A Data-Adaptive Principal Component Analysis: Use of Composite Asymmetric Huber Function

Yaeji Lim and Hee-Seok Oh

This article considers a new type of principal component analysis (PCA) that adaptively reflects the information of data. The ordinary PCA is useful for dimension reduction and identifying important features of multivariate data. However, it uses the second moment of data only, and consequently, it is not efficient for analyzing real observations in the case that these are skewed or asymmetric data. To extend the scope of PCA to non-Gaussian distributed data that cannot be well represented by the second moment, a new approach for PCA is proposed. The core of the methodology is to use a composite asymmetric Huber function defined as a weighted linear combination of modified Huber loss functions, which replaces the conventional square loss function. A practical algorithm to implement the data-adaptive PCA is discussed. Results from numerical studies including simulation study and real data analysis demonstrate the promising empirical properties of the proposed approach. Supplementary materials for this article are available online.

Key Words: Composite asymmetric Huber function; High-dimensional data; Principal component analysis; Pseudo data; Robustness; Skewed data.

1. INTRODUCTION

1.1 BACKGROUND

Principal component analysis (PCA) has been most widely used for dimension reduction and feature extraction of high-dimensional data. Suppose we have a $p \times n$ dimensional data matrix $M$ that approximately lies on a low-dimensional subspace. Then the data matrix $M$ can be expressed as

$$M = L + Z,$$

where $L$ is a low-rank matrix that represents the subspace and $Z$ is a matrix that models a small noisy perturbation of each entry of $L$. PCA finds the best rank-$q$ estimate of $L$ in the
ℓ_2 sense, which can be derived by singular value decomposition (SVD) of the data matrix and thresholding. Thus, it yields an optimal estimate of the subspace when the perturbation of Z is an iid Gaussian random variable.

By following the projection point of view, we consider a “self-contained” regression-type criterion to pose the PCA problem. Let \( \mathbf{x} = (x_1, \ldots, x_p)^T \) be a p-dimensional random vector with mean vector \( \mu \) and covariance matrix \( \Sigma_{XX} \). PCA finds a linear approximation to represent the data \( \mathbf{x} \) as

\[
g(\xi) = \mu + A\xi, \tag{2}\]

where \( A \) is a \( p \times q \) matrix and \( \xi \) is a q-dimensional vector of parameters, which is a linear projection given by \( \xi = \mathbf{B}\mathbf{x} \) with a \( q \times p \) weight matrix \( \mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_q)^T \) (\( q \leq p \)). Then PCA can be defined as a solution of the following least-square criterion,

\[
\min_{\mu, A, \xi} E\{(\mathbf{x} - \mu - A\xi)^T(\mathbf{x} - \mu - A\xi)\}. \tag{3}\]

For a sample version of (3), given the observations \( \mathbf{x}_1, \ldots, \mathbf{x}_n \), PCA fits the model (2) to the data by minimizing the squares of reconstruction errors,

\[
\min_{\mu, A, \xi_j} \sum_{j=1}^{n} \| \mathbf{x}_j - \mu - A\xi_j \|^2. \tag{4}\]

The solutions of the optimization above can be summarized as

\[
\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j, \quad \hat{A} = \mathbf{V}_q = (\mathbf{v}_1, \ldots, \mathbf{v}_q) = \hat{\mathbf{B}}^T, \quad \text{and} \quad \hat{\xi}_j = \mathbf{V}_q^T(\mathbf{x}_j - \hat{\mu}),
\]

where \( \mathbf{v}_j \) is the \( p \)-dimensional eigenvector associated with the \( j \)th largest eigenvalue of \( \Sigma_{XX} \). Hence, the best rank-q approximation to the data is given by

\[
\hat{\mathbf{x}}_j = \hat{\mu} + \mathbf{V}_q \mathbf{V}_q^T(\mathbf{x}_j - \hat{\mu}). \tag{5}\]

An interpretation of (5) is that \( \mathbf{V}_q \mathbf{V}_q^T \) is a projection matrix that maps \( \mathbf{x}_j \) onto the rank \( q \) subspace with minimum squared reconstruction error.

As one can observe, the conventional PCA explained by the above approach uses up to the second moment of the data, and hence, it may overlook various aspects of the data structure (Caussinus 1992). In literature, various extensions of PCA have been studied. A popular approach is to replace a conventional covariance matrix with a robust scatter matrix (Campbell 1980; Devlin, Gnanadesikan, and Kettenring 1981). However, this method shows low breakdown points in high dimensions. The projection pursuit method is also used to find directions at which a certain robust criterion is maximized (Li and Chen 1985). Hubert, Rousseeuw, and Vanden Branden (2005) had recently proposed a robust method, named ROBPCA, that combines projection pursuit with robust scatter matrix estimation. However, existing methods focus on a robustness to outliers. More recently, Hubert, Rousseeuw, and Verdonck (2009) proposed a modified ROBPCA for skewed data, termed S-ROB.
1.2 Motivation and Main Idea

Along with the linear projection formulation (“regression-type”) of PCA in (3) and (4), Caussinus (1986, 1992) considered a fixed effect model for PCA to obtain a generalized PCA. In our context, the model can be expressed as

\[ x = g(\xi) + e, \quad (6) \]

where \( e \) denotes a vector of independent random errors with \( E(e) = 0 \) and a certain covariance, typically \( I \). Tipping and Bishop (1999) also presented a probabilistic model for PCA, which is the same as (6), based on a Gaussian latent variable model that is closely relevant to factor model. Additionally, they considered the assumption of Gaussian distributed error for statistical inferences and further analysis.

However, if the structure of errors cannot be dealt with the second order statistic, that is, the distribution of the errors is skewed or asymmetric, then the linear projection-type PCA may not perform dimension reduction properly. It can be conjectured that although the conventional PCA represents the major patterns of data, some minor patterns might be distorted due to the structure of the errors.

For a particular example, we consider summer precipitation by averaging three summer months (June-July-August) data for the period of years 1979–2007 that are taken from Climate Prediction Center merged analysis of precipitation (CMAP) data (Xie and Arkin 1997). For understanding regional activities of the precipitation field, we focus on the East Asia region spanning 20°N–50°N and 120°E–150°E, which only includes 169 grid points. We use anomalies of the CMAP data for our analysis. The anomaly is computed by simply subtracting the sample mean of the raw data. We perform the conventional PCA to analyze the precipitation data, and obtain the first four leading PCs and the corresponding loading matrices. Note that the cumulated variances by four PCs are 31.7%, 49.1%, 60.0%, and 68.4%, respectively.

Using the four loading matrices and PCs, we generate a linear approximation \( g(\xi) \) that represents the precipitation data. Figure 1(a) shows the first three loading matrices that are used for generating \( g(\xi) \). We now simulate synthetic data \( x \) by adding Gaussian distributed errors \( e \) into \( g(\xi) \), and then apply the conventional PCA and the proposed data-adaptive PCA (DPCA), which will be discussed in Section 2. The resultant loading matrices are displayed in Figure 1(b) and 1(c). As shown, the loading matrices obtained by both PCAs are almost identical to the true loading matrices in Figure 1(a). It seems that the results of PCA and DPCA are not affected by Gaussian distributed noises.

On the other hand, for reflecting structured noises, we consider a skewed \( t \)-distribution \( St(\xi = 0, \omega = 1, \alpha = 5, df = 3) \), where \( \xi \) is a location parameter, \( \omega \) is a scale parameter, \( \alpha \) is a shape parameter, and \( df \) is a degree of freedom (Azzalini and Capitanio 2014). We apply the conventional PCA and DPCA to the synthetic data, which are contaminated by the skewed \( t \)-distributed errors. Figure 2 shows the first three true loading matrices and the estimated loading matrices by PCA and DPCA, respectively. To further assess the quality of the method, we consider a discrepancy between the true loading matrices and the loading
Figure 1. (a) The first three true loading matrices from the precipitation data. (b) and (c) Loading matrices obtained by PCA an DPCA for precipitation data contaminated by Gaussian distributed errors, respectively.

matrices estimated by each PCA as

$$\text{RMSE}(\hat{v}) = \sqrt{\frac{1}{169} \sum_{i=1}^{13} \sum_{j=1}^{13} (v_{ij} - \hat{v}_{ij})^2},$$

where $v_{ij}$ and $\hat{v}_{ij}$ denote the true loading value and its estimate at the grid $(i, j)$, respectively. The results are listed in Table 1, where the DPCA provides much smaller root mean squared error (RMSE) values than PCA. From this example, we make an important observation that the conventional PCA is sensitive to the noise effect, $e$. More specifically, PCA can be easily distorted by structured errors, while it is not affected by Gaussian distributed errors.

Although we consider only positively skewed errors across observations in the experiment above, it is plausible that some real data can be contaminated by various structured

|       | 1st  | 2nd  | 3rd  | 4th  | Average |
|-------|------|------|------|------|---------|
| PCA   | 0.025| 0.029| 0.035| 0.038| 0.032   |
| DPCA  | 0.021| 0.027| 0.032| 0.035| 0.029   |
errors such as skewed, heavy-tailed, or bimodal ones. The primary goal of this study is to propose a new data-adaptive PCA that does not require prior information of the data. That is to say, it automatically reflects the distributional features of the data. For this purpose, on the basis of the regression-type view of PCA in (3) and (4), we propose a new PCA by replacing the least-square loss function in the conventional PCA criterion with a composite asymmetric Huber function to be introduced in Section 2. The key components of the proposed method are two-fold: (i) PCA can be expressed as a least-square framework and (ii) the quadratic loss function is replaced by a weighted linear combination of asymmetric Huber functions. The weight is set to reflect the heavy-tailed, bimodal, or skewed form of the distribution of the data. Hence, it can be expected that the proposed method works well for various types of data.

Before closing this section, we remark that the conventional PCA focuses on the extraction of major features of the data, whereas both ROBPCA and S-ROB pay attention to detect outliers when the data are heavy-tailed or skewed, respectively. As mentioned earlier, the proposed approach is designed for representing the important patterns of the data as well as for identifying some minor activities properly when the data are contaminated by structured errors. Thus, it seems that our approach strikes a middle ground between two objectives.
2. DATA-ADAPTIVE PCA

2.1 COMPOSITE ASYMMETRIC HUBER FUNCTION

We now discuss a connection between skewed distributions and asymmetric Huber-type loss functions, and consider its utility for developing a new PCA.

To represent skewed data or bimodal data, we consider the following asymmetric density function,

\[ f_{\tau,c}(u; \mu, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{ -\rho_{\tau,c}\left(\frac{u-\mu}{\sigma}\right) \right\}, \]

where \( \mu \) and \( \sigma \) are location and scale parameters, respectively, and the function \( \rho_{\tau,c}(u) \) is given by

\[
\rho_{\tau,c}(u) = \begin{cases} 
(\tau - 1)(u + 0.5c) & \text{for } u < -c \\
0.5(1-\tau)u^2/c & \text{for } -c \leq u < 0 \\
0.5\tau u^2/c & \text{for } 0 \leq u < c \\
\tau(u - 0.5c) & \text{for } c \leq u,
\end{cases}
\]

with \( c = 1.345 \). The function \( \rho_{\tau,c}(u) \) with \( \tau = 0.5 \) is equivalent to the Huber loss function (Huber 1964). By introducing the skewness parameter \( \tau \), this function can be termed “asymmetric Huber function.” Furthermore, we consider a weighted average of the density \( f_{\tau,c}(u; \mu, \sigma) \) according to different \( \tau \)'s. Figure 3(a) shows five density functions \( f_{\tau,c}(u; \mu, \sigma) \) with several \( \tau = 0.05, 0.25, 0.5, 0.75, 0.95 \), respectively, and three different weighted averages of \( f_{\tau,c}(u; \mu, \sigma) \) are displayed in Figure 3(b)–3(d). The weights for Figure 3(b) is 1 for \( f_{\tau,c}(u; \mu, \sigma) \) with \( \tau = 0.1 \) and all other density functions have 0 weights. The uniform weight in Figure 3(c) is set to 1/19 for all 19 density functions, where the skewness parameter varies from \( \tau = 0.05 \) to \( \tau = 0.95 \) with equal space. The weights in Figure 3(d) are 0 for all density functions except ones with \( \tau = 0.5, 0.9, 0.95 \), in which the weights are 0.125, 0.625, 0.25. As shown, the weighted averages of \( f_{\tau,c}(u; \mu, \sigma) \) generate skewed or multi-modal density functions that might reflect various noise distributions. We observe that the weighted average of \( f_{\tau,c}(u) \) corresponds to the function \( \sum_{k=1}^{K} w_k \rho_{\tau_k,c}(u) \), termed “composite asymmetric Huber function.” In other words, given observations, the minimization of the composite asymmetric Huber function is equivalent to the maximization of the likelihood from the weighted average of \( f_{\tau,c}(u) \).

2.2 PCA BASED ON COMPOSITE ASYMMETRIC HUBER FUNCTION

Here, we propose a new PCA that produces data-adaptive loadings (“regression coefficients”) under the regression-type optimization view of PCA. For this purpose, we consider the following optimization problem,

\[
\min_{\mu, A, \xi} \mathbb{E} \sum_{k=1}^{K} w_k \|x - \mu - A\xi\|_{\rho_{\tau_k}} \text{ subject to } A^T A = I_q, \tag{7}
\]

where \( \|u\|_{\rho_{\tau_k}} := \sum_{i=1}^{p} \rho_{\tau_k,c}(u_i) \) with \( u = (u_1, \ldots, u_p)^T \), and the weights \( w_1, \ldots, w_K \) are positive constants, sum up to 1 (\( \sum w_k = 1 \)), which control the contribution of each func-
Figure 3. (a) Asymmetric density functions with different $\tau$'s, and (b)-(d) the weighted averages of 19 asymmetric density functions.

Employing the composite asymmetric Huber functions is capable of reflecting distributional features of data such as skewness as well as bounding outliers effectively, compared to using a single skewed function.

We now consider an empirical version of (7) with $n$ independent random vectors $x_1, \ldots, x_n$. Then the data-adaptive loadings of $q$ principal components can be expressed as the solution of

$$
\min_{\mu, A, \xi} \sum_{j=1}^{n} \sum_{k=1}^{K} w_{j,k} \| x_j - \mu - A \xi_j \|_{\rho_{\tau,c}} \quad \text{subject to} \quad A^T A = I_q.
$$

(8)

The above criterion can be rewritten as

$$
\sum_{j=1}^{n} \left[ \sum_{k=1}^{K} w_{j,k} \left\{ \sum_{i=1}^{p} \rho_{\tau,c} \left( (x_j)_i - (\mu)_i - (A \xi_j)_i \right) \right\} \right],
$$

where $(u)_i$ denotes the $i$th element of $u$. As one can see from the criterion, the weights $\{w_{j,k}\}_{k=1}^{K}$ for a particular observation vector $x_j$ control the contribution of the func-
tions \(\{\rho_{t,c}(\cdot)\}_{k=1}^{K}\). We want to point out a distinction between the above criterion for the proposed method and a criterion with a weighted skewness parameter \(\tau^* = \sum_{k=1}^{K} w_k \tau_k\). For simplicity of discussion, we assume that all weights are 0 except \(w_1 = w_{19} = 1/2\).

Then the criterion with the weighted skewness parameter (not composite one) becomes
\[
\frac{1}{2} \sum_{j=1}^{n} \left[ \sum_{i=1}^{p} \rho_{t,c}(\cdot) ((x_j)_i - (\mu)_i - (A \xi_j)_i) \right] \quad \text{with} \quad \tau^* = (\tau_1 + \tau_{19})/2 ,
\]
which focuses on the average skewness of \(\tau_1\) and \(\tau_{19}\) only. On the other hand, the cost function for the proposed composite one is
\[
\frac{1}{2} \sum_{j=1}^{n} \left[ \sum_{i=1}^{p} \rho_{t,c}(\cdot) ((x_j)_i - (\mu)_i - (A \xi_j)_i) + \sum_{i=1}^{p} \rho_{t,c}(\cdot) ((x_j)_i - (\mu)_i - (A \xi_j)_i) \right],
\]
and hence, the proposed criterion can reflect the skewness information for both \(\tau_1\) and \(\tau_{19}\).

We note that for the number of \(K\), it should be predetermined for implementing the method. The function \(\rho_{t,c}(\cdot)\) is almost identical to the quantile function when \(c\) is sufficiently small. Thus, the choice of quantile levels in the setting of composite quantile regression (CQR) is similar to our setting in some sense. Zou and Yuan (2008) investigated that a large \(K\) leads to an oracle-like estimator in variable selection framework under some conditions including an assumption that the errors follow a symmetric distribution. Recently, Kai, Li, and Zou (2010) suggested that a substantial gain in efficiency of the local CQR estimator can be obtained with a relatively small \(K\) such as \(K = 9\). In computational aspects, the computational cost increases dramatically with the number of quantiles. In this study, we set \(K = 19\) with equally spaced skewness parameters, \(\tau_1 = 0.05, \ldots, \tau_{19} = 0.95\). As an objective way of selecting \(K\), we can consider a data-adaptive method such as cross-validation. This topic is left for future study.

