Journal of Machine Learning Research 5 (2004) 453–471 Submitted 4/03; Revised 10/03; Published 5/04

Robust Principal Component Analysis with Adaptive Selection for Tuning Parameters

Isao Higuchi HIGUCHI@AMATH.HIROSHIMA-U.AC.JP
Department of Applied Mathematics
Hiroshima University
1-4-1, Kagamiyama
Higashi-Hiroshima 739-8527, Japan

Shinto Eguchi EGUCHI@ISM.AC.JP
Institute of Statistical Mathematics and
Graduate University of Advanced Studies
4-6-7, Minami-Azabu, Minato-ku
Tokyo 106-8569, Japan

Editor: David Madigan

Abstract

The present paper discusses robustness against outliers in a principal component analysis (PCA). We propose a class of procedures for PCA based on the minimum psi principle, which unifies various approaches, including the classical procedure and recently proposed procedures. The reweighted matrix algorithm for off-line data and the gradient algorithm for on-line data are both investigated with respect to robustness. The reweighted matrix algorithm is shown to satisfy a desirable property with local convergence, and the on-line gradient algorithm is shown to satisfy an asymptotical stability of convergence. Some procedures in the class involve tuning parameters, which control sensitivity to outliers. We propose a shape-adaptive selection rule for tuning parameters using K-fold cross validation.

Keywords: K-fold cross validation, on-line algorithm, reweighted matrix algorithm, influence function, data contamination

1. Introduction

In both neural network and statistical studies, PCA is one of the most fundamental tools of dimensionality reduction for extracting effective features from high-dimensional vectors of input data. See Croux and Haesbroeck (2000) and De la Torre and Black (2001) for recent discussions. PCA is implemented by projecting input data onto the most informative subspace of lower dimension so that the hidden structure behind the input data may be clarified. The procedure of detecting principal components from input data in an on-line manner is related to the mechanism of a single neuron by the Hebbian adaptation rule, for which the learning theory has been discussed (Amari, 1977; Oja, 1982).

One of the frequently occurring difficulties in PCA is that a few outliers give disturbance in finding the effective features in a bulk of input vectors. The usual PCA satisfies the statistical optimality only under the assumption of a Gaussian distribution for all of the input data. A small departure from the assumption produces a gross error in the performance of the principal component vector or subspace in the PCA. This motivates our study of robust PCA procedures. In what follows, we discuss an ε-contamination model for a data distribution F_ε defined by

$$F_\varepsilon = (1-\varepsilon)N(\mu,V) + \varepsilon H, \tag{1}$$

where $N(\mu,V)$ is a Gaussian distribution with mean vector μ and covariance matrix V, and H is a distribution of possible outliers. In this modeling, it is assumed that ε, or the probability of outliers, is small and that the distribution H is unspecified but qualitatively different from the supposed distribution $N(\mu,V)$. Thus, the model (1) lies in a kind of ε-neighborhood surrounding $N(\mu,V)$ taking into account all of the possible distributions H. If H has the mean vector μ_H and the covariance matrix V_H, then the data distribution F_ε has the covariance matrix

$$V_\varepsilon = (1-\varepsilon)V + \varepsilon V_H + \varepsilon(1-\varepsilon)(\mu-\mu_H)(\mu-\mu_H)^T. \tag{2}$$

Thus, the classical procedure works properly as long as $V_H \simeq V$ and $\mu_H \simeq \mu$, because the procedure basically searches for the dominant eigenvectors of V_ε by learning from input data generated by the assumed distribution F_ε. However, even if the probability ε is quite small, the classical procedure often breaks down when input data have a distribution F_ε, such that μ_H or V_H is far from the assumed μ or V, as observed from (2).

A variety of outlier distributions H have infinite dimensionality, and the simplest candidate is a point-mass distribution $\delta_\xi(x)$ degenerated at $x = \xi$. This choice corresponds to a situation in which the outliers occur deterministically in a singleton ξ with probability ε. The influence function of a procedure in the PCA is defined as the derivative at $\varepsilon = 0$ of the procedure under F_ε in (1) with $H = \delta_\xi$. This concept will be more explicitly explored in a subsequent discussion. See Higuchi and Eguchi (1998) for the case of the PCA, and see Hampel (1974) and Hampel *et al.* (1986) for the general case.

We discuss a class of principal component analyzers defined using generic functions which contain tuning parameters. For example if we adopt a log-sigmoidal function as a generic function, the tuning parameters are the inverse temperature and saturation value parameters, as will be discussed in detail. In general the tuning parameter set makes a delicate trade between loss of information and degree of insensitivity to outliers. The main objective in the present paper is to provide a reasonable selection of tuning parameters of principal component analyzers. The basic idea is to craft a loss function that reflects as appropriate trade off between loss of information and robustness to outliers. We introduce K-fold cross validation for estimating the expected loss based on a given data set. As a result we build a method of data-adaptive selection of tuning parameters. In a simulation study, we examine the performance of the adaptive selection under three types of outlier distributions H displaying deterministic, structural and distributional contaminations based on (2). The three types of outliers are simulated in a numerical experiment, and we test the performance in a few cases of principal component analyzers. We provide an S implementation of the basic robust PCA at `http://home.hiroshima-u.ac.jp/oxbow/RobustPCA/`.

The present paper is organized as follows. Section 2 introduces a class of procedures in PCA derived by the minimum psi method. In Section 3 we discuss the robustness of the procedure in the class, and in Section 4 we present an adaptive method of selecting tuning parameters. Finally, in Section 5 we provide the results of a simple simulation study to validate our theoretical discussion in previous sections and tests numerical behavior in three types of departures from the Gaussian distributional assumption.

2. A Class of Principal Component Vectors

In this section we propose a class of principal component vectors. In general, PCA aims to extract the most informative k-dimensional output vector y from an input vector x of p-dimension. This is achieved by learning the matrix Γ which connects x to $y = \Gamma^{\mathrm{T}}(x - \mu)$ based on input data $\{x_t; t = 1, 2, \cdots\}$, where μ is a vector of center of the input data and Γ is a $p \times k$ orthonormal matrix, or $\Gamma^{\mathrm{T}}\Gamma = I$ (the k-identity matrix). In neural networks, Γ is interpreted as the matrix of coefficients connecting p neurons to k neurons, where a learning process works by renewing Γ according to a batch of inputs in an off-line manner or sequential input vectors in an on-line manner (Oja, 1982, 1989 and §8 in Haykin, 1999). By combining these approaches, we propose a certain class of procedures for PCA.

We present a concise review of the classical PCA for detecting the principal k-subspace. Let

$$z(x, \mu, \Gamma) = \frac{1}{2}\{\|x - \mu\|^2 - \|\Gamma^{\mathrm{T}}(x - \mu)\|^2\} \tag{3}$$

be half the square of the residual distance of $x - \mu$ from the subspace spanned by the columns of Γ. We note that $z(x, \mu, \Gamma) = \frac{1}{2}\min_{\beta \in R^k} \|x - \mu - \Gamma\beta\|^2$. See Hotelling (1933) for the original derivation. The classical PCA is simply given by minimizing

$$\frac{1}{n}\sum_{t=1}^{n} z(x_t, \mu, \Gamma)$$

with respect to μ and Γ, which reduces to solving k dominant eigenvectors of the sample covariance matrix

$$S = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{\mu})(x_t - \bar{\mu})^{\mathrm{T}}, \tag{4}$$

where the centralized vector $\bar{\mu}$ is given by $\sum_{t=1}^{n} x_t / n$. Thus, we obtain a solution Γ by stacking the k dominant eigenvectors of S, which we write in the form

$$\Gamma = \mathrm{eigen}(S).$$

We propose a variant of this classical procedure for PCA obtained by minimizing an objective function

$$\mathcal{E}(\mu, \Gamma) = \frac{1}{n}\sum_{t=1}^{n} \Psi(z(x_t, \mu, \Gamma)), \tag{5}$$

where $\Psi(z)$ is assumed to be a monotonic increasing function of $z > 0$. Various Ψ yield various procedures for PCA. As typical examples, the identity function $\Psi_0(z) = z$ reduces to the classical PCA and

$$\Psi_1(z) = \log \frac{1}{1 + \exp\{-\beta(z - \eta)\}} \tag{6}$$

defines Xu and Yuille's self-organizing rule, where β and η are tuning parameters, referred to as the inverse temperature and saturation value, respectively (Xu and Yuille, 1995). Another possible function is

$$\Psi_2(z) = \frac{1 - \exp(-\beta z)}{\beta}. \tag{7}$$

In general, Ψ is interpreted as the generic function which gives the total function $\mathcal{E}$, and we refer to the minimization of $\mathcal{E}$ in (5) as the "minimum psi principle generated by Ψ".

Based on an argument similar to that of the classical PCA, we observe that the minimizer $(\tilde{\mu}, \tilde{\Gamma})$ of $\mathcal{E}(\mu, \Gamma)$ satisfies the stationary equations

$$\tilde{\mu} = \sum_{t=1}^{n} p_t(\tilde{\mu}, \tilde{\Gamma}) x_t, \tag{8}$$

$$\tilde{\Gamma} = \mathrm{eigen}(S(\tilde{\mu}, \tilde{\Gamma})), \tag{9}$$

where

$$p_t(\mu, \Gamma) = \frac{\psi(z(x_t, \mu, \Gamma))}{\sum_{s=1}^{n} \psi(z(x_s, \mu, \Gamma))},$$

$$S(\mu, \Gamma) = \sum_{t=1}^{n} p_t(\mu, \Gamma)(x_t - \mu)(x_t - \mu)^{\mathrm{T}}, \tag{10}$$

with $\psi(z) = (\partial/\partial z)\Psi(z)$. Thus, the equilibrium point $(\tilde{\mu}, \tilde{\Gamma})$ is expressed by the weighted mean and the covariance matrix, where the weight function p_t depends upon $\tilde{\mu}$ and $\tilde{\Gamma}$, except for the case of $\psi(z) = 1$, which yields the classical procedure. In effect, (8) is determined only up to the addition of a vector in the subspace associated with $\tilde{\Gamma}$. Thus $\mu^* = \tilde{\mu} + \gamma$ is also a solution of (8) if $\gamma \in \mathrm{Im}(\tilde{\Gamma}) \equiv \{\tilde{\Gamma} c | c \in R^k\}$, because $\mu^* + \mathrm{Im}(\tilde{\Gamma}) = \tilde{\mu} + \mathrm{Im}(\tilde{\Gamma})$. However, (9) is independent of the choice of possible solutions, so we adopt the expression (8) for convenience.

In PCA research, centralization of input data by a vector other than a sample mean has not been considered. In the present paper, we investigate the problem of centralization and explore the usefulness of a method using a vector such as (8). In conventional robust statistics, the estimation of location has been studied extensively. (See, for example, Huber, 1981.) Our estimator $\tilde{\mu}$ can be viewed as one of several variants for robust estimation. However, our main objective concerning $\tilde{\mu}$ is not the location estimation itself, but rather the data centralization for the extraction of principal components. Thus, $\tilde{\mu}$ is naturally linked to $\tilde{\Gamma}$ in the optimization of $\mathcal{E}(\mu, \Gamma)$.

For a batch of data $\{x_t : 1 \le t \le n\}$, we propose a fixed-point algorithm. See Hyvarinen and Oja (1997) for the related discussion on a fixed-point algorithm for ICA. This algorithm alternates two steps associated with the stationary equations (8) and (9) in the following:

Step 1: Given (μ_1, Γ_1), calculate

$$p_t^{(1)} = \frac{\psi(z(x_t, \mu_1, \Gamma_1))}{\sum_{s=1}^{n} \psi(z(x_s, \mu_1, \Gamma_1))}.$$

Step 2: Using the estimated $\{p_t^{(1)}\}$ in step 1, perform the same task as in the classical PCA:

$$\mu_2 = \sum_{t=1}^{n} p_t^{(1)} x_t \quad \text{and}$$

$$\Gamma_2 = \mathrm{eigen}(S^{(1)}), \tag{11}$$

where $S^{(1)}$ is a weighted matrix defined by $\sum p_t^{(1)}(x_t - \mu_2)(x_t - \mu_2)^{\mathrm{T}}$. In this way, the algorithm alternates between two steps, and we refer to this as the reweighted matrix (RM) algorithm.

Assuming hereafter that the generic function $\Psi(z)$ is strictly concave in z, we have

$$\mathcal{E}(\mu_2,\Gamma_2) - \mathcal{E}(\mu_1,\Gamma_1) < \frac{\sum_{t=1}^{n} \psi(z(x_t,\mu_1,\Gamma_1))}{n}\{\sum_{t=1}^{n} p_t^{(1)} z(x_t,\mu_2,\Gamma_2) - \sum_{t=1}^{n} p_t^{(1)} z(x_t,\mu_1,\Gamma_1)\}$$

because, by assumption, $\Psi(z_2) - \Psi(z_1) < \psi(z_1)(z_2 - z_1)$ for $z_1 < z_2$. In step 2, the procedure is equivalent to minimization of

$$\sum p_t^{(1)} z(x_t,\mu,\Gamma)$$

with respect to (μ,Γ). Therefore, we conclude that the RM algorithm generated by $\{(\mu_j,\Gamma_j) : j \geq 1\}$ is responsible for the strict decrease of the objective function

$$\mathcal{E}(\mu_1,\Gamma_1) > \cdots > \mathcal{E}(\mu_j,\Gamma_j) > \cdots.$$

This desirable property is mathematically the same as that of the EM algorithm. The possible region of (μ,Γ) that the algorithm works in is $\mathcal{X} \times O_{p,k}$, where $\mathcal{X}$ is the convex hull of data and $O_{p,k}$ is the space of $p \times k$ orthonormal matrices. We can easily check a condition for the convergence to the solution of the equations such that a set

$$\{(\mu,\Gamma) \in \mathcal{X} \times O_{p,k} : \mathcal{E}(\mu,\Gamma) \leq c\}$$

is compact for any fixed $c \leq \mathcal{E}(\mu_1,\Gamma_1)$. This is referred to as the regularity condition of Wu (1983), found in §3.4.2 of McLachlan and Krishnan (1997), and this condition implies that the sequence $\{(\mu_j,\Gamma_j) : j \geq 1\}$ is convergent to a set of solutions of equations (8) and (9). However, even when this regularity condition holds, the computational complexity with high dimensional data may be prohibitive since each iteration requires the solution of (11).

2.1 Stability of On-line Gradient Algorithm

We next discuss the on-line gradient algorithm. The gradient vector of the objective function is the sum of

$$G(\mu,\Gamma)(x_t) = \psi(z(x_t,\mu,\Gamma))G_1(\mu,\Gamma)(x_t)$$

over $t = 1,2,\cdots$, where

$$G_1(\mu,\Gamma)(x) = \begin{bmatrix} x-\mu \\ (x-\mu)(x-\mu)^{\mathrm{T}}\Gamma - \Gamma\,\mathcal{LT}[yy^{\mathrm{T}}] \end{bmatrix}$$

with $y = \Gamma^{\mathrm{T}}(x-\mu)$, where the operator $\mathcal{LT}[\cdot]$ sets all of the elements above the diagonal of its matrix argument to zero. Hence, the on-line gradient algorithm is given by

$$\begin{bmatrix} \mu_{t+1} \\ \Gamma_{t+1} \end{bmatrix} = \begin{bmatrix} \mu_t \\ \Gamma_t \end{bmatrix} + r_t G(\mu_t,\Gamma_t)(x_t) \tag{12}$$

for $t = 1,2,\cdots$ with a learning rate r_t. If we apply the classical procedure this algorithm reduces to the Oja algorithm (1982). See §8 in Haykin (1999) for the related algorithmic developments in PCA. The gradient algorithm (12) is different from the Oja algorithm only with respect to the factor $\psi(z)$, which depends on the t-step (μ_t,Γ_t) and the t-th example x_t through $z = z(x_t,\mu_t,\Gamma_t)$ defined in (3). In the classical procedure, the μ-part of the algorithm (12) reduces to the usual centralization and has no connection to Γ. However, in the case of a non-constant weight function $\psi(z)$, the μ-part is essentially connected not only to μ itself, but also to Γ through the z-variable. For the case of non-constant $\psi(z)$, we confine ourselves to the Xu-Yuille rule, which is generated by $\psi_1(z) = \beta/(1+e^{\beta(z-\eta)})$. Xu and Yuille implemented the on-line algorithm in the same fashion as (12) for the Γ-part, but the centralizing mean $\bar{\mu}$ was used for the μ-part. We will make a simple comparison between the two methods for the μ-part in a subsequent discussion.

The on-line gradient algorithm does not satisfy the property of uniform decrease of the objective function possessed by the reweighted algorithm as shown above. We first discuss the asymptotic convergence of (12) for the case of $k=1$, $\Gamma=\gamma$. See §8.4 in Haykin (1999) for the proof for the classical procedure. The on-line gradient algorithm (12) is a special case of the generic stochastic approximation algorithm

$$\begin{bmatrix} \mu_{t+1} \\ \gamma_{t+1} \end{bmatrix} = \begin{bmatrix} \mu_t \\ \gamma_t \end{bmatrix} + r_t h(\gamma_t, \mu_t, x_t), \tag{13}$$

where

$$h(\gamma,\mu,x) = \psi(z(x,\mu,\gamma)) \begin{bmatrix} x-\mu \\ (x-\mu)(x-\mu)^{\mathrm{T}}\gamma - \{\gamma^{\mathrm{T}}(x-\mu)(x-\mu)^{\mathrm{T}}\gamma\}\gamma \end{bmatrix}.$$

If $\psi(z) \equiv 1$, then (13) leads to the classical PCA. It is assumed that ψ is a finite function, so the convergence is proved using an argument similar to that used in the case of classical PCA.

Take the expectation of $h(\gamma_t,\mu_t,x_t)$ over x, and then in the limit we have

$$\begin{aligned} \bar{h}(\gamma_\infty,\mu_\infty) &= \lim_{t\to\infty} E[h(\gamma_t,\mu_t,x_t)] \\ &= \begin{bmatrix} m(\gamma_\infty,\mu_\infty) - \kappa(\gamma_\infty,\mu_\infty)\mu_\infty \\ R(\gamma_\infty,\mu_\infty)\gamma_\infty - \{\gamma_\infty^{\mathrm{T}} R(\gamma_\infty,\mu_\infty)\gamma_\infty\}\gamma_\infty \end{bmatrix}, \end{aligned}$$

where

$$\kappa(\gamma,\mu) = E\{\psi(z(x,\mu,\gamma))\}, \quad m(\gamma,\mu) = E\{\psi(z(x,\mu,\gamma))x\}$$

and

$$R(\gamma,\mu) = E[\psi(z(x,\mu,\gamma))(x-\mu)(x-\mu)^{\mathrm{T}}].$$

Thus, our differential equation is

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} \mu_t \\ \gamma_t \end{bmatrix} &= \bar{h}(\gamma_t) \\ &= \begin{bmatrix} m(\gamma_t,\mu_t) - \kappa(\gamma_t,\mu_t)\mu_t \\ R(\gamma_t,\mu_\infty)\gamma_t - \{\gamma_t^{\mathrm{T}} R(\gamma_t,\mu_\infty)\gamma_t\}\gamma_t \end{bmatrix}. \end{aligned} \tag{14}$$

In the μ-part, we observe that μ_t behaves asymptotically as $e^{-\kappa_\infty t}a + m_\infty/\kappa_\infty$, which implies that m_∞/κ_∞ is the stable-point, where $\kappa_\infty = \kappa(\gamma_\infty,\mu_\infty)$ and $m_\infty = m(\gamma_\infty,\mu_\infty)$. Therefore, we consider only

$$\frac{d}{dt}\gamma_t = R(\gamma_t,\mu_\infty)\gamma_t - \{\gamma_t R(\gamma_t,\mu_\infty)\gamma_t\}\gamma_t.$$

We expand γ_t in terms of the set of eigenvectors $\{q_k(\infty) : k = 1,\cdots,p\}$ of $R(\gamma_\infty,\mu_\infty)$ with the dominant eigenvector $q_1(\infty)$ as follows:

$$\gamma_t = \sum \theta_k(t) q_k(\infty),$$

Let us decompose $\bar{h}(\gamma)$ into $\bar{h}_1(\gamma,\gamma_\infty)+\bar{h}_2(\gamma,\gamma_\infty)$, where

$$\begin{aligned}\bar{h}_1(\gamma,\gamma_\infty) &= R(\gamma_\infty,\mu_\infty)\gamma-\{\gamma^{\mathrm{T}}R(\gamma_\infty,\mu_\infty)\gamma\}\gamma \\ \bar{h}_2(\gamma,\gamma_\infty) &= \{R(\gamma,\mu_\infty)-R(\gamma_\infty,\mu_\infty)\}\gamma \\ &\quad -[\gamma^{\mathrm{T}}\{R(\gamma,\mu_\infty)-R(\gamma_\infty,\mu_\infty)\}\gamma]\gamma.\end{aligned} \tag{15}$$

Then the equilibrium condition $\bar{h}_1(\gamma_\infty,\gamma_\infty)=0$ implies that γ_∞ reduces to one of k eigenvectors of $R(\gamma_\infty,\mu_\infty)$; $\bar{h}_2(\gamma_\infty,\gamma_\infty)=0$ holds identically. The differential equation

$$\frac{d}{dt}\gamma_t=\bar{h}_1(\gamma_t,\gamma_\infty)$$

has an asymptotically stable point $q_1(\infty)$ through the same discussion established in §8.4 in Haykin (1999). Hence the differential equation (14) leads to stable convergence to $q_1(\infty)$.

Secondly, we observe that the case of k principal component vectors also satisfies the stable convergence, noting that our differential equation is

$$\frac{d}{dt}\Gamma_t=R(\Gamma_t,\mu_\infty)\Gamma_t-\mathcal{LT}[\Gamma_t^{\mathrm{T}}R(\Gamma_t,\mu_\infty)\Gamma_t]\Gamma_t,$$

where

$$R(\Gamma,\mu)=E\{\psi(z(x,\mu,\Gamma))(x-\mu)(x-\mu)^{\mathrm{T}}\}.$$

In effect, the RM algorithm is applicable to on-line data by solving the eigen problem for batch data with a new observation incorporated in each step. The computational burden is quite heavy relative to the on-line gradient algorithm, but we will pursue more rapid convergence property in a simulation study.

In the statistical literature another type of PCA methods has been proposed by minimizing

$$\frac{1}{n}\sum_{t=1}^{n}\Psi(d(x_t,\mu,V))$$

with respect to (μ,V), where d is Mahalanobis squared distance, that is,

$$d(x_t,\mu,V)=\frac{1}{2}(x_t-\mu)^{\mathrm{T}}V^{-1}(x_t-\mu).$$

See Campbell(1980), Devlin *et al.* (1981), Caussinus and Ruiz (1990), and Croux and Haesbroeck (2000). The use of the nonlinear generic function Ψ is the same, but the essential difference is that our method aims at estimating the principal component vectors rather than estimating the scatter matrix V. One advantage of our method is that it does not need all the information of V. In fact only the first k dominant eigenvalues and the corresponding eigenvectors are needed in the algorithm, which is easily implemented by the singular-value decomposition algorithm even if the data set is of high dimension.

3. Robustness of the Proposed Principal Component Vectors

Data analysts have frequently found that the classical PCA breaks down in the presence of outliers. It can happen that a single outlier changes the principal component subspace into the orthogonal complement. As a result, the PCA fails to capture an important feature of the bulk of the data, which will be observed from a simple simulation study in Section 5. In the statistical literature, the robustification of the classical likelihood-based procedures has been discussed and well established: see Huber (1981) for some notions on robustness. Such contamination is typically expressed by the ε -contamination model,

$$(1-\varepsilon)N(\mu,V)(x)+\varepsilon\delta_\xi(x),$$

with the point-mass distribution δ_ξ as discussed in the Introduction. In the expression, ε is undetectably small; nevertheless, the likelihood procedure based on the density function of $N(\mu,V)$ may sometimes break down for an extreme vector ξ. We next explore which first principal component vector or principal subspace is robust against outliers in our class. First, we consider the case of the first principal component vector, or $k=1$. In Higuchi and Eguchi (1998), the influence function of the Xu and Yuille rule defined by Ψ_1 in (6) is given by

$$\mathrm{IF}_{\Psi_1}(\xi) = -\psi_1(z(\xi,\mu,\gamma_1))\tilde{\gamma}_1^{\mathrm{T}}(\xi-\mu)\sum_{j=2}^{p}\frac{\hat{\lambda}_j\tilde{\gamma}_j^{\mathrm{T}}(\xi-\mu)}{\tilde{\lambda}_j(\hat{\lambda}_j-\hat{\lambda}_1)}\tilde{\gamma}_j,$$

where $(\hat{\lambda}_j,\hat{\gamma}_j)$ is the pair of the j-th dominant eigenvalue and its associated eigenvector of S defined at (4) and $(\tilde{\lambda}_j,\tilde{\gamma}_j)$ is that of $S(\mu,\Gamma)$ defined at (10), and

$$\psi_1(z) = \frac{\partial\Psi_1(z)}{\partial z} = \frac{\beta}{1+\exp\{\beta(z-\eta)\}}. \tag{16}$$

The influence function assesses the effect on the principal component subspace of the contamination of the data $\{x_t|1\le t\le n\}$ by the outlier ξ.

Secondly, we discuss a general case of k principal component vectors $\Gamma=(\gamma_1,\ldots,\gamma_k)$. We consider the matrix

$$P=\Gamma\Gamma^{\mathrm{T}}.$$

The matrix P is the projection operator onto the subspace spanned by the eigenvectors γ_I, see Tanaka (1988). Our estimator $\tilde{P}=\tilde{\Gamma}\tilde{\Gamma}^{\mathrm{T}}$ has the influence function

$$\mathrm{IF}_{\Psi_1}(\xi;\tilde{P}) = -\psi_1(z(\xi,\tilde{\mu},\tilde{\Gamma}))\sum\frac{\hat{\lambda}_j\tilde{\gamma}_I{}^{\mathrm{T}}(\xi-\tilde{\mu})(\xi-\tilde{\mu})^{\mathrm{T}}\tilde{\gamma}_j}{\tilde{\lambda}_j(\hat{\lambda}_j-\hat{\lambda}_I)}(\tilde{\gamma}_I\tilde{\gamma}_j^{\mathrm{T}}+\tilde{\gamma}_j\tilde{\gamma}_I^{\mathrm{T}}), \tag{17}$$

where the summation is taken over $\{(I,j): I=1,\cdots,k, j=1,\cdots,p, I\neq j\}$, and we will also use this summation convention in a subsequent discussion.

The formula is valid for the minimum psi principle for a general Ψ, see Kamiya and Eguchi (2001) for a detailed discussion, as well as a discussion of relative efficiency under a Gaussian distribution. The influence function, as a function of ξ, assesses the smoothness of the principal component subspace around the supposed distribution $N(\mu,V)$. The boundedness of the influence function in ξ qualitatively guarantees robustness for the target principal component subspace.

In PCA, μ needs to be estimated in order to centralize the data before extracting the principal component subspace. This is usually estimated as $\sum x_t/n$, which can be expressed in the form of a functional $\int x dF(x)$ with $F=F_n$ (empirical distribution). This usual estimation is Fisher consistent since $\int x dF(x)=\mu$ when $F=N(\mu,V)$, but is quite sensitive to the outlier ξ, since the functional evaluated at F is $(1-\varepsilon)\mu+\varepsilon\xi$, and so the influence function is $\xi-\mu$. In contrast, we will show that our PCA procedure automatically leads to a robust centralization of data. We typically observe an unbounded case for the usual principal component subspace, and a bounded case for Xu and Yuille's principal component subspace.

The boundedness of IF_Ψ conditional on

$$\|\tilde{\Gamma}^{\mathrm{T}}(\xi-\tilde{\mu})\|^2 \leq d^2 \tag{18}$$

with a positive constant d guarantees robustness against any outliers ξ satisfying (18) in the contamination model, cf. Higuchi and Eguchi (1998) for the justification for this robustness. Using the formula (17) with general Ψ, we find a sufficient condition for the robustness

$$\sup_{z>0} \sqrt{z}\psi(z) < \infty \tag{19}$$

with $\psi(z) = \partial\Psi(z)/\partial z$. The proof is given as follows. First, we obtain

$$\begin{aligned}
\|\mathrm{IF}_\Psi(\xi;\tilde{P})\| &\leq \psi(z(\xi,\tilde{\mu},\tilde{\Gamma}))\sum\left|\frac{\hat{\lambda}_j\tilde{\gamma}_I{}^{\mathrm{T}}(\xi-\tilde{\mu})(\xi-\tilde{\mu})^{\mathrm{T}}\tilde{\gamma}_j}{\tilde{\lambda}_j(\hat{\lambda}_j-\hat{\lambda}_I)}\right|\|\tilde{\gamma}_I\tilde{\gamma}_j^{\mathrm{T}}+\tilde{\gamma}_j\tilde{\gamma}_I^{\mathrm{T}}\| \\
&\leq d\sum\left|\frac{\hat{\lambda}_j\|\tilde{\gamma}_I\tilde{\gamma}_j^{\mathrm{T}}+\tilde{\gamma}_j\tilde{\gamma}_I^{\mathrm{T}}\|}{\tilde{\lambda}_j(\hat{\lambda}_j-\hat{\lambda}_I)}\right||(\xi-\tilde{\mu})^{\mathrm{T}}\tilde{\gamma}_j|\psi(z(\xi,\tilde{\mu},\tilde{\Gamma}))
\end{aligned}$$

from the assumption of (18). Since

$$z(\xi,\tilde{\mu},\tilde{\Gamma}) = \frac{1}{2}\sum_{j=k+1}^{p}\{\tilde{\gamma}_j^{\mathrm{T}}(\xi-\tilde{\mu})\}^2 \geq \frac{1}{2}\{\tilde{\gamma}_j^{\mathrm{T}}(\xi-\tilde{\mu})\}^2$$

for any j, $k+1 \leq j \leq p$ by the definition of z at (3), we obtain

$$\|\mathrm{IF}_\Psi(\xi;\tilde{P})\| \leq \sum\left|\frac{\hat{\lambda}_j\|\tilde{\gamma}_I\tilde{\gamma}_j^{\mathrm{T}}+\tilde{\gamma}_j\tilde{\gamma}_I^{\mathrm{T}}\|}{\tilde{\lambda}_j(\hat{\lambda}_j-\hat{\lambda}_I)}\right|\left(d\sqrt{2}\sqrt{z(\xi,\tilde{\mu},\tilde{\Gamma})}\psi(z(\xi,\tilde{\mu},\tilde{\Gamma}))+d^2\psi(z(\xi,\tilde{\mu},\tilde{\Gamma}))\right).$$

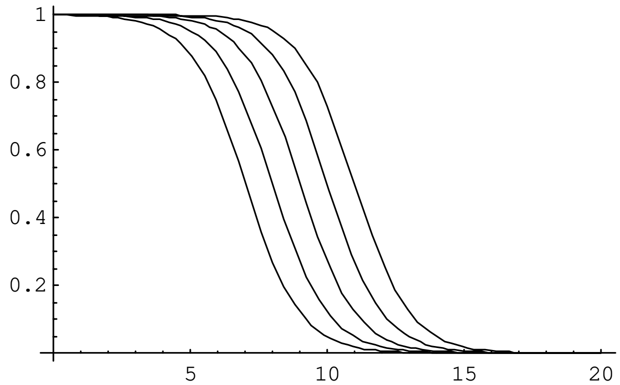
Finally, we conclude that

$$\|\mathrm{IF}_\Psi(\xi;\tilde{P})\| \leq C\left(d\sqrt{2}\sup_{z>0}\sqrt{z}\psