International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 10, 2010, pp 19-35

Robust Principal Component Analysis Based on Robust Estimation of Multivariate Normal Distribution

Md. Manir Hossain Mollah
Laboratory of Biometry and Bioinformatics,
Graduate School of Agricultural and Life Sciences,
The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku,
Tokyo 113-8657, Japan, E-mail: mhmollah06@yahoo.com

Md. Golam Hossain
Department of Statistics,
University of Rajshahi, Rajshahi-6205,
Bangladesh, E-mail: hossain95@yahoo.com

Md. Nurul Haque Mollah
Department of Statistics,
University of Rajshahi, Rajshahi-6205,
Bangladesh, E-mail: mnhmollah@yahoo.co.in

[Received August 5, 2008; Revised November 16, 2009; Accepted June 13, 2010]

Abstract

This paper proposes a new adaptive algorithm for robust principal component analysis (PCA). The proposed method is formulated based on robust estimators of the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ of multivariate normal distribution. The robust estimators for $\boldsymbol{\mu}$ and Σ are obtained by the minimum β-divergence method (Mollah et al. , 2008). An appropriate value of the tuning parameter β controls the trade-off between robustness and efficiency of the estimators. Therefore, we discuss the selection procedure for the tuning parameter β in this paper also. Finally, we demonstrate the performance of the proposed method in a comparison of the classical method and a most recent adaptive robust PCA algorithm based on the minimum Ψ-principle (Higuchi and Eguchi , 2004). Simulation results show that the proposed method improves the performance over the classical PCA algorithm as well as the minimum Ψ-principle based adaptive robust PCA algorithm.

Keywords and Phrases: Principal component analysis, Multivariate normal distribution, Minimum β-divergence method, β-selection by cross validation and Robustness.

AMS Classification: Primary 62H25, 62G07; Secondary 62G05, 93D21

1 Introduction

Principal component analysis (PCA) is one of the most popular statistical technique for processing, compressing, visualizing and reducing dimensionality of multivariate data. It is widely used in statistical signal processing, neural computing and social science. Principal components (PCs) are also used as the inputs of several statistical procedures including regression analysis, cluster analysis, factor analysis and independent component analysis. It depends solely on the covariance matrix Σ of sample observations. Therefore, it does not require any distributional assumption. However, principal components derived for multivariate normal populations have useful interpretations in terms of the constant density ellipsoids. Further, inferences can be made from the sample components when the population is multivariate normal (Jolliffe, 2002; Johnson and Wichern, 2002). In general, PCA aims to extract the most informative q-dimensional output vector $\boldsymbol{y}_j = (y_{1j}, y_{2j}, ..., y_{qj})^T$ from input vector $\boldsymbol{x}_j = (x_{1j}, x_{2j}, ..., x_{mj})^T$ of dimension $m \geq q$ whose components are assumed to be linearly correlated of each other. This is achieved by learning the $m \times q$ orthogonal matrix $W = [\boldsymbol{w}_1\ \boldsymbol{w}_2\ ...\ \boldsymbol{w}_q]_{m\times q}$ which relates $\boldsymbol{x}_j$ to $\boldsymbol{y}_j$ by

$$\boldsymbol{y}_j = W^T(\boldsymbol{x}_j - \boldsymbol{\mu}), \quad j = 1, 2, \ldots, n \tag{1.1}$$

such that components of $\boldsymbol{y}_j$ are mutually uncorrelated satisfying the order of the variances according to the component number of $\boldsymbol{y}_j$ (Higuchi and Eguchi , 2004).
In the context of off-line learning, $\boldsymbol{\mu}$ and W are directly computed respectively by the sample mean vector $\widehat{\boldsymbol{\mu}}$ and the q dominant eigenvectors of the sample covariance matrix $\widehat{\Sigma}$ as defined by

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j, \tag{1.2}$$

and

$$W = \text{eigen}(\widehat{\Sigma}). \tag{1.3}$$

where

$$\widehat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n} (\boldsymbol{x}_j - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_j - \widehat{\boldsymbol{\mu}})^T, \tag{1.4}$$

such that

$$W^T\widehat{\Sigma}W = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_q) \tag{1.5}$$

satisfying $\lambda_1 > \lambda_2 > ... > \lambda_q$, where λ_i is the variance of ith principle component. It should be noted here that the sample mean vector $\widehat{\boldsymbol{\mu}}$ and the covariance matrix $\widehat{\Sigma}$ as

defined by equations (1.2) and (1.4) are also the minimizer of Kullback-Leibler divergence or equivalently the maximizer of likelihood function based on the multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$. However, it is well known that the estimation of the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ as defined by equations (1.2) and (1.4) are very much sensitive to outliers. Therefore, classical PCA based on (1.2) and (1.4) produces misleading results in the presence of outliers.

There are several robust PCA algorithms based on robust estimation of the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ including Campbell (1980) and Croux and Haesbroeck (2000). However, most of them are not adaptive algorithms. It is well known that the performance of classical PCA algorithm is better than any robust PCA algorithms if input vectors come from multivariate normal distribution and the input dataset is not contaminated by outliers. If dataset is contaminated by outliers, then any robust estimation procedure would be better than classical estimating procedure for PCA. On the other hand, it is very difficult to know in advance whether a dataset is contaminated by outliers or not. In this situation, an adaptive robust PCA method would be better than the non-adaptive robust method, since the adaptive method reduces to the classical PCA algorithm by selecting the tuning parameter using cross-validation in the absence of outliers (Higuchi and Eguchi, 2004). In this paper, an attempt is made to propose a new adaptive robust PCA algorithm based on robust estimation of the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ of multivariate normal distribution by the minimum β-divergence method. We observe that the proposed method is more robust than the most recent adaptive robust PCA algorithm based on minimum Ψ-principle (Higuchi and Eguchi, 2004).

We discuss the adaptive PCA algorithms in section 2, where we summarize the minimum Ψ-principle based adaptive robust PCA algorithm in the sub section 2.1 and introduce the proposed method in the sub section 2.2. The performance of the proposed method is investigated using both synthetic and real datasets in section 3 and make a conclusion of this study in section 4.

2 Adaptive Robust PCA

An adaptive robust PCA algorithm controls the trade-off between robustness and efficiency of the estimators based on the value of the tuning parameters. It reduces to the classical PCA algorithm in the absence of outliers. For comparison of our proposed method with the recently proposed adaptive robust PCA algorithm based on minimum Ψ-principle (Higuchi and Eguchi, 2004), we introduce this method and our proposed method in the sub sections 2.1 and 2.2, respectively.