For the weights of the composite asymmetric Huber functions, it is important to obtain optimal weights that reflect the distribution (or skewness) information of the data well. The ideal weight vector is one that gives high values on the meaningful skewness parameters and near zeros on the meaningless ones. A simple choice is to take uniform weights, which provide the following criterion
\[
\min_{\mu, A, \xi} \sum_{j=1}^{n} \sum_{k=1}^{K} ||x_j - \mu - A \xi_j||_{\rho_k} \quad \text{subject to} \quad A^T A = I_g.
\]
In this study, for a better performance, we consider a \(K\)-dimensional (marginal) weight vector for each observation vector \(x_j = (x_{1j}, \ldots, x_{pj})^T\), \(j = 1, \ldots, n\) as \(w_j = (w_{j,1}, \ldots, w_{j,K})^T = (f_j(F_j^{-1}(\tau_1)), \ldots, f_j(F_j^{-1}(\tau_K)))^T\), where \(f_j\) and \(F_j\) are the marginal probability density function and cumulative distribution function of the observation \(x_j\), respectively. In practice, we estimate these \(f_j\) and \(F_j\) using a kernel density estimation with \(x_j = (x_{1j}, \ldots, x_{pj})^T, \tilde{f}_j(x) = \frac{1}{ph} \sum_{i=1}^{p} K(\frac{1-w_i}{h})\), where \(K(\cdot)\) is a Gaussian kernel function and \(h\) denotes a bandwidth.

### 2.3 Pseudo Data-Based Data-Adaptive PCA

Here, we introduce a pseudo data approach and show that the solution of (8) can be interpreted as a conventional PCA of the pseudo data. The pseudo data were introduced by Huber (1964), and used by Cox (1983) for M-type smoothing splines and by Oh, Nychka, and Lee (2007) for penalized robust smoothing including wavelet shrinkage.

We modify the pseudo data for PCA as
\[
\tilde{x} = g(\xi) + \frac{1}{2} \sum_{k=1}^{K} w_k \psi_{t,c}(x - g(\xi)), \quad (9)
\]
where $\psi_{t,c}(u) := (\psi_{t,c}(u_1), \ldots, \psi_{t,c}(u_p))^T$ with $u = (u_1, \ldots, u_p)^T$ and $\psi_{t,c}(u) = \rho'_{t,c}(u)$. The least-square optimization of (4) for PCA coupled with the pseudo data becomes

$$\min_{\xi} \sum_{j=1}^{n} \|\tilde{x}_j - g(\xi_j)\|^2 = \min_{\xi} \sum_{j=1}^{n} \sum_{i=1}^{p} (\tilde{x}_{ij} - g_{ij})^2,$$

where $\tilde{x}_j = (\tilde{x}_{1j}, \ldots, \tilde{x}_{pj})^T$ and $g(\xi_j) = (g_{1j}, \ldots, g_{pj})^T$. The above solution can be characterized by its score equations. Let $\Psi(g)$ be the gradient of the least-square criterion in (10). Then, one obtains

$$\Psi(g)_{ij} = -2(\tilde{x}_{ij} - g_{ij}), \quad i = 1, \ldots, p; j = 1, \ldots, n.$$  

(11)

and the PCA solution $\hat{g}$ is given by $\Psi(\hat{g}) = 0$. We restate the optimization for the data-adaptive PCA in (8) as

$$\min_{\xi} \sum_{j=1}^{n} \sum_{k=1}^{K} w_{j,k} \|x_j - g(\xi_j)\|_{\rho_{iq}} = \min_{\xi} \sum_{j=1}^{n} \sum_{k=1}^{K} w_{j,k} \left[ \sum_{i=1}^{p} \rho_{t,c}(x_{ij} - g_{ij}) \right].$$

(12)

Then, it follows that, from the definition of pseudo data, the score equations of (11) can be expressed as

$$\Psi(g)_{ij} = -2(\tilde{x}_{ij} - g_{ij}) = - \left[ \sum_{k=1}^{K} w_{j,k} \psi_{t,c}(x_{ij} - g_{ij}) \right] = \Phi(g)_{ij},$$

which is the component-wise score equations of $\Phi(g)$ that is the gradient of the criterion in (12). The solution of the data-adaptive PCA will be the root of $\Phi(\hat{g}) = 0$.

An important question we now pose is that the minimization of (8) is equivalent to the minimization of the conventional PCA criterion with pseudo data $\tilde{x}$ as

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

(13)

To answer this question, we investigate some theoretical results. Let $q \times n$ matrix $\tilde{\xi} = (\tilde{\xi}_1, \ldots, \tilde{\xi}_n)$ with $\tilde{\xi}_j = (\tilde{\xi}_{1j}, \ldots, \tilde{\xi}_{pj})^T$ be a solution of (13). Then we show that the solution $\tilde{\xi}$ is asymptotically equivalent to $\hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_n)$ with $\hat{\xi}_j = (\hat{\xi}_{1j}, \ldots, \hat{\xi}_{pq})^T$, which is the minimizer of

$$\sum_{j=1}^{n} \sum_{k=1}^{K} w_{j,k} \|x_j - \mu - A\hat{\xi}_j\|_{\rho_{iq}} \text{ subject to } A^T A = I_q.$$  

(14)

Theorem 1. Assume that the diagonal elements of the projection matrix $\Gamma = A(A^T A)^{-1}A^T$ are uniformly small, that is, $\max_{1 \leq i \leq p} \gamma_{ii} = \epsilon_{n,p} \ll 1$. Under a further assumption that $n \epsilon_{n,p} q^2 \to 0$, we have

$$\sum_{i=1}^{q} \left| \sum_{j=1}^{n} \tilde{\xi}_{ij} - \sum_{j=1}^{n} \hat{\xi}_{ij} \right| \to 0 \text{ in probability as } n \to \infty,$$

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

Theorem 1 implies that $\hat{\xi}$ inherits the same asymptotic properties as $\tilde{\xi}$. In fact, the above result can be considered as an extension to the PCA framework of Huber (1973). A detailed proof can be found in the online supplement.

