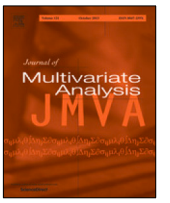
Journal of Multivariate Analysis 184 (2021) 104754



Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva



Principal loading analysis



Jan O. Bauer *, Bernhard Drabant

Baden-Wuerttemberg Cooperative State University Mannheim, Coblitzallee 1-9, 68163 Mannheim, Germany

ARTICLE INFO

Article history:
Received 6 July 2020
Received in revised form 7 March 2021
Accepted 7 March 2021
Available online 18 March 2021

AMS 2010 subject classifications:
62H25
62J10
15A18
15A42
65F15

Keywords:
Component loading
Dimensionality reduction
Matrix perturbation theory
Principal component analysis

ABSTRACT

This paper proposes a tool for dimension reduction where the dimension of the original space is reduced: the principal loading analysis. Principal loading analysis is a tool to reduce dimensions by discarding variables. The intuition is that variables are dropped which distort the covariance matrix only by a little. Our method is introduced and an algorithm for conducting principal loading analysis is provided. Further, we give bounds for the noise arising in the sample case.

© 2021 Elsevier Inc. All rights reserved.

1. Introduction

When data is of high dimension it is often beneficial to reduce dimension. Two ways of reducing dimension exist: either by transforming the variables to a reduced set of variables or by selecting a subset of the existing variables [2,10]. We propose a new approach pursuing the latter one despite adopting ideas from principal component analysis (PCA).

Classical PCA was formulated by [4,14] and has been extended over the years while still being an active field of research. The basic goal is to transform a set of variables into a subspace spanned by orthogonal variables containing most of its original variance (see for example [7] for an overview). One extension to a non-linear approach using kernels was given by [17], to the so called kernel PCA. Further, [20] proposed a sparse principal component analysis (sparse PCA) to reduce the dimension as well as the number of used variables. This is done by implementing a further restriction in the underlying maximization problem. Regarding the purpose of discarding variables, [5,6] presented four methods and provided examples for artificial as well as real data respectively. The methods are based on the idea to select variables that are highly present in the largest eigenvectors i.e. in the eigenvectors associated with the largest eigenvalues, or to discard variables that are highly present in the smallest eigenvectors i.e. in the eigenvectors associated with the smallest eigenvalues. In a regression context, [1] propose an iterative method to select the covariates using PCA and [11] to discard variables considering the respective increase in the residual sum of squares.

[9] provided a broad overview of existing theories regarding the asymptotics of eigenvectors and eigenvalues of sample correlation and covariance matrices respectively. [3] pointed out that the assumption of simple eigenvalues is needed

* Corresponding author.
E-mail address: jan.bauer@dhbw-mannheim.de (J.O. Bauer).

https://doi.org/10.1016/j.jmva.2021.104754
0047-259X/© 2021 Elsevier Inc. All rights reserved.

because the orthonormal basis of the eigenmanifold of eigenvalues with a multiplicity greater than one can be obtained by rotation hence asymptotics for the corresponding eigenvalues are problematic to obtain.

However, to our knowledge, there has been no effort in developing a technique where dimension is reduced by selecting a subset of the observed variables based on non-impact in the eigenvectors. Our main contribution is a method as such: the principal loading analysis (PLA). We investigate the underlying form of the sample covariance matrix needed to conduct PLA where we take the presence of perturbations caused by a small sample size and due to the fuzziness of PLA into account. We further provide an algorithm to conduct PLA in practice.

The rest of the paper is organized as follows: Section 2 provides assumptions needed for the remainder of this work. In Section 3, we recap the methodology of PCA and explain the idea of PLA. Our focus lies on Section 4, where we introduce the PLA method and the underlying covariance matrix structure needed for PLA. We will provide bounds for the sample counterparts essential for PLA: for the sample covariance, sample eigenvectors and sample eigenvalues. We suggest using a cut-off threshold for application and Section 5 complements this step of PLA. We give recommendations for threshold values in Section 6 based on simulation studies. We give simulated examples in Section 7 and take a resume as well as suggest extensions in Section 8.

2. Setup

We first state some notation, assumptions and lemmas used throughout this work. We consider $\boldsymbol{x} = \left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\right) \in \mathbb{R}^{N\times M}$ to be an independent and identically distributed (IID) sample containing $n \in \{1, \dots, N\}$ observations of a random vector $\boldsymbol{X} = \left(X_1, \dots, X_M\right)^\top$ with covariance matrix $\boldsymbol{\Sigma} = (\sigma_{i,j})$ for some $i, j \in \{1, \dots, M\}$. We also consider the case when the covariance matrix is slightly perturbed with

$$\tilde{\boldsymbol{\Sigma}} \equiv \boldsymbol{\Sigma} + \mathbf{E}$$

where $\mathbf{E} = (\varepsilon_{i,j})$ is a sparse matrix. $\mathbf{E}$ is a technical construction and contains small components we want to extract from $\boldsymbol{\Sigma}$. Hence, $\varepsilon_{i,j} \neq 0 \Rightarrow \sigma_{i,j} = 0$. The sample counterpart $\hat{\tilde{\boldsymbol{\Sigma}}}$ is of the form

$$\hat{\tilde{\boldsymbol{\Sigma}}} \equiv (\hat{\tilde{\sigma}}_{i,j}) \equiv \boldsymbol{\Sigma} + \mathbf{E} + \mathbf{H}_N$$

where $\mathbf{H}_N = (\eta_{i,j|N})$ is a perturbation in the form of a random noise matrix. The noise is due to having only a finite number of observations in the sample. We consider the eigendecomposition of $\tilde{\boldsymbol{\Sigma}}$ to be given by

$$\tilde{\boldsymbol{\Sigma}} \equiv \tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{V}}^\top \tag{1}$$

with $\tilde{\boldsymbol{V}}^\top\tilde{\boldsymbol{V}} = \boldsymbol{I}$, $\tilde{\boldsymbol{\Lambda}} = \operatorname{diag}(\tilde{\lambda}_1, \dots, \tilde{\lambda}_M)$ and $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_M$. The eigenvectors $\tilde{\boldsymbol{V}} = \left(\tilde{\boldsymbol{v}}_1, \dots, \tilde{\boldsymbol{v}}_M\right)$ are ordered according to the respective eigenvalues. The eigendecomposition for $\hat{\tilde{\boldsymbol{\Sigma}}}$ is denoted analogously. The corresponding sample eigenvalues and eigenvectors are given by

$$\hat{\tilde{\lambda}}_j \equiv \lambda_j + \varepsilon_j^\lambda + \eta_{j|N}^\lambda$$

and

$$\hat{\tilde{\boldsymbol{v}}}_j \equiv \boldsymbol{v}_j + \boldsymbol{\varepsilon}_j + \boldsymbol{\eta}_{j|N} = \begin{pmatrix} v_j^{(1)} \\ \vdots \\ v_j^{(M)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{j|N}^{(1)} \\ \vdots \\ \varepsilon_{j|N}^{(M)} \end{pmatrix} + \begin{pmatrix} \eta_{j|N}^{(1)} \\ \vdots \\ \eta_{j|N}^{(M)} \end{pmatrix},$$

respectively, for some $j \in \{1, \dots, M\}$. For a vector $\boldsymbol{v} = (v^{(1)}, \dots, v^{(M)})^\top \in \mathbb{R}^M$, we use the $\mathcal{L}_2$ and $\mathcal{L}_\infty$ vector norms as $\|\boldsymbol{v}\|_2 \equiv \left(\sum_m (v^{(m)})^2\right)^{1/2}$ and $\|\boldsymbol{v}\|_\infty \equiv \max_m |v^{(m)}|$, respectively. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_F$ denotes the Frobenius norm. "$\xrightarrow{P}$" denotes convergence in probability where we always consider the limit for $N \to \infty$. $\mathcal{O}_p$ denotes stochastic boundedness and $\boldsymbol{v} = \mathcal{O}_p(\cdot) \Leftrightarrow \|\boldsymbol{v}\|_\infty = \mathcal{O}_p(\cdot)$. When two blocks of random variables $X_{i_1}, \dots, X_{i_I}$ and $X_{j_1}, \dots, X_{j_J}$ are uncorrelated, we write $(X_{i_1}, \dots, X_{i_I}) \overset{\text{c}}{\perp\!\!\!\perp} (X_{j_1}, \dots, X_{j_J})$. Due to the sparse perturbation $\mathbf{E}$, we introduce the following definition:

Definition 1. We say that two blocks of random variables $X_{i_1}, \dots, X_{i_I}$ and $X_{j_1}, \dots, X_{j_J}$ are ε-uncorrelated if $\tilde{\sigma}_{\bar{i},\bar{j}} = \varepsilon_{\bar{i},\bar{j}}$ for $\bar{i} \in \{1, \dots, I\}$ and $\bar{j} \in \{1, \dots, J\}$. We then write $(X_{i_1}, \dots, X_{i_I}) \overset{\varepsilon}{\perp\!\!\!\perp} (X_{j_1}, \dots, X_{j_J})$.

Given the notation above, the following assumption are made for $i, j \in \{1, \dots, M\}$:

Assumption 1. $\tilde{\boldsymbol{\Sigma}}$ has distinct eigenvalues $\tilde{\lambda}_1 > \cdots > \tilde{\lambda}_M$.

Assumption 1 is needed to apply Theorem 6 in [9]. In presence of eigenvalues with an algebraic order larger than one it is problematic to obtain asymptotic results since the corresponding orthonormal basis of the eigenmanifold can

be obtained by rotation. Hence we rule out this case for $\tilde{\boldsymbol{\Sigma}}$ in Assumption 1. We refer to [9] for a detailed overview of asymptotic results and to [3] for a more elaborate explanation why an algebraic order of one is needed for the eigenvalues of concern.

Assumption 2. $E\{(\boldsymbol{X} - E\boldsymbol{X})^4\} < \infty$.

We assume $\boldsymbol{X}$ to be non-explosive. The following two lemmata provide the methodology for the noise. The estimator $\hat{\tilde{\boldsymbol{\Sigma}}}$ is distorted by noise but provides consistent estimation in the limit. The intuition behind this is trivial: we cannot assume that our sample represents $\boldsymbol{X}$ perfectly in the first place. The deviation between the sample values and the population values is described by the noise terms. However, the deviation vanishes when the sample size increases.

Lemma 1. *For $i, j \in \{1, \dots, M\}$ it holds that $\eta_{i,j|N} = \mathcal{O}_p\left(1/\sqrt{N}\right)$.*

Proof. $\eta_{i,j|N} = \mathcal{O}_p\left(1/\sqrt{N}\right)$ is an immediate result of Theorem 1 in [13]. □

We then can further provide bounds for the sample eigenvalues:

Lemma 2. *For $j \in \{1, \dots, M\}$ it holds that $\eta^{\lambda}_{j|N} = \mathcal{O}_p\left(1/N^w\right)$, $w \in [\frac{1}{2}, \infty)$.*

Proof. From Weyl's inequality (Corollary 4.10 in [18]) and Lemma 1 we can conclude that $\max_j |\hat{\tilde{\lambda}}_j - \tilde{\lambda}_j| \leq \|\mathbf{H}_N\|_F = \mathcal{O}_p\left(1/\sqrt{N}\right)$ hence $\eta^{\lambda}_{j|N} = \mathcal{O}_p\left(1/N^w\right)$ with $w \in [\frac{1}{2}, \infty)$. □

3. Background and methodology

Firstly, we recap PCA in this section because some steps of PCA correspond with PLA. Afterwards, we introduce the intuition of PLA which is complemented by its applied parts in Section 4.

PCA is a tool for dimension reduction while containing most of the variance within the data. Every observation n is projected into a K-dimensional subspace, with $K \leq M$, spanned by the eigenvectors of the covariance matrix denoted loadings. The projected observations then are represented by the so called PCs. Clearly, it is most appealing when $K \ll M$ and $K \leq 3$ for graphic purposes. The underlying decomposition is the eigendecomposition $\tilde{\boldsymbol{\Sigma}} = \tilde{\boldsymbol{V}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{V}}^{\top}$ in (1). The identity of PCA is then based on

$$\boldsymbol{X} = \tilde{\boldsymbol{V}}\tilde{\boldsymbol{\xi}} = \tilde{\boldsymbol{V}}_K\tilde{\boldsymbol{\xi}}_K + \tilde{\boldsymbol{V}}_{M-K}\tilde{\boldsymbol{\xi}}_{M-K} \ ,$$

see [12] among others. Here, $\tilde{\boldsymbol{\xi}} = \left(\tilde{\xi}^{(1)}, \dots, \tilde{\xi}^{(M)}\right)^{\top}$ is the vector of PCs. $\tilde{\boldsymbol{\xi}}_K = \left(\tilde{\xi}^{(1)}, \dots, \tilde{\xi}^{(K)}\right)^{\top}$ denotes the PCs corresponding to the K largest eigenvectors and $\tilde{\boldsymbol{\xi}}_{M-K} = \left(\tilde{\xi}^{(K+1)}, \dots, \tilde{\xi}^{(M)}\right)^{\top}$ denotes the PCs corresponding to the $M-K$ smallest eigenvectors, respectively. The notation for $\tilde{\boldsymbol{V}}_K = \left(\tilde{\boldsymbol{v}}_1, \dots, \tilde{\boldsymbol{v}}_K\right)$ and $\tilde{\boldsymbol{V}}_{M-K}$ goes in an analogue manner.

Considering the geometric shape of a covariance matrix, the components of the eigenvectors reflect the distortion in each dimension. When the, say, jth component of an eigenvector contains a small value, we can conclude that the jth variable affects the observations projected into a subspace containing this very eigenvector only a little (if at all) for this particular eigenvector-axis. Further, the importance of each eigenvector-axis is given by the size of the corresponding eigenvalue. This is due to the eigendecomposition

$$\mathrm{Var}\left(\tilde{\xi}^{(j)}\right) = \mathrm{Var}\left(\tilde{\boldsymbol{v}}_j^{\top}\boldsymbol{X}\right) = \tilde{\boldsymbol{v}}_j^{\top}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{v}}_j = \tilde{\lambda}_j \ .$$

Hence, the explained variance by the first K PCs is percentaged given by

$$\frac{\sum_{m=1}^{K}\tilde{\lambda}_m}{\sum_{m=1}^{M}\tilde{\lambda}_m} \ . \tag{2}$$

When the jth component of each of the eigenvectors (except for one) is small, the jth variable affects the projection into the subspace containing most of the variance marginally. Our motivation for PLA is then to discard this very variable instead of projecting all observations into the subspace spanned by the K largest eigenvectors also containing the jth variable. Therefore, PLA is a tool to detect variables that do not account for any or only for little distortion of the variance and to discard them.

PLA proceeds as follows: Assuming that the data lies in the Euclidean space spanned by the unit vectors $\mathbf{e}_1, \dots, \mathbf{e}_M$ and that the rows $j_1, \dots, j_{M^*}$ of $M - M^*$ eigenvectors contain only values below a threshold (in absolute terms) hence do not distort the covariance matrix, we consider to discard the corresponding random variables $X_{j_1}, \dots, X_{j_{M^*}}$ hence transform the data into a subspace spanned by $\mathbf{e}_j$ with $j \neq j_1, \dots, j_{M^*}$. Therefore, we do not merge the original variables by transforming them into a subspace spanned by the different eigenvectors as done in PCA but rather contain the original

variables except the discarded one(s). However, the size of the eigenvalues corresponding to the $M - M^*$ eigenvectors has to be considered according to (2) since the $X_{j_1}, \ldots, X_{j_{M^*}}$ might explain a fair proportion of the variance.

Note that we link variables to eigenvectors and their corresponding eigenvalues hence link a set of variables to a set of eigenvectors with corresponding eigenvalues. Keeping this in mind, a further interpretation of $\mathbf{E}$ is outlined in the following remark.

Remark 1. The eigenvalues can be interpreted as estimators of the explained variance biased by $\mathbf{E}$ since

$$\max_j |\hat{\tilde{\lambda}}_j - \lambda_j| \overset{\text{Weyl}}{\leq} \|\mathbf{E} + \mathbf{H}_N\|_F \xrightarrow{P} \|\mathbf{E}\|_F$$

where convergence in probability results from Lemma 1 and the Continuous Mapping Theorem. The bias is, however, small due to the sparseness of $\mathbf{E}$ and vanishes when the variables are uncorrelated.

4. Principal loading analysis

In this section we provide an explicit algorithm to execute PLA and we investigate the underlying structure of the eigendecomposition of the sample covariance matrix needed to conduct PLA when the sample covariance matrix is of the form $\hat{\tilde{\boldsymbol{\Sigma}}} = \boldsymbol{\Sigma} + \mathbf{E} + \mathbf{H}_N$ meaning in the presence of sparse perturbation and noise. We propose that variables can be discarded when small components occur in the population eigenvectors because small entries reflect that the corresponding variable does not distort the covariance matrix in this very direction. In fact, we do look for ε-uncorrelated blocks of variables which only deform towards a few dimensions. However, those variables corresponding to the blocks are only discarded if the sum of the respective eigenvalues is small hence explains only little of the overall variance within the covariance matrix. Considering $\mathbf{E} = \mathbf{0}$, non-zero components of concern can be due to noise within the sample or if the components are in fact different from zero. To get an intuition, we provide bounds for the relevant eigenvectors which have to hold under PLA, as well as bounds for the eigenvalues reflecting the explained variance.

To get a first intuition, we start by considering $\mathbf{E} = \mathbf{0}$ again. It holds that variables cause zero components in the eigenvectors when the variables are uncorrelated in blocks, in a way that

$$\underbrace{(X_1, \ldots, X_{M_1})}_{\kappa_1\text{-many}} \overset{c}{\perp\!\!\!\perp} \underbrace{(X_{M_1+1}, \ldots, X_{M_2})}_{\kappa_2\text{-many}} \overset{c}{\perp\!\!\!\perp} \cdots \overset{c}{\perp\!\!\!\perp} \underbrace{(X_{M_{L-1}+1}, \ldots, X_{M_L})}_{\kappa_L\text{-many}}, \tag{3}$$

for an arbitrary $L > 1$. Considering $\mathbf{E} \neq \mathbf{0}$, $\mathbf{E}$ captures small correlations among the random variables which results in a covariance matrix of the form

$$\hat{\tilde{\boldsymbol{\Sigma}}} = \boldsymbol{\Sigma} + \mathbf{E} + \mathbf{H}_N = \operatorname{diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_L) + \mathbf{E} + \mathbf{H}_N \,.$$

We can assume that the covariance matrix behaves in this convenient way because we can always obtain this structure using a permutation matrix. Therefore, (3) with $\mathbf{E} \neq \mathbf{0}$ is equivalent to

$$\underbrace{\tilde{\boldsymbol{\Sigma}}_1}_{\kappa_1 \times \kappa_1} \overset{\varepsilon}{\perp\!\!\!\perp} \cdots \overset{\varepsilon}{\perp\!\!\!\perp} \underbrace{\tilde{\boldsymbol{\Sigma}}_L}_{\kappa_L \times \kappa_L} \,.$$

Due to the block-structure, the eigenvectors of $\boldsymbol{\Sigma}$ are of shape

$$\underbrace{\begin{pmatrix} *_{\kappa_1} \\ \mathbf{0} \end{pmatrix}, \ldots, \begin{pmatrix} *_{\kappa_1} \\ \mathbf{0} \end{pmatrix}}_{\kappa_1\text{-many}}, \underbrace{\begin{pmatrix} \mathbf{0} \\ *_{\kappa_2} \\ \mathbf{0} \end{pmatrix}, \ldots, \begin{pmatrix} \mathbf{0} \\ *_{\kappa_2} \\ \mathbf{0} \end{pmatrix}}_{\kappa_2\text{-many}}, \ldots, \underbrace{\begin{pmatrix} \mathbf{0} \\ *_{\kappa_L} \end{pmatrix}, \ldots, \begin{pmatrix} \mathbf{0} \\ *_{\kappa_L} \end{pmatrix}}_{\kappa_L\text{-many}}, \tag{4}$$

where $*_{\kappa_\ell}$ with $\ell \in \{1, \ldots, L\}$ are vectors of length κ_ℓ and $\mathbf{0}$ are vectors of suitable dimension containing zeros. The first κ_{b_1} eigenvectors have (at least) $M - \kappa_{b_1}$ zero-components, the following κ_{b_2} eigenvectors have (at least) $M - \kappa_{b_2}$ zero-components and so on. The eigenvectors of $\hat{\tilde{\boldsymbol{\Sigma}}}$ follow the same shape however being slightly perturbed due to $\mathbf{E}$ and distorted by the noise $\mathbf{H}_N$.