2.1 Adaptive Robust PCA by Minimum Ψ-Principle

The classical PCA is characterized by minimizing the empirical loss function

$$\frac{1}{n}\sum_{j=1}^{n} z(\boldsymbol{x}_j, \boldsymbol{\mu}, W) \tag{2.1}$$

with respect to $\boldsymbol{\mu}$ and W, where

$$z(\boldsymbol{x}_j, \boldsymbol{\mu}, W) = \frac{1}{2}\Big\{\|\boldsymbol{x}_j - \boldsymbol{\mu}\|^2 - \|W^T(\boldsymbol{x}_j - \boldsymbol{\mu})\|^2\Big\} \tag{2.2}$$

or half the squared residual distance of $(\boldsymbol{x}_j - \boldsymbol{\mu})$ projected onto the subspace spanned by the columns of W (Hotelling, 1933).
Higuchi and Eguchi (2004) proposed a variant of this classical procedure for adaptive robust PCA by minimizing the empirical loss function

$$L_\Psi(\boldsymbol{\mu}, W) = \frac{1}{n}\sum_{j=1}^{n} \Psi(z(\boldsymbol{x}_j, \boldsymbol{\mu}, W)) \tag{2.3}$$

where $\Psi(z)$ is assumed to be a monotonically increasing. Various choices of Ψ's yield various procedures for PCA. As typical examples, the identity function $\Psi_0(z) = z$ reduces to the classical PCA and the sigmoid function

$$\Psi_1(z) = \log \frac{1}{1 + \exp\{-\lambda(z - \eta)\}} \tag{2.4}$$

formulate the self-organizing rule for robust PCA, where λ and η are tuning parameters, referred to as the inverse temperature and saturation value, respectively (Xu and Yuille, 1995). In general, Ψ is interpreted as a generic function to give the loss function L_Ψ. The minimization of L_Ψ in equation (2.3) is referred as *minimum psi principle generated by* Ψ which we call minimum Ψ-principle for convenience of presentation. Using minimum Ψ-principle, Higuchi and Eguchi (2004) found that the minimizer $(\tilde{\boldsymbol{\mu}}, \tilde{W})$ of $L_\Psi(\boldsymbol{\mu}, W)$ satisfies the stationary equations

$$\tilde{\boldsymbol{\mu}} = \sum_{j=1}^{n} h_j(\tilde{\boldsymbol{\mu}}, \tilde{W})\boldsymbol{x}_j, \tag{2.5}$$

and

$$\tilde{W} = \mathrm{eigen}(S(\tilde{\boldsymbol{\mu}}, \tilde{W})) \tag{2.6}$$

where

$$h_j(\boldsymbol{\mu}, W) = \frac{\psi(z(\boldsymbol{x}_j, \boldsymbol{\mu}, W))}{\sum_{j=1}^{n} \psi(z(\boldsymbol{x}_j, \boldsymbol{\mu}, W))} \tag{2.7}$$

and

$$S\left(\boldsymbol{\mu}, W\right)=\sum_{j=1}^{n} h_j\left(\boldsymbol{\mu}, W\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}\right)^T \tag{2.8}$$

with $\psi(z)=(\partial/\partial z)\Psi(z)$. The equilibrium point $(\tilde{\boldsymbol{\mu}}, \tilde{W})$ is expressed by the weighted mean and the covariance matrix, where the weight function h_j depends upon $\tilde{\boldsymbol{\mu}}$ and $\tilde{W}$, except for the case of $\psi(z)=1$, which yields the classical PCA.

2.2 Adaptive Robust PCA by Minimum β-Divergence (Proposed)

The β-divergence between two probability density functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is defined by

$$D_\beta(p,q)=\int\Big[\frac{1}{\beta}\Big\{p^\beta(\boldsymbol{x})-q^\beta(\boldsymbol{x})\Big\}p(\boldsymbol{x})-\frac{1}{\beta+1}\Big\{p^{\beta+1}(\boldsymbol{x})-q^{\beta+1}(\boldsymbol{x})\Big\}\Big]d\boldsymbol{x},\ \text{for } \beta>0,$$

which is non-negative, that is $D_\beta(p,q)\geq 0$, equality holds iff p = q, (Minami and Eguchi , 2002). It measures the discrepancy between two probability density functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ For $\beta\longrightarrow 0$, it reduces to Kullback Leibler (KL) divergence. That is

$$\lim_{\beta\downarrow 0} D_\beta(p,q)=\int p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}d\boldsymbol{x}=D_{\mathrm{KL}}(p,q). \tag{2.9}$$

The minimum β-divergence method (Minami and Eguchi , 2002; Mollah et al., 2006) minimizes the discrepancy between the parametric and non-parametric (empirical) distributions of a random variable $\boldsymbol{x}$ with respect to the parameters. Mollah et al. (2007) used minimum β-divergence method to estimate the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ to robustify pre-whitening algorithm for independent component analysis (ICA) based on unusual multivariate normal distribution $\kappa N(\boldsymbol{\mu},\Sigma)$ with $\kappa>0$. It is defined as

$$(\kappa,\boldsymbol{\mu},\Sigma) \;=\; \underset{\kappa',\boldsymbol{\mu}',\Sigma'}{\operatorname{argmin}}\; D_(p(\boldsymbol{x}),\ \kappa'\varphi_{\boldsymbol{\mu}',\Sigma'}(\boldsymbol{x})),$$

where $\varphi_{\boldsymbol{\mu},\Sigma}(\boldsymbol{x})=N(\boldsymbol{\mu},\Sigma)$ is the usual multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. For robust estimation of the usual multivariate normal distribution, Mollah et al. (2008) also used the minimum β-divergence method. They minimized the β-divergence between $\varphi_{\boldsymbol{\mu},\Sigma}(\boldsymbol{x})$ and the empirical pdf $p(\boldsymbol{x})$. It is defined as

$$\begin{aligned}(\boldsymbol{\mu},\Sigma) &= \underset{\boldsymbol{\mu}',\Sigma'}{\operatorname{argmin}}\; D_\beta(p(\boldsymbol{x}),\ \varphi_{\boldsymbol{\mu}',\Sigma'}(\boldsymbol{x}))\\ &= \underset{\boldsymbol{\mu}',\Sigma'}{\operatorname{argmin}}\; L_\beta\left(\boldsymbol{\mu}',\Sigma'\right),\end{aligned}$$

where

$$L_\beta\left(\boldsymbol{\mu},\Sigma\right)=\frac{1}{\beta+1}\int\left\{\varphi_{\boldsymbol{\mu},\Sigma}^{\beta+1}(\boldsymbol{x})\right\}d\boldsymbol{x}-\frac{1}{\beta}\int\left\{p(\boldsymbol{x})\varphi_{\boldsymbol{\mu},\Sigma}^{\beta}(\boldsymbol{x})\right\}d\boldsymbol{x}.$$