3. PRACTICAL ALGORITHM

Here, we develop a practical algorithm based on the pseudo data. Ideally, if one knew $A$, one could form the pseudo data and apply the least-square method. In practice, the pseudo data are not available since $A$ and $B$ are unknown. Thus, we consider a fixed point analogy to the pseudo data, “empirical pseudo data” defined as

$$z = \hat{\mu} + \hat{A}\hat{\xi} + \sum_{k=1}^{K} w_k \psi_{\tau_k,c} (x - \hat{\mu} - \hat{A}\hat{\xi}) = \hat{x} + \sum_{k=1}^{K} w_k \psi_{\tau_k,c} (x - \hat{x}),$$

where $\hat{\mu}$, $\hat{A}$, and $\hat{\xi}$ are computed by the following iterative algorithm. Suppose that we observe a $p \times n$ data matrix $X = (x_1, \ldots, x_n) = [X_{ij}]_{i=1,\ldots,p; j=1,\ldots,n}$. For initial $m \times p$ matrix $\hat{B}(0)$ and $p \times n$ matrix $\hat{M}(0) = (\hat{\mu}_{1}^{(0)}, \ldots, \hat{\mu}_{n}^{(0)}) = [\hat{\mu}_{ij}^{(0)}]_{i=1,\ldots,p; j=1,\ldots,n}$,

1. Regress $X$ on the rank $m$ subspace $\hat{B}(0)X$, and obtain a $p \times m$ coefficient matrix $\hat{A}(0)$ and $\hat{X}(0) = \hat{M}(0) + \hat{A}(0)\hat{B}(0)(X - \hat{M}(0))$. That is, estimate a reconstruction data matrix $\hat{X}(0)$ with the linear projection $\tilde{\xi}(0) := \hat{B}(0)(X - \hat{M}(0))$ by mean squared error criterion.

2. Iterate, until convergence, the following two steps for $L = 0, 1, \ldots$

(a) Compute the empirical pseudo data matrix $Z$ in the element-wise manner,

$$Z_{ij}^{(L)} = \hat{X}_{ij}^{(L)} + \frac{1}{2} \sum_{k=1}^{K} w_{j,k} \psi_{\tau_k,c} (X_{ij} - \hat{X}_{ij}^{(L)})$$

for given weights $(w_{1,j}, \ldots, w_{K,j})$ at each $j$.

(b) Obtain next iterative estimates $\hat{A}^{(L+1)}$ and $\hat{B}^{(L+1)}$ by minimizing the conventional least-square criterion (13) with $\hat{x}_j = z_{j}^{(L)}$ for all $j$. Then estimate $\hat{X}^{(L+1)} = \hat{M}^{(L+1)} + \hat{A}^{(L+1)}\hat{B}^{(L+1)}(Z^{(L)} - \hat{M}^{(L+1)}).$

The first $q$ eigenvectors and eigenvalues of covariance matrix of the converged matrix $\hat{X}^{(L)}$ are our final estimators, respectively. Here, the number of principal components $q$ is chosen to explain more than 80% of total variance.

A few remarks are in order.

1. Initial values $\hat{M}^{(0)}$ and $\hat{B}^{(0)}$: For $\hat{M}^{(0)} = (\hat{\mu}_1^{(0)}, \ldots, \hat{\mu}_n^{(0)})$, each $\hat{\mu}_j^{(0)}$ is set as univariate minimum covariance determinant estimator of location, which has been used by Hubert, Rousseeuw, and Vanden Branden (2005) for ROBPCA implementation. If $\hat{B}^{(0)}$ is close to true matrix, then the above algorithm performs well. However, in practice, a proper initial estimate $\hat{B}^{(0)}$ is not trivial. In this study, we employ a loading
Figure 4. (a) Histogram of positively skewed data and a normal distribution, which has the same mean and variance (red line). (b) Transformed pseudo data from (a) and normal distribution that has the same mean and variance (red line). (c) and (d) Negatively skewed data and the rest are the same as (a) and (b).

matrix computed from the ROBPCA for $\hat{B}^{(0)}$. We note that the dimension $m$ is set to be 5.

2. The effect of pseudo data transformation: To evaluate the effect of the pseudo data transformation, we generate positively (negatively) skewed data and compute the converged pseudo data $z^{(L)}$ from the above iterative algorithm. Figure 4(a) shows a histogram of positively skewed original data and a normal density (red line) with the same mean and variance as a contrast. The converged pseudo data $z^{(L)}$ of the positively skewed data are displayed in Figure 4(b). As shown, the distribution of the transformed pseudo data tends to be close to a bell shape distribution. From Figure 4(c) and 4(d), we have similar results for the negatively skewed data.

We now investigate the convergence property of the above algorithm based on the empirical pseudo data. Let $G := (g(\xi_1), \ldots, g(\xi_n)) = [G_{ij}]_{i=1,\ldots,p; j=1,\ldots,n}$, where $g(\xi_j) = (g(\xi_j)_1, \ldots, g(\xi_j)_p)^T$ and $G_{ij} := g(\xi_j)_i).$ Here, $\xi_j$ is a $q$-dimensional vector.

Theorem 2. Let

$$R(G) = \sum_{i=1}^{p} \sum_{j=1}^{n} \sum_{k=1}^{K} w_{j,k} \rho_{t_{i,k}}(X_{ij} - G_{ij}).$$
Under the assumption that $\psi_{t_k,c}(u) = \rho_{t_k,c}(u)$ is continuous and $\psi_{t_k,c}(u) \leq 2$ almost everywhere, we obtain

$$R(\hat{X}^{(L+1)}) \leq R(\hat{X}^{(L)})$$

where $\hat{X}^{(L)}$ is the solution of the proposed algorithm at the stage $L$.

A proof of Theorem 2 can be found in the online supplement. For further analysis, letting $\Omega$ be the set of $p \times n$ matrices of rank $q$, we define a stationary point of the function $R(G)$ as a point of $G \in \Omega$ at which the gradient vector is zero. We further assume two conditions on the subset $\Omega_0$ and the function $R$: (i) $\Omega_0 := \{ G : R(G) \leq R(\hat{G}^{(0)}) \}$ is compact for any $R(\hat{G}^{(0)}) > -\infty$ and (ii) $R$ is continuous in $\Omega$ and differentiable in the interior of $\Omega$.

In addition, let $\mathcal{M}(G)$ denote the map defined by the algorithm that takes a point $\hat{G}^{(L)}$ to $\hat{G}^{(L+1)}$. Then the limit points of the set $\{ \hat{G}^{(L)} : L = 0, 1, 2, \ldots \}$ are characterized by the following result.