PLA for discarding, say, A blocks $\boldsymbol{\Sigma}_{b_1}, \ldots, \boldsymbol{\Sigma}_{b_A}$ with $\boldsymbol{\Sigma}_{b_a} \overset{c}{\perp\!\!\!\perp} \boldsymbol{\Sigma}_{b_l} \forall l \neq a$ for $a \in \{1, \ldots, A\}$ is provided in the following algorithm:

Algorithm 1 (*PLA*)**.** Discard the variables corresponding to $\boldsymbol{\Sigma}_{b_1}, \ldots, \boldsymbol{\Sigma}_{b_A}$ according to PLA proceeds as follows:

1. Check if the eigenvectors of $\boldsymbol{\Sigma}$ satisfy the required structure in (4) to discard $\boldsymbol{\Sigma}_{b_1}, \ldots, \boldsymbol{\Sigma}_{b_A}$.

2. Decide if $\boldsymbol{\Sigma}_{b_1}, \ldots, \boldsymbol{\Sigma}_{b_A}$ are relevant according to the explained variance (2) of the realizations of their contained random variables by calculating

$$\frac{\sum_j \hat{\tilde{\lambda}}_j}{\sum_m \hat{\tilde{\lambda}}_m} \tag{5}$$

where $m \in \{1, \ldots, M\}$ and j indexes all eigenvalues associated with $\boldsymbol{\Sigma}_{b_1}, \ldots, \boldsymbol{\Sigma}_{b_A}$.

3. Discard $\boldsymbol{\Sigma}_{b_1}, \ldots, \boldsymbol{\Sigma}_{b_A}$.

In the spirit of PCA, during the first step of Algorithm 1 we check if all relevant components of an eigenvector are below a given threshold τ in absolute terms. The purposes of this cut-off value are twofold: we want to detect the required structure despite the presence of noise and, due to the fuzziness of τ, we detect the ε-uncorrelated variables. In Section 6 we provide recommendations for τ. The convergence rate for the sample eigenvectors and sample eigenvalues is provided in the following theorem.

Theorem 1. *Let Assumptions 1 and 2 hold. When dropping the variables corresponding to $\boldsymbol{\Sigma}_{b_a}$ according to PLA it holds that there are κ_{b_a} eigenvectors of $\hat{\tilde{\boldsymbol{\Sigma}}}$ of the form*

$$\hat{\tilde{\boldsymbol{v}}}_j = \boldsymbol{v}_j + \boldsymbol{\varepsilon}_j + \boldsymbol{\eta}_{j|N} = \begin{pmatrix} \boldsymbol{0} & v_j^{(M_{a-1}+1)} & \cdots & v_j^{(M_a)} & \boldsymbol{0} \end{pmatrix}^\top + \boldsymbol{\varepsilon}_j + \boldsymbol{\eta}_{j|N}$$

where j indexes all the eigenvalues associated with $\boldsymbol{\Sigma}_{b_a}$ and with corresponding eigenvalues

$$\hat{\tilde{\lambda}}_j = \tilde{\lambda}_j + \eta^{\lambda}_{j|N} \, .$$

Then it holds that

$$\boldsymbol{\eta}_{j|N} = \mathcal{O}_p\left(\frac{1}{\sqrt{N}}\right), \; \eta^{\lambda}_{j|N} = \mathcal{O}_p\left(\frac{1}{N^w}\right), \; w \in [\tfrac{1}{2}, \infty) \, .$$

Proof. $\eta^{\lambda}_{j|N} = \mathcal{O}_p\left(1/N^w\right)$ with $w \in [\frac{1}{2}, \infty)$ is due to Lemma 2 and $\boldsymbol{\eta}_{j|N} = \mathcal{O}_p\left(1/\sqrt{N}\right)$ is a conclusion of Theorem 6 in [9]. □

Sufficient bounds for the perturbations that have to hold in order to satisfy correct discarding are given in the following theorem. It provides an intuition of the possible magnitude of the perturbations that results in a drop. Note that the bounds are sufficient and not necessary and sufficient.

Theorem 2. *Denote $\tilde{\lambda}_0 = \lambda_0 \equiv \infty$ and $\tilde{\lambda}_{M+1} = \lambda_{M+1} \equiv -\infty$. For $j \in \{1, \ldots, M\}$ it holds that*

$$\frac{2^{3/2}\|\boldsymbol{E} + \boldsymbol{H}_N\|_F}{\min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})} < \tau \;\Rightarrow\; \|\boldsymbol{\varepsilon}_j + \boldsymbol{\eta}_{j|N}\|_\infty < \tau \, .$$

Proof. From Corollary 1 in [19] we can conclude that $\|\boldsymbol{\varepsilon}_j + \boldsymbol{\eta}_{j|N}\|_2 \leq 2^{3/2}\|\mathbf{E} + \mathbf{H}_N\|_F / \min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$ which yields our desired result since $\|\boldsymbol{\varepsilon}_j + \boldsymbol{\eta}_{j|N}\|_\infty \leq \|\boldsymbol{\varepsilon}_j + \boldsymbol{\eta}_{j|N}\|_2$. □

Hence, discarding is ensured if the perturbations are little conditioning on that the eigenvalues are not close. A next step is naturally to investigate what assures discarding in case of large perturbations and what causes close eigenvalues. We provide a brief intuition to answer those questions, however it is subject to future research. We cover the population case with $\mathbf{H}_N = \mathbf{0}$ and consider that $\lambda_{\ell-1}$, λ_ℓ and $\lambda_{\ell+1}$ are eigenvalues of $\boldsymbol{\Sigma}_{\overline{\ell}}$, $\boldsymbol{\Sigma}_{\ell-}$ and $\boldsymbol{\Sigma}_{\underline{\ell}}$ respectively. If needed, we ensure the validity of subtraction of quadratic matrices with different dimensions by padding the matrix with smaller dimension of concern, say, $\boldsymbol{A}$ in a way that $\boldsymbol{A} \rightsquigarrow \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Delta} \end{pmatrix}$ with $\boldsymbol{\Delta} = \operatorname{diag}(\delta, \ldots, \delta)$ and $\delta \to 0$ with $\delta > 0$ is small.