Then it is obtained that

$$\boldsymbol{\mu}=\frac{\mathrm{E}_p\left[\phi_\beta(\boldsymbol{X}|\boldsymbol{\mu},\Sigma)\boldsymbol{X}\right]}{\mathrm{E}_p\left[\phi_\beta(\boldsymbol{X}|\boldsymbol{\mu},\Sigma)\right]},\tag{2.10}$$

and

$$\Sigma=\frac{\mathrm{E}_p\left[\phi_\beta(\boldsymbol{X}|\boldsymbol{\mu},\Sigma)(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\right]}{\mathrm{E}_p\left[\phi_\beta(\boldsymbol{X}|\boldsymbol{\mu},\Sigma)\right]-\beta(1+\beta)^{-(m+2)/2}},\tag{2.11}$$

where p denotes the empirical distribution $p(\boldsymbol{x})$ of $\boldsymbol{x}$ and

$$\phi_\beta(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)=\exp\left\{-\frac{\beta}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}\tag{2.12}$$

which we call β-weight function. The detail derivation of equations (2.10) and (2.11) have been discussed in Mollah et al. (2008). The equations (2.10) & (2.11) are solved iteratively as follows

$$\boldsymbol{\mu}_{t+1}=\frac{\sum_{j=1}^{n}\phi_\beta(\boldsymbol{x}_j|\boldsymbol{\mu}_t,\Sigma_t)\boldsymbol{x}_j}{\sum_{j=1}^{n}\phi_\beta(\boldsymbol{x}_j|\boldsymbol{\mu}_t,\Sigma_t)}\tag{2.13}$$

and

$$\Sigma_{t+1}=\frac{\sum_{j=1}^{n}\phi_\beta(\boldsymbol{x}_j|\boldsymbol{\mu}_t,\Sigma_t)(\boldsymbol{x}_j-\boldsymbol{\mu}_t)(\boldsymbol{x}_j-\boldsymbol{\mu}_t)^T}{\sum_{j=1}^{n}\phi_\beta(\boldsymbol{x}_j|\boldsymbol{\mu}_t,\Sigma_t)-\beta(1+\beta)^{-(m+2)/2}}.\tag{2.14}$$

For $\beta=0$, iterative solutions of (2.13) and (2.14) reduces to the classical non-iterative solutions as defined by equations (1.2) and (1.4), respectively.
To obtain adaptively robust principal components using equation (1.1), we compute $\boldsymbol{\mu}$ and W by the minimum β-divergence estimators of $\boldsymbol{\mu}$ and Σ as obtained iteratively by equations (2.13) and (2.14). However, the robustness performance of the proposed method depends on the value of the tuning parameter β. Therefore, we discuss an adaptive selection procedure for the tuning parameter β by cross validation in the next sub section 2.2.1. We also discuss the robustness of the of the proposed method in the sub section 2.2.2.

2.2.1 β-Selection by K-Fold Cross Validation

To find an appropriate β for the minimum β-divergence estimators, Mollah et al. (2007) have been used β-divergence with a fixed value β_0 of β as a measure for evaluation of the minimum β-divergence estimators. In this paper, we also use the same measure for β selection using cross validation (Hastie et al., 2001). To define the measure for β selection using cross validation, let us split the entire dataset $\mathscr{D}$ into K subsets; $\mathscr{D}(1)$, $\cdots, \mathscr{D}(K)$ and let $\mathscr{D}^{-k} = \{\boldsymbol{x} | \boldsymbol{x} \notin \mathscr{D}(k)\}$. Then the measure for β selection can be defined by

$$D_{\beta_0}(\beta) = \frac{1}{n}\sum_{k=1}^{K} L_{\beta_0}\left(\widehat{\boldsymbol{\mu}}_\beta, \widehat{\Sigma}_\beta | \boldsymbol{x} \in \mathscr{D}(k)\right), \tag{2.15}$$

where $\widehat{\boldsymbol{\mu}}_\beta$ and $\widehat{\Sigma}_\beta$ are obtained by solving equations (2.13) and (2.14) iteratively and

$$L_{\beta_0}\left(\widehat{\boldsymbol{\mu}}_\beta, \widehat{\Sigma}_\beta | \boldsymbol{x} \in \mathscr{D}(k)\right) = (\beta_0+1)^{-(m+2)/2}\left\{\det(2\pi\widehat{\Sigma}_\beta)\right\}^{-\beta_0/2} - \frac{1}{n^{(k)}\beta_0}\sum_{\boldsymbol{x}\in\mathscr{D}(k)} \varphi^{\beta_0}_{\widehat{\boldsymbol{\mu}}_\beta, \widehat{\Sigma}_\beta}(\boldsymbol{x}),$$

where $n^{(k)}$ is the number of observation in the subset $\mathscr{D}(k)$. Then we estimate an appropriate β by

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}}\ \widehat{D}_{\beta_0}(\beta)$$

using 'one standard error' rule (Hastie et al., 2001). For more discussion about β-selection by cross validation, please see Mollah et al. (2007).

2.2.2 Robustness

The robustness of minimum β-divergence estimators of $\boldsymbol{\mu}$ and Σ of multivariate normal distribution has been investigated using the influence function (Hampel et al., 1986) by Mollah et al. (2008). In both equations (2.13) and (2.14), the β-weight function $\phi_\beta(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ provides almost zero weight for outlying observations, because outlying observations are usually far from the the center of the original data point. Therefore, estimating equations as well as the influence functions becomes bounded for $\beta > 0$, while they are unbounded for $\beta = 0$. Thus minimum β-divergence estimators are B-robust. For detail discussion about the B-robustness of minimum β-divergence estimators based on influence function, please see Mollah et al. (2008).

