Robust Principal Component Analysis of Data with Missing Values

Tommi Kärkkäinen and Mirka Saarela

University of Jyväskylä, Department of Mathematical Information Technology
40014, Jyväskylä, Finland

Abstract. Principal component analysis is one of the most popular machine learning and data mining techniques. Having its origins in statistics, principal component analysis is used in numerous applications. However, there seems to be not much systematic testing and assessment of principal component analysis for cases with erroneous and incomplete data. The purpose of this article is to propose multiple robust approaches for carrying out principal component analysis and, especially, to estimate the relative importances of the principal components to explain the data variability. Computational experiments are first focused on carefully designed simulated tests where the ground truth is known and can be used to assess the accuracy of the results of the different methods. In addition, a practical application and evaluation of the methods for an educational data set is given.

Keywords: PCA, Missing Data, Robust Statistics

1 Introduction

Principal component analysis (PCA) is one of the most popular methods in machine learning (ML) and data mining (DM) of statistical origin [12]. It is typically introduced in all textbooks of ML and DM areas (e.g., [1, 10]) and is used in numerous applications [15]. It seems that the versatile line of utilization has also partly redefined the original terminology from statistics: in ML&DM, the computation of principal components and their explained variability of data, many times together with dimension reduction, is referred to as PCA, even if the term *analysis*, especially historically, refers to statistical hypothesis testing [12]. However, nowadays the use of the term PCA points to the actual computational procedure. Certainly one of the appealing facets of PCA is its algorithmic simplicity with a supporting linear algebra library: a) create covariance matrix, b) compute eigenvalues and eigenvectors, c) compute data variability using eigenvalues, and, if needed, transform data to the new coordinate system determined by the eigenvectors. This is also the algorithmic skeleton underlying this work.

Even if much researched, the use of PCA for sparse data with missing values (not to be mixed with sparse PCA referring to the sparsity of the linear model [6]) seems not to be a widely addressed topic, although [27] provides a comparison of a set of second-order (classical) methods. We assume here that there is no

further information on the sparsity pattern so that the non-existing subset of data is *missing completely at random* (MCAR) [18]. As argued in [24, 25], a missing value can, in principle, represent any value from the possible range of an individual variable so that it becomes difficult to justify assumptions on data or error normality, which underlie the classical PCA that is based on second-order statistics. Hence, we also consider the so-called nonparametric, robust statistical techniques [13, 11], which allow deviations from normality assumptions while still producing reliable and well-defined estimators.

The two simplest robust estimates of location are median and spatial median. The median, a middle value of the ordered univariate sample (unique only for odd number of points, see [16]), is inherently one-dimensional, and with missing data uses only the available values of an individual variable from the marginal distribution (similarly to the mean). The spatial median, on the other hand, is truly a multidimensional location estimate and utilizes the available data pattern as a whole. These estimates and their intrinsic properties are illustrated and more thoroughly discussed in [16]. The spatial median has many attractive statistical properties; particularly that its breakdown point is 0.5, that is, it can handle up to 50% of the contaminated data, which makes it very appealing for high-dimensional data with severe degradations and outliers, possibly in the form of missing values. In statistics, robust estimation of data scattering (i.e., covariability) has been advanced in many papers [19, 28, 7], but, as far as we know, sparse data have not been treated in them.

The content of this work is as follows: First, we briefly derive and define basic and robust PCA and unify their use to coincide with the geometrical interpretation. Then, we propose two modifications of the basic robust PCA for sparse data. All the proposed methods are then compared using a sequence of carefully designed test data sets. Finally, we provide one application of the most potential procedures, i.e., dimension reduction and identifying the main variables, for an educational data set, whose national subset was analyzed in [24].

2 Methods

Assume that a set of observations $\{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbf{R}^n$, is given, so that N denotes the number of observations and n the number of variables, respectively. To avoid the low-rank matrices by the form of the data, we assume that $n < N$. In the usual way, define the data matrix $\mathbf{X} \in \mathbf{R}^{N \times n}$ as $\mathbf{X} = \left(\mathbf{x}_i^T\right), i = 1, \ldots, N$.

2.1 Derivation and interpretation of the classical PCA

We first provide a compact derivation underlying classical principal component analysis along the lines of [4]. For the linear algebra, see, for example, [8]. In general, the purpose of PCA is to derive a linear transformation to reduce the dimension of a given set of vectors while still retaining their information content (in practice, their variability). Hence, the original set of vectors $\{\mathbf{x}_i\}$ is to be

transferred to a set of new vectors $\{\mathbf{y}_i\}$ with $\mathbf{y}_i \in \mathbf{R}^m$, such that $m < n$ but also $\mathbf{x}_i \sim \mathbf{y}_i$ in a suitable sense. Note that every vector $\mathbf{x} \in \mathbf{R}^n$ can be represented using a set of orthonormal basis vectors $\begin{bmatrix}\mathbf{u}_1 \ldots \mathbf{u}_n\end{bmatrix}$ as $\mathbf{x} = \sum_{k=1}^{n} z_k \mathbf{u}_k$, where $z_k = \mathbf{u}_k^T \mathbf{x}$. Geometrically, this rotates the original coordinate system.

Let us consider a new vector $\tilde{\mathbf{x}} = \sum_{k=1}^{m} z_k \mathbf{u}_k + \sum_{k=m+1}^{n} b_k \mathbf{u}_k$, where the last term represents the residual error $\mathbf{x} - \tilde{\mathbf{x}} = \sum_{k=m+1}^{n} (z_k - b_k)\, \mathbf{u}_k$. In case of the classical PCA, consider the minimization of the least-squares-error:

$$\mathcal{J} = \frac{1}{2} \sum_{i=1}^{N} \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|^2 = \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \tilde{\mathbf{x}}_i)^T\, (\mathbf{x}_i - \tilde{\mathbf{x}}_i) = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=m+1}^{n} (z_{i,k} - b_k)^2. \quad (1)$$

By direct calculation, one obtains $b_k = \mathbf{u}_k^T\, \bar{\mathbf{x}}$, where $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$ is the sample mean. Then (1) can be rewritten as $((\mathbf{u}^T\, \mathbf{v})^2 = \mathbf{u}^T\, (\mathbf{v}\, \mathbf{v}^T)\, \mathbf{u}$ for vectors $\mathbf{u}, \mathbf{v}$) so that

$$\mathcal{J} = \frac{1}{2} \sum_{k=m+1}^{n} \sum_{i=1}^{N} \left(\mathbf{u}_k^T (\mathbf{x}_i - \bar{\mathbf{x}})\right)^2 = \frac{1}{2} \sum_{k=m+1}^{n} \mathbf{u}_k^T\, \Sigma\, \mathbf{u}_k, \quad (2)$$

where Σ is the sample covariance matrix

$$\Sigma = \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})\, (\mathbf{x}_i - \bar{\mathbf{x}})^T. \quad (3)$$

Note that the standard technique (e.g., in Matlab) for sparse data is to compute (3) only for those data pairs where both values $(\mathbf{x}_i)_j$ and $(\mathbf{x}_i)_k$ exist. By setting $\mathbf{v}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, we have for the quadratic form, with an arbitrary vector $\mathbf{x} \neq 0$:

$$\mathbf{x}^T\, \Sigma \mathbf{x} = \mathbf{x}^T \left[\mathbf{v}_1\, \mathbf{v}_1^T + \ldots + \mathbf{v}_N\, \mathbf{v}_N^T\right] \mathbf{x} = (\mathbf{x}^T \mathbf{v}_1)^2 + \ldots + (\mathbf{x}^T \mathbf{v}_N)^2 \geq 0. \quad (4)$$

This shows that any matrix of the form of (3) is always at least positive semidefinite, with positive eigenvalues if $\mathbf{v}_i$'s span $\mathbf{R}^n$, that is, if $\text{rank}[\mathbf{v}_1 \ \ldots \mathbf{v}_N] \geq n$. The existence of missing values clearly increases the possibility of semidefiniteness.

Now, let $\{\lambda_k, \mathbf{u}_k\}$ be the kth eigenvalue and eigenvector of Σ satisfying

$$\Sigma \mathbf{u}_k = \lambda_k \mathbf{u}_k, \quad k = 1, \ldots, n. \quad (5)$$

This identity can be written in the matrix form as $\Sigma \mathbf{U} = \mathbf{U}\mathbf{D}$, where $\mathbf{D} = \text{Diag}\{\lambda_1, \ldots, \lambda_n\}$ (vector $\boldsymbol{\lambda}$ as the diagonal matrix) and $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1\, \mathbf{u}_2 \ldots \mathbf{u}_n\end{bmatrix}$. Using (5) shows that (2) reduces to $\mathcal{J} = \frac{1}{2} \sum_{k=m+1}^{n} \lambda_k$. This means that the reduced representation consists of those m eigenvectors that correspond to the m largest eigenvalues of matrix Σ. For the unbiased estimate of the sample covariance matrix $\Sigma \simeq \frac{1}{N-1} \Sigma$, one can use scaling such as in (3) because it does not affect eigenvectors or the relative sizes of the eigenvalues. Finally, for any $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} = \mathbf{U}^T \mathbf{x}$, we have

$$\mathbf{x}^T\, \Sigma \mathbf{x} = \mathbf{y}^T \mathbf{D} \mathbf{y} = \sum_{k=1}^{n} \lambda_k \mathbf{y}_k^2 = \sum_{k=1}^{n} \frac{\mathbf{y}_k^2}{\left(\lambda_k^{-\frac{1}{2}}\right)^2}. \quad (6)$$

Geometrically, this means that in the transformed coordinate system $\mathbf{U}^T\mathbf{e}_k$ ($\mathbf{e}_k$s are the base vectors for the original coordinates), the data define an n-dimensional hyperellipsoid for which the lengths of the principal semi-axis are proportional to $\sqrt{\lambda_k}$.

To this end, we redefine the well-known principle (see, e.g., [15]) for choosing a certain number of principal components in dimension reduction. Namely, the derivations above show that eigenvalues of the sample covariance matrix Σ represent *the variance* along the new coordinate system, $\lambda_k = \sigma_k^2$, whereas the geometric interpretation related to (6) proposes to use the standard deviation $\sigma_k = \sqrt{\lambda_k}$ to assess the variability of data.

Proposition 1. *The relative importance RI_k (in percentages) of a new variable y_k for the principal component transformation based on the sample covariance matrix is defined as $RI_k = 100\frac{\sqrt{\lambda_k}}{\sum_{i=1}^{n}\sqrt{\lambda_i}}$, where λ_k satisfy (5). We refer to $\sqrt{\lambda_i}$ as the estimated variability of the ith (new) variable.*

2.2 Derivation of robust PCA for sparse data

Formally, a straightforward derivation of the classical PCA as given above is obtained from the optimality condition for the least-squares problem (1). Namely, assume that instead of the reduced representation, the problem $\min_{\mathbf{x}} \mathcal{J}(\mathbf{x})$ as in (1) is used to estimate the location of the given data $\{\mathbf{x}_i\}$. In second-order statistics, this provides the sample mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$, whose explicit formula can be obtained from the optimality condition (see [16]):

$$\frac{d\mathcal{J}(\bar{\mathbf{x}})}{d\mathbf{x}} = \frac{d}{d\mathbf{x}}\frac{1}{2}\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{x}\|^2 = \sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{0}. \tag{7}$$

The covariate form of this optimality condition $\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ readily provides us the sample covariance matrix up to the constant $\frac{1}{N-1}$.