Remark 2. $\|\boldsymbol{\varepsilon}_j\|_\infty < \tau$ is satisfied if

$$\|\mathbf{E}\|_F < \tau \cdot D_{\bar{l},\underline{l}} \tag{6}$$

where $D_{\bar{\ell},\underline{\ell}}$ is the largest element lying in the Gershgorin discs of $\boldsymbol{\Sigma}_{\overline{\ell}} - \boldsymbol{\Sigma}_{\ell-}$ or $\boldsymbol{\Sigma}_{\ell-} - \boldsymbol{\Sigma}_{\underline{\ell}}$.

Hence, the more the correlation structure of $\boldsymbol{\Sigma}_{\ell-}$ differs from the other blocks, the more likely we are to ensure correct dropping.

5. Decision about discarding: The choice of thresholds

In this section we take a deeper look at the second step of Algorithm 1. We recap the idea behind (5) and the choice of the cut-off value τ respectively. Further, we provide an extension of (5) used in Algorithm 1.

Some steps of PLA correspond with PCA by construction, the choice of optimal thresholds is such an intersection. Therefore, we can utilize established approaches from PCA for this particular step. An elaborate overview of common techniques to evaluate the importance of each principal component (PC) in PCA is provided in [7,16]. In this work, we evaluate the explained variance provided by the blocks we consider to discard and proceed to drop if the explained variance is small. This corresponds to the PCA method when checking the cumulative percentage of total variation explained by the PCs one wants to keep. While the contribution of the PCs is desired to be large in PCA, however, we want the contribution of the blocks in PLA to be small. The threshold in PCA depends on practical details of a particular data set as well as on the intention of the applicant and a rule of thumb is to choose the share of the explained variance by the PCs to be between $70\% - 90\%$ [7]. We suggest this procedure for PLA to decide in practice if the variables of consideration explain little of the overall variance, that is to say if the set of the reduced variables explains an acceptable amount of the overall variance.

The threshold τ regarding the cut-off of elements vary among applications. [15] recaps some cut-off values from published studies which lie between 0.3 and 0.5. However, in Section 6 we discovered that PLA performs well even for smaller thresholds than the ones mentioned above.

It is worth mentioning that we can obtain a measure of the contribution to the explained variance by combining both methods. Therefore, we consider again the case to discard A blocks $\boldsymbol{\Sigma}_{b_1}, \ldots, \boldsymbol{\Sigma}_{b_A}$ with corresponding eigenvectors $\hat{\hat{\boldsymbol{v}}}_j$ and eigenvalues $\hat{\hat{\lambda}}_j$ respectively. For convenience purposes, we assume that the contained random variables are X_i with $j \in \{1, \ldots, \kappa_A\}$. The explained variance of $\sum_j \hat{\hat{\lambda}}_j$ is then given by (5). However, note that $\hat{\hat{v}}_j^{(k)}$, $k \notin \{1, \ldots, \kappa_A\}$, does not necessarily equal zero due to the perturbation. The eigenvectors are dominated by the random variables we consider to discard, but the remaining random variables also distort to some degree. Further, the elements of the complementary eigenvectors $\hat{\hat{v}}_k^{(j)}$ are not necessarily equal to zero either. Hence, the X_j might also slightly influence the covariance matrix in further directions. Since the eigenvectors are normalized, we can consider each squared eigenvector element as the portion of the corresponding random variable towards this eigenvector direction. Hence, the measure of the contribution to the explained variance is in spirit of (5) given by

$$\left(\sum_m \hat{\hat{\lambda}}_m\right)^{-1} \left(\sum_{j\in\{1,\ldots,\kappa_A\}} \hat{\hat{\lambda}}_j \sum_{i=1}^{\kappa_A} (\hat{\hat{v}}_j^{(i)})^2 + \sum_{k\notin\{1,\ldots,\kappa_A\}} \hat{\hat{\lambda}}_k \sum_{i=1}^{\kappa_A} (\hat{\hat{v}}_k^{(i)})^2\right) .$$

In practice, we can then adjust τ in order to increase or decrease this explained variance sufficient for application.

6. Simulation study

Finding an optimal threshold τ theoretically is rather difficult due to the fuzziness of τ. However, we conducted two simulation studies where the population $\boldsymbol{X}$ consisting of M variables with 100 000 realizations was simulated $S = 10\,000$ times each. By construction, either $k \in \{1, \ldots, 5\}$ variables were uncorrelated or a $\kappa \times \kappa$ block was uncorrelated with $\kappa \in \{2, \ldots, 6\}$. Then, for each S a sample $\boldsymbol{x}$ of size $N = 10\,000$ had been drawn and we conducted PLA for $\tau \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. For τ, we considered cut-off values used in published studies as an orientation [15]. However, we also focused on tighter values because PLA appeared to perform well for small thresholds.

Since the covariance provided uncorrelated patterns by construction, we were able to calculate the type I errors as the share of iterations where PLA did not lead to a consideration of a drop i.e., when step 1 in Algorithm 1 was not fulfilled. The results can be found in Appendix B in Table 1 for the variable case and in Table 2 for the block case respectively. In a type I error sense, PLA performs best for the cut-off values in the center. However, tighter thresholds diminish the likelihood of conducting a type II error. Hence, we recommend to use a threshold of 0.3 when the blocks are of dimension one and a threshold of 0.4 otherwise because they still yield satisfying results while being a tight choice. Note that the explained variance has not been considered in this simulation since we were only interested in the correct detection of uncorrelation.

As indicated in (6), the threshold for ε-uncorrelated blocks depends on many factors and hence the type I errors are hard to simulate. As mentioned previously, we provide a starting point for further research regarding the search for optimal cut-off values in Section 8. Still, the given values for τ do work as we demonstrate in Section 7.

7. Example

We provide an example of PLA for the case of an uncorrelated block and for the case of a single ε-uncorrelated variable in this section. The examples are based on simulated data sets and all values are rounded to two decimal places.

