Bootstrap confidence intervals for principal covariates regression

Giordani, Paolo; Kiers, Henk A. L.

Published in:
British journal of mathematical & statistical psychology

DOI:
10.1111/bmsp.12238

Publication date:
2021

the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2021), 74, 541–566

www.wileyonlinelibrary.com

Bootstrap confidence intervals for principal covariates regression

Paolo Giordani[1]* and Henk A. L. Kiers[2]

[1]Department of Statistical Sciences, Sapienza University of Rome, Italy
[2]Department of Psychology, University of Groningen, The Netherlands

Principal covariate regression (PCOVR) is a method for regressing a set of criterion variables with respect to a set of predictor variables when the latter are many in number and/or collinear. This is done by extracting a limited number of components that simultaneously synthesize the predictor variables and predict the criterion ones. So far, no procedure has been offered for estimating statistical uncertainties of the obtained PCOVR parameter estimates. The present paper shows how this goal can be achieved, conditionally on the model specification, by means of the bootstrap approach. Four strategies for estimating bootstrap confidence intervals are derived and their statistical behaviour in terms of coverage is assessed by means of a simulation experiment. Such strategies are distinguished by the use of the varimax and quartimin procedures and by the use of Procrustes rotations of bootstrap solutions towards the sample solution. In general, the four strategies showed appropriate statistical behaviour, with coverage tending to the desired level for increasing sample sizes. The main exception involved strategies based on the quartimin procedure in cases characterized by complex underlying structures of the components. The appropriateness of the statistical behaviour was higher when the proper number of components were extracted.

1. Introduction

There exist several methods for analysing the dependence between two sets of variables, say $\mathbf{Y}$ and $\mathbf{X}$, observed on the same group of units. The simplest strategy is to perform multivariate linear regression. However, such an approach may lead to poor results, especially when the number of predictor variables in $\mathbf{X}$ is large, due to a high risk of multicollinearity of the predictor variables. If so, $\mathbf{X}$ is (close to) rank deficient, implying that the estimates of the regression coefficients tend to be large as well as the variances of the corresponding estimators. For this reason, alternative methods have been developed. At least two main classes of methods can be distinguished (see, for example, Hastie, Tibshirani, & Friedman, 2009, Chapter 3). The first class involves regularization methods such as the well-known ridge regression, proposed by Hoerl and Kennard (1970) following Tikhonov (1943). These techniques consider different regularizations of the regression coefficients in order to prevent the estimates and the variances from growing. The second class involves the use of dimension reduction techniques in order to reduce the space of the predictor variables by means of a limited number of underlying components. For

*Correspondence should be addressed to Paolo Giordani, P.le Aldo Moro, 5, 00185 Rome, Italy (email: paolo.giordani@uniroma1.it).

DOI:10.1111/bmsp.12238

an overview, the interested reader may refer to De Jong and Kiers (1992), Kiers and Smilde (2007), and Vervloet (2017). A popular technique belonging to this class is principal component regression (PCR; Jolliffe, 1982), where components are first extracted from $\mathbf{X}$ and then the variables in $\mathbf{Y}$ are regressed on the scores of these components. On the one hand, this solves the multicollinearity problem because the components are orthogonal. On the other hand, it is not guaranteed that the components explain the criterion variables in $\mathbf{Y}$ reasonably well, hence the prediction error of $\mathbf{Y}$ may be large. An alternative strategy is represented by reduced-rank regression (RRR; Anderson, 1951, 1958; Izenman, 1975). The primary goal of RRR is the prediction of $\mathbf{Y}$ by looking for the components of $\mathbf{X}$ such that the prediction error of $\mathbf{Y}$ is minimized, even if they may synthesize $\mathbf{X}$ in an unsatisfactory way. Therefore, both PCR and RRR extract components, but differ in the way the components are found. The former approach gives components accounting for the variance of $\mathbf{X}$ at best (regardless of the correlation with $\mathbf{Y}$), while the latter gives components correlated with $\mathbf{Y}$ as much as possible (regardless the explained variance of $\mathbf{X}$). In order to take into account both the objectives, principal covariates regression (PCOVR; De Jong & Kiers, 1992) has been proposed. In PCOVR, the components are found by minimizing a linear combination of the two previously described criteria. In this way, the PCOVR components are able to summarize $\mathbf{X}$ and predict $\mathbf{Y}$.

The effectiveness of PCOVR, also in comparison with its potential competitors, has been shown in several papers; see, for example, Kiers and Smilde (2007) and Vervloet, Van Deun, Van den Noortgate, and Ceulemans (2016). To judge to what extent the PCOVR estimates can be generalized to the population, inferential statistics are needed. In this paper, we examine a bootstrap procedure to estimate confidence intervals for PCOVR parameters. The use of bootstrapping is motivated by the fact that it works well in principal component analysis and, therefore, we expect the same for PCOVR. In principal component methods, several variants of the bootstrap are available; see, for example, Timmerman, Kiers, and Smilde (2007). The most relevant ones are recalled and applied to the PCOVR domain; furthermore, a simulation experiment is carried out in order to assess their quality. The paper is structured as follows. In the next section PCOVR is presented. Section 3 focuses on some strategies for estimating confidence intervals of the PCOVR parameter estimates. The results of the simulation experiment are reported in Section 4. A final discussion in Section 5 concludes the paper.

2. Principal covariates regression

Let $\mathbf{Y}$ and $\mathbf{X}$ be two matrices of order $N \times K$ and $N \times J$, respectively, where N denotes the number of units and K and J (usually, but not necessarily, $K < J$) are the number of criterion variables and predictor variables, respectively. In PCOVR, R ($<J$) components, expressed as a linear combination of $\mathbf{X}$, are sought:

$$\mathbf{T} = \mathbf{X}\mathbf{W}, \tag{1}$$

where $\mathbf{T}$ is the component score matrix of order $N \times R$ and $\mathbf{W}$ is the component weight matrix of order $J \times R$. The component scores in $\mathbf{T}$ play the role of explaining both $\mathbf{Y}$ and $\mathbf{X}$, namely,

$$\mathbf{Y} = \mathbf{T}\mathbf{P}_{\mathbf{Y}} + \mathbf{E}_{\mathbf{Y}}, \tag{2}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}_{\mathbf{X}} + \mathbf{E}_{\mathbf{X}}, \tag{3}$$

where $\mathbf{P}_{\mathbf{Y}}$ is the $R \times K$ matrix of the regression weights for the K criterion variables on the R components and $\mathbf{P}_{\mathbf{X}}$ is the $R \times J$ matrix of the component loadings for the J predictor variables on the components. Finally, $\mathbf{E}_{\mathbf{Y}}$ and $\mathbf{E}_{\mathbf{X}}$ are the error matrices for $\mathbf{Y}$ and $\mathbf{X}$, respectively.

The parameter matrices $\mathbf{P}_{\mathbf{Y}}$, $\mathbf{P}_{\mathbf{X}}$ and $\mathbf{T}$ (and implicitly $\mathbf{W}$ according to (1)) are estimated by minimizing a linear combination of the sum of squares of $\mathbf{E}_{\mathbf{Y}}$ and $\mathbf{E}_{\mathbf{X}}$, expressed as

$$f(\mathbf{P}_{\mathbf{Y}}, \mathbf{P}_{\mathbf{X}}, \mathbf{T}) = (1-\alpha)\frac{\| \mathbf{Y} - \mathbf{T}\mathbf{P}_{\mathbf{Y}} \|^2}{\| \mathbf{Y} \|^2} + \alpha\frac{\| \mathbf{X} - \mathbf{T}\mathbf{P}_{\mathbf{X}} \|^2}{\| \mathbf{X} \|^2}, \tag{4}$$

where $\| \cdot \|$ denotes the Frobenius norm of matrices and $\alpha \in [0, 1]$ is a tuning parameter for the variances of $\mathbf{X}$ and $\mathbf{Y}$ explained by $\mathbf{T}$. When $\alpha = 0$, PCOVR reduces to RRR, while setting $\alpha = 1$ leads to PCR. The automatic selection of α has been widely investigated in several papers; see De Jong and Kiers (1992), Kiers and Smilde (2007), Vervloet, Van Deun, Van den Noortgate, and Ceulemans (2013), and Vervloet, Van Deun, Van den Noortgate, and Ceulemans (2016).

Given α and R, closed-form estimates of the parameter matrices can be found. For this purpose, we observe that (4) can be rewritten as

$$f(\mathbf{P}_{\mathbf{Y}}, \mathbf{P}_{\mathbf{X}}, \mathbf{T}) = \| \begin{bmatrix} \frac{(1-\alpha)^{1/2}}{||\mathbf{Y}||}\mathbf{Y} \\ \frac{\alpha^{1/2}}{||\mathbf{X}||}\mathbf{X} \end{bmatrix} - \mathbf{T} \begin{bmatrix} \frac{(1-\alpha)^{1/2}}{||\mathbf{Y}||}\mathbf{P}_{\mathbf{Y}} \\ \frac{\alpha^{1/2}}{||\mathbf{X}||}\mathbf{P}_{\mathbf{X}} \end{bmatrix} \|^2. \tag{5}$$

De Jong and Kiers (1992) showed that $\mathbf{T}$ contains in its columns the first R eigenvectors of

$$(1-\alpha)\frac{\mathbf{H}_{\mathbf{X}}\mathbf{Y}\mathbf{Y}'\mathbf{H}_{\mathbf{X}}}{||\mathbf{Y}||^2} + \alpha\frac{\mathbf{X}\mathbf{X}'}{||\mathbf{X}||^2}, \tag{6}$$

where $\mathbf{H}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the role of which is to project $\mathbf{Y}$ on the space spanned by the columns of $\mathbf{X}$. Once the estimate of $\mathbf{T}$ is found, $\mathbf{P}_{\mathbf{Y}}$ and $\mathbf{P}_{\mathbf{X}}$ are estimated by

$$\mathbf{P}_{\mathbf{Y}} = \mathbf{T}'\mathbf{Y}, \tag{7}$$

$$\mathbf{P}_{\mathbf{X}} = \mathbf{T}'\mathbf{X}, \tag{8}$$

Finally, the component weights are estimated by

$$\mathbf{W} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{T}. \tag{9}$$

For further details, see De Jong and Kiers (1992).

The columns of $\mathbf{T}$ are usually normalized so that, in each column, the elements have variance fixed at 1. In this case, when $\mathbf{X}$ contains standardized data and the extracted components are orthogonal, the component loadings in $\mathbf{P_X}$ are equal to the correlations between components and variables. Such a normalization of $\mathbf{T}$ is not sufficient to fully identify the solution. In fact, equally fitting solutions can be found by premultiplying $\mathbf{P_Y}$ and $\mathbf{P_X}$ by a rotation matrix $\mathbf{B}$, provided that this rotation is compensated by postmultiplying $\mathbf{T}$ by $\mathbf{B}^{-1}$. Letting $\mathbf{P_Y^R} = \mathbf{BP_Y}$ $\mathbf{P_X^R} = \mathbf{BP_X}$, and $\mathbf{T^R} = \mathbf{TB}^{-1}$, we have

$$\mathbf{Y} = \mathbf{T^R P_Y^R} + \mathbf{E_Y} = \mathbf{TB}^{-1}\mathbf{BP_Y} + \mathbf{E_Y} = \mathbf{TP_Y} + \mathbf{E_Y}, \tag{10}$$

$$\mathbf{X} = \mathbf{T^R P_X^R} + \mathbf{E_X} = \mathbf{TB}^{-1}\mathbf{BP_X} + \mathbf{E_X} = \mathbf{TP_X} + \mathbf{E_X}. \tag{11}$$

The rotational freedom can be exploited in order to simplify the interpretation of the components by means of, for example, the (normalized or non-normalized) varimax procedure (Kaiser, 1958). Throughout the present paper, we opt for the non-normalized version.

For practical purposes, the applicability of PCOVR is enhanced by the availability of the R (R Core Team, 2020) package *PCovR* (Vervloet et al., 2015). This was used to perform the PCOVR analysis in the following example as well as in the simulation experiment in Section 4.

2.1. Example

A PCOVR analysis is applied to the Rohwer data (Timm, 1975) for illustrative purposes. The data refer to an experiment on $N = 69$ children. The research interest lies in assessing whether and how a set of $J = 5$ paired-associate (PA) tasks predicts the performance on $K = 3$ measures of aptitude and achievement. The PA tasks, labelled named (N), still (S), named still (NS), named action (NA), and sentence still (SS), vary in how the stimuli are presented. The three measures are a student achievement test (SAT), the Peabody picture vocabulary test (PPVT) and the Raven progressive matrices test (RPMT). The need for PCOVR arises because several predictor variables are highly correlated (see Table 1).

PCOVR is run on such data, varying R in the set $\{2, 3, 4\}$ and setting α following the sequential procedure proposed by Vervloet, Van Deun, Van den Noortgate, and Ceulemans (2013). First, the value of α is determined on the basis of maximum likelihood principles. Then, by using the selected weighting value, a generalization of the well-known scree test is used to determine R. We get the optimal values $R = 3$ and $\alpha = .91$. In order to interpret the solution, we inspect the varimax rotated component loading matrix reported in Table 2.

Table 1. Correlation matrix for the PA tasks

	N	S	NS	NA	SS
N	1.00	.25	.51	.49	.46
S		1.00	.34	.55	.43
NS			1.00	.68	.66
NA				1.00	.72
SS					1.00

Table 2. Varimax rotated component loading matrix for the predictor variables ($\mathbf{P}_{\mathbf{X}'}$). Component loadings higher than .30 in absolute sense are in bold

	Component 1	Component 2	Component 3
N	.25	.11	**.96**
S	.20	**.96**	.10
NS	**.83**	.09	**.32**
NA	**.74**	**.45**	.27
SS	**.84**	.26	.21

Table 3. Regression weight matrix for the criterion variables on the components ($\mathbf{P}_{\mathbf{Y}'}$)

	Component 1	Component 2	Component 3
SAT	.15	.23	.24
PPVT	.47	.17	.25
RPMT	.23	.26	.16

Table 4. Regression weight matrix for the criterion variables on the predictor variables ($\mathbf{WP}_{\mathbf{y}}$)

	SAT	PPVT	RPMT
N	.20	.11	.09
S	.18	.01	.18
NS	.00	.19	.04
NA	.08	.17	.11
SS	.02	.19	.07

We can see that component 1 has high positive loadings for all the tasks with particular reference to SS, NS and NA. Components 2 and 3 are mainly related to the remaining two tasks. Specifically, component 2 is positively related to S and component 3 to N. In order to assess how the components predict the criterion variables, we observe the regression weight matrix given in Table 3.

We see that positive relations emerge. PPVT is well predicted by component 1. Therefore, good performances of SS, NS, and NA are predictors of good performances in PPVT. The predictions of SAT and RPMT are more complex. SAT appears to be related most to components 2 and 3. Hence, good performances of SAT primarily reflects good performances for S and N. Concerning RPMT, to a certain extent a connection with components 1 and 2 is evident. To explicitly inspect the relationship between the criterion and predictor variables, the matrix $\mathbf{WP}_{\mathbf{y}}$ can be analysed. This matrix gives the regression weight for estimating $\mathbf{Y}$ from $\mathbf{X}$, bypassing the interpretation of the components. We have the weights reported in Table 4.

Obviously, by inspecting Table 4 we can draw conclusions consistent with those based on Tables 2 and 3. Furthermore, we can highlight the role of S in predicting RPMT. The matrix $\mathbf{T}$ (not reported here) may also be considered to study how the units take different scores with respect to the extracted components.

3. Bootstrap methodology in principal covariates regression

Once we have the PCOVR solution, it would be interesting to assess the inferential properties of the estimates obtained. In particular, a relevant point to address is whether the solution obtained represents a good estimate of the population solution. This can be done by computing either standard errors (SEs) or confidence intervals (CIs) for the parameter matrices. In the particular domain of PCOVR, this problem has not yet been covered. Therefore, we will use the results in the PCA framework as a starting point, taking into account the existing connection between PCOVR and PCA reported in (5). For this purpose, we rely strongly on the paper by Timmerman et al. (2007). We must note that our results are developed conditionally on the PCOVR model specification. Specifically, on the basis of the available data, we first choose the value of α and the number of components R. Then, given this specification, we study the uncertainty of the parameter estimates.

The computation of SEs or CIs in PCA can be done following at least two strategies depending on the multivariate normality assumption. If this holds, assuming that the scores are independently randomly sampled from identical distributions, analytic results can be derived (for an overview, see Anderson, 1984). Refinements for particular cases are also available. For instance, Ogasawara (2000) studies asymptotic SEs for component loadings when data are standardized. The same author investigates SEs for rotated component loadings (Ogasawara, 1999, 2002).

3.1. Bootstrap

When the multivariate normality assumption is violated, as is often the case, analytic results are no longer available. We note that there are several studies on the problem of misspecification and how it affects parameter estimates and standard errors (see, e.g., White, 1982). An alternative approach, adopted in this paper, is represented by the use of resampling techniques such as the bootstrap, originally introduced by Efron (1979). Let θ be a population parameter of interest. It can be estimated by the sample statistic t that represents the realization of the estimator T. The bootstrap leads to an assessment of the uncertainty associated with T by computing SEs or CIs. This is done by mimicking the sampling process by using the observed sample. Let t_b be the estimate of θ obtained by using the bth bootstrap sample ($b = 1, \ldots, B$), where B is the number of bootstrap samples. The bootstrap SE of T is

$$SE(T) = \sqrt{\sum_{b=1}^{B} \left(t_b - \frac{\sum_{b=1}^{B} t_b}{B} \right)^2 / B}. \tag{12}$$

The use of the bootstrap is motivated by at least two main considerations. First of all, there is no need to specify the unknown distribution for correct inference. Moreover, the bootstrap distribution can be expected to be a better approximation to the small-sample distribution than the standard asymptotic approximation.

Given the confidence level $1-\gamma$ (usually $\gamma = .05$), the bootstrap CI can be computed in at least three ways. The bootstrap Wald CI can be constructed by using the bootstrap SE defined in (11) as, for instance, $t \pm z_{1-\gamma/2} SE(T)$, where $z_{1-\gamma/2}$ denotes the $100(1-\gamma/2)$th percentile of the standard normal distribution. The Wald CI is not invariant under monotonic transformations of the sample statistic t, and its values are not confined to the range of what is to be estimated. The bootstrap percentile CI is based on the percentiles of

the bootstrap distribution of t_b, $b = 1, \ldots, B$. The CI is then defined as $(t_{\gamma/2}, t_{1-\gamma/2})$, where $t_{\gamma/2}$ and $t_{1-\gamma/2}$ are the 100(γ/2)th and 100(1 – γ/2)th percentiles of the bootstrap distribution, respectively. Unlike the Wald CI, the percentile CI interval is transformation respecting and range preserving. The bias-corrected and accelerated (BCa) percentile bootstrap CI, proposed by Efron (1987), represents an enhancement of the percentile version. In detail, Efron (1987) improves the accuracy of the percentile CI by using adaptive percentiles for bias corrections and acceleration adjustments in order to correct the standard percentile CI for bias and skewness in the bootstrap distribution. A description of the procedure is provided in the Appendix. The BCa percentile bootstrap CI, is transformation respecting, range preserving, and has a lower coverage error in comparison with percentile CI (see, for example, Efron & Tibshirani, 1993).

3.2. Bootstrap confidence intervals in principal covariates regression

Timmerman et al. (2007) illustrate different strategies for estimating confidence intervals in PCA. They show that the bootstrap approach should be preferred in comparison with the asymptotic approach and, among the various alternatives for estimating CIs by bootstrapping, the BCa percentile method performed best by far. Since the PCOVR solution coincides with the PCA one of a particular matrix, we reasonably expect that BCa percentile CI works well also in the PCOVR domain and hence we have used it in the present paper. As described by Timmerman et al. (2007), bootstrapping can be fruitfully applied in order to estimate CIs of the component loadings. By means of these estimated CIs, we hope to find intervals of values that contain the true population parameter with probability 1 – γ.

The main difficulty lies in the non-identifiability of the PCA solution in terms of sign and axis position because the bootstrap component loadings must be made consistent with the loadings from the observed sample. Timmerman et al. (2007) discuss the non-identifiability of the PCA solution by presenting three different cases related to different interpretations of the component loadings. The first involves limiting the attention to the principal axes, that is, the component loadings corresponding to the eigenvectors associated with the largest eigenvalues of the covariance/correlation data matrix. In this case, the non-uniqueness of the PCA solution is related to the sign indeterminacy. In practice, the bootstrap component loadings must be multiplied by ± 1 so that they optimally resemble those obtained from the observed sample, and the same must be done for the population component loadings. In this case, the interpretation involves the unrotated component loadings. This is rarely done in psychology, where one usually applies rotations of the component loadings in order to simplify their interpretation. For this reason, we do not consider such a strategy in the present paper.

