A Data-Adaptive Principal Component Analysis

Yaeji Lim and Hee-Seok Oh

Department of Statistics, Seoul National University, Seoul, Korea

Corresponding author: Hee-Seok Oh, E-mail: heeseok.oh@gmail.com

Abstract

This paper studies a data-adaptive principal component analysis (PCA) that does not require prior information of data distribution. The ordinary PCA is useful for dimension reduction and for identifying important features of data that are consist of a large number of interrelated variables. However, it is stringent to the Gaussian assumption of the data, and therefore may not be efficient for analyzing real observations that may be non-Gaussian distributed, such as skewed or heavy-tailed data. To extend the scope of PCA to non-Gaussian distributed data, a new approach for PCA is proposed. The core of the methodology is the use of a composite quantile, which is a weighted linear combination of convex loss functions instead of the square loss function, and the weights are determined data-adaptively. In addition, a practical algorithm to implement the data-adaptive PCA is derived. Moreover, a penalized version of the proposed composite quantile PCA with a penalty is considered. Results from a simulation study and a real data analysis demonstrate the promising empirical properties of the proposed approach.

Keywords: High-dimensional data; Principal component analysis; Pseudo data; Quantile estimation; Robustness.

1 Introduction

In multivariate data analysis, principal component analysis (PCA) has been most widely used for dimension reduction and feature extraction of high-dimensional data. It is essentially based on the computation of eigenvalues and eigenvectors of the sample covariance or correlation matrix. Therefore, the results may be extremely sensitive to the presence of outliers. Taking a different point of view, geometrically, PCA finds a subspace with smaller dimensions obtained from a linear combination of variables that minimizes the mean squared orthogonal distances of the data points. Thus, PCA induced by this approach is depend on the Gaussian assumption of the data.

Figure 1 shows the density functions of maximum precipitation data in August of four selected years. We focus on the East Asia region that covers 30–50°N and 120–140°E. As shown in the figure, the maximum precipitations do not follow only Gaussian distribution. The data are rather positively skewed or bimodal, and this feature cannot be fully reflected by ordinary PCA. Thus, it is necessary to develop a new PCA method that can accommodate these abnormal distributions.

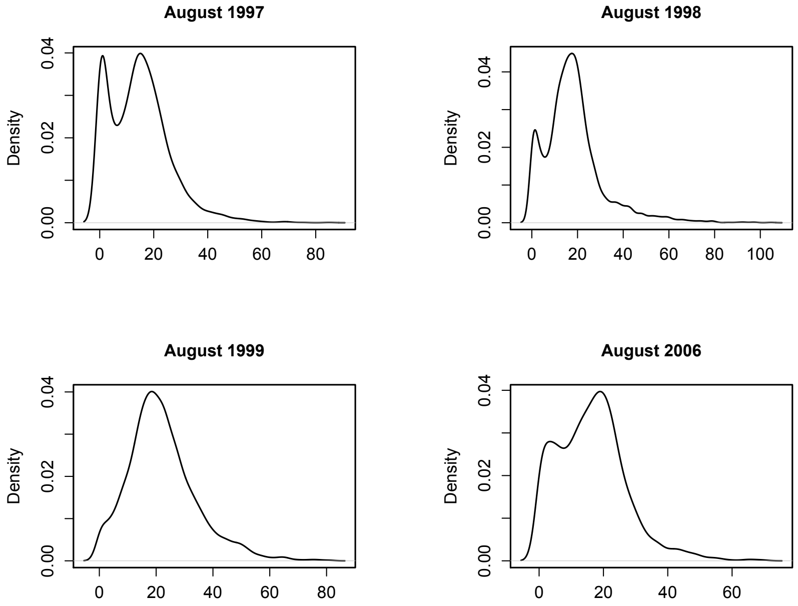

Figure 1: Density functions of maximum daily precipitations in August of four selected years.

In this paper, we propose a new data-adaptive PCA that does not require prior information of data distribution, which implies that it automatically reflects the distributional features of data. The key components of the proposed method are two-folds: (1) the PCA can be expressed as a least squares framework, and (2) the quadratic loss function is replaced with a weighted linear combination of convex loss functions, termed "composite quantile functions". The weight is set to reflect the heavy-tailed, bimodal or skewed form of the distribution of the data, and hence, it can be expected that the proposed method works well for various types of data.

We also develop an efficient computational algorithm for the proposed PCA. To that end, we employ the concept of pseudo data by Oh *et al.* (2007). The pseudo data transforms the composite quantile non-linear setting into a conventional least squares framework in which theoretical properties and efficient algorithms have been well developed.

2 Composite Quantile PCA

Let $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ be a p-dimensional random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma_{\boldsymbol{x}\boldsymbol{x}}$. We consider to find a linear approximation to represent the data $\boldsymbol{x}$ as

$$\boldsymbol{g}(\boldsymbol{\xi}) = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\xi}, \tag{1}$$

where $\boldsymbol{A}$ is a $p \times q$ matrix and $\boldsymbol{\xi}$ is a q-dimensional vector of parameters, which is a linear projection given by

$$\boldsymbol{\xi} = \boldsymbol{B}\boldsymbol{x}$$

with a $q \times p$ weight matrix $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q)^T$ $(q \leq p)$. We now consider the following least squares criterion for a solution of PCA,

$$\min_{\boldsymbol{\mu}, \mathbf{A}, \boldsymbol{\xi}} \mathrm{E}\big\{(\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\xi})^T(\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\xi})\big\}. \tag{2}$$

For a sample version of (2), given the observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we would like to fit the model (1) to the data by minimizing the squares amounts of reconstruction errors, that is,

$$\min_{\boldsymbol{\mu}, \mathbf{A}, \boldsymbol{\xi}_j} \sum_{j=1}^{n} \|\boldsymbol{x}_j - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\xi}_j\|^2. \tag{3}$$

The solutions for the above optimization can be summarized as

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \bar{\boldsymbol{x}}, \\ \hat{\boldsymbol{A}} &= \boldsymbol{V}_q = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q) = \hat{\boldsymbol{B}}^T, \quad \text{and} \\ \hat{\boldsymbol{\xi}}_i &= \boldsymbol{V}_q^T(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}), \end{aligned}$$

where $\boldsymbol{v}_j$ is the p-dimensional eigenvector associated with the jth largest eigenvalue of $\Sigma_{\boldsymbol{x}\boldsymbol{x}}$. Hence, the best rank-$q$ approximation to the data is given by

$$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \boldsymbol{V}_q\boldsymbol{V}_q^T(\boldsymbol{x} - \bar{\boldsymbol{x}}). \tag{4}$$

An interpretation of (4) is that $\boldsymbol{V}_q\boldsymbol{V}_q^T$ is a projection matrix that maps $\boldsymbol{x}$ onto the rank q subspace with minimum squared reconstruction error.

In this paper, along with the above regression-type optimization view of PCA, we propose a new PCA that produces data-adaptive loadings ("regression coefficients"), which reflect the data distribution well. For this purpose, we adopt a weighted linear combination of quantile loss functions that replaces the square loss function of (2), and consider the following optimization problem,

$$\min_{\boldsymbol{\mu}, \mathbf{A}, \boldsymbol{\xi}} \mathrm{E} \sum_{k=1}^{K} w_k \rho_k \|\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\xi}\| \quad \text{subject to} \quad \boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}_q, \tag{5}$$

where $\rho_k\|\boldsymbol{u}\| := \sum_{\ell=1}^{p} \rho_k(u_\ell)$ with $\boldsymbol{u} = (u_1, \ldots, u_p)^T$, the check function defined as $\rho_k(u) := u\{\tau_k - I(u < 0)\}$ with K different quantiles, $0 < \tau_1 < \cdots < \tau_K < 1$, and the weights $w_1, \ldots, w_K$ are positive constants, sum up to 1 ($\sum w_k = 1$), which control the contribution of quantile functions. Employing the composite quantile functions is capable of reflecting distributional features of data such as skewness as well as bounding outliers effectively, compared to using a single quantile function.