3 Numerical Examples

To demonstrate the performance of the proposed method in a comparison of classical PCA and the adaptive robust PCA based on minimum Ψ-principle (Higuchi and Eguchi , 2004), we generate two and five dimensional datasets from Gaussian distribution as follows:

- **Two dimensional datasets**: We draw a random sample of size 400 from bivariate normal distribution $N(\mathbf{0}, \Sigma_1)$, where

$$\Sigma_1 = \begin{pmatrix} 1.15 & 0.50 \\ 0.50 & 0.30 \end{pmatrix}.$$

  Ḟigure 1a1 represent the scatter plot of this dataset. Then 20, 40, 60, 80, 100 and 120 data points from the random positions are replaced by 20, 40, 60, 80, 100 and 120 outliers '+' so that data contamination rates are 5%, 10%, 15%, 20%, 25% and 30%, respectively. Outlying observations in each case are also generated from $N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} \neq \mathbf{0}$ and $\Sigma \neq \Sigma_1$. Figures 1(a2-a7) show the scatter plot of these contaminated datasets with contamination rates 5%, 10%, 15%, 20%, 25% and 30%, respectively.

- **Five dimensional datasets**: We draw a random sample of size 200 from five-variate normal distribution $N(\mathbf{0}, \Sigma_2)$, where $\Sigma_2 = \text{diag}(8, 6, 1, 0.5, 0.1)$. Then 30 data points are replaced randomly from the original 200 data points by 30 outliers '+' so that data contamination rate is 15%. Figure 2(upper-triangular) represents the scatter plot of this contaminated dataset. Again we replace by 60 data points randomly from the original 200 data points by 60 outliers '+' so that data contamination rate is 30%. Figure 2(lower-triangular) represents the scatter plot of this contaminated dataset. Here also note that outlying observations in each case are generated from $N(\boldsymbol{\mu}, \Sigma)$ as before, where $\boldsymbol{\mu} \neq \mathbf{0}$ and $\Sigma \neq \Sigma_2$.

Table 1: PCA results in the presence of outliers. The notation r_{12} represents the correlation coefficient between PC_1 and PC_2 corresponding to the uncontaminated observation. The notations λ_1 and λ_2 in parentheses are the variances of PC_1 and PC_2, respectively. The symbol '*' indicates the significant correlation at 5% level.

Contamination rate (%)	Classical PCA r_{12} (λ_1, λ_2)	ψ-Principle Based PCA r_{12} (λ_1, λ_2)	Proposed PCA r_{12} (λ_1, λ_2)
0%	0.00 (1.36, 0.06)	0.00 (1.35, 0.05)	0.000 (1.36, 0.06)
5%	0.78* (1.10, 0.28)	0.05 (1.20, 0.07)	0.000 (1.36, 0.06)
10%	-0.74* (1.07, 0.64)	0.08 (1.30, 0.08)	-0.001 (1.36, 0.06)
15%	-0.82* (0.90, 1.03)	0.10 (1.32, 0.10)	0.01 (1.30, 0.09)
20%	-0.82* (0.70, 1.11)	-0.57* (1.02, 0.61)	0.07 (1.22, 0.09)
25%	-0.83* (0.23, 1.19)	-0.87* (0.33, 1.08)	0.08 (1.20, 0.11)
30%	-0.84* (0.23, 1.19)	-0.85* (0.26, 1.16)	0.11 (1.14, 0.15)

In order to investigate the performance of the proposed method in a comparison of the classical PCA and a recently developed adaptive robust PCA based on minimum Ψ-principle, we consider two dimensional datasets in presence of 0%, 5%, 10%, 15%,