**Proposition 1.** For an initial value $\hat{G}^{(0)}$, let $\hat{G}^{(L)} = \mathcal{M}^L(\hat{G}^{(0)})$, where $\mathcal{M}^L$ denotes an $L$ times map composition. If $R(G) = R(\mathcal{M}(G))$ only for stationary points $G$ of $R$ and if $\hat{G}^*$ is a limit point of the sequence $\{ \hat{G}^{(L)} \}$ such that $\mathcal{M}(G)$ is continuous at $\hat{G}^*$, then $R(\hat{G}^{(L)})$ monotonically converges to $R(\hat{G}^*)$ for some stationary point $\hat{G}^*$.

A proof of Proposition 1 can be found in the online supplement.

### 4. PRACTICAL PERFORMANCE

In this section, simulation studies and a real data analysis are conducted to evaluate the practical performance of the proposed methods. Throughout this section, we compare the proposed method, the data-adaptive PCA (DPCA) with existing ones, the ordinary PCA, ROBPCA, and S-ROB. Note that the codes of the R statistical package used to implement the methods and to carry out all the experiments are available at http://stat.snu.ac.kr/heeseok/dpca in order that one can reproduce the same results.

#### 4.1 A SIMULATED EXAMPLE

We consider three orthonormal sine functions as true eigenvectors, which are shown in Figure 5(a) where the length of each curve is 200 (Bailey 2012). Thus, the dimension of the loading matrix $A$ is $200 \times 3$. To simulate 30 data vectors with the length 200, we generate a $3 \times 30$ dimensional principal component matrix from a multivariate normal distribution $\text{MVN}(\mu, \Sigma)$ with $\mu = (0, \ldots, 0)^T$ and $\Sigma = \text{diag}(1, 25, 50)$. We obtain a $200 \times 30$ dimensional low-rank matrix $L$ of the model (1). Note that each column of the matrix $L$ represents a noiseless data $g$. Figure 5(b) shows five projection data vectors among 30. We finally generate a noisy data matrix $M$ by adding noise component, that is, each column of the matrix $M$ is $x_i = g_i + e_i, i = 1, \ldots, 30$, where the noise term is considered as two cases: (i) normal noise, $e_i \sim \mathcal{N}(\mu, \sigma^2)$ and (ii) skewed $t$-distributed case, $e_i \sim \text{St}(\xi, \omega, \alpha, df)$ with a location parameter $\xi$, a scale parameter $\omega$, a shape parameter
\( \alpha \), and a degree of freedom \( \text{df} \). The corresponding five noisy data according to two types of noise are displayed in Figure 5(c) and 5(d). The detailed descriptions of the skewed distributions are found in Azzalini and Capitanio (2014). In the simulation, we set \( \mu = 0 \) and \( \xi = 0 \) for all cases. For the skewed case, the shape parameter \( \alpha \) is set to be 15 for the positively skewed distribution, or \(-15\) for the negatively skewed one, and the \( \text{df} \) is set as \( \text{df} = 3 \). Note that, in the skewed \( t \)-distribution, the mean can be computed as \( \xi + \alpha \frac{\text{df}}{\text{df} - 2} \).

For our simulation, we consider two cases for \( \omega \) and \( \sigma \), \( \omega = \sigma = 8 \) and \( \omega = \sigma = 11 \).

To compare the PCA methods, we evaluate a measure of discrepancy, \( \text{RMSE} \) between three true eigenvectors and the corresponding estimated eigenvectors \( \text{RMSE}(\hat{v}) = \sqrt{\|v_j - \hat{v}_j\|^2} \), \( j = 1, 2, 3 \), where \( v_j \) denotes the \( p \)-dimensional eigenvector associated with the \( j \)th largest eigenvalue. Note that since the dimension of the subspace is 3, we only consider the first three ones. In addition, we consider a reconstruction error (RE),

\[
\text{RE}(\hat{g}) = \sqrt{\frac{1}{N} \sum_i \sum_j (g_{ij} - \hat{g}_{ij})^2},
\]

where \( g_{ij} \) is the \((i, j)\)th component of the noiseless data and \( \hat{g}_{ij} \) is the reconstructed one by a PCA method based on up to three principal components.
Table 2. RMSE values of the first four eigenvectors obtained by four PCA methods. The standard errors are in the parentheses.

| | PCA | ROBPCA | S-ROB | DPCA |
|---|---|---|---|---|
| 1st eigenvector | | | | |
| Normal | 0.021 (0.003) | 0.029 (0.014) | 0.022 (0.006) | 0.022 (0.003) |
| Positively skewed $t$ | 0.035 (0.019) | 0.032 (0.015) | 0.027 (0.006) | **0.026 (0.006)** |
| Negatively skewed $t$ | 0.033 (0.017) | 0.030 (0.010) | 0.027 (0.007) | **0.026 (0.007)** |
| 2nd eigenvector | | | | |
| Normal | **0.030 (0.004)** | 0.039 (0.012) | 0.032 (0.006) | 0.032 (0.004) |
| Positively skewed $t$ | 0.053 (0.022) | 0.044 (0.013) | 0.040 (0.008) | **0.038 (0.008)** |
| Negatively skewed $t$ | 0.054 (0.022) | 0.042 (0.009) | 0.041 (0.009) | **0.039 (0.007)** |
| 3rd eigenvector | | | | |
| Normal | 0.092 (0.006) | 0.093 (0.005) | 0.092 (0.006) | 0.093 (0.005) |
| Positively skewed $t$ | 0.097 (0.003) | 0.095 (0.004) | 0.096 (0.003) | 0.095 (0.004) |
| Negatively skewed $t$ | 0.096 (0.003) | 0.095 (0.004) | 0.095 (0.004) | 0.095 (0.004) |
| 1st eigenvector | | | | |
| Normal | 0.028 (0.004) | 0.036 (0.013) | 0.030 (0.007) | 0.029 (0.004) |
| Positively skewed $t$ | 0.050 (0.023) | 0.042 (0.015) | 0.038 (0.012) | **0.036 (0.007)** |
| Negatively skewed $t$ | 0.054 (0.024) | 0.041 (0.012) | 0.039 (0.011) | **0.037 (0.008)** |
| 2nd eigenvector | | | | |
| Normal | **0.040 (0.005)** | 0.050 (0.014) | 0.042 (0.007) | 0.042 (0.005) |
| Positively skewed $t$ | 0.072 (0.019) | 0.058 (0.013) | 0.058 (0.014) | **0.053 (0.011)** |
| Negatively skewed $t$ | 0.072 (0.019) | 0.056 (0.01) | 0.057 (0.014) | **0.052 (0.009)** |
| 3rd eigenvector | | | | |
| Normal | 0.095 (0.004) | 0.095 (0.004) | 0.095 (0.004) | 0.095 (0.004) |
| Positively skewed $t$ | 0.097 (0.002) | 0.097 (0.003) | 0.097 (0.003) | 0.097 (0.003) |
| Negatively skewed $t$ | 0.097 (0.002) | 0.096 (0.003) | 0.097 (0.003) | 0.096 (0.003) |