In order to achieve simplicity, the second strategy naturally arises. This involves applying the same rotation method to the sample component loadings as to the bootstrap ones. For instance, if the sample component loadings are varimax rotated, the varimax rotation is also applied to the bootstrap component loadings. For comparative purposes, the sign indeterminacy and the ordering of the components should be checked. The previously described approach is usually referred to as the bootstrap fixed rotation method, in order to emphasize that a fixed rotation method (e.g., the varimax) is employed for all the (bootstrap) samples. In this case, particular emphasis in the interpretation is put on the rotation method selected. This strategy will be referred to as 'fixed criterion'.

Finally, the third approach involves first rotating the original sample component loading matrix to simple structure, and then rotating the bootstrap component loading

matrices in such a way that they resemble the (simple structure rotated) original one as much as possible. The idea is that each bootstrap solution just spans the space of a class of infinitely many rotated loadings matrices, and from these we wish to identify one that is interpreted (as much as possible) similarly to the (simple structure rotated) sample loading matrix. This can be done by means of Procrustes rotation of each bootstrap loading matrix towards the (simple structure rotated) sample loading matrix as a target. Hence, the strategy can be called the bootstrap target rotation method. This case implies that the rotation method as such is no longer relevant for interpretative purposes of the bootstrap solutions. In fact, all attention is paid to the *outcome* of the simple structure rotation procedure for the sample loading matrix only, and this sample solution is taken very seriously, and serves as the target for the other ones. The rationale is that this sample loading matrix supposedly is easy to interpret, and therefore serves as a good reference basis. This strategy will be referred to as 'Procrustes rotation'.

The performances of alternative strategies for computing CIs in PCA have been investigated in depth by Timmerman et al. (2007), who focused on CIs for varimax rotated component loadings. In the study, the authors considered both bootstrap and asymptotic CIs and analysed their quality in terms of coverage. Specifically, the coverage of a CI is appropriate when the probability of a $100(1 - \gamma)\%$ CI not covering the true population parameter θ from above and below is equal to $\gamma/2$. Setting $\gamma = .05$, this means that the true population parameter θ belongs to the CI with a probability equal to .95 (the confidence level). With probability equal to .025, θ is lower (higher) than the lower (higher) bound of the interval.

It is important to note that the non-identifiability of the PCA solution implies that it is not clear what the population parameters are. In fact, the different interpretations of the component loadings recalled above and the related rotation variants also imply a different stance on the population component loadings, because the population loading matrix actually spans the space of a class of infinitely many rotated loadings matrices. Because, in the current approach, we have decided that the (simple structure rotated) sample loading matrix serves as a reference matrix, the population loading matrix must be transformed optimally towards the sample loading matrix in order to identify it, and in order to see to what extent the CIs we set up as estimates of confidence intervals cover comparable population values. This must be done in the same way as for the bootstrap component loadings because the rotation applied to the population component loadings must comply with the rotation applied to the bootstrap component loadings. For a more in-depth discussion, see Kiers (2004).

Timmerman et al. (2007) found that the bootstrap CIs performed better than asymptotic CIs. Furthermore, among the different methods to estimate CIs by means of bootstrapping, the BCa percentile CIs generally performed best.

The results of Timmerman et al. (2007) represent the starting point for estimating CIs in PCOVR. However, such results cannot be straightforwardly applied to PCOVR due to its higher level of complexity. With respect to PCA, the scope of the bootstrap is broadened. In fact, it is reasonable to estimate CIs for several parameter matrices, namely:

- the component loading matrix for the predictor variables on the components ($\mathbf{P_X}$);
- the regression weight matrix for the criterion variables on the components ($\mathbf{P_Y}$);
- the component weight matrix ($\mathbf{W}$);
- the regression weight matrix for estimating the criterion variables from the prediction variables ($\mathbf{WP_Y}$).

We decided to limit our attention to the bootstrap approach and, in particular, to the BCa percentile method due to its useful performance. Moreover, we extended the analysis by considering not only the (orthogonal) varimax rotation, but also the (oblique) quartimin rotation (Carroll, 1953). Such criteria will be considered for both the fixed criterion and the Procrustes rotation strategies.

In order to compute bootstrap CIs, the following ordered steps for managing the indeterminacy of the PCOVR solution should be taken. Let us consider the varimax case.

1. The sample component loading matrix, say $\mathbf{P}_{\mathbf{X}S}$, is rotated to simple structure according to the varimax criterion and the rotation is compensated in the remaining parameter matrices.
2. The bootstrap samples solution, setting R and α as for the sample case, is rotated to be consistent with the rotated sample solution. This can be done by means of two alternative strategies.

   a. In the fixed criterion strategy, the component loading matrix for the bth bootstrap sample ($b = 1, \ldots, B$), say $\mathbf{P}_{\mathbf{X}b}$, is varimax rotated. It is not yet guaranteed that $\mathbf{P}_{\mathbf{X}S}$ and $\mathbf{P}_{\mathbf{X}b}$ are comparable because of the signs and the ordering of the components (in the rows), but this can be fixed easily by adjusting the signs and order of the rows of $\mathbf{P}_{\mathbf{X}b}$ such that they optimally resemble those of $\mathbf{P}_{\mathbf{X}S}$ in terms of the Tucker congruence coefficient (Tucker, 1951). All these transformations are finally compensated in the remaining parameter matrices.
   b. In the Procrustes rotation strategy, $\mathbf{P}_{\mathbf{X}b}$ is rotated by means of an orthogonal rotation matrix $\mathbf{B}$ so as to resemble the sample component loading matrix $\mathbf{P}_{\mathbf{X}}$ as much as possible. Since the rows refer to the different components, the Procrustes rotation is applied to $\mathbf{P}_{\mathbf{X}b'}$ with respect to $\mathbf{P}_{\mathbf{X}'}$. Hence, the function

$$\| \mathbf{P}_{\mathbf{X}b'}\mathbf{B} - \mathbf{P}_{\mathbf{X}'} \|^2 \tag{13}$$

is minimized over $\mathbf{B}$. The minimization problem in (13) is usually referred to as orthogonal Procrustes rotation (Cliff, 1966). Once $\mathbf{B}$ is found, the rotation is compensated in the remaining parameter matrices by postmultiplying $\mathbf{T}$ by $(\mathbf{B}')^{-1}$ provided that $\mathbf{P}_{\mathbf{Y}b}$ is premultiplied by $\mathbf{B}'$.

In the quartimin case, the previously described steps still hold provided that the varimax rotation is replaced by the quartimin one in the fixed criterion strategy and the minimum of (13) is achieved with respect to an oblique rotation matrix $\mathbf{B}$ (oblique Procrustes rotation problem; Jennrich, 2002) in the Procrustes rotation strategy.

In this paper we will compute CIs based on either the varimax or quartimin rotation and either the fixed criterion or Procrustes rotation strategy. This leads to the following four variants:

- varimax rotated sample and bootstrap solutions (VSB);
- quartimin rotated sample and bootstrap solutions (QSB);
- varimax rotated sample and Procrustes rotated bootstrap solutions (VSPB);
- quartimin rotated sample and Procrustes rotated bootstrap solutions (QSPB).

These four methods have been implemented in R (R Core Team, 2020) and are available upon request from the corresponding author. It must be underlined that the above list is not exhaustive. Many other variants could be used, adopting different rotational criteria.

3.3. Example

In order to assess the statistical uncertainty of the PCOVR solution for the Rohwer data reported in Section 2.1, we estimate the BCa percentile CIs for the parameter matrices given in Tables 2–4. Since the sample solution is varimax rotated, we now apply the variants VSB and VSPB. Setting $\gamma = .05$ and $B = 1{,}000$, we get the CIs given in Tables 5–7.

From Table 5, the inspection of the CIs for $\mathbf{P_X}$ highlights that the variability of the estimates is quite large. This especially occurs for component 1 and for the low loadings in absolute sense of components 2 and 3. The variables playing a relevant role in interpreting the components are well captured by the observed sample. In fact, all these estimates are clearly above zero, but for component 1, these values are highly uncertain. The most stable component loadings appear to be that of S on component 2 and that of N on component 3. Such findings hold for both the fixed varimax and the Procrustes varimax strategies, although the bounds of the CIs differ considerably. The analysis of the CIs for $\mathbf{P_Y}$ and $\mathbf{WP_Y}$ reported in Tables 6 and 7 can be carried out in a similar way as for $\mathbf{P_X}$.

Apart from the CIs for $\mathbf{WP_Y}$ (any rotation does not alter $\mathbf{WP_Y}$: $\mathbf{WP_Y = WB^{-1}BP_Y}$), the use of VSB and VSPB leads to different CIs. This occurs because the former CIs consider

Table 5. BCa percentile CIs at the 95% confidence level for $\mathbf{P_X}$

VSB				VSPB		
Component 1	Component 2	Component 3		Component 1	Component 2	Component 3
(.16, .39)	(–.02, .22)	(.94, .98)	N	(.15, .36)	(–.04, .22)	(.94, .99)
(.11, .30)	(.93, .99)	(–.01, .20)	S	(.02, .34)	(.91, 1.00)	(–.02, .21)
(.35, .94)	(–.16, .27)	(.14, .66)	NS	(.46, .96)	(–.10, .36)	(.02, .57)
(.47, .87)	(.24, .68)	(.11, .45)	NA	(.54, .84)	(.30, .60)	(.12, .46)
(.48, .94)	(.10, .56)	(–.02, .59)	SS	(.61, .95)	(.11, .46)	(–.08, .54)

Table 6. BCa percentile CIs at the 95% confidence level for $\mathbf{P_Y}$

VSB				VSPB		
Component 1	Component 2	Component 3		Component 1	Component 2	Component 3
(–.15, .44)	(.01, .45)	(–.18, .47)	SAT	(–.09, .41)	(.00, .44)	(–.19, .48)
(.17, .65)	(–.04, .41)	(.04, .45)	PPVT	(.23,.64)	(–.04, .38)	(.04, .50)
(.05, .44)	(.00, .44)	(–.04, .35)	RPMT	(.05, .43)	(.00, .45)	(–.03, .35)

Table 7. BCa percentile CIs at the 95% confidence level for $\mathbf{WP_y}$

VSB and VSPB			
	SAT	PPVT	RPMT
N	(–.23, .44)	(–.15, .32)	(–.12, .27)
S	(–.08, .38)	(–.22, .19)	(–.10, .37)
NS	(–.36, .13)	(.06, .30)	(–.18, .17)
NA	(–.07, .13)	(.06, .25)	(.02, .18)
SS	(–.18, .10)	(.08, .33)	(–.13, .15)

sampling fluctuations due to the rotation criterion (varimax) and the latter CIs express sampling fluctuations related to the chosen target loading matrix. As such, CIs estimated by VSB are expected to be wider than those estimated by VSPB. This comment also holds for QSB and QSPB, with the difference involving the use of the (oblique) quartimin criterion.