Next we assume that there are missing values in the given data. To define their pattern, let us introduce the projection vectors $\mathbf{p}_i$, with $i = 1\ldots, N$ (see [17, 2, 24, 25]), which capture the availability of the components:

$$(\mathbf{p}_i)_j = \begin{cases} 1, \text{if } (\mathbf{x}_i)_j \text{ exists,} \\ 0, \text{otherwise.} \end{cases} \tag{8}$$

We also define the corresponding matrix $\mathbf{P} \in \mathbf{R}^{N\times n}$ that contains these projections in the rows, being of compatible size with the data matrix $\mathbf{X}$.

The spatial median $\mathbf{s}$ with the so-called available data strategy can be obtained as the solution of the projected Weber problem

$$\min_{\mathbf{v}\in\mathbf{R}^n} \mathcal{J}(\mathbf{v}), \quad \text{where } \mathcal{J}(\mathbf{v}) = \sum_{i=1}^{n_j}\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{x}_i - \mathbf{v})\|. \tag{9}$$

As described in [16], this optimization problem is nonsmooth, that is, it is not classically differentiable at zero. Instead, the so-called subgradient of $\mathcal{J}(\mathbf{v})$ always exists and is characterized by the condition

$$\partial\mathcal{J}(\mathbf{v}) = \sum_{i=1}^{N} \boldsymbol{\xi}_i \text{ for } \begin{cases} (\boldsymbol{\xi}_i)_j = \dfrac{\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{v}-\mathbf{x}_i)_j}{\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{v}-\mathbf{x}_i)\|}, \text{if } \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{u}-\mathbf{x}_i)\| \neq 0, \\ \|\boldsymbol{\xi}_i\| \leq \ 1, \text{ when } \|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{u}-\mathbf{x}_i)\| = 0. \end{cases} \tag{10}$$

Then, the minimizer $\mathbf{s}$ of (9) satisfies $\mathbf{0} \in \partial\mathcal{J}(\mathbf{s})$. In [20] it is shown, for the complete data case, that if the sample $\{\mathbf{x}_i\}$ belongs to a Euclidean space and is not concentrated on a line, the spatial median $\mathbf{s}$ is unique. In practice (see [2]), one can obtain an accurate approximation for the solution of the nonsmooth problem by solving the following equation corresponding to the regularized form

$$\sum_{i=1}^{N} \frac{\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\max\{\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|, \varepsilon\}} = \mathbf{0} \quad \text{for } \varepsilon > 0. \tag{11}$$

This can be solved using the SOR (Sequential Overrelaxation) algorithm [2] with the overrelaxation parameter $\omega = 1.5$. For simplicity, define $\|\mathbf{v}\|_\varepsilon = \max\{\|\mathbf{v}\|, \varepsilon\}$.

To this end, the comparison of (7) and (11) allows us to define the *robust covariance matrix* corresponding to the spatial median s:

$$\Sigma_R = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|_\varepsilon} \right) \left(\frac{\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|_\varepsilon} \right)^T. \tag{12}$$

This form can be referred to as the *multivariate sign covariance matrix* [5, 28, 7]. By construction, the nonzero covariate vectors have a unit length, so that they only accumulate the deviations of angles and not the sizes of the available variables. Such an observation is related to one perspective on statistical robustness that can be formalized using the so-called influence function [9]. Using Σ_R as the sample covariance matrix, one can, by again solving the corresponding eigenvalue problem (5), recover a new basis $\{\mathbf{u}_k\}$ for which the corresponding eigenvalues $\{\lambda_k\}$, again, explain the amount of variability along the new coordinates. Because Σ_R is based on the first-order approximation, the nonnegative eigenvalues readily correspond to the geometric variability represented by the standard deviation in the second-order statistics, and, then, we do not need to take any square roots when computing the relative importances of the robust procedure as in Proposition 1. Hence, the two PCA approaches are comparable to each other.

2.3 Projection using PCA-based transformation

In the matrix form, the existence of a new basis in the columns of the given unitary matrix $\mathbf{U}$, and given a complete location estimate for the sparse data $\mathbf{s} \in \mathbf{R}^n$ (i.e., the spatial median), for which we define the corresponding matrix $\mathbf{S} \in \mathbf{R}^{N\times n}$ by replication of $\mathbf{s}^T$ in N rows, yields the transformed data matrix

$$\mathbf{Y} = (\mathbf{P} \circ (\mathbf{X} - \mathbf{S}))\,\mathbf{U}, \tag{13}$$

where $\circ$ denotes the Hadamard product. When $\mathbf{U}$ is ordered based on RI_k's, the dimension reduction is obtained by selecting only m of the n coordinates (columns) in $\mathbf{Y}$. Hence, we see that even if there are missing values in the original data, the resulting new data vectors become complete. We also know from the basic linear algebra that, for complete data, both the length of the original vectors and the angle between any two vectors are preserved in (13) because $\mathbf{U}$ is unitary. However, in the case of missing data, some of the coordinate values of the original vectors are not present, and then, presumably, the transformed vectors in $\mathbf{Y}$ are of smaller length, i.e., closer to the origin in the transformed space. Moreover, the angles might also become degraded. These simple observations readily raise some doubts concerning the available data strategy in the form of incomplete data vectors as proposed in (12).

2.4 Two modifications of the robust PCA procedure

Let us define two modifications of the robust PCA procedure that are based on the similar form of the covariance matrix as defined in (12). As discussed above, both the amount of variability of data and/or the main directions of variability might be underestimated due to sparse data vectors, that is, missing coordinate values. Our suggested modifications are both based on a simple idea: use only the "almost complete" data in estimation (cf. the cascadic initializations of robust clustering in [24, 25]). Note that this is one step further than the typical way of using only the complete pairs or complete observations in the computation of a covariance matrix.