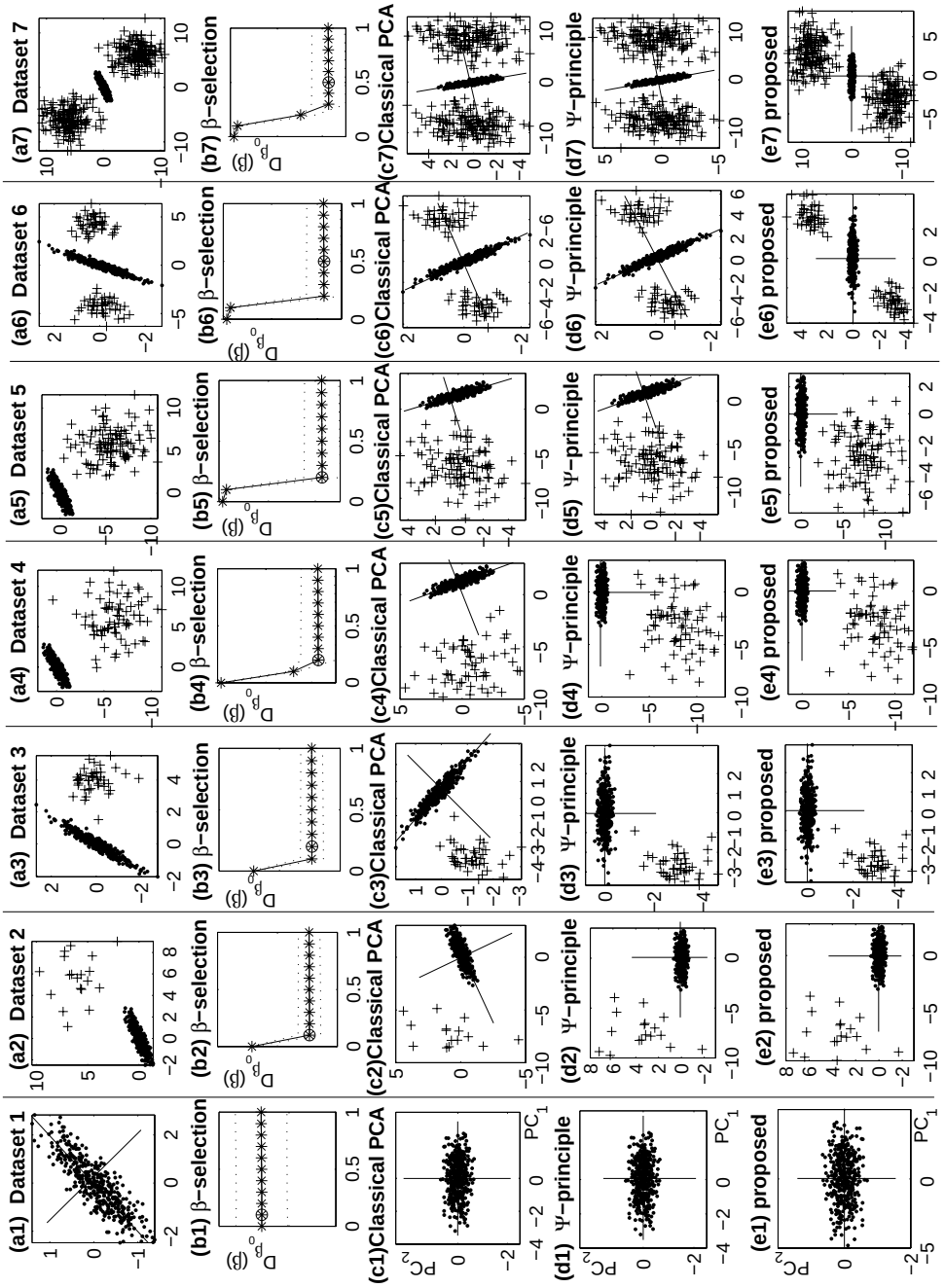

Figure 1: a(1-7) Scatter plot of two dimensional dataset in the presence of 0%, 5%, 10%, 15%, 20%, 25% and 30% outliers, respectively. b(1-7) β selection for the proposed method using cross validation for datasets as shown in figures a(1-7), respectively. c(1-7) Scatter plots of first principle component (PC_1) and second principle component (PC_2) obtained by classical method with datasets as shown in figures a(1-7), respectively. d(1-7) Scatter plots of PC_1 and PC_2 obtained by the minimum Ψ-principle with the same datasets as shown in figures a(1-7), respectively. e(1-7) Scatter plots of PC_1 and PC_2 obtained by the proposed method with the same datasets as shown in figures a(1-7), respectively.

20%, 25% and 30% (or, 0%-30%) outliers as describe in Figures 1(a1-a7), respectively. For each of these datasets, we select β by cross validation. Figures 1(b1-b7) show the cross validation results. We see that appropriate values of β=0, 0.1, 0.1, 0.2, 0.2 and 0.3 for the proposed method in the presence of 0%-30% outliers, respectively. It may be remind here that the proposed PCA with β=0 is equivalent to the classical PCA. Figures 1(c1-c7) represent the scatter plot of classical PC_1 and PC_2 in the presence of 0%-30% outliers, respectively. Figures 1(d1-d7) represent the scatter plot of PC_1 and PC_2 based on ψ-principle in the presence of 0%-30% outliers, respectively. Figures 1(e1-e7) represent the scatter plot of PC_1 and PC_2 based on the proposed method in the presence of 0%-30% outliers as before, respectively. Table 1 shows the correlation coefficient (r_{12}) between PC_1 and PC_2, and their variances λ_1 and λ_2 for each of three methods mentioned above in the presence of 0%-30% outliers, respectively. To observe the PCA results with only uncontaminated observations from a contaminated dataset, we compute r_{12}, λ_1 and λ_2 using PC_1 and PC_2 scores corresponding to uncontaminated observations '.' only. From figures 1(c1-e1) in the first column in the absence of outliers, we observe that there is no significant change in the PC_1 scores over the whole range of PC_2, and no change in the PC_2 scores over the whole range of PC_1 for all three methods. Also these graphical presentation are satisfied by $r_{12} = 0.000$ in the absence of outliers as given in the second row of table 1 for each of three methods mentioned above. Thus uncorrelatedness property of PCA is satisfied by PC_1 and PC_2 with each of three methods in the absence of outliers. Again, we observe that $\text{Var}(PC_1) = \lambda_1 > \lambda_2 = \text{Var}(PC_2)$ in the absence of outliers as given in the second row of table 1 for each of three methods. Thus PC_1 and PC_2 satisfy both uncorrelatedness and variance properties of PCA with each of three methods in the absence of outliers. However, in a similar fashion using figures 1(c2-c7), 1(d2-d7), 1(e2-e7) and the values of r_{12}, λ_1 and λ_2 as given in the (3-8)th rows of table 1, we observe that PC_1 and PC_2 corresponding to the uncontaminated observations based on (1) classical method don't satisfy either uncorrelatedness or variance properties of PCA in the presence of 5% or more outliers, (2) minimum ψ-principle don't satisfy either uncorrelatedness or variance properties of PCA in the presence of more than 15% outliers, and (3) the proposed method satisfy both uncorrelatedness and variance properties of PCA in the presence of all cases (0%-30%) of outliers under consideration. Thus the performance of the proposed method is better than both classical method and the method of minimum ψ-principle in the presence of huge amount of data contamination.

To examine the performance of the proposed method for high dimensional datasets, we consider five dimensional datasets in the presence of 0%, 15% and 30% outliers as viewed in figure 2 based two-dimensional coordinate systems. To obtain principal components by orthogonal transformation, the true orthogonal matrix is $W = (\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_5) = I$ (the identity matrix) for each of five dimensional datasets, since the true covariance matrix for each these datasets is $\Sigma = \text{diag}(8, 6, 1, 0.5, 0.1)$, the diagonal matrix. Therefore, an estimate $\hat{W} = (\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2, ..., \hat{\boldsymbol{w}}_5)$ of W will be good