Summing up, since different CIs are obtained when considering the four strategies summarized in Table 5, it appears to be important to assess their statistical behaviour. For this purpose, we carried out a simulation experiment, the results of which are reported in the next section.

4. Simulation experiment

The aim of the simulation experiment was to assess how the statistical behaviour of the bootstrap CIs was affected by the data structure. For this purpose, we explored different situations to see whether good results were obtained and also to discover whether under such conditions differences in the quality of the CIs emerged. These situations seemed reasonably representative of what one might encounter in practice and, therefore, the simulation experiment offers practical recommendations and suggestions on the computation of bootstrap CIs for PCOVR parameter matrices. These recommendations and suggestions concern, among others, the required sample size and, in general, highlight particular data structures leading to CIs of poorer quality.

The simulation experiment was split into two studies. In the first study, we considered the case with more than one criterion variable. Data were generated according to different scenarios with a given number of components, which were expected to have similar relevance and strength. PCOVR solutions and bootstrap CIs were estimated by using the true number of components. In the second study, taking inspiration from Vervloet et al. (2016), a different set-up was considered where the relevance and strength of the components varied and PCOVR solutions and bootstrap CIs were estimated by using fewer components than those used for generating the data. The R (R Core Team, 2020) code used for the simulation experiment is available upon request from the corresponding author.

4.1. Simulation study 1

4.1.1. Set-up

In order to assess the quality of the different strategies for computing bootstrap BCa percentile CIs for the PCOVR solutions, we assumed a population of 10,000 units on which K criterion and J predictor variables were observed. We simulated population data according to the PCOVR model in (2) and (3). Denoting by R the number of components, we had

$$\mathbf{Y}^{\text{POP}} = \mathbf{T}^{\text{POP}}\mathbf{P}_{\mathbf{Y}}^{\text{POP}} + \varepsilon_{\mathbf{Y}}\mathbf{E}_{\mathbf{Y}}^{\text{POP}}, \tag{14}$$

$$\mathbf{X}^{\text{POP}} = \mathbf{T}^{\text{POP}}\mathbf{P}_{\mathbf{X}}^{\text{POP}} + \varepsilon_{\mathbf{X}}\mathbf{E}_{\mathbf{X}}^{\text{POP}}, \tag{15}$$

where $\mathbf{Y}^{\text{POP}}$ (of order $10{,}000 \times K$) and $\mathbf{X}^{\text{POP}}$ (of order $10{,}000 \times J$) were the population data matrices for the criterion and predictor variables, respectively, $\mathbf{T}^{\text{POP}}$ (of order $10{,}000 \times R$) was the population component score matrix, $\mathbf{P}_{\mathbf{Y}}^{\text{POP}}$ (of order $R \times K$) was the population matrix of the regression weights for the criterion variables on the

components, and $\mathbf{P}_{\mathbf{X}}^{\text{POP}}$ (of order $R \times J$) was the population matrix of the component loading for the predictor variables on the components. Furthermore, $\mathbf{E}_{\mathbf{Y}}^{\text{POP}}$ and $\mathbf{E}_{\mathbf{X}}^{\text{POP}}$ were the population error matrices for $\mathbf{Y}^{\text{POP}}$ (of order $10{,}000 \times K$) and $\mathbf{X}^{\text{POP}}$ (of order $10{,}000 \times J$), respectively. Thus, the population data matrices can be split into a structural part (either $\mathbf{T}^{\text{POP}}\mathbf{P}_{\mathbf{Y}}^{\text{POP}}$ or $\mathbf{T}^{\text{POP}}\mathbf{P}_{\mathbf{Y}}^{\text{POP}}$) and an error part (either $\mathbf{E}_{\mathbf{Y}}^{\text{POP}}$ or $\mathbf{E}_{\mathbf{X}}^{\text{POP}}$ The elements of $\mathbf{E}_{\mathbf{Y}}^{\text{POP}}$ and $\mathbf{E}_{\mathbf{X}}^{\text{POP}}$ were randomly generated from the standard normal distribution and multiplied by a scalar (one for each matrix) such that $\| \mathbf{T}^{\text{POP}} \| = \| \mathbf{E}_{\mathbf{Y}}^{\text{POP}} \|$ and $\| \mathbf{T}^{\text{POP}}\mathbf{P}_{\mathbf{X}}^{\text{POP}} \| = \| \mathbf{E}_{\mathbf{X}}^{\text{POP}} \|$. The scalars $\epsilon_{\mathbf{Y}}$ and $\epsilon_{\mathbf{X}}$ tuned the amount of noise in the population data matrices. We considered two levels of noise for $\mathbf{Y}^{\text{POP}}$ ($\epsilon_{\mathbf{Y}} = 0.1$, low noise $\mathbf{Y}$, and $\epsilon_{\mathbf{Y}} = 0.3$, high noise $\mathbf{Y}$) and for $\mathbf{X}^{\text{POP}}$ ($\epsilon_{\mathbf{X}} = 0.1$, low noise $\mathbf{X}$, and $\epsilon_{\mathbf{X}} = 0.3$, high noise $\mathbf{X}$) in order to assess whether a larger amount of noise deteriorated the statistical behaviour of the CIs. Concerning the structural part, we operated as follows. We set $K = 6$ and we considered two levels for the ratio between criterion and prediction variables: in the small ratio case, $J/K = 2$, hence $J = 12$; in the large ratio case, $J/K = 3$, hence $J = 18$. This choice was made to assess whether differences in the numbers of prediction and criterion variables affected the quality of CIs. The number of components was set to $R = 2$ (two-component case) or $R = 3$ (three-component case) in order to inspect whether the selection of the number of components had an impact on the CIs obtained. The elements of $\mathbf{T}^{\text{POP}}$ were randomly drawn from the standard normal distribution. Two levels of simplicity for $\mathbf{P}_{\mathbf{Y}}^{\text{POP}}$ and $\mathbf{P}_{\mathbf{X}}^{\text{POP}}$ were assumed. In the simple case the matrices were constructed in such a way that every (criterion or prediction) variable was related to only one component and every row of $\mathbf{P}_{\mathbf{Y}}^{\text{POP}}$ and $\mathbf{P}_{\mathbf{X}}^{\text{POP}}$ contained 50% (if $R = 2$) or 67% (if $R = 3$) zero elements. For instance, when $K = 6$, the structure of $\mathbf{P}_{\mathbf{Y}}^{\text{POP}}$ with, respectively, $R = 2$ and $R = 3$, was

$$\mathbf{P}_{\mathbf{Y}}^{\text{POP}} = \begin{bmatrix} x & x & x & 0 & 0 & 0 \\ 0 & 0 & 0 & x & x & x \end{bmatrix}, \tag{16}$$

$$\mathbf{P}_{\mathbf{Y}}^{\text{POP}} = \begin{bmatrix} x & x & 0 & 0 & 0 & 0 \\ 0 & 0 & x & x & 0 & 0 \\ 0 & 0 & 0 & 0 & x & x \end{bmatrix}, \tag{17}$$

where the symbols x denote randomly generated values from the uniform distribution on $[0.5, 1]$. Such a case for $\mathbf{P}_{\mathbf{Y}}^{\text{POP}}$ and $\mathbf{P}_{\mathbf{X}}^{\text{POP}}$ was labelled *simple structure*. In the complex case, variables were related to at least one component (exactly two when $R = 3$) and every row had 33% zero elements. Thus, (16) and (17) were replaced by

$$\mathbf{P}_{\mathbf{Y}}^{\text{POP}} = \begin{bmatrix} x & x & x & x & 0 & 0 \\ 0 & 0 & x & x & x & x \end{bmatrix}, \tag{18}$$

$$\mathbf{P}_{\mathbf{Y}}^{\text{POP}} = \begin{bmatrix} x & x & x & x & 0 & 0 \\ 0 & 0 & x & x & x & x \\ x & x & 0 & 0 & x & x \end{bmatrix}, \tag{19}$$

The complex structure case for $\mathbf{P}_{\mathbf{Y}}^{\text{POP}}$ was referred to as *complex structure*. Hence, by considering simple and complex structures for the component loadings and the

regression weights we were interested in assessing whether and how the quality of the estimated CIs was affected by situations where some criterion and prediction variables are influenced by more than one component. In fact, we expected this to lead to less accurate estimated CIs.

Once the population data were available, we randomly generated the sample data by varying the sample size: $N = 50$ (small), $N = 100$ (medium) and $N = 500$ (large). In this way, we aimed to study whether a limited number of units would prevent us from obtaining good results in the CIs. For each sample size, 1,000 sample data sets were randomly generated from the population data. The design was fully crossed, leading to 2 (low noise $\mathbf{Y}$ and high noise $\mathbf{Y}$) × 2 (low noise $\mathbf{X}$ and high noise $\mathbf{X}$) × 2 (small ratio and large ratio) × 2 (two-component and three-component) × 2 (simple structure and complex structure) × 3 (small size, medium size, and large size) × 1,000 (sample replications) = 96,000 sample data sets.

4.1.2. Quality criteria

The aim of the simulation study was to assess the statistical behaviour of the estimated CIs in terms of coverage of the proportion of observations outside them.

For coverage, we checked whether the probabilities of the $100(1 - \gamma)\%$ CIs covering the true population parameter θ were equal to $1 - \gamma$ or not. This was assessed for the estimated CIs of $\mathbf{P_X}$, $\mathbf{P_Y}$, $\mathbf{W}$, and $\mathbf{WP_Y}$. The true population parameter matrices were those estimated by applying PCOVR to the population data by using the proper number of components and selecting α according to Vervloet et al. (2013). Note that the overall average value of α was .70. Following Kiers (2004), the population solutions were transformed optimally towards the sample ones in order to interpret the CIs we set up as estimates of confidence intervals. This was done by either the fixed criterion (varimax or quartimin) or the Procrustes rotation strategy applied to $\mathbf{P_X}$ consistently with the four examined strategies for computing CIs (VSB, QSB, VSPB, QSPB) so that the same type of rotations was applied to the population and bootstrap solutions. Note that, as pointed out by Timmerman et al. (2007), the coverage was assessed in terms of the estimated parameters using the population data instead of the parameters used to generate the population data, because the bootstrap is not adopted for compensating a model misspecification and the related systematic bias.

For each sample data set, we performed the PCOVR analysis by using the same values of R and α found for the corresponding population data and the solution was rotated to simple structure by applying either the varimax or the quartimin procedure to $\mathbf{P_X}$ consistently with the choice made for the population solution. Then we estimated the BCa percentile bootstrap CIs by means of the above-mentioned four strategies. This was done by considering $B = 1{,}000$ bootstrap samples. Such a value should be chosen in order to reduce the simulation error to an acceptable level. In doing so, we followed the choice made by Timmerman et al. (2007). However, as a reviewer observed, Dufour and Kiviet (1998) showed that the number of bootstrap samples should be such that $(B + 1)\gamma$ is an integer. However, since $B = 1{,}000$, differences are negligible. For instance, in the QSPB case, $\mathbf{P_X}$ of the sample solution was quartimin rotated and the component loadings for the predictor variables of the population and bootstrap solutions were obliquely Procrustes rotated towards the sample ones. Obviously, in each solution, the rotation was compensated in the remaining parameter matrices to obtain an equally fitting solution.