The first suggested modification, for the computation of the relative importances of the principal components, is related to using the actual projections along the new coordinate axis for this purpose. Similar to the alpha-trimmed mean [3], which presumably neglects outlying observations, we use (see the tests in [26]) the 10% and 90% percentiles, denoted as $\mathrm{prc}_{10}(\cdot)$ and $\mathrm{prc}_{90}(\cdot)$, related to the transformed data matrix $\mathbf{Y}$ in (13). Namely, for the each new variable $\{y_k\}$, its estimated variability is computed as

$$RI_k = 100(\mathrm{prc}_{90}(\{y_k\}) - \mathrm{prc}_{10}(\{y_k\})). \tag{14}$$

Moreover, because it is precisely the sparsity that diminishes the lengths and angles of the transformed data vectors, we restrict the computation of (14) to that subset of the original data, where at most one variable is missing from an observation $\mathbf{x}_i$. This subset satisfies $\sum_{j=1}^{n}(\mathbf{p}_i)_j \geq n-1$.

Our second suggested modification uses a similar approach, but already directly for the robust covariance matrix (12), by taking into account only those observations of which at most one variable is missing. Hence, we define the following subsets of the original set of indices $\mathcal{N} = \{1, 2, \ldots, N\}$:

$$I_c = \{i \in \mathcal{N} \mid \mathbf{x}_i \text{ is complete}\},$$
$$I_j = \{i \in \mathcal{N} \mid \text{variable j is missing from } \mathbf{x}_\mathrm{i}\}.$$

We propose computing a reduced, robust covariance matrix $\widetilde{\Sigma}_R$ as

$$\widetilde{\Sigma}_R = \frac{1}{\widetilde{N}-1}\left(\sum_{i\in I_c}\mathbf{v}_i\mathbf{v}_i^T + \sum_{j=1}^{n}\sum_{i\in I_j}\mathbf{v}_i\mathbf{v}_i^T\right), \quad \mathbf{v}_i = \frac{\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)}{\|\mathrm{Diag}\{\mathbf{p}_i\}(\mathbf{s}-\mathbf{x}_i)\|_\varepsilon},$$

with $\widetilde{N} = |I_c| + \sum_j |I_j|$. Hence, only that part of the first-order covariability that corresponds to the almost complete observations is used.

3 Computational results

Computational experiments in the form of simulated test cases, when knowing the target result, are given first. The parametrized test is introduced in Section 3.1, and the computational results for the different procedures are provided in Section 3.2. Finally, we apply the best methods to analyze the educational data of PISA in Section 3.3. As a reference method related to the classical, second-order statistics as derived in Section 2.1, with sparse data, we use the Matlab's PCA routine with the '*pairwise*' option.

3.1 The simulated test cases

For simplicity, we fix the number of observations as $N = 1000$. For the fixed size of an observation n, let us define a vector of predetermined standard deviations as $\boldsymbol{\sigma} = \begin{bmatrix}\sigma_1 & \sigma_2 & \dots & \sigma_n\end{bmatrix}$. Moreover, let $\mathbf{R}_{a,b}(\theta) \in \mathbf{R}^{n\times n}$ be an orthonormal (clockwise) rotation matrix of the form

$$\mathbf{R}_{ab}(\theta) = \{\mathbf{M} = \mathbf{I}_n \wedge \mathbf{M}_{aa} = \mathbf{M}_{bb} = \cos(\theta),\ \mathbf{M}_{ab} = -\mathbf{M}_{ba} = -\sin(\theta)\},$$

where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. Then, the simulated data $\{\mathbf{d}_i\}_{i=1}^N$ is generated as

$$\begin{aligned}\mathbf{d}_i^T \sim & \frac{\boldsymbol{\sigma}}{2} + \left[\mathcal{N}(0,\sigma_1)\,\mathcal{N}(0,\sigma_2)\dots\mathcal{N}(0,\sigma_n)\right] \\ & + \eta_i\left[\mathbf{R}_n\left[\mathcal{U}([-\sigma_1,\sigma_1])\,\mathcal{U}([-\sigma_2,\sigma_2])\dots\mathcal{U}([-\sigma_n,\sigma_n])\right]^T\right]^T,\end{aligned} \tag{15}$$

where $\mathcal{N}(0,\sigma)$ denotes the zero-mean normal distribution with standard deviation σ and $\mathcal{U}([-c,c])$ the uniform distribution on the interval $[-c,c]$, respectively. $\mathbf{R}_n$ defines the n-dimensional rotation that we use to orientate the latter noise term in (15) along the diagonal of the hypercube, that is, we always choose $\theta = \frac{\pi}{4}$ and take, for the actual tests in $2D, 3D, 4D$, and $6D$,

$$\begin{aligned}&\mathbf{R}_2 = \mathbf{R}_{12}(\theta),\quad \mathbf{R}_3 = \mathbf{R}_{23}(\theta)\mathbf{R}_{12}(\theta),\quad \mathbf{R}_4 = \mathbf{R}_{14}(\theta)\mathbf{R}_{23}(\theta)\mathbf{R}_{34}(\theta)\mathbf{R}_{12}(\theta),\\ &\mathbf{R}_6 = \mathbf{R}_{36}(\theta)\mathbf{R}_{45}(\theta)\mathbf{R}_{56}(\theta)\mathbf{R}_{14}(\theta)\mathbf{R}_{23}(\theta)\mathbf{R}_{34}(\theta)\mathbf{R}_{12}(\theta).\end{aligned}$$

Finally, a random sparsity pattern of a given percentage of missing values represented by the matrix $\mathbf{P}$ as defined in (8) is attached to data.

To conclude, the simulated data are parametrized by the vector $\boldsymbol{\sigma}$, which defines the true data variability. Moreover, the target directions of the principal components are just the original unit vectors $\mathbf{e}_k,\ k = 1, \ldots, n$. Their estimation is disturbed by the noise, which comes from the uniform distribution whose width coordinatewise coincides with the clean data. Because the noise is rotated towards the diagonal of the hypercube, its maximal effect is characterized by $\frac{\max_k \sigma_k}{\min_k \sigma_k}$. By choosing σ_k's as the powers of two and three for $n = 2, 3, 4, 6$, we are then gradually increasing the effect of the error when the dimension of the data is increasing. Finally, we fix the amount of noise to 10% so that $\eta_i = 1$ with a probability of 0.1 in (15). In this way, testing up to 40% of missing values randomly attached to $\{\mathbf{d}_i\}$ will always contain less than 50% of the degradations (missing values and/or noise) as a whole.