In order to conduct PLA on an uncorrelated block, we simulated the population $\boldsymbol{X}$ consisting of 100 000 realizations of ten variables $X_1, \ldots, X_{10}$ such that $\text{Corr}(X_j, X_i) = 0$ for $j \in \{1, 2\}$, $i \in \{3, \ldots, 10\}$ and $\text{Corr}(X_1, X_2) \neq 0$. The sample $\boldsymbol{x}$ with

corresponding sample correlation matrix $\hat{\boldsymbol{\Sigma}}$ was constructed by drawing 10 000 observations without replacement from $\boldsymbol{X}$. The estimator $\hat{\boldsymbol{\Sigma}}$ can be found in Appendix A with

$$\hat{\boldsymbol{V}} = \begin{pmatrix} 0.00 & 0.06 & -0.69 & -0.04 & -0.01 & 0.01 & 0.00 & -0.01 & 0.72 & 0.01 \\ 0.00 & 0.06 & -0.72 & -0.04 & -0.01 & 0.00 & 0.00 & 0.01 & -0.69 & 0.00 \\ -0.18 & -0.11 & -0.01 & -0.03 & 0.18 & 0.08 & -0.50 & 0.79 & 0.02 & -0.20 \\ -0.37 & 0.05 & 0.05 & -0.75 & -0.35 & -0.31 & 0.00 & -0.07 & 0.00 & -0.26 \\ -0.32 & -0.35 & -0.02 & 0.03 & -0.31 & 0.78 & -0.05 & -0.20 & 0.00 & -0.17 \\ -0.45 & 0.53 & 0.04 & 0.20 & -0.35 & 0.09 & 0.23 & 0.32 & 0.00 & 0.44 \\ -0.23 & 0.17 & 0.00 & 0.26 & -0.10 & -0.21 & -0.77 & -0.43 & -0.01 & 0.12 \\ -0.58 & -0.54 & -0.07 & 0.31 & 0.18 & -0.41 & 0.27 & -0.02 & 0.00 & 0.03 \\ -0.23 & 0.49 & 0.02 & 0.26 & 0.28 & 0.05 & 0.17 & -0.12 & 0.00 & -0.72 \\ -0.30 & 0.13 & 0.02 & -0.40 & 0.71 & 0.27 & -0.02 & -0.18 & -0.01 & 0.37 \end{pmatrix}$$

and

$$\operatorname{diag}(\hat{\boldsymbol{\Lambda}}) = \begin{pmatrix} 154.97 & 26.70 & 23.93 & 19.74 & 9.05 & 6.24 & 2.48 & 1.68 & 0.99 & 0.22 \end{pmatrix}.$$

$\hat{\boldsymbol{V}}$ reveals that X_1 and X_2 are dominant in $\hat{\boldsymbol{v}}_3$ and $\hat{\boldsymbol{v}}_9$ however influence the remaining eigenvectors only marginally hence primarily distort $\boldsymbol{\Sigma}$ in only two directions. We then consider to drop the block containing X_1 and X_2 denoted $\boldsymbol{\Sigma}_1$ with $\kappa_1 = 2$ represented by $\hat{\boldsymbol{v}}_3$ and $\hat{\boldsymbol{v}}_9$. We choose $\tau = 0.3$ following our argument in Section 6 and conduct PLA.

1. Check if the eigenvectors of $\boldsymbol{\Sigma}$ satisfy the required structure in (4) to discard $\boldsymbol{\Sigma}_1$.
   For $\hat{\boldsymbol{v}}_3$ and $\hat{\boldsymbol{v}}_9$, it holds that all components in absolute terms are below the given threshold except the first two components.
2. Decide if $\boldsymbol{\Sigma}_1$ is relevant according to the explained variance (2) of the realizations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of their contained random variables X_1 and X_2.
   It holds that $\hat{\boldsymbol{\Sigma}}_1$ explains $(\hat{\lambda}_3 + \hat{\lambda}_9)/\sum_m \hat{\lambda}_m \approx 0.1013$ hence 10.13% of the overall variance which is a fairly small amount.
3. Discard $\boldsymbol{\Sigma}_1$.
   We drop X_1 and X_2.

To provide an example for a single ε-uncorrelated variable, we simulated the population $\boldsymbol{X}$ consisting of 100 000 realizations of ten variables $X_1, \ldots, X_{10}$ such that $\operatorname{Corr}(X_1, X_i) = \textit{small}$ for $i \in \{2, \ldots, 10\}$. The sample $\boldsymbol{x}$ was constructed by drawing 10 000 observations without replacement from $\boldsymbol{X}$. The first row of $\bar{\boldsymbol{\Sigma}}$ displaying $\operatorname{Cov}(X_1, X_i)$ is

$$\begin{pmatrix} 11.01 & -1.22 & 0.26 & -0.91 & 0.51 & 0.04 & 0.53 & -1.33 & -0.29 & -0.93 \end{pmatrix}$$

and the corresponding sample counterpart is given by

$$\begin{pmatrix} 10.94 & -1.22 & 0.50 & -0.83 & 0.52 & 0.30 & 0.48 & -1.05 & -0.28 & -0.89 \end{pmatrix}.$$

The whole matrices can be found in Appendix A. This example emphasizes the difficulty, whether small values in the estimator $\hat{\bar{\boldsymbol{\Sigma}}}$ are due to uncorrelation or due to ε-uncorrelation. The eigendecomposition of $\hat{\bar{\boldsymbol{\Sigma}}}$ yields

$$\hat{\bar{\boldsymbol{V}}} = \begin{pmatrix} 0.00 & 0.00 & 0.10 & 0.10 & 0.94 & 0.29 & -0.12 & -0.04 & 0.01 & 0.03 \\ -0.30 & 0.33 & -0.31 & -0.13 & -0.06 & 0.46 & 0.42 & -0.38 & 0.04 & 0.39 \\ -0.35 & -0.57 & 0.22 & 0.50 & -0.02 & -0.19 & 0.17 & -0.23 & 0.03 & 0.36 \\ -0.38 & -0.01 & -0.69 & 0.24 & 0.13 & -0.19 & -0.01 & 0.37 & -0.35 & -0.11 \\ -0.29 & 0.24 & 0.52 & 0.03 & -0.12 & 0.26 & 0.00 & 0.57 & -0.34 & 0.26 \\ -0.31 & -0.20 & 0.23 & -0.34 & 0.10 & -0.02 & 0.53 & -0.09 & -0.21 & -0.60 \\ -0.30 & 0.58 & 0.16 & 0.35 & 0.07 & -0.34 & 0.11 & -0.01 & 0.49 & -0.22 \\ -0.43 & -0.16 & -0.04 & -0.63 & 0.11 & -0.26 & -0.25 & 0.16 & 0.38 & 0.27 \\ -0.18 & 0.28 & 0.15 & -0.11 & 0.03 & -0.34 & -0.38 & -0.53 & -0.55 & 0.06 \\ -0.39 & -0.14 & -0.02 & 0.15 & -0.24 & 0.52 & -0.52 & -0.15 & 0.16 & -0.39 \end{pmatrix}$$

and

$$\operatorname{diag}(\hat{\bar{\boldsymbol{\Lambda}}}) = \begin{pmatrix} 267.69 & 43.74 & 25.22 & 19.76 & 11.00 & 8.79 & 5.87 & 1.04 & 0.32 & 0.10 \end{pmatrix}.$$

From $\hat{\bar{\boldsymbol{V}}}$ it appears that X_1 distorts $\boldsymbol{\Sigma}$ primarily via $\boldsymbol{v}_5$ and hardly influences the other directions. Hence, we consider to drop X_1 where we choose $\tau = 0.3$ according to Section 6. PLA proceeds as follows:

1. Check if the eigenvectors of $\boldsymbol{\Sigma}$ satisfy the required structure in (4) to discard X_1.
   It holds that all components except the first one of $\hat{\bar{\boldsymbol{v}}}_5$ are below the given threshold in absolute terms.
2. Decide if X_1 is relevant according to the explained variance (2) of the realizations $\boldsymbol{x}_1$.
   It holds that $\boldsymbol{x}_1$ explains $\hat{\bar{\lambda}}_5/\sum_m \hat{\bar{\lambda}}_m \approx 0.0287$ hence 2.87% of the overall variance which is a fairly small amount.