Once the solutions were rotated, we could properly evaluate the coverage because the rotated population solution was fully comparable with the bootstrap CIs. Exactly as in

Timmerman et al. (2007), the coverage was calculated as the percentage of times for which the true rotated parameter was inside the $100(1-\gamma)\%$ estimated CIs. In practice, setting $\gamma = .05$, we calculated the percentages of times for which the true parameter was lower than the 2.5% bound or higher than the 97.5% bound of the CI. Denoting these exceeding percentages by 2.5% EP and 97.5% EP, respectively, the coverage was determined as

$$\text{Coverage} = 100\% - 2.5\%\ \text{EP} - 97.5\%\ \text{EP}. \tag{20}$$

Note that this percentage was computed for every condition, that is, for every combination of all the levels of all the design variables, by using the 1,000 replications. Coverage close to 95% indicates good statistical behaviour.

4.1.3. Results

The results on the coverage of the BCa percentile CIs for $\mathbf{P_X}$, $\mathbf{P_Y}$, $\mathbf{W}$, and $\mathbf{WP_Y}$ are reported by distinguishing with respect to the levels of the design variables and to the four strategies. Moreover, concerning $\mathbf{P_X}$ and $\mathbf{P_Y}$, we also distinguished the coverage with respect to the 'high' and 'low' values (i.e., those respectively denoted by x and 0 in (17)–(20).) The mean 95% coverage levels computed with respect to all 96,000 data sets and distinguished by parameter matrix are displayed in Figure 1.

First of all, we can see that a more or less pronounced undercoverage was registered on average. Such a comment especially holds for the CIs estimated by QSPB. For the

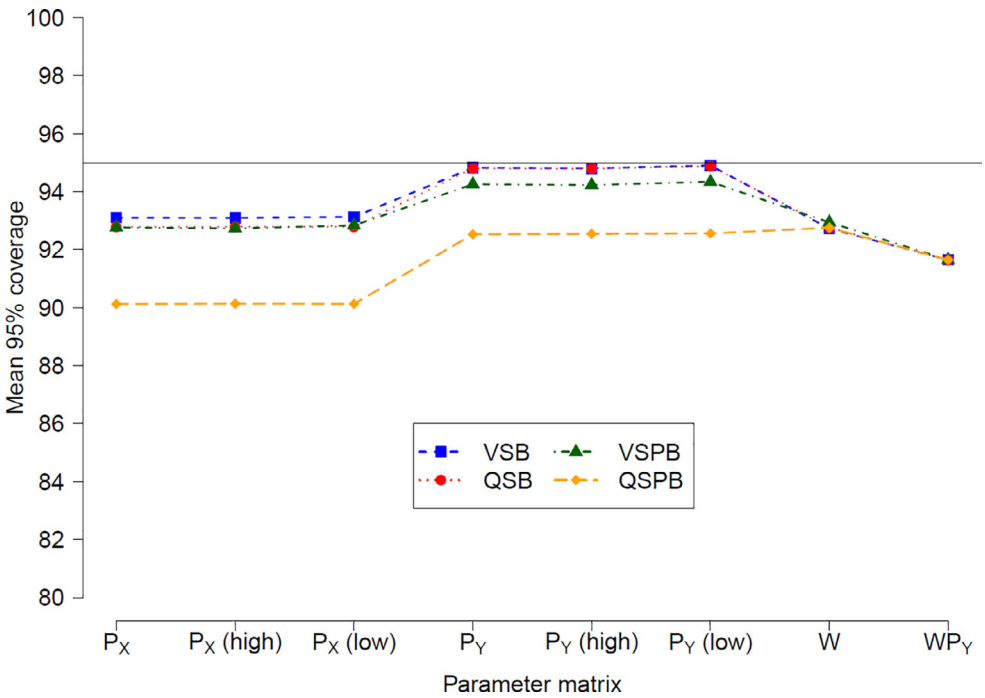

Figure 1. Mean 95% coverage levels of the CIs for the parameter matrices (simulation study 1). The horizontal line indicates the desired coverage level.

remaining three strategies, mean coverage levels were very similar. The estimated CIs for $\mathbf{P_Y}$ were, on average, the best in terms of coverage. For VSB and QSB, these were almost equal to the desired level (94.83 and 94.81%, respectively). With respect to $\mathbf{P_X}$, VSB (93.10% on average), QSB (92.78%) and VSPB (92.77%) had essentially the same mean coverage. Virtually no differences emerged for the mean coverage levels of the CIs estimated for $\mathbf{W}$, ranging from 92.73% for VSB to 92.94% for VSPB. Finally, the estimated CIs for $\mathbf{WP_Y}$ (the same for all strategies) were characterized by a quite large undercoverage (91.65% on average).

A deeper insight into the assessment of the statistical behaviour of the estimated CIs was obtained by analysing the mean coverage levels distinguished by the levels of the design variables. In particular, we checked whether, for each scenario, the mean coverage became closer to 95% for increasing values of N. There were in total $2^5 = 32$ scenarios corresponding to the combinations of the levels of noise Y, noise X, ratio, component, and structure. Figure 2 shows the results for $\mathbf{P_X}$, while outcomes for the remaining parameter matrices are reported in Figures S1–S7 in the online supplementary material. Note that the mean coverage levels for $\mathbf{P_X}$ and $\mathbf{P_Y}$ are also distinguished by high and low loadings. By inspecting Figure 2, we can see an appropriate statistical behaviour, that is, the mean coverage tends to 95% when the sample size N increases. Except for QSPB, for all the remaining three variants, this is often visible when $N = 500$ and, sometimes, even when $N = 100$. Specifically, in scenarios characterized by large ratio, three-component and complex structure, the mean coverage levels for VSB, QSB and VSPB were slightly lower than 95% for samples of size $N = 100$. Conversely, in such scenarios, CIs computed by QSPB presented undercoverage, even more severe when $N = 500$. In general, the statistical behaviour of CIs constructed according to QSPB was not satisfactory in complex structure scenarios. This may be because the simple structure rotated sample loadings in the complex condition were rather unstable, because in these cases different rotations may lead to fairly similar quartimin values. Since these matrices were taken as references to which population and bootstrap loading matrices were rotated, it is conceivable that this instability led to weaker quality confidence intervals for QSPB.

The previous comments also hold for the mean coverage levels of CIs referring to the high and low loadings in $\mathbf{P_X}$ (Figures S1 and S2, respectively). Some minor differences were visible in the mean coverage levels, but, as far as we saw, no general trends emerged. The mean coverage levels for $\mathbf{P_Y}$ are displayed in Figures S3–S5, where the last two figures concern the high and low regression weights. All in all, as we already observed, for each variant, the mean coverage levels of the CIs for $\mathbf{P_Y}$ were by far the best in comparison with the other parameter matrices. Consistently with the results for $\mathbf{P_X}$, appropriate statistical behaviours were observed for all the variants except for QSPB. Specifically, as for $\mathbf{Px}$, a pattern of undercoverage occurred. The mean coverage levels for $\mathbf{W}$ are reported in Figures S6. The differences in performance of the four criteria were less noticeable with respect to $\mathbf{P_X}$. However, similar statistical behaviours to $\mathbf{P_X}$ emerged. Finally, the mean coverage levels for $\mathbf{WP_Y}$ are plotted in Figures S7. As already observed, the same CIs were found for all the criteria since $\mathbf{WP_Y}$ is not affected by rotations. In general, we observed undercoverage for the CIs independently of the design variables.

4.2. Simulation study 2

A smaller simulation study was also implemented with the aim of assessing the quality of the estimated CIs when the components underlying the data had different relevance and strength. In doing so, we were interested in studying what happened when fewer

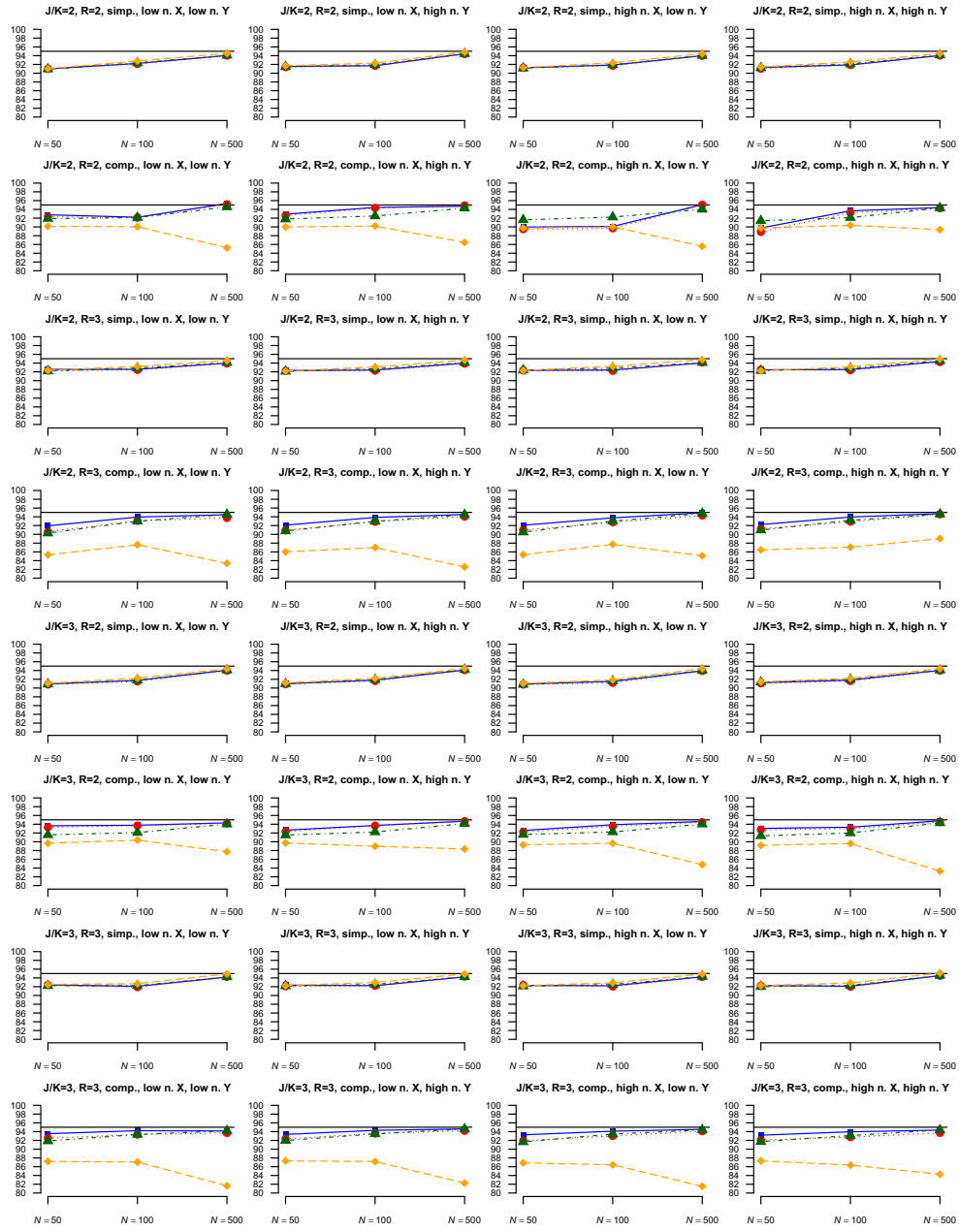

Figure 2. Mean 95% coverage levels of the CIs for $\mathbf{P_X}$ (simulation study 1) for increasing sample sizes distinguished by all combinations of levels of the design variables. Levels for VSB, QSB, VSPB and QSPB are denoted by, respectively, blue filled squares, red filled circles, green filled triangles and orange filled diamonds. The horizontal line indicates the desired coverage level.