We now consider an empirical version of (5) with n independent random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Then, the data-adaptive loadings of q principal components can be expressed as the solution of

$$\min_{\boldsymbol{\mu}, \mathbf{A}, \boldsymbol{\xi}_j} \sum_{j=1}^{n} \sum_{k=1}^{K} w_{j,k} \rho_k \|\boldsymbol{x}_j - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\xi}_j\| \quad \text{subject to} \quad \boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}_q. \tag{6}$$

The ideal weight vector is one that gives high values on the meaningful quantiles and near zeros on the meaningless ones. In this study, we consider the following length K vector for each observation vector $\boldsymbol{x}$ as weight,

$$\boldsymbol{w}_{opt} = \big(f\{F^{-1}(\tau_1)\}, \ldots, f\{F^{-1}(\tau_K)\}\big)^T,$$

where f and F are the probability density function and cumulative distribution function of $\boldsymbol{x}$, respectively. In practice, we estimate these f and F using kernel density estimation with $\boldsymbol{x} = (x_1, \ldots, x_p)^T$,

$$\hat{f}(x) = \frac{1}{ph}\sum_{i=1}^{p} \mathcal{K}\Big(\frac{x - x_i}{h}\Big),$$

where $\mathcal{K}(\cdot)$ is a kernel function and h denotes a bandwidth.

3 Proposed Algorithm

3.1 Theoretic Properties of the Proposed Algorithm

Owing to the nonlinearity of the optimization problem, finding the solution of (5) is not trivial. Here, we employ the concept of pseudo data, which gives a relatively straightforward way to obtain a simple solution of (5) and further facilitates the derivation of asymptotic results. The pseudo data was introduced by Cox (1983) for M-type smoothing splines and used by Oh *et al.* (2007) for penalized robust smoothing including wavelet shrinkage.

We visit the pseudo data again and redefine it for the practical algorithm of PCA as

$$\tilde{\boldsymbol{x}} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{\xi} + \sum_{k=1}^{K} w_k \psi_{k,c} \|\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\xi}\|, \tag{7}$$

where $\psi_{k,c}\|\boldsymbol{u}\| := \sum_{\ell=1}^{p} \psi_{k,c}(u_\ell)$ with $\boldsymbol{u} = (u_1, \ldots, u_p)^T$. Here $\psi_{k,c}(u) = \rho'_{k,c}(u)$, where $\rho_{k,c}(u)$ is defined as

$$\rho_{k,c}(u) = \begin{cases} (\tau_k - 1)(u + 0.5c) & \text{for } u < -c \\ 0.5(1 - \tau_k)u^2/c & \text{for } -c \le u < 0 \\ 0.5\tau_k u^2/c & \text{for } 0 \le u < c \\ \tau_k(u - 0.5c) & \text{for } c \le u, \end{cases}$$

for τ_k. This is called a modified check function which is differentiable at zero. As c goes to zero, the function $\rho_{k,c}$ converges to ρ_k. In practice, we set $c = 10^{-6}$.

We now consider a $q \times n$ matrix $\tilde{\boldsymbol{\Xi}} = (\tilde{\boldsymbol{\xi}}_1, \ldots, \tilde{\boldsymbol{\xi}}_n)$ with $\tilde{\boldsymbol{\xi}}_j = (\tilde{\xi}_{1j}, \ldots, \tilde{\xi}_{qj})^T$ that is a solution of the following conventional least squares criterion coupled with the pseudo data $\tilde{\boldsymbol{x}}_j$ $(j = 1, \ldots, n)$ of (7),

$$\sum_{j=1}^{n} \|\tilde{\boldsymbol{x}}_j - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\xi}_j\|^2. \tag{8}$$

Then we show that the solution $\tilde{\boldsymbol{\Xi}}$ is asymptotically equivalent to $\hat{\boldsymbol{\Xi}} = (\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_n)$ with $\hat{\boldsymbol{\xi}}_j = (\hat{\xi}_{1j}, \ldots, \hat{\xi}_{qj})^T$ which is the minimizer of

$$\sum_{j=1}^{n}\sum_{k=1}^{K} w_{j,k}\rho_{k,c}\|\boldsymbol{x}_j - \boldsymbol{\mu} - \boldsymbol{A}\boldsymbol{\xi}_j\| \quad \text{subject to} \quad \boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}_q. \tag{9}$$

Theorem 3.1. *Assume that the diagonal elements of the projection matrix* $\mathbf{\Gamma} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$ *are uniformly small, that is*

$$\max_{1\le i\le p} \gamma_{ii} = \epsilon_{n,p} \ll 1.$$

Under a further assumption that $n\epsilon_{n,p}q^2 \to 0$, *we have*

$$\sum_{t=1}^{q}\Big|\sum_{j=1}^{n}\hat{\xi}_{tj} - \sum_{j=1}^{n}\tilde{\xi}_{tj}\Big| \to 0 \quad \textit{in probability as } n \to \infty,$$

where $\tilde{\xi}_{tj}$ *is the* (t, j)*th component of* $\tilde{\mathbf{\Xi}}$, *and* $\hat{\xi}_{tj}$ *is the* (t, j)*th component of* $\hat{\mathbf{\Xi}}$.