3. Discard X_1.
   We drop X_1.

8. Concluding remarks and outlook

PLA is a tool for dimension reduction. We have shown the different covariance matrix structures needed to conduct PLA and provided bounds for the sample covariance, sample eigenvectors and sample eigenvalues. Based on our simulation study, we suggest using a cut-off threshold of 0.3 for the eigenvectors to detect if variables or blocks of variables can be discarded.

As an extension, we will compare the regression performance of PLA reduced data not only with ordinary least squares (OLS) regression but also with PCA regression as well as with PCA-methods reduced regression in an upcoming work. For instance, the natural link between OLS regression and PCA is given since PCA can be seen as a minimization problem in OLS sense:

$$\min_{\boldsymbol{\beta}} \|\tilde{\xi}^{(i)} - \boldsymbol{\beta}^\top \boldsymbol{X}\|_2 = \min_{\boldsymbol{\beta}} \|\tilde{\xi}^{(i)} - \boldsymbol{\beta}^\top \tilde{\boldsymbol{V}} \tilde{\boldsymbol{\xi}}\|_2$$

is minimized for $\boldsymbol{\beta} = \tilde{v}_i$ hence the loadings minimize the distance between the subspace spanned by the K PCs and the original space. As an appetizer, we will address the link between OLS and PLA in this outlook however for the special case that $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$ hence with $\mathbf{E} = \mathbf{0}$. In a general regression problem

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{u}$$

where $\boldsymbol{y}$ denotes the $N \times 1$ vector of dependent observations of a random variable Y and $\boldsymbol{u}$ denotes a $N \times 1$ vector of unobserved random errors, the well known $M \times 1$ OLS estimator is given by

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} \equiv (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{y} = \hat{\boldsymbol{\Sigma}}^{-1} \begin{pmatrix} \hat{\text{Cov}}(X_1, Y) \\ \vdots \\ \hat{\text{Cov}}(X_M, Y) \end{pmatrix} \xrightarrow{P} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \text{Cov}(X_1, Y) \\ \vdots \\ \text{Cov}(X_M, Y) \end{pmatrix} .$$

If we consider a sample $\boldsymbol{x}_{\boldsymbol{y}} \equiv \begin{pmatrix} \boldsymbol{y} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_M \end{pmatrix}$ containing $\boldsymbol{x}$ and $\boldsymbol{y}$ of a random vector $\boldsymbol{X}_Y \equiv \begin{pmatrix} Y, X_1, \ldots, X_M \end{pmatrix}^\top$ and assume that, say, X_1 is discarded by PLA, it holds that

$$\boldsymbol{\beta}_{\text{OLS}} = \begin{pmatrix} \sigma_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & * \end{pmatrix} \begin{pmatrix} 0 \\ * \end{pmatrix} = \begin{pmatrix} 0 \\ * \end{pmatrix}$$

hence the first component of $\boldsymbol{\beta}_{\text{OLS}}$ is zero. With $\mathbf{E} \neq \mathbf{0}$ however, the first component is different from zero which is caused by $\mathbf{E}$. To find an optimal threshold subject to $\mathbf{E}$, we investigate the regression performance of $\boldsymbol{\beta}_{\text{OLS}}$ depending on τ and $\mathbf{E}$.

Further, we consider different mechanics for the choice of τ. This may imply a combination of an upper and lower threshold or a flexible cut-off value depending on the ratio of discarded variables with respect to the overall amount of variables M, etc. We also take into account that uncorrelated variables cause zero components in the rows as well as in the columns of $\boldsymbol{V}$ and investigate if checking both, rows and columns, might increase PLA performance. Checking for type II errors corresponding to τ is also part of this work.

One step beyond: instead of choosing a certain cut-off value one could implement a penalty term, such as an elastic net or lasso as a special case, to diminish the absolute value of the loading elements. By construction, this shrinks the loadings towards zero [8,20].

We will also address the case if the dimensionality of $\boldsymbol{x}$ is such that $N/M \to \theta \in (0, \infty)$ hence that N is not necessarily larger than M and that M is possibly large in general. We will investigate if PLA is still a feasible tool for dimension reduction.

CRediT authorship contribution statement

Jan O. Bauer: Methodology, Software, Formal analysis Writing - original draft. **Bernhard Drabant:** Verification, Writing - review & editing, Supervision.

Appendix A. Complementary results for Section 7

In the following, we provide complement material for the examples in Section 7. All values are rounded to two decimal places. The sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \mathbf{H}$ for the example of dropping a single uncorrelated block hence with

$\mathbf{E} = \mathbf{0}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 12.06 & 11.46 & 0.01 & -0.08 & -0.12 & 0.11 & 0.08 & 0.12 & 0.20 & 0.10 \\ 11.46 & 12.86 & -0.03 & -0.21 & -0.23 & -0.01 & 0.01 & -0.05 & 0.13 & 0.01 \\ 0.01 & -0.03 & 7.51 & 9.82 & 9.83 & 10.41 & 6.04 & 17.51 & 4.90 & 9.29 \\ -0.08 & -0.21 & 9.82 & 33.77 & 16.85 & 23.99 & 10.46 & 27.48 & 8.86 & 20.14 \\ -0.1 & 2 - 0.23 & 9.83 & 16.85 & 24.25 & 18.89 & 9.84 & 31.79 & 6.52 & 12.73 \\ 0.11 & -0.01 & 10.41 & 23.99 & 18.89 & 40.59 & 19.17 & 32.93 & 22.72 & 18.64 \\ 0.08 & 0.01 & 6.04 & 10.46 & 9.84 & 19.17 & 12.77 & 20.05 & 11.23 & 8.55 \\ 0.12 & -0.05 & 17.51 & 27.48 & 31.79 & 32.93 & 20.05 & 63.25 & 15.29 & 22.63 \\ 0.20 & 0.13 & 4.90 & 8.86 & 6.52 & 22.72 & 11.23 & 15.29 & 16.67 & 12.04 \\ 0.10 & 0.01 & 9.29 & 20.14 & 12.73 & 18.64 & 8.55 & 22.63 & 12.04 & 22.26 \end{pmatrix}.$$