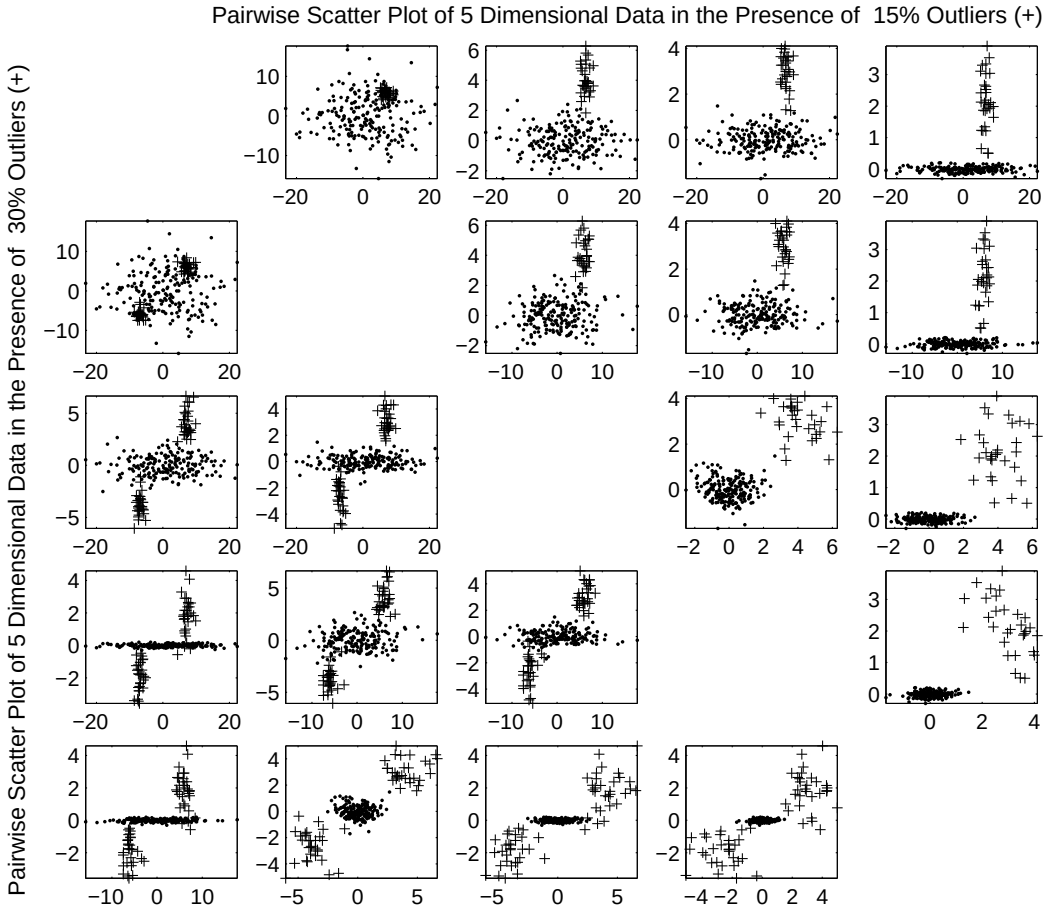

Figure 2: Pairwise scatter plot of five dimensional data. (Upper triangular) Scatter plot of five dimensional data in the presence of 15% outliers '+'. (Lower triangular) Scatter plot of five dimensional data in the presence of 30% outliers '+'.

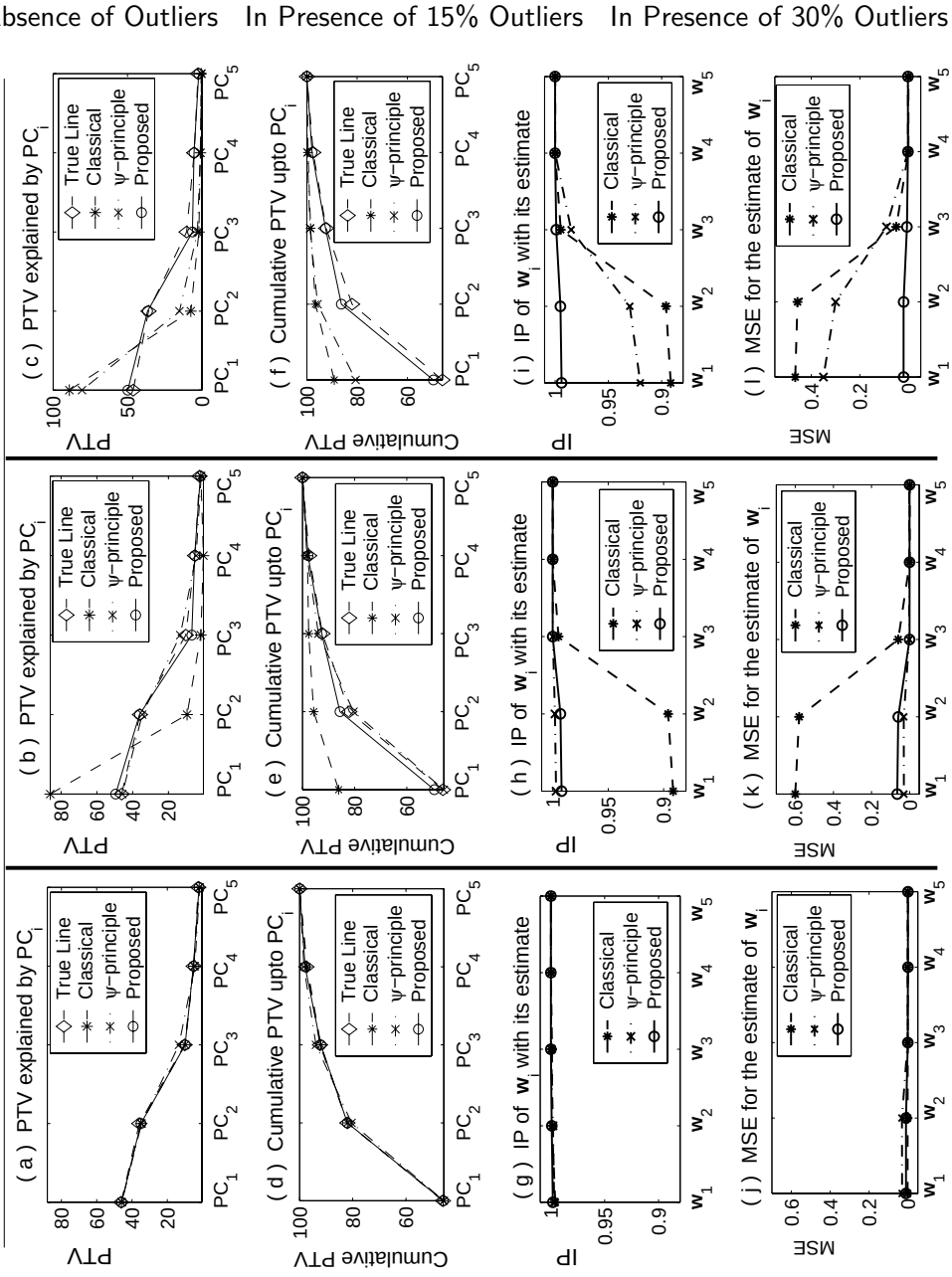

Figure 3: (a-c) Percentage of total variation (PTV) contributed by the principal components in the presence of 0%, 15% and 30% outliers, respectively. (d-f) Cumulative PTV contributed by the principal components in the presence of 0%, 15% and 30% outliers, respectively. (g-i) Inner product between true PC vector $\boldsymbol{w}_i$ and its estimates $\widehat{\boldsymbol{w}}_i$ in the presence of 0%, 15% and 30% outliers, respectively. (j-l) MSE of $\widehat{\boldsymbol{w}}_i$ in the presence of 0%, 15% and 30% outliers, respectively.