NOTE: Minimum value for each simulation case is shown in boldface.

For each combination of noise scenario (normal, positively skewed $\alpha = 15$, negatively skewed $\alpha = -15$), $\sigma$, and $\omega$, we generate 200 sets of samples. For each simulated dataset, the three PCA methods are applied to obtain principal components, and then RMSE($\hat{e}$) and RE($\hat{g}$) are computed and listed in Tables 2 and 3. We observe that (i) the results of the DPCA are comparable to those of PCA with regard to normal noise, (ii) ROBPCA, S-ROB, and DPCA approaches outperform PCA for skewed noisy data, and (iii) for all cases, DPCA gives identical results with PCA, or DPCA is superior to ROBPCA and S-ROB. We note that it seems that the third PCs are distorted by noise components, so that the RE values are increased when the third PCs are used for reconstruction.

4.2 Real Data Analysis

We consider the daily maximum precipitations in August from CMAP during the period of year 1997 to 2008 ($n = 12$). These are the maximum precipitations on 360 $\times$ 180 grids that cover the entire globe with an $1^\circ$ interval. We focus on the East Asia region that covers 30–50$^\circ$N and 120–140$^\circ$E, and hence, the number of variables is $p = 441$. We apply PCA, ROBPCA, S-ROB, and DPCA to the maximum precipitation data, and then obtain loading
Table 3. RE values obtained by four PCA methods. The standard errors are in the parentheses

| Method | 1st PC | 2nd PC | 3rd PC | 4th PC |
|--------|--------|--------|--------|--------|
| PCA    |        |        |        |        |
| Normal | 3.791 (0.405) | 3.939 (0.461) | 3.812 (0.416) | 3.809 (0.410) |
| Positively skewed | 4.827 (2.079) | 4.152 (0.479) | 4.085 (0.418) | 4.062 (0.439) |
| Negatively skewed | 4.546 (1.471) | 4.071 (0.427) | 4.026 (0.403) | 3.987 (0.394) |

Using 1st and 2nd PCs

| Method | 1st PC | 2nd PC | 3rd PC | 4th PC |
|--------|--------|--------|--------|--------|
| Normal | 3.056 (0.068) | 3.097 (0.085) | 3.071 (0.087) | 3.094 (0.083) |
| Positively skewed | 4.361 (2.209) | 3.139 (0.276) | 3.221 (0.397) | 3.106 (0.319) |
| Negatively skewed | 4.273 (1.715) | 3.112 (0.220) | 3.226 (0.443) | 3.065 (0.235) |

Using up to 3rd PC

| Method | 1st PC | 2nd PC | 3rd PC | 4th PC |
|--------|--------|--------|--------|--------|
| Normal | 2.386 (0.074) | 2.548 (0.163) | 2.414 (0.101) | 2.448 (0.100) |
| Positively skewed | 4.091 (0.339) | 3.221 (0.397) | 3.106 (0.319) | 3.044 (0.337) |
| Negatively skewed | 3.112 (0.220) | 3.226 (0.443) | 3.065 (0.235) | 3.044 (0.337) |

Using up to 3rd PC

| Method | 1st PC | 2nd PC | 3rd PC | 4th PC |
|--------|--------|--------|--------|--------|
| Normal | 3.292 (0.113) | 3.578 (0.355) | 3.331 (0.152) | 3.379 (0.135) |
| Positively skewed | 4.492 (0.388) | 4.843 (0.810) | 4.48 (0.474) | 4.48 (0.474) |
| Negatively skewed | 4.724 (5.017) | 4.503 (0.657) | 4.924 (1.369) | 4.513 (0.735) |

Using up to 3rd PC

| Method | 1st PC | 2nd PC | 3rd PC | 4th PC |
|--------|--------|--------|--------|--------|
| Normal | 4.224 (0.096) | 4.277 (0.109) | 4.237 (0.111) | 4.274 (0.109) |
| Positively skewed | 5.669 (0.435) | 6.259 (0.955) | 5.632 (0.436) | 5.632 (0.436) |
| Negatively skewed | 5.679 (0.674) | 6.294 (1.340) | 5.643 (0.680) | 5.643 (0.680) |

NOTE: Minimum value for each simulation case is shown in boldface.

matrices of four leading PCs. As shown in Figure 6, the loadings obtained by DPCA are different from those of the conventional PCA, especially in the third and fourth PCs. The main features of the maximum precipitation data are explained in the first and second PCs, and the differences of the noise treatment between each PCA method are represented in the third and fourth PCs. We note that the four PCs explain almost 90% of the variance, and the data satisfy the assumption in Theorem 1; \( \Gamma = A(A^T A)^{-1} A^T \) are assumed to be uniformly small.