3.2 Results for the simulated tests

The test data generation was repeated 10 times, and the means and standard deviations (in parentheses) over these are reported. As the error measure for the directions of $\{\mathbf{u}_k\}$, we use their deviation from being parallel to the target unit vectors. Hence, we take $\mathrm{DirE} = \max_k\{1 - |\mathbf{u}_k^T \mathbf{e}_k|\},\ k = 1, \ldots, n$, such that $\mathrm{DirE} \in [0, 1]$. In the result tables below, we report the relative importances of RI_k in the order of their importance. 'Clas' refers to the classical PCA, 'Rob' to the original robust formulation, 'RobP' to the modification using percentiles for the importances, and 'RobR' to the use of the reduced covariance matrix $\tilde{\Sigma}_R$. The real relative importances ('True') by generation are provided in the third column.

From all simulated tests (Tables 1-4), we see that the the classical method and 'RobP' show the closest relative importances of the principal components to the true geometric variability in the data. Moreover, both of these approaches show a very stable behavior, and the results for the relative importances do not change that much, even when a high number of missing data is present. The results for the other two approaches, the basic robust and 'RobR', on the other hand, are much less stable, and particularly the basic robust procedure starts to underestimate the relative importances of the major components when the amount of missing data increases.

The directions remain stable for all the simulated test cases, even when a large amount of missing data is present. Over all the simulated tests, the 'RobP' with the original robust covariance bears the closest resemblance to the true directions. It can tolerate more noise compared to 'Clas', as shown in Table 3. We also conclude that the missing data do not affect the results of the PCA procedures as much as the noise. Tables 3 and 4 show that, for a large noise, the increase in sparsity can actually improve the performance of the robust method because it decreases the absolute number of noisy observations. Interestingly, as can be seen from Table 4, the geometric variability was estimated accurately, even if the directions were wrong.

Table 1. Results for $\boldsymbol{\sigma} = [3\ 1]$

Missing	**PC**	**True(Std)**	**Clas(Std)**	**Rob(Std)**	**RobP(Std)**	**RobR(Std)**
0%	1	75.0(0.00)	73.7(0.8)	73.0(1.1)	73.0(1.3)	73.0(1.1)
	2	25.0(0.00)	26.3(0.8)	27.0(1.1)	27.0(1.3)	27.0(1.1)
	DirE	-	0.001	0.004		0.004
10%	1	75.0(0.00)	73.9(0.9)	68.9(1.2)	73.2(1.3)	68.9(1.2)
	2	25.0(0.00)	26.1(0.9)	31.1(1.2)	26.8(1.3)	31.1(1.2)
	DirE	-	0.001	0.005		0.005
20%	1	75.0(0.00)	73.5(1.2)	65.1(1.0)	72.5(1.7)	65.1(1.0)
	2	25.0(0.00)	26.5(1.2)	34.9(1.0)	27.5(1.7)	34.9(1.0)
	DirE	-	0.001	0.009		0.009
30%	1	75.0(0.00)	73.8(1.0)	62.4(0.9)	73.0(1.4)	62.4(0.9)
	2	25.0(0.00)	26.2(1.0)	37.6(0.9)	27.0(1.4)	37.6(0.9)
	DirE	-	0.001	0.003		0.003
40%	1	75.0(0.00)	74.0(0.8)	60.3(1.6)	73.1(1.4)	60.3(1.6)
	2	25.0(0.00)	26.0(0.8)	39.7(1.6)	26.9(1.4)	39.7(1.6)
	DirE	-	0.002	0.008		0.008

3.3 Results for PISA data set

Next, we apply the different PCA methods tested in the previous section to a large educational data set, namely the latest data from the Programme for International Student Assessment[1] (PISA 2012). The data contain 485490 observations, and as variables we use the 15 scale indices [24] that are known to explain the student performance in mathematics, the main assessment area in PISA 2012. The scale indices are derived variables that summarize information from student background questionnaires [22], and are scaled so that their mean is zero with a standard deviation of one. Due to the rotated design of PISA (each student answers only one of the three different background questionnaires), this data set has 33.24% of missing data by design, a special case of MCAR.

In Table 5, the relative importances $\{RI_k\}$ are depicted. The table also shows the variance-based view for the classical method, denoted as 'ClsVar'. As can be seen from the table, the first principal component is much higher for 'ClsVar' than for the other approaches. In consequence, fewer principal components would be selected with 'ClsVar' when a certain threshold of how much the principal components should account for is given. As illustrated in Fig. 1, if the threshold is set to 90%, we would select 11 components with 'ClsVar' but 13 for both the classical PCA and for the 'RobP'.

In Fig. 2, the loadings of the first two principal components are visualized for the classical and for the robust version. We see that for both versions, the three scale indices ANXMAT, FAILMAT, and ESCS are the most distinct from the others. However, the robust version is able to distinguish this finding more clearly. That *index of economic, social and cultural status* (ESCS) accounts for

[1] Available at `http://www.oecd.org/pisa/pisaproducts/`.