if the inner product (IP)

$$\boldsymbol{w}_i^T \hat{\boldsymbol{w}}_i = 1, \tag{3.1}$$

or equivalently, the mean square error (MSE)

$$\text{MSE}(\hat{\boldsymbol{w}}_i) = ||\hat{\boldsymbol{w}}_i - \boldsymbol{w}_i||^2/m = 0, \tag{3.2}$$

for all $i = 1, 2, ..., 5$, where $\hat{\boldsymbol{w}}_i$ is the estimate of $\boldsymbol{w}_i$ and m=5 is the length of $\boldsymbol{w}_i$. It should be noted here that MSE($\hat{\boldsymbol{w}}_i$) measures the distance between $\hat{\boldsymbol{w}}_i$ and $\boldsymbol{w}_i$, and hence MSE($\hat{\boldsymbol{w}}_i$) > 0 for $\hat{\boldsymbol{w}}_i \neq \boldsymbol{w}_i$. The criterion MSE($\hat{\boldsymbol{w}}_i$) can only be used for the performance evaluation of PCA algorithms when the true principal eigenvector $\boldsymbol{w}_i$ is known. This criterion cannot be used in the case of real data analysis due to the unknown principal eigenvectors. To apply the proposed method in each case of five dimensional datasets, we select β by cross validation as before. We obtained that appropriate values of β= 0, 0.05 and 0.1 for the proposed method in the presence of 0%, 15% and 30% outliers, respectively. Figures 3(a-c) represent the percentage of total variation (PTV) contributed by each PC in the presence of 0%, 15% and 30% outliers, respectively. The dashed line with marker style ($\diamond$) represent the true PTV for each PC, while the dashed line with marker style (*), the dash-dot line with marker style ($\times$) and the solid line with marker style (o) represent the estimated PTV by classical method, minimum Ψ-principle and the proposed method, respectively. Similarly, Figures 3(d-f) represent the cumulative PTV explained by the PC in the presence of 0%, 15% and 30% outliers, respectively. The dashed line with marker style ($\diamond$) represent the true cumulative PTV by each PC, while the dashed line with marker style (*), the dash-dot line with marker style ($\times$) and the solid line with marker style (o) represent the estimated cumulative PTV by classical method, minimum Ψ-principle and the proposed method as before, respectively. Figures 3(g-i) represent the inner product (IP) between the true PC vector ($\boldsymbol{w}_i$) with its estimate ($\hat{\boldsymbol{w}}_i$; i=1,2,...,5) in the presence of 0%, 15% and 30% outliers, respectively. The dash-dot line with marker style ($\times$) and the solid line with marker style (o) represent the IP with the estimates of classical method, minimum Ψ-principle and the proposed method, respectively. Similarly, Figures 3(j-l) represent the MSE for the estimates of PC vector in the presence of 0%, 15% and 30% outliers, respectively. The dash-dot line with marker style ($\times$) and the solid line with marker style (o) represent the MSE with the estimates of classical method, minimum Ψ-principle and the proposed method as before, respectively. Figure 3(a,d) shows that the percentage of total variation (PTV) as well as cumulative PTV by the estimated PC_1,..., PC_5 based on classical, minimum Ψ-principle and proposed methods are almost same as the PTV as well as cumulative PTV obtained by the true PC_1... PC_5, respectively in the absence of outliers. We have also investigated the performance of these methods in the absence of outliers using IP and MSE as defined in (3.1) and (3.2), respectively. From figures 3(g,j), we observe that IP is almost close to 1 (i.e., IP$\approx$1) and smaller MSE occurs (i.e., MSE$\approx$0) for each of PC_1,..., PC_5 based on each of classical, minimum Ψ-principle and proposed methods.

Therefore, performance of all of three methods are good in the absence of outliers. Now comparing figures, 3(a) with 3(b-c), 3(d) with 3(e-f), 3(g) with 3(h-i), and 3(j) with 3(k-l), clearly we see that performance of classical PCA is not good at all in the presence of outliers. In the presence of 15% outliers, performance of both minimum Ψ-principle and the proposed method are good and almost equivalent, however, in the presence of 30% outliers, the performance of minimum Ψ-principle is not good; while the performance of the proposed method is good in this case also.

3.1 Real Data Analysis

In a study of size and shape relationships for painted turtles, Jolicoeur and Mosimann measured carapace length, width and height of 24 male turtles. Their data in terms of logarithms is analyzed using classical PCA in pages 441-443 of Johnson and Wichern (2002). To demonstrate the performance of our proposed method for real data analysis in a comparison of the classical PCA algorithm, we consider their data in terms of logarithms as given in table 1, where last 6 data points (bold) are newly included as outliers. First we perform PCA by the proposed method in absence outliers for comparison with the existing classical PCA results that is discussed in Johnson and Wichern (2002). We have presented the existing results (plain) along with our proposed results (bold) in table 2 for convenience of comparison between two methods in the absence of outliers.

Table 1: Carapace Measurements (in Millimeters) for Painted Turtles

ln(Length) (X_1)	ln(Width) (X_2)	ln(Height) (X_3)	ln(Length) (X_1)	ln(Width) (X_2)	ln(Height) (X_3)
4.532599	4.304065	3.610918	4.779123	4.532599	3.713572
4.543295	4.356709	3.555348	4.787492	4.488636	3.688879
4.564348	4.382027	3.555348	4.787492	4.532599	3.784190
4.615121	4.430817	3.663562	4.795791	4.553877	3.737670
4.624973	4.442651	3.637586	4.828314	4.532599	3.806662
4.634729	4.394449	3.610918	4.844187	4.564348	3.806662
4.644391	4.418841	3.663562	4.852030	4.553877	3.806662
4.663439	4.418841	3.663562	4.875197	4.553877	3.828641
4.672829	4.406719	3.637586	4.905275	4.663439	3.850148
4.718499	4.488636	3.688879	**10.596635**	**20.126631**	**9.903488**
4.727388	4.477337	3.688879	**15.596635**	**9.903488**	**19.615805**
4.736198	4.454347	3.688879	**9.903488**	**21.289782**	**11.407565**
4.753590	4.499810	3.761200	**10.463103**	**3.401197**	**13.710150**
4.762174	4.499810	3.713572	**10.308953**	**2.302585**	**22.429216**
4.762174	4.510860	3.713572	**10.596635**	**19.903488**	**9.615805**