Next, we provide the population covariance matrix $\tilde{\boldsymbol{\Sigma}}$ and its sample counterpart $\hat{\tilde{\boldsymbol{\Sigma}}}$ for the example when discarding a single ε-uncorrelated variable.

$$\tilde{\boldsymbol{\Sigma}} = \begin{pmatrix} 11.01 & -1.22 & 0.26 & -0.91 & 0.51 & 0.04 & 0.53 & -1.33 & -0.29 & -0.93 \\ -1.22 & 34.19 & 16.81 & 33.71 & 23.65 & 22.54 & 29.05 & 32.69 & 15.30 & 30.21 \\ 0.26 & 16.81 & 54.58 & 34.41 & 24.94 & 33.69 & 19.44 & 39.46 & 9.99 & 41.10 \\ -0.91 & 33.71 & 34.41 & 51.63 & 20.69 & 26.62 & 29.92 & 42.55 & 15.44 & 39.61 \\ 0.51 & 23.65 & 24.94 & 20.69 & 33.63 & 25.59 & 31.17 & 31.54 & 17.77 & 31.17 \\ 0.04 & 22.54 & 33.69 & 26.62 & 25.59 & 33.90 & 19.92 & 41.76 & 13.19 & 31.69 \\ 0.53 & 29.05 & 19.44 & 29.92 & 31.17 & 19.92 & 42.91 & 27.96 & 22.01 & 27.61 \\ -1.33 & 32.69 & 39.46 & 42.55 & 31.54 & 41.76 & 27.96 & 61.35 & 21.28 & 44.78 \\ -0.29 & 15.30 & 9.99 & 15.44 & 17.77 & 13.19 & 22.01 & 21.28 & 14.93 & 16.50 \\ -0.93 & 30.21 & 41.10 & 39.61 & 31.17 & 31.69 & 27.61 & 44.78 & 16.50 & 47.81 \end{pmatrix},$$

$$\hat{\tilde{\boldsymbol{\Sigma}}} = \begin{pmatrix} 10.94 & -1.22 & 0.50 & -0.83 & 0.52 & 0.30 & 0.48 & -1.05 & -0.28 & -0.89 \\ -1.22 & 34.20 & 16.60 & 33.46 & 23.53 & 22.17 & 29.01 & 31.95 & 15.04 & 30.01 \\ 0.50 & 16.60 & 54.60 & 34.43 & 24.43 & 33.21 & 19.13 & 38.56 & 9.57 & 40.77 \\ -0.83 & 33.46 & 34.43 & 51.43 & 20.24 & 26.09 & 29.62 & 41.52 & 14.96 & 39.28 \\ 0.52 & 23.53 & 24.43 & 20.24 & 33.52 & 24.97 & 31.31 & 30.51 & 17.62 & 30.73 \\ 0.30 & 22.17 & 33.21 & 26.09 & 24.97 & 33.29 & 19.29 & 40.74 & 12.69 & 31.08 \\ 0.48 & 29.01 & 19.13 & 29.62 & 31.31 & 19.29 & 43.60 & 26.90 & 22.07 & 27.27 \\ -1.05 & 31.95 & 38.56 & 41.52 & 30.51 & 40.74 & 26.90 & 59.78 & 20.54 & 43.69 \\ -0.28 & 15.04 & 9.57 & 14.96 & 17.62 & 12.69 & 22.07 & 20.54 & 14.84 & 16.03 \\ -0.89 & 30.01 & 40.77 & 39.28 & 30.73 & 31.08 & 27.27 & 43.69 & 16.03 & 47.33 \end{pmatrix}.$$

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2021.104754.

The supplementary material contains simulated type I error rates for PLA, both for discarding single uncorrelated variables and for discarding an uncorrelated block. As specified in Section 6, the probabilities are calculated as the share of iterations where PLA did not lead to a consideration of a drop despite it should have been considered by construction.

References

[1] S. Boneh, G.R. Mendieta, Variable selection in regression models using principal components, Commun. Stat. - Theory Methods 23 (1) (1994) 197–213.
[2] M. Cord, P. Cunningham, Machine Learning Techniques for Multimedia, Springer, New York, 2008.
[3] J. Dauxois, A. Pousse, Y. Romain, Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference, J. Multivariate Anal. 12 (1) (1982) 136–154.
[4] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417–441.
[5] I. Jolliffe, Discarding variables in a principal component analysis. I: Artificial data, J. R. Stat. Soc. C 21 (2) (1972) 160–173.
[6] I. Jolliffe, Discarding variables in a principal component analysis. II: Real data, J. R. Stat. Soc. C 21 (1) (1973) 21–31.
[7] I. Jolliffe, Principal Component Analysis, second ed., Springer Series in Statistics, New York, 2002.
[8] I. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Phil. Trans. R. Soc. A 374 (2016).
[9] T. Kollo, H. Neudecker, Asymptotics of eigenvalues and unit-length eigenvectors of sample variance and correlation matrices, J. Multivariate Anal. 47 (2) (1993) 283–300.
[10] L. van der Maaten, E. Postma, J. van den Herik, Dimensionality reduction: A comparative review, J. Mach. Learn. Res. 10 (2009) 66–71.
[11] E.R. Mansfield, J.T. Webster, R.F. Gunst, An analytic variable selection technique for principal component regression, J. R. Stat. Soc. C 26 (1) (1977) 34–40.
[12] C. Morana, PC-VAR estimation of vector autoregressive models, Open J. Stat. 2 (2012) 251–259.
[13] H. Neudecker, A.M. Wesselman, The asymptotic variance matrix of the sample correlation matrix, Linear Algebra Appl. 127 (1990) 589–599.
[14] K. Pearson, On lines and planes of closest fit to systems of points in space, Phil. Mag. 2 (1901) 559–572.

[15] P.R. Peres-Neto, D.A. Jackson, K.M. Somers, Giving meaningful interpretation to ordination axes: Assessing loading significance in principal component nalysis, Ecology 84 (9) (2003) 2347–2363.
[16] P.R. Peres-Neto, D.A. Jackson, K.M. Somers, How many principal components? Stopping rules for determining the number of non-trivial axes revisited, Comput. Stat. Data Anal. 49 (4) (2005) 974–997.
[17] B. Schölkopf, A. Smola, K. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 20 (5) (1998) 1299–1319.
[18] G.W. Stewart, J. Sun, Matrix Perturbation Theory, Academic Press, Boston, 1990.
[19] Y. Yu, T. Wang, R.J. Samworth, A useful variant of the Davis–Kahan theorem for statisticians, Biometrika 102 (2) (2015) 315–323.
[20] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2006) 265–286.