Table 2. Results for $\boldsymbol{\sigma} = [4\ 2\ 1]$

Missing	**PC**	**True(Std)**	**Clas(Std)**	**Rob(Std)**	**RobP(Std)**	**RobR(Std)**
0%	1	57.1(0.00)	56.3(0.7)	58.6(1.3)	55.8(1.0)	58.6(1.3)
	2	28.6(0.00)	28.6(0.8)	28.9(1.2)	28.7(1.0)	28.9(1.2)
	3	14.3(0.00)	15.2(0.3)	12.6(0.4)	15.5(0.4)	12.6(0.4)
	DirE	-	0.005	0.017		0.017
10%	1	57.1(0.00)	56.3(0.8)	55.7(1.3)	55.8(1.0)	54.5(1.6)
	2	28.6(0.00)	28.6(0.9)	30.0(1.2)	28.6(0.9)	30.7(1.2)
	3	14.3(0.00)	15.1(0.4)	14.3(0.6)	15.5(0.5)	14.8(0.8)
	DirE	-	0.008	0.015		0.015
20%	1	57.1(0.00)	56.2(0.8)	51.7(1.4)	55.6(1.1)	51.7(1.4)
	2	28.6(0.00)	28.6(0.9)	30.7(1.2)	28.8(1.3)	31.5(1.5)
	3	14.3(0.00)	15.2(0.3)	17.6(0.8)	15.6(0.5)	16.7(0.7)
	DirE	-	0.005	0.020		0.014
30%	1	57.1(0.00)	56.0(0.7)	49.2(0.7)	55.3(0.9)	50.9(0.7)
	2	28.6(0.00)	28.8(0.8)	31.6(0.8)	29.0(0.9)	32.1(1.3)
	3	14.3(0.00)	15.2(0.4)	19.2(1.0)	15.7(0.5)	17.1(1.3)
	DirE	-	0.006	0.013		0.012
40%	1	57.1(0.00)	56.2(0.9)	46.2(1.4)	55.8(1.2)	49.9(1.6)
	2	28.6(0.00)	28.7(1.2)	32.0(1.7)	28.5(1.3)	32.5(1.7)
	3	14.3(0.00)	15.1(0.4)	21.8(1.1)	15.6(0.7)	17.6(1.0)
	DirE	-	0.010	0.014		0.013

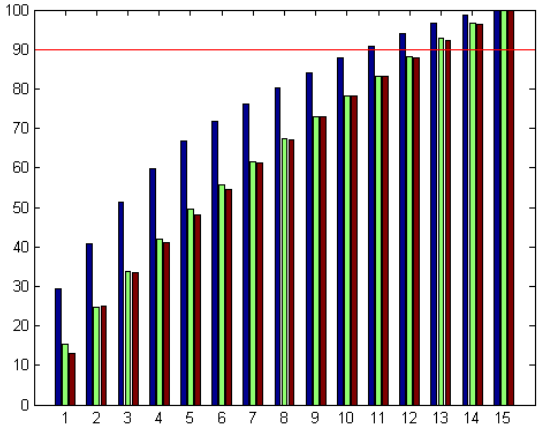

Fig. 1. Cumulative sum of the relative importances for the classical PCA using variance, the classical PCA, and the robust PCA using percentiles (from left to right).

much of the variability in the data, being the "strongest single factor associated with performance in PISA" [21], is always highlighted in PISA documentations and can be clearly seen in Fig. 2, especially from the robust PC 1.

Table 3. Results for $\boldsymbol{\sigma} = [27\ 9\ 3\ 1]$

Missing	**PC**	**True(Std)**	**Clas(Std)**	**Rob(Std)**	**RobP(Std)**	**RobR(Std)**
0%	1	67.5(0.00)	62.5(0.8)	66.9(0.9)	63.6(1.1)	66.9(0.9)
	2	22.5(0.00)	22.1(0.6)	23.6(0.8)	22.9(0.8)	23.6(0.8)
	3	7.5(0.00)	10.1(0.2)	6.9(0.4)	9.1(0.4)	6.9(0.4)
	4	2.5(0.00)	5.3(0.2)	2.6(0.2)	4.5(0.2)	2.6(0.2)
	DirE	-	0.168	0.080		0.080
10%	1	67.5(0.00)	62.5(0.8)	62.3(1.3)	64.0(1.2)	60.3(2.3)
	2	22.5(0.00)	22.1(0.6)	25.7(0.9)	23.0(0.9)	26.9(1.7)
	3	7.5(0.00)	10.0(0.2)	8.6(0.5)	9.0(0.4)	9.2(0.6)
	4	2.5(0.00)	5.3(0.2)	3.5(0.3)	4.0(0.2)	3.6(0.4)
	DirE	-	0.157	0.045		0.046
20%	1	67.5(0.00)	62.5(1.0)	56.9(1.2)	64.2(1.2)	58.3(1.7)
	2	22.5(0.00)	22.1(0.7)	27.4(1.0)	22.9(0.9)	28.1(1.1)
	3	7.5(0.00)	10.1(0.3)	11.0(0.6)	8.9(0.5)	9.8(1.0)
	4	2.5(0.00)	5.4(0.3)	4.7(0.4)	3.9(0.2)	3.7(0.4)
	DirE	-	0.164	0.032		0.031
30%	1	67.5(0.00)	62.7(0.8)	52.1(1.6)	64.4(0.9)	57.7(1.3)
	2	22.5(0.00)	22.0(0.6)	28.2(1.4)	23.2(0.6)	28.2(1.5)
	3	7.5(0.00)	10.0(0.4)	13.2(1.0)	8.6(0.4)	10.2(0.5)
	4	2.5(0.00)	5.3(0.3)	6.5(0.5)	3.8(0.2)	3.9(0.4)
	DirE	-	0.177	0.023		0.038
40%	1	67.5(0.00)	62.7(0.8)	46.9(0.8)	64.2(1.4)	55.9(1.2)
	2	22.5(0.00)	22.2(0.6)	28.7(1.0)	23.5(1.3)	29.8(1.5)
	3	7.5(0.00)	9.9(0.3)	15.7(0.5)	8.8(0.3)	10.8(1.1)
	4	2.5(0.00)	5.2(0.3)	8.6(0.8)	3.6(0.4)	3.5(0.5)
	DirE	-	0.189	0.016		0.040

4 Conclusions

Although PCA is one of the most widely used ML and DM techniques, systematic testing and assessment of PCA in the presence of missing data seem to still be an important topic to study. In this article, we have proposed a robust PCA method and two modifications (one using percentiles for the importance and one with a reduced covariance matrix) of this method. The testing of these three approaches was done in comparison with the classical, reference PCA for sparse data. First, we illustrated the results for carefully designed simulated data and then for a large, real educational data set.