components than those used to generate the data were extracted. In fact, in the previous simulation study, we used the true number of components for fitting PCOVR to the data. In the current simulation study, we checked whether and possibly how an incorrect choice of R affected the quality of the estimated CIs.

4.2.1. Set-up

The set-up of this simulation study took inspiration from that of simulation study 1 in Vervloet et al. (2016). We assumed a population of 10,000 units, on which $K = 1$ criterion and $J = 24$ or $J = 48$ predictor variables were observed. Following the usual notation, we simulated population data as:

$$\mathbf{y}^{\text{POP}} = \mathbf{T}^{\text{POP}}\mathbf{p}_{\mathbf{Y}}^{\text{POP}} + \varepsilon_{\mathbf{Y}}\mathbf{e}_{\mathbf{Y}}^{\text{POP}}, \tag{21}$$

$$\mathbf{X}^{\text{POP}} = \mathbf{T}^{\text{POP}}\mathbf{P}_{\mathbf{X}}^{\text{POP}} + \varepsilon_{\mathbf{X}}\mathbf{E}_{\mathbf{X}}^{\text{POP}}, \tag{22}$$

where $\mathbf{y}^{\text{POP}}$ and $\mathbf{e}_{\mathbf{Y}}{}^{\text{POP}}$ are now vectors of length 10,000 and $\mathbf{p}_{\mathbf{Y}}^{\text{POP}}$ is a vector of length 4, that is, $R = 4$ components were used to generate data. The vector $\mathbf{p}_{\mathbf{Y}}^{\text{POP}}$ was constructed according to three cases. Namely, $R/2$ (= 2) elements (the second and fourth) were always set to 0 and the remaining $R/2$ values (the first and the third, respectively) were equal to 0.71 and 0.71 (case labelled 'no difference'), 0.60 and 0.80 ('small difference') and 0.44 and 0.90 ('large difference'). As in Vervloet et al. (2016), this structure of $\mathbf{p}_{\mathbf{Y}}^{\text{POP}}$ allowed for varying the relevance of the components. In the current study, we were interested in assessing whether and how this affected the quality of the CIs obtained. The matrix $\mathbf{P}_{\mathbf{X}}^{\text{POP}}$ was such that, for every variable, one loading was equal to 1 and the others were 0. For every component, the numbers of loadings equal to 1 differed in order to consider various levels of strength of the components. Such numbers were chosen in the same way as in Vervloet et al. (2016) and reported in Table 8, leading to six cases labelled '4 vs. 46%', '8 vs. 42%', '13 vs. 38%', '17 vs. 33%', '21 vs. 29%' and '25 vs. 25%'. For instance, '4 vs. 46%' means that, when $J = 24$, two components had (1/24)×100% ≈ 4% of loadings equal to 1 and the other two components (11/24)×100% ≈ 46% loadings equal to 1. When $J = 48$, such percentages corresponded to about (2/48)×100% and (22/48)×100%, respectively. Such a design variable allowed us to see whether under such conditions differences between the quality of the CIs could be discerned. The elements of $\mathbf{T}^{\text{POP}}$ were randomly drawn from the standard normal distribution and the same was done for the elements of $\mathbf{e}_{\mathbf{Y}}{}^{\text{POP}}$ and $\mathbf{E}_{\mathbf{X}}^{\text{POP}}$. Next, these were normalized such that $\| \mathbf{e}_{\mathbf{Y}}{}^{\text{POP}} \| = \| \mathbf{T}^{\text{POP}}\mathbf{p}_{\mathbf{Y}}^{\text{POP}} \|$, and $\| \mathbf{E}_{\mathbf{X}}^{\text{POP}} \| = \|\mathbf{T}^{\text{POP}}\mathbf{P}_{\mathbf{X}}^{\text{POP}} \|$, respectively. In addition, to tune the amount of noise in the population data matrices, $\mathbf{e}_{\mathbf{Y}}^{\text{POP}}$ and $\mathbf{E}_{\mathbf{X}}^{\text{POP}}$ were also multiplied by the scalars $\epsilon_{\mathbf{Y}}$ and $\epsilon_{\mathbf{X}}$. These two scalars took three levels (0.05, 0.25 and 0.45).

Once the population data were available, we randomly generated sample data with $N = 200$. For each condition, 25 sample data sets were randomly generated from the population data. The design was fully crossed, leading to 3 (levels of noise for $\mathbf{Y}$, $\epsilon_{\mathbf{Y}}$) × 3 (levels of noise for $\mathbf{X}$, $\epsilon_{\mathbf{X}}$) × 2 (numbers of predictor variables, J) × 3 (levels of relevance: no difference, small difference, large difference) × 6 (levels of strength: '4 vs. 46%', '8 vs. 42%', '13 vs. 38%', '17 vs. 33%', '21 vs. 29%', '25 vs. 25%') × 25 (sample replications) = 8,100 sample data sets.

Table 8. Number of component loadings equal to 1 per component (simulation study 2)

	Component 1	Component 2	Component 3	Component 4
$J = 24$				
4 vs. 46%	1	1	11	11
8 vs. 42%	2	2	10	10
13 vs. 38%	3	3	9	9
17 vs. 33%	4	4	8	8
21 vs. 29%	5	5	7	7
25 vs. 25%	6	6	6	6
$J = 48$				
4 vs. 46%	2	2	22	22
8 vs. 42%	4	4	20	20
13 vs. 38%	6	6	18	18
17 vs. 33%	8	8	16	16
21 vs. 29%	10	10	14	14
25 vs. 25%	12	12	12	12

4.2.2. Results

The PCOVR analysis (including the selection of α) and the assessment of the coverage of the CIs of $\mathbf{P_X}$, $\mathbf{p_Y}$, $\mathbf{W}$, and $\mathbf{Wp_Y}$ was done as described in Section 4.1.2. The most relevant difference is that the true population parameter matrices and the sample ones were the ones estimated by applying PCOVR to the population data by extracting not only the proper number of components ($R = 4$), but also a lower number ($R = 3$) to assess whether this affected the coverage.

The mean 95% coverage levels computed with respect to all the 8,100 data sets and distinguished by parameter matrix are displayed in Figure 3. The results were consistent with those registered for the previous simulation study. Namely, once again, we observed undercoverage on average with respect to the CIs for all the parameter matrices except for $\mathbf{p_Y}$, for which the mean coverage was around the desired level. Although some differences occurred, this comment largely holds for all the estimation strategies. An interesting finding was that for misspecified models (i.e., fewer components extracted than the true number) undercoverage was more severe on average. This was mainly observed for $\mathbf{W}$ and $\mathbf{Wp_Y}$ and, to a limited extent, for $\mathbf{P_x}$ and $\mathbf{p_Y}$. Given the impact of an incorrect number of components on the coverage, we further inspected this situation. For this purpose, we considered a subset of $3 \times 3 \times 2 \times 3 \times 6 = 324$ randomly generated data sets corresponding to the first replication of every scenario and studied the width of the CIs for $\mathbf{p_Y}$. Note that the components were ordered with respect to the CI widths for the elements of $\mathbf{p_Y}$, in increasing order. Figure 4 contains the boxplots of the CI widths for $\mathbf{p_Y}$ distinguished by the number of extracted components R and the criterion for estimating CIs.

We discovered that extracting a smaller number of components led to rather unstable solutions and, therefore, the CIs were relatively wide when $R = 3$ compared to when $R = 4$. In particular, one component (C3) appeared to be extremely unstable. This phenomenon occurred for all the criteria. Now one might think that such wider intervals would not lead to undercoverage, but rather to overcoverage (as wider intervals cover more values). However, the instability of the parameters can be expected to also make the estimates of the intervals less stable, and could therefore lead to more generally less

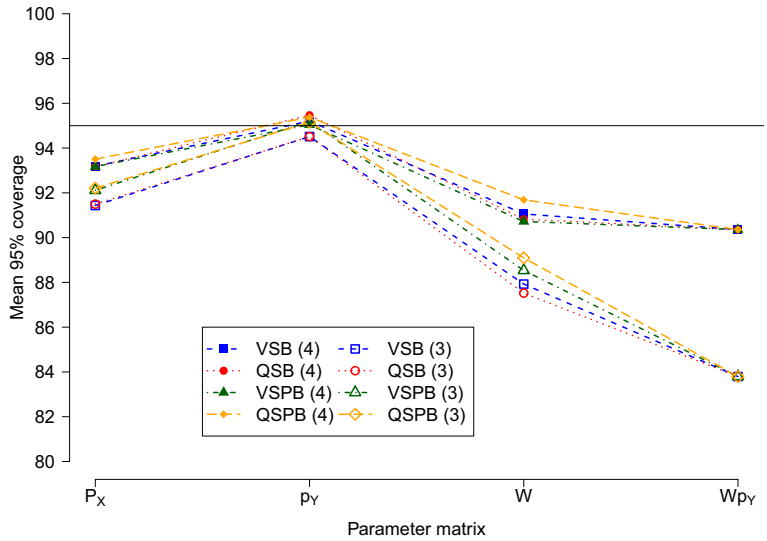

Figure 3. Mean 95% coverage levels of the CIs for the parameter matrices (simulation study 2). The horizontal line indicates the desired coverage level.

predictable and desired behaviour. Why this then led to undercoverage and not to overcoverage is not fully clear to us, but obviously, given that the nominal coverage should be 95%, it cannot easily get higher, whereas it can easily get lower. Furthermore, as expected, the use of fixed criterion strategies led to CIs wider than the corresponding Procrustes rotation strategies (i.e., VSB vs. VSPB and QSB vs. QSPB). This was observed for cases with $R = 3$ and $R = 4$.