To investigate the robustness of the proposed method in comparison of the classical PCA for real data analysis, we perform PCA by both the classical and the proposed methods. We have presented the classical PCA results (plain) along with our proposed results (bold) in table 3 for convenience of comparison as before between two methods in the presence of outliers. In both tables 2 and 3, the notation $\hat{\boldsymbol{w}}_i$ denotes the orthogonal vector for computing ith principal components and the notation $r_{\hat{y}_1,x_k}$ in parentheses represents the correlation coefficient between PC1 (first PC) and kth input variable (X_k). Comparing the values of $r_{\hat{y}_1,x_k}$, $\hat{\boldsymbol{w}}_i$'s, λ_i's and cumulative PTV from table 2, we see that performance of both classical PCA and the proposed methods are almost same in absence of outliers. Therefore the proposed PCA results are good in the absence of outliers, since the classical PCA results are good in the absence of outliers. Again we see that classical PCA results in the presence of outliers (from table 3) are completely different from the classical PCA results in the absence of outliers (from table 2), while the proposed PCA results in the presence of outliers (from table 3) are almost same as the classical PCA results in the absence of outliers (from table 2). Now if we fix cumulative PTV 97 as the threshold to select the PC's for further investigation, we can select only PC_1 having contribution rate around 96% in the total variation by both classical and proposed methods in the absence of outliers. However, in the presence of outliers, we need to select classical PC_1 and PC_2 having cumulative contribution rate around 96% in the total variation, while we can select only proposed PC_1 having contribution rate around 96% in the total variation as before. Thus classical PCA produces misleading results for dimensionality reduction in the presence of outliers, while the proposed PCA produces appropriate results in presence of outliers also. Therefore, the proposed method improves the performance over the classical method in the presence of outliers; otherwise it keeps almost equal performance in the case of real data analysis also.

The proposed algorithm converges within 15 iterations. The computational time of the proposed method as shown in figures 1 & 3, and in tables 2 & 3 are needed around 190, 220 and 170 seconds, respectively. For computation, I used MATLAB programming version 6.5 in my Laptop with system Intel Pentium M-processor 1.64GHz, 1.25 GB of RAM.

Table 2: PCA results in the absence of outliers. Results in parentheses represent the correlation coefficients. Bold numbers represent the results of the proposed method

Variables	$\hat{\boldsymbol{w}}_1(r_{\hat{y}_1,x_k})$	$\hat{\boldsymbol{w}}_2$	$\hat{\boldsymbol{w}}_3$
ln(Length)	0.683 (0.99)	-0.159	-0.713
	0.680 (0.99)	**-0.157**	**-0.711**
ln(Width)	0.510 (0.97)	-0.594	0.622
	0.512 (0.97)	**-0.592**	**0.620**
ln(Height)	0.523 (0.97)	0.788	0.324
	0.524 (0.98)	**0.778**	**0.321**
Variance (λ_i)	23.30×10^{-3}	0.60×10^{-3}	0.36×10^{-3}
	$\mathbf{23.10\times10^{-3}}$	$\mathbf{0.59\times10^{-3}}$	$\mathbf{0.37\times10^{-3}}$
Cumulative PTV	96.1	98.5	100
	96.2	**98.8**	**100**

Table 3: PCA results in the presence of outliers. Results in parentheses represent the correlation coefficients. Bold numbers represent the results of the proposed method

Variables	$\hat{\boldsymbol{w}}_1(r_{\hat{y}_1,x_k})$	$\hat{\boldsymbol{w}}_2$	$\hat{\boldsymbol{w}}_3$
ln(Length)	-0.550 (-0.96)	0.034	0.841
	0.683 (0.99)	**-0.154**	**-0.709**
ln(Width)	-0.330 (-0.77)	-0.902	-0.305
	0.511 (0.99)	**-0.591**	**0.624**
ln(Height)	-0.783 (-0.93)	0.395	-0.499
	0.519 (0.98)	**0.785**	**0.318**
Variance (λ_i)	18048.62×10^{-3}	3079.42×10^{-3}	811.29×10^{-3}
	$\mathbf{24.14\times10^{-3}}$	$\mathbf{0.61\times10^{-3}}$	$\mathbf{0.40\times10^{-3}}$
Cumulative PTV	82.22	96.30	100
	96.05	**98.50**	**100**

4 Conclusion

This paper discusses the robust principal component analysis based on the robust estimation of multivariate normal distribution. The minimum β-divergence method is used for robust estimation of the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ of the multivariate normal distribution. The performance of this method depends on the value of the tuning parameter β. It is equivalent to the classical PCA algorithm for $\beta = 0$. A cross-validation technique is discussed as an adaptive selection procedure for the tuning parameter β in the subsection (2.2.1). Simulation results show that the performance of the the proposed method is equivalent to the classical PCA method in the absence of outliers. In the presence of few outliers, the performance of the proposed method is almost equivalent to the adaptive robust PCA algorithm based on the minimum Ψ-principle (Higuchi and Eguchi, 2004). However it shows better performance in the presence of huge amount of outliers.

References

Campbell, N. A. (1980). Robust procedures in multivariate analysis 1: Robust covariance estimation. *Appl. Statist.*, 29, 231-237.

Croux, C. and Haesbroeck, G. (2000): Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87, 603-618.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W.A. (1986): *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2001): *The Elements of Statistical Learning*. New York: Springer.

Higuchi, I. and Eguchi, S. (2004): Robust Principal Component Analysis With Adaptive Selection for Tuning Parameters. *J. Machine Learning Research* 5, 453-471. 82(397):249-266.

Hotelling, H. (1933): Analysis of a complex of statistical variables into principal components. *Jounnal of Educational Psychology*, 24, 417-441.

Johnson, R. A., Wichern, D.W. (2002): "Applied multivariate statistical analysis "Fifth edition, Prentice-Hall.

Jolliffe, I. T. (2002): *Principal Component Analysis*. Springer-Verlag.

Minami, M. and Eguchi, S. (2002): Robust Blind Source Separation by beta-Divergence. *Neural Computation* **14**, 1859-1886.

Mollah, M. N. H., Minami, M. and Eguchi, S. (2006): Exploring Latent Structure of Mixture ICA Models by the Minimum β-Divergence Method, *Neural Computation*, **18(1)**, pp. 166-190.

Mollah, M. N. H., Minami, M. and Eguchi, S. (2007): Robust prewhitening for ICA by minimizing β-divergence and its application to FastICA. *Neural Processing Letters*, **25(2)**, pp. 91-110.

Mollah, M. M. H., Hossain, M. G. and Mollah, M. N. H. (2008): Robust Estimation for Multivariate Normal Distribution. *Journal of Applied Statistical Science*, Vol. 16(3), pp. 377-386.

Xu, L. and Yuille, A. (1995): Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Trans. on Neural Networks,* **6**, 131-143.