From the simulated tests, we concluded that the percentiles-based robust method and the classical PCA showed the best results, especially when the relative importance of the principal components were compared with the true variability of the data. The basic robust approach started to underestimate the relative importance of the major components when the amount of missing data increased. The results of the simulated tests were stable, and the variance be-

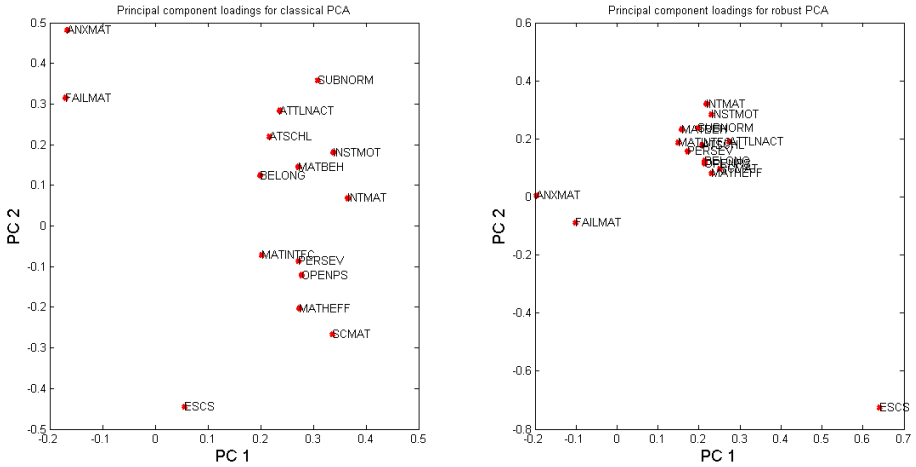

Fig. 2. Principal component loadings for PISA data for the classical (left) and robust (right) approaches.

tween repeated test runs was very small. Likewise, the estimated directions remained also stable even with a large amount of missing data. Tests with PISA data showed that the proposed robust methods are applicable for large, real data sets with one-third of the values missing, where the interpretation of the robust result yielded clearer known discrimination of the original variables compared to the classical PCA.

The classical PCA uses variance to estimate the importance of the principal components, which highlights (as demonstrated in Table 5 and Figure 1) the major components. As shown by the simulated results, it is more prone to nongaussian errors in the data. These points might explain some of the difficulties the classical method faced in applications [23]. In [14], seven distinctions of the PCA problem in the presence of missing values were listed: 1) no analytical solution since even the estimation of the data covariance matrix is nontrivial, 2) the optimized cost function typically has multiple local minima, 3) no analytical solution even for the location estimate, 4) standard approaches can lead to overfitting, 5) algorithms may require heavy computations, 6) the concept of the PCA basis in the principal subspace is not easily generalized, and 7) the choice of the dimensionality of the principal subspace is more difficult than in classical PCA. We conclude that the proposed robust methods successfully addressed all these distinctions: 1) well-defined covariance matrix, 2) being positive semidefinite, 3) a unique location estimate in the form of the spatial median, 4) resistance to noise due to robustness, 5) the same linear algebra as in the classical approach, and 6)–7) a geometrically consistent definition of the principal subspace and its dimension related to the data variability.

Acknowledgments. The authors would like to thank Professor Tuomo Rossi for many helpful discussions on the contents of the paper.

Table 4. Results for $\boldsymbol{\sigma} = [32\ 16\ 8\ 4\ 2\ 1]$

Missing	**PC**	**True(Std)**	**Clas(Std)**	**Rob(Std)**	**RobP(Std)**	**RobR(Std)**
0%	1	50.8(0.00)	48.1(0.5)	55.3(1.0)	48.6(0.8)	55.3(1.0)
	2	25.4(0.00)	24.3(0.4)	26.7(0.6)	24.4(0.4)	26.7(0.6)
	3	12.7(0.00)	12.6(0.2)	10.6(0.4)	12.7(0.3)	10.6(0.4)
	4	6.3(0.00)	7.5(0.2)	4.7(0.3)	7.1(0.2)	4.7(0.3)
	5	3.2(0.00)	4.4(0.1)	1.6(0.1)	4.3(0.1)	1.6(0.1)
	6	1.6(0.00)	3.2(0.1)	1.0(0.1)	2.8(0.2)	1.0(0.1)
	DirE	-	0.298	0.374		0.374
10%	1	50.8(0.00)	48.0(0.6)	51.6(1.1)	48.5(1.1)	51.0(1.9)
	2	25.4(0.00)	24.3(0.5)	27.5(0.8)	24.8(0.6)	28.1(1.1)
	3	12.7(0.00)	12.6(0.2)	12.0(0.5)	12.8(0.4)	12.0(0.9)
	4	6.3(0.00)	7.5(0.2)	5.5(0.2)	7.0(0.2)	5.5(0.4)
	5	3.2(0.00)	4.4(0.1)	2.2(0.2)	4.2(0.1)	2.2(0.2)
	6	1.6(0.00)	3.2(0.2)	1.3(0.1)	2.7(0.2)	1.3(0.2)
	DirE	-	0.318	0.277		0.358
20%	1	50.8(0.00)	48.2(0.5)	48.6(1.0)	48.9(1.6)	51.3(1.4)
	2	25.4(0.00)	24.2(0.5)	27.3(0.8)	24.6(0.6)	27.3(0.7)
	3	12.7(0.00)	12.7(0.3)	13.2(0.8)	13.0(0.6)	12.3(1.2)
	4	6.3(0.00)	7.4(0.2)	6.4(0.3)	7.0(0.3)	5.5(0.6)
	5	3.2(0.00)	4.4(0.2)	2.7(0.3)	4.0(0.2)	2.2(0.2)
	6	1.6(0.00)	3.2(0.2)	1.7(0.2)	2.4(0.2)	1.3(0.3)
	DirE	-	0.372	0.090		0.137
30%	1	50.8(0.00)	48.1(0.6)	43.8(1.2)	48.6(1.4)	49.4(2.4)
	2	25.4(0.00)	24.3(0.5)	27.5(0.8)	25.0(0.8)	28.5(1.7)
	3	12.7(0.00)	12.6(0.1)	15.0(0.6)	12.9(0.5)	12.5(0.8)
	4	6.3(0.00)	7.5(0.2)	7.6(0.5)	7.1(0.4)	5.8(0.8)
	5	3.2(0.00)	4.3(0.1)	3.8(0.5)	4.0(0.2)	2.2(0.2)
	6	1.6(0.00)	3.2(0.2)	2.3(0.3)	2.4(0.2)	1.5(0.4)
	DirE	-	0.335	0.092		0.468
40%	1	50.8(0.00)	48.0(0.6)	39.7(1.5)	48.3(1.7)	50.2(2.9)
	2	25.4(0.00)	24.3(0.4)	26.6(1.0)	25.1(1.1)	28.3(2.4)
	3	12.7(0.00)	12.6(0.3)	15.8(1.0)	13.0(0.7)	11.9(1.3)
	4	6.3(0.00)	7.5(0.2)	9.5(0.8)	7.3(0.3)	6.0(0.8)
	5	3.2(0.00)	4.4(0.3)	5.1(0.5)	3.9(0.3)	2.2(0.3)
	6	1.6(0.00)	3.1(0.2)	3.3(0.4)	2.3(0.2)	1.3(0.2)
	DirE	-	0.516	0.078		0.518

Table 5. Results for PISA data

	RI_1	RI_2	RI_3	RI_4	RI_5	RI_6	RI_7	RI_8	RI_9	RI_{10}	RI_{11}	RI_{12}	RI_{13}	RI_{14}	RI_{15}
ClsVar	29.5	11.4	10.4	8.6	6.8	5.0	4.4	4.1	3.8	3.7	3.2	3.0	2.8	2.0	1.3
Cls	15.3	9.5	9.1	8.3	7.3	6.3	5.9	5.7	5.5	5.4	5.0	4.8	4.7	4.0	3.3
RobP	13.1	11.9	8.6	7.5	7.2	6.5	6.5	5.9	5.9	5.2	4.8	4.8	4.5	3.9	3.7

Bibliography

[1] E. Alpaydin. *Introduction to Machine Learning*. The MIT Press, Cambridge, MA, USA, 2nd edition, 2010.

[2] S. Äyrämö. *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing*. University of Jyväskylä, 2006.

[3] J. Bednar and T. Watt. Alpha-trimmed means and their relationship to median filters. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(1):145–153, 1984.

[4] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[5] C. Croux, E. Ollila, and H. Oja. Sign and rank covariance matrices: statistical properties and application to principal components analysis. In *Statistical data analysis based on the L1-norm and related methods*, pages 257–269. Springer, 2002.

[6] A. d'Aspremont, F. Bach, and L. E. Ghaoui. Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research*, 9:1269–1294, 2008.

[7] D. Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600, 2008.

[8] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[9] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

[10] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[11] T. P. Hettmansperger and J. W. McKean. *Robust nonparametric statistical methods*. Edward Arnold, London, 1998.

[12] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[13] P. J. Huber. *Robust Statistics*. John Wiley & Sons Inc., New York, 1981.

[14] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000, 2010.

[15] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[16] T. Kärkkäinen and E. Heikkola. Robust formulations for training multilayer perceptrons. *Neural Computation*, 16:837–862, 2004.

[17] T. Kärkkäinen and J. Toivanen. Building blocks for odd-even multigrid with applications to reduced systems. *Journal of Computational and Applied Mathematics*, 131:15–33, 2001.

[18] R.J.A. Little and D.B. Rubin. *Statistical analysis with missing data*, volume 4. Wiley New York, 1987.

[19] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, K.L. Cohen, G. Boente, R. Fraiman, B. Brumback, C. Croux, et al. Robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

[20] P. Milasevic and G. R. Ducharme. Uniqueness of the spatial median. *Ann. Statist.*, 15(3):1332–1333, 1987.

[21] OECD. *PISA Data Analysis Manual: SPSS and SAS, Second Edition*. OECD Publishing, 2009.

[22] OECD. *PISA 2012 Results: Ready to Learn - Students' Engagement, Drive and Self-Beliefs (Volume III)*. PISA, OECD Publishing, 2013.

[23] H. Ringberg, A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. In *ACM SIGMETRICS Performance Evaluation Review*, volume 35, pages 109–120. ACM, 2007.

[24] M. Saarela and T. Kärkkäinen. Discovering gender-specific knowledge from Finnish basic education using PISA scale indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 60–68, 2014.

[25] M. Saarela and T. Kärkkäinen. Analysing student performance using sparse data of core bachelor courses. *To appear in JEDM-Journal of Educational Data Mining*, 2015.

[26] S. M. Stigler. Do robust estimators work with real data? *The Annals of Statistics*, pages 1055–1098, 1977.

[27] J. R. Van Ginkel, P. M. Kroonenberg, and H. A. Kiers. Missing data in principal component analysis of questionnaire data: a comparison of methods. *Journal of Statistical Computation and Simulation*, (ahead-of-print):1–18, 2013.

[28] S. Visuri, V. Koivunen, and H. Oja. Sign and rank covariance matrices. *Journal of Statistical Planning and Inference*, 91(2):557–575, 2000.