To gather more information, we investigated whether and how the coverage differed with respect to the levels of the design variables. From the previous simulation study, we observed a slight impact on the estimated CIs by the number of (predictor) variables and the levels of noise. For this reason, we checked how such design variables affected the coverage. The mean coverage levels for each parameter matrix are reported in Figures S8–S11. All in all, the mean coverage levels remained stable, even if a few minor differences were found especially for $\mathbf{W}$ and $\mathbf{W}\mathbf{p}_\mathbf{Y}$. Based on this result, we then studied whether coverage depended on the different relevance and strength of the components.

Figure 5 contains the mean coverage levels for all the parameter matrices distinguished by all combinations of levels of the design variables relevance and strength. First, we noticed a more pronounced undercoverage on average when $R = 3$ components were extracted in comparison with the $R = 4$ case. This occurred for all the parameter matrices except $\mathbf{p}_\mathbf{Y}$. As far as we could see, the relevance of the components had a negligible impact on the mean coverage levels. However, they appeared to be related to the strength. This holds especially for $\mathbf{W}$ and $\mathbf{W}\mathbf{p}_\mathbf{Y}$. In particular, the mean differences in coverage between the $R = 3$ and $R = 4$ cases were larger when the components had almost the same strength ('17 vs. 33%', '21 vs. 29%', '25 vs. 25%'). If so, the model misspecification remarkably increased the risk of estimating CIs with undercoverage. Differences were smaller for the remaining levels of strength, in particular for '4 vs. 46%'. In this case, the mean coverage levels were essentially the same, apart from $\mathbf{W}\mathbf{p}_\mathbf{Y}$ in the no-difference case. We also observed that, for the '4 vs. 46%' scenario, the worst mean coverage (severe

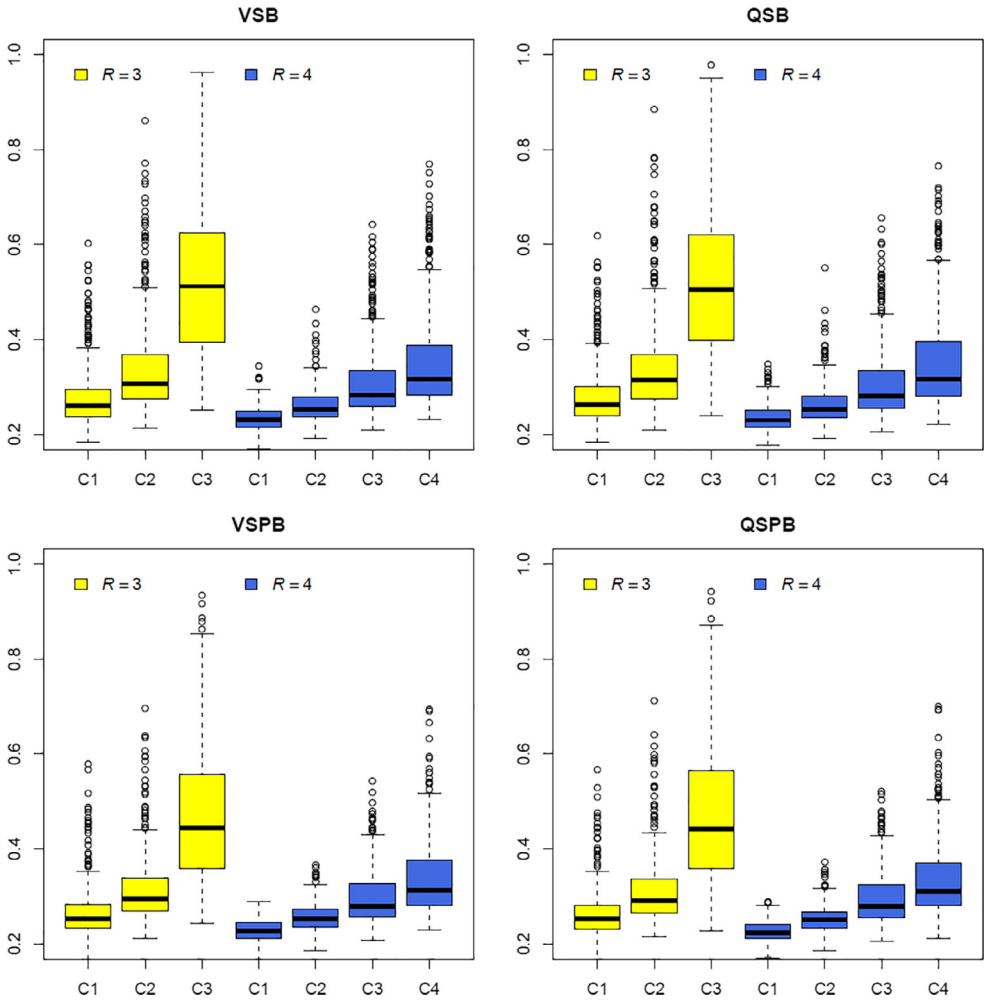

Figure 4. Boxplots of the CI widths for $\mathbf{p_Y}$ (subset of 324 randomly generated data sets in simulation study 2) distinguished by the number of extracted components R.

undercoverage) of the CIs for $\mathbf{W}$ and $\mathbf{Wp_Y}$ was registered. Hence, such CIs should be taken with a pinch of salt if underlying components are expected to have rather different strength. Note that the previous comments hold for all four criteria for estimating CIs.

5. Discussion

This paper has discussed some variants for computing CIs for the parameter matrices of the PCOVR solution. In the PCA framework, this point has been investigated and there is a certain consensus towards the use of bootstrap CIs for making inference on the PCA solution. Among the various alternatives for computing CIs, those based on the bootstrap BCa percentile method appear to be the most useful choice (see Timmerman et al., 2007). In the PCOVR framework, as far as we know, no studies

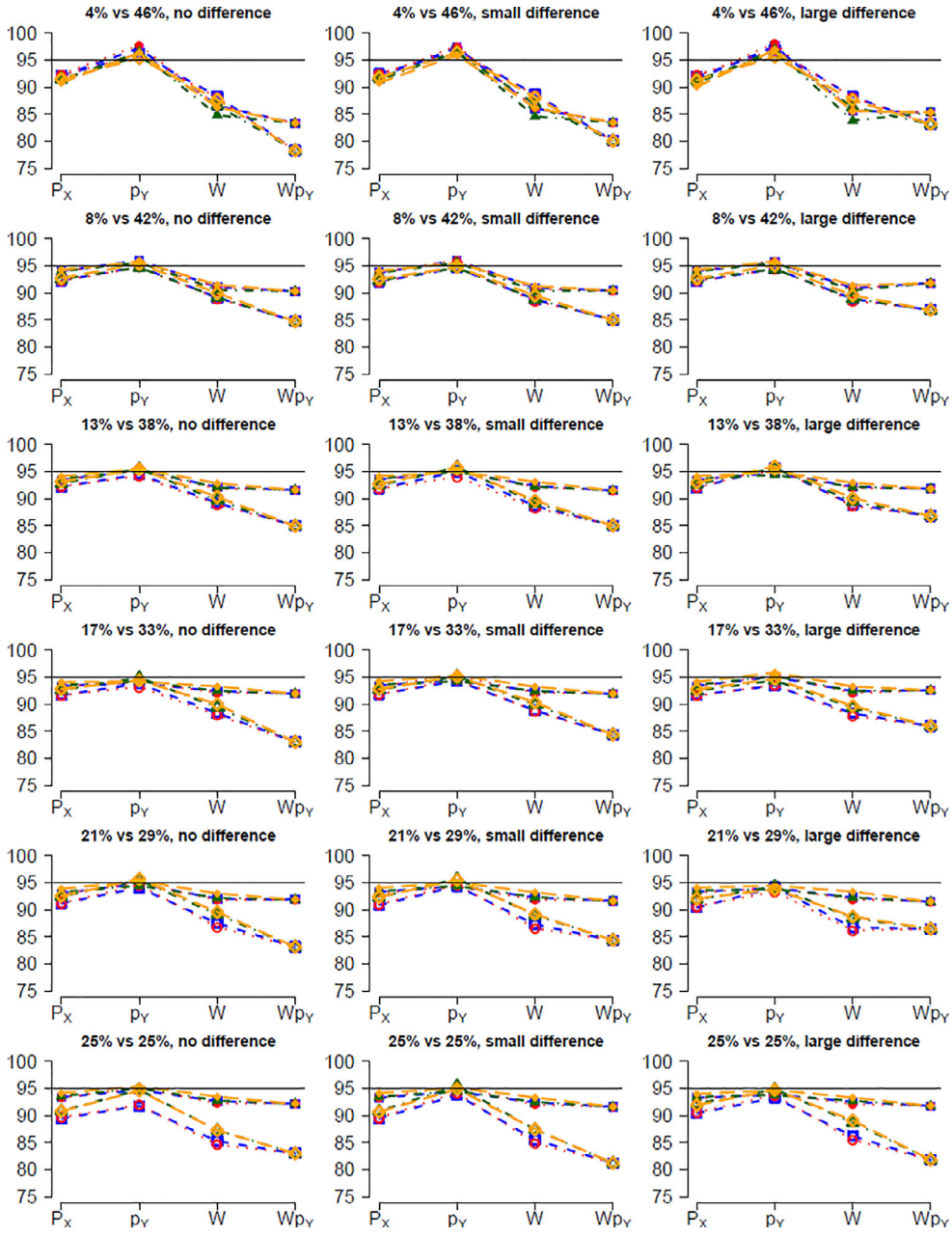

Figure 5. Mean 95% coverage levels of the CIs for all the parameter matrices (simulation study 2) distinguished by levels of relevance and strength of the components. Levels for VSB, QSB, VSPB and QSPB are denoted by, respectively, blue filled squares, red filled circles, green filled triangles and orange filled diamonds when $R = 4$ and by the corresponding empty symbols when $R = 3$. The horizontal line indicates the desired coverage level.

have yet been carried out. This stimulated us to study the performance of the BCa percentile bootstrap method for computing CIs for the parameter matrices of a PCOVR solution, namely, $\mathbf{P_X}$, $\mathbf{P_Y}$, $\mathbf{W}$, and the combined matrix $\mathbf{WP_Y}$. To handle the non-uniqueness of the PCOVR solutions, we considered four strategies involving varimax and quartimin rotations of $\mathbf{P_X}$ and distinguishing the use of a fixed criterion and the Procrustes rotation.

