005

Wedel, M., & DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification, 12,* 21-55.

Wilderjans, T. F., & Ceulemans, E. (2013). Clusterwise Parafac to identify heterogeneity in three-way data. *Chemometrics and Intelligent Laboratory Systems, 129*, 87-97. doi:10.1016/j.chemolab.2013.09.010

Wilderjans, T. F., Ceulemans, E., & Kuppens, P. (2012). Clusterwise HICLAS: A generic modeling strategy to trace similarities and differences in multi-block binary data. *Behavior Research Methods, 44*, 532-545. doi:10.3758/s13428-011-0166-9

Wilderjans, T. F., Ceulemans, E., & Van Mechelen, I. (2009). Simultaneous analysis of coupled data blocks differing in size: A comparison of two weighting schemes. *Computational Statistics and Data Analysis, 53*, 1086-1098. doi:10.1016/j.csda.2008.09.031

Wilderjans, T. F., Ceulemans, E., Van Mechelen, I., & van den Berg, R. A. (2011). Simultaneous analysis of coupled data matrices subject to different amounts of noise. *British Journal of Mathematical and Statistical Psychology, 64*, 277-290. doi:10.1348/000711010X513263

Wold, H. (1966). Estimation of principal component and related methods by iterative least squares. In P. R. Krishnaiah (ed.), *Multivariate analysis* (pp. 391-420). New York, USA: Academic Press.