Theorem 3.1 implies that $\hat{\mathbf{\Xi}}$ inherits the same asymptotic squared error properties as $\tilde{\mathbf{\Xi}}$. In fact, the above result can be considered as an extension to the PCA framework of Huber (1973).

4 Practical Performance

4.1 Real Data Analysis

We consider an application of the proposed methods to analyze the maximum daily precipitation data in August from the CPC Merged Analysis of Precipitation (CMAP), which is analyzed by the Climate Research Unit, UK during year 1997–2008. For analysis, we generate yearly data by averaging daily data at each weather station, so that the number of observation is $n = 12$. These are the maximum precipitation on 360 × 180 grids that cover the entire globe with a 1° interval. We first focus on the East Asia region that covers 30–50°N and 120–140°E, and hence, the number of variables is $p = 441$.

The maximum precipitation data is positively skewed or bimodal, and hence, the ordinary PCA cannot accommodate such data since it considers only up to the second moments of the data. To evaluate the performance of the methods, we reconstruct data by one to four PCs and see how well they approximate the original data in the sense of root mean square error (RMSE). As shown in Figure 2, the proposed QPCA provides the smallest RMSE_{rec} values.

5 Conclusions

In this paper, we have proposed a new data-adaptive PCA method. The distribution assumption of data for PCA is extended from Gaussianity to a more general distribution family including contaminated and skewed distributions. We believe that this by itself is an important step and the main contribution of the paper. In approaching this problem, we have found it useful to focus on a composite quantile function that replaces the classical least squares loss function.

Furthermore, we have proposed a practical algorithm to implement the data-adaptive PCA based on the concept of pseudo data. As a generalization, we have developed a penalized composite quantile PCA that provides sparse representations. We have also discussed some theoretical properties related to the algorithm such as the convergence property of the proposed algorithm.

References

Cox, D.D. (1983). Asymptotics for M-type smoothing splines. *The Annals of Statistics.* **11**, 530–551.

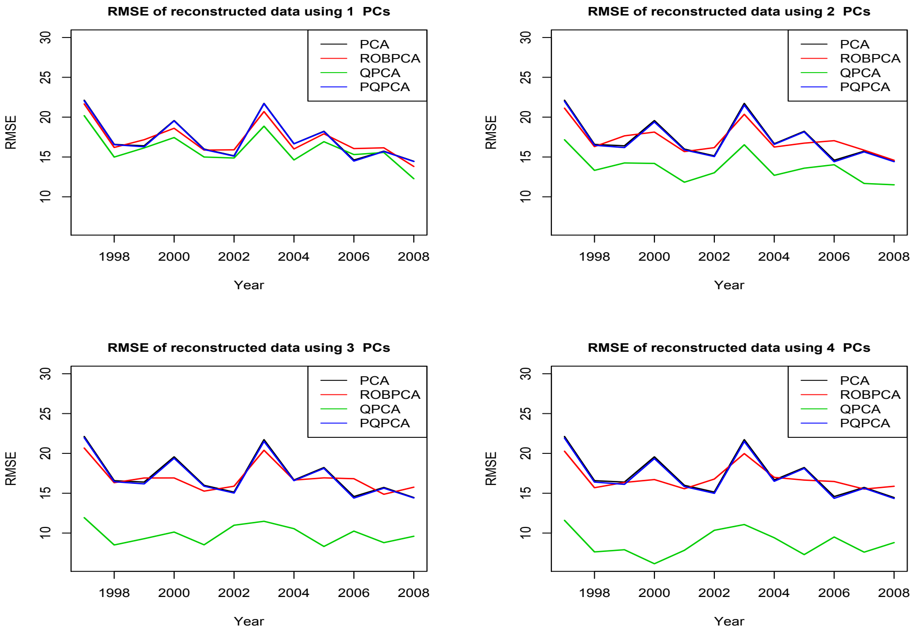

Figure 2: RMSE values between the original maximum precipitation data on the region and the reconstructions by one to four PCs.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

Oh, H.-S., Nychka D.W. and Lee, T.C.M. (2007). The role of pseudo data for robust smoothing with application to wavelet regression. *Biometrika*. **94**, 893–904.