We conducted a simulation experiment to study whether the coverage of the CIs obtained by using the four above-mentioned strategies tended to expected levels as the sample size increased. In doing so, we did not make a comparative assessment among the variants because the variants determine CIs expressing different types of uncertainty and, hence, are not comparable. The simulation experiment offered recommendations on the computation of CIs for the PCOVR parameters when doing empirical research. In fact, some differences in the statistical behaviour of the CIs emerged with respect to the variant adopted, the parameter matrix, and the characteristics of the data. In some cases the CIs obtained appeared to be reasonably good estimates, while in others the quality degraded. In detail, the simulation experiment was split into two parts. In the first, we considered the case with more than one criterion variable and studied the statistical behaviour of the estimated CIs in different scenarios. We found that, in general, a relatively small number of units (about 100) allows for building CIs that are reasonably good in terms of coverage. When the underlying structure of the components is complex (i.e., with predictor and criterion variables related to more than one component), a larger number of units should be used for computing CIs. This seems to hold especially for those variants based on the quartimin rotation. In particular, in such complex scenarios, the statistical behaviour of CIs constructed according to QSPB was quite poor. This may be because the simple structure rotated sample loadings in the complex condition were rather unstable, because in these cases different rotations may lead to fairly similar quartimin values. Since these matrices were taken as references to which population and bootstrap loading matrices were rotated, it is conceivable that this instability led to weaker-quality CIs for QSPB.

Apart from the adopted variant, some differences in the quality of the CIs for the parameter matrices emerged. In particular, we can state that the CIs for $\mathbf{P_X}$ and, especially, $\mathbf{P_Y}$ are more accurate than those for $\mathbf{W}$ and $\mathbf{WP_Y}$. The lowest quality of the CIs for $\mathbf{WP_Y}$ can possibly be explained by the fact that such CIs are affected by a double source of uncertainty, that is, those concerning the CIs for $\mathbf{W}$ and $\mathbf{P_Y}$. The poorer quality of the CIs for $\mathbf{W}$ in comparison with those for $\mathbf{P_X}$ and $\mathbf{P_Y}$ is less easy to understand. Future research will be needed to understand this result in detail.

The previous results were observed when PCOVR was applied by setting the true number of components and avoiding considering components having different relevance and strength. This point was addressed in the second part of the simulation study where data sets involving one criterion variable were generated. We discovered that, to some extent, the relevance and strength of the components affected the coverage of the CIs, in particular, for $\mathbf{W}$ and $\mathbf{WP_Y}$. A much more relevant role is played by the choice of the number of components. In fact, the quality of the CIs deteriorated when fewer components are used than those used to generate the data. This suggests that, if the number of components used is too low, one should be careful in interpreting CIs, because they may actually have coverage levels considerably lower than their nominal value.

Acknowledgements

We thank the three anonymous reviewers whose comments and suggestions greatly improved the quality of the paper.

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Paolo Giordani (Writing – original draft; Writing – review & editing) Henk A.L. Kiers (Writing – original draft; Writing – review & editing).

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, *22*, 327–351. https://doi.org/10.1214/aoms/1177729580

Anderson, T. W. (1958). *Introduction to multivariate statistical analysis*. New York, NY: Wiley.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: Wiley.

Carroll, J. B. (1953). An analytic solution for approximating simple structure in factor analysis. *Psychometrika*, *18*, 23–38. https://doi.org/10.1007/BF02289025

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika*, *31*, 33–42. https://doi.org/10.1007/BF02289455

De Jong, S., & Kiers, H. A. L. (1992). Principal covariates regression: part I. *Theory. Chemometrics and Intelligent Laboratory Systems*, *14*, 155–164. https://doi.org/10.1016/0169-7439(92)80100-I

Dufour, J.-M., & Kiviet, J. F. (1998). Exact inference methods for first-order autoregressive distributed lag models. *Econometrica*, *66*, 79–104. https://doi.org/10.2307/2998541

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Efron, E. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*, 171–185. https://doi.org/10.2307/2289144

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer-Verlag.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67. https://doi.org/10.1080/00401706.1970.10488634

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, *5*, 248–264. https://doi.org/10.1016/0047-259X(75)90042-1

Jennrich, R. I. (2002). A simple general method for oblique rotation. *Psychometrika*, *67*, 7–20. https://doi.org/10.1007/BF02294706

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, *31*, 300–303. https://doi.org/10.2307/2348005

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23* (3), 187–200. https://doi.org/10.1007/BF02289233

Kiers, H. A. L. (2004). Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics*, *18*, 22–36. https://doi.org/10.1002/cem.841

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

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence intervals for factor loadings. *Multivariate Behavioral Research*, *26*, 421–434. https://doi.org/10.1207/s15327906mbr2603_3

Ogasawara, H. (1999). Standard errors for Procrustes solutions. *Japanese Psychological Research*, *41*, 121–130. https://doi.org/10.1111/1468-5884.00111

Ogasawara, H. (2000). Standard errors of the principal component loadings for unstandardized and standardized variables. *British Journal of Mathematical and Statistical Psychology*, *53*, 155–174. https://doi.org/10.1348/000711000159277

Ogasawara, H. (2002). Concise formulas for the standard errors of component loading estimates. *Psychometrika*, *67*, 289–297. https://doi.org/10.1007/BF02294847

R Core Team (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tikhonov, A. N. (1943). *On the stability of inverse problems*. Doklady Akademii Nauk SSSR, 39, 195–198.

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Belmont, CA: Brooks/Cole.

Timmerman, M. E., Kiers, H. A. L., & Smilde, A. K. (2007). Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology*, *60*, 295–314. https://doi.org/10.1348/000711006X109636

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Princeton, NJ: Educational Testing Service.

Vervloet, M. (2017). *Unraveling and unlocking the assets of principal covariates regression*, (PhD thesis). Leuven, Belgium: KU Leuven.

Vervloet, M., Kiers, H. A. L., Van den Noortgate, W., & Ceulemans, E. (2015). PCovR: An R package for principal covariates regression. *Journal of Statistical Software*, *65*, 1–14. https://doi.org/10.18637/jss.v065.i08

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. https://doi.org/10.1016/j.chemolab.2013.02.005

Vervloet, M., Van Deun, K., Van den Noortgate, W., & Ceulemans, E. (2016). Model selection in principal covariates regression. *Chemometrics and Intelligent Laboratory Systems*, *151*, 26–33. https://doi.org/10.1016/j.chemolab.2015.12.004

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25. https://doi.org/10.2307/1912526

Received 18 September 2019; revised version received 23 October 2020

Supporting Information

The following supporting information may be found in the online edition of the article:

Figure S1. Mean 95% coverage levels of the CIs for $\mathbf{P_X}$ ('high' values, Simulation study n. 1) for increasing sample sizes distinguished by all combinations of levels of the design variables.
Figure S2. Mean 95% coverage levels of the CIs for $\mathbf{P_X}$ ('low' values, Simulation study n. 1) for increasing sample sizes distinguished by all combinations of levels of the design variables.
Figure S3. Mean 95% coverage levels of the CIs for $\mathbf{P_Y}$ (Simulation study n. 1) for increasing sample sizes distinguished by all combinations of levels of the design variables.
Figure S4. Mean 95% coverage levels of the CIs for $\mathbf{P_Y}$ ('high' values, Simulation study n. 1) for increasing sample sizes distinguished by all combinations of levels of the design variables.
Figure S5. Mean 95% coverage levels of the CIs for $\mathbf{P_Y}$ ('low' values, Simulation study n. 1) for increasing sample sizes distinguished by all combinations of levels of the design variables.
Figure S6. Mean 95% coverage levels of the CIs for $\mathbf{W}$ (Simulation study n. 1) for increasing sample sizes distinguished by all combinations of levels of the design variables.
Figure S7. Mean 95% coverage levels of the CIs for $\mathbf{WP_Y}$ (Simulation study n. 1) for increasing sample sizes distinguished by all combinations of levels of the design variables.
Figure S8. Mean 95% coverage levels of the CIs for $\mathbf{P_X}$ (Simulation study n. 2) distinguished by the design variables. The horizontal line indicates the desired coverage level.
Figure S9. Mean 95% coverage levels of the CIs for $\mathbf{p_Y}$ (Simulation study n. 2) distinguished by the design variables. The horizontal line indicates the desired coverage level.
Figure S10. Mean 95% coverage levels of the CIs for $\mathbf{W}$ (Simulation study n. 2) distinguished by the design variables. The horizontal line indicates the desired coverage level.
Figure S11. Mean 95% coverage levels of the CIs for $\mathbf{Wp_Y}$ (Simulation study n. 2) distinguished by the design variables. The horizontal line indicates the desired coverage level.

Appendix :

The lower and upper bounds of the BCa percentile CIs depend on the bootstrap distribution and on two parameters, z_0 and a. The bias-correction parameter, z_0, is determined by considering the proportion of bootstrap estimates that are less than the observed sample statistic t. It can be estimated as

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\sum_{b=1}^{B} I(t_b < t)}{B}\right), \tag{A1}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and $I(t_b < t)$ is the indicator function equal to 1 if $t_b < t$, and 0 otherwise. The acceleration parameter, a, is proportional to the skewness of the bootstrap distribution. A common way to estimate a is based on the jackknife method. The jackknife is another resampling method that estimates a parameter of interest by using the observed sample and adding or removing single units (positive and negative jackknife, respectively). According to Lambert, Wildt, and Durand (1991), we can estimate a as

$$\hat{a} = \frac{\sum_{i=1}^{N}[(t^{+i} - t)/(N+1)]^3}{6\left\{\sum_{i=1}^{N}[(t^{+i} - t)/(N+1)]^2\right\}^{3/2}} = \frac{\sum_{i=1}^{N}(t^{+i} - t)^3}{6\left\{\sum_{i=1}^{N}(t^{+i} - t)^2\right\}^{3/2}} \tag{A2}$$

where t^{+i} is the positive jackknife estimate of θ computed by considering the original sample and adding unit i, $i = 1, \ldots, N$. By using (A1) and (A2), the lower and upper bounds of the BCa percentile CIs at the confidence level $1 - \gamma$ are given by

$$\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\gamma/2}}{1 - \hat{a}(\hat{z}_0 + z_{\gamma/2})}\right) \tag{A3}$$

and

$$\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\gamma/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\gamma/2})}\right), \tag{A4}$$

respectively. For further details on the bootstrap see, for example, Efron and Tibshirani (1993).