For analysis, we compute the one-year-out cross-validated prediction by several PCA methods as follows: (i) A particular year \( t^* \) is excluded from the total \( n \) years. (ii) Apply a PCA method to the remaining \((n - 1) \times p\) dimensional data, and obtain the \( p \times q \) dimensional eigenvector matrix \( A \). (iii) Compute the principal components for prediction as \( \xi = A^T x_{t^*} \) with the excluded year data \( x_{t^*} \). (iv) Obtain the prediction values as \( \hat{x}_{t^*} = \hat{A} \xi \). Then, we compute prediction errors of each method, \( PE(\hat{x}) = \sqrt{\sum_{t^*=1}^{n} ||x_{t^*} - \hat{x}_{t^*}||^2} \). The prediction errors are sequentially computed by up to four PCs. Figure 7 shows the prediction errors over years computed by different methods. As one can see, for all years, the proposed DPCA produces the smallest PE values and other PCA methods give similar results. We remark that ROBPCA and S-ROB are suitable for identification of outliers, but
they provide almost identical results to those of PCA in this prediction problem. To be specific, the prediction results from PCA are almost identical to those from ROBPCA.

5. CONCLUDING REMARKS

In this article, we have proposed a new data-adaptive PCA method that represents skewed or heavy-tailed distributed data, which may not be well reflected by the ordinary PCA. We believe that this is the main contribution of the article and by itself is an important step. We have found it useful to focus on a composite asymmetric Huber function that replaces the conventional least-square loss function. Furthermore, by introducing the concept of pseudo data, the proposed data-adaptive PCA method can be interpreted as another version of the ordinary PCA, and a practical algorithm based on the pseudo data has been developed. We have also discussed some theoretical properties related to the algorithm such as the convergence property.

Finally, the problems we bring up here can be analyzed by independent component analysis (ICA) that considers higher-order statistics of data (Comon 1994; Hyvarinen and Oja 2000). As PCA, by constraining the number of independent components (ICs), ICA can be used to carry out dimension reduction while retaining the useful information. Figure 8 shows the patterns obtained from both methods, ICA and DPCA, and we observe the differences between them. These differences may occur because ICA cannot identify a correct ordering of the ICs nor the proper scaling (including sign) of the ICs. Moreover, ICA method has been developed to find hidden signals from mixed ones, and hence, it may
Figure 7. The prediction errors by several methods over year $t$. Note that four PCs are sequentially used for the reconstruction.

Figure 8. The first four loading matrices from DPCA for regional maximum daily precipitation (top row), and four IC patterns from ICA (bottom row).

be difficult to interpret extracted patterns as we do in PCA methods. On the other hand, the proposed methods can be understood in the PCA framework, which makes them easy to interpret.

SUPPLEMENTARY MATERIALS

Supplementary materials: Proofs of Theorems 1 and 2 and Proposition 1.
ACKNOWLEDGMENTS

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (Nos. 20110030811, 2012R1A2A2A01002717, and 2014R1A1A3049447).

[Received January 2014. Revised June 2015.]

REFERENCES

Azzalini, A., and Capitanio, A. (2014), The Skew-Normal and Related Families, Cambridge: Cambridge University Press. [1232,1241]

Bailey, S. (2012), “Principal Component Analysis With Noisy and/or Missing Data,” Publications of the Astronomical Society of the Pacific, 124, 1015–1023. [1241]

Campbell, N. A. (1980), “Procedures in Multivariate Analysis I: Robust Covariance Estimation,” Applied Statistics, 29, 231–237. [1231]

Caussinus, H. (1986), “Models and Uses of Principal Component Analysis” (with discussion), in Multidimensional Data Analysis, eds. J. De Leeuw, W. J. Heiser, J. J. Meulman, and F. Critchley, Leiden: DSWO Press, pp. 149–178. [1231]

——— (1992), “The Use of Probabilistic Models to Produce Optimal Graphical Displays of High-Dimensional Data Sets,” Journal of the Italian Statistical Society, 1, 51–65. [1231]

Comon, P. (1994), “Independent Component Analysis, a New Concept?” Signal Processing, 36, 287–314. [1246]

Cox, D. D. (1983), “Asymptotics for M-Type Smoothing Splines,” The Annals of Statistics, 11, 530–551. [1237]

Devlin, S. J., Gnanadesikan, R., and Kettenring, J. R. (1981), “Robust Estimation of Dispersion Matrices and Principal Components,” Journal of the American Statistical Association, 76, 354–362. [1231]

Huber, P. J. (1964), “Robust Estimation of a Location Parameter,” The Annals of Mathematical Statistics, 35, 73–101. [1235,1237]

——— (1973), “Robust Regression: Asymptotics, Conjectures and Monte Carlo,” The Annals of Statistics, 1, 799–821. [1238]

Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005), “ROBPCA: A New Approach to Robust Principal Components Analysis,” Technometrics, 47, 64–79. [1231,1239]

Hubert, M., Rousseeuw, P. J., and Verdonck, T. (2009), “Robust PCA for Skewed Data and Its Outlier Map,” Computational Statistics and Data Analysis, 53, 2264–2274. [1231]

Hyvarinen, A., and Oja, E. (2000), “Independent Component Analysis: Algorithms and Application,” Neural Networks, 13, 411–430. [1246]

Kai, B., Li, R., and Zou, H. (2010), “Local Composite Quantile Regression Smoothing: An Efficient and Safe Alternative to Local Polynomial Regression,” Journal of the Royal Statistical Society, Series B, 71, 49–69. [1237]

Li, G., and Chen, Z. (1985), “Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo,” Journal of the American Statistical Association, 80, 759–766. [1231]

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. [1237]

Tipping, M. E., and Bishop, C. M. (1999), “Probabilistic Principal Component Analysis,” Journal of the Royal Statistical Society, Series B, 61, 611–622. [1232]

Xie, P., and Arkin, P. A. (1997), “Global Precipitation, A 17-year Monthly Analysis Based on Gauge Observation, Satellite Estimates, and Numerical Model Outputs,” Bulletin of the American Meteorological Society, 78, 2539–2588. [1232]

Zou, H., and Yuan, M. (2008), “Composite Quantile Regression and the Oracle Model Selection Theory,” The Annals of Statistics, 36, 1108–1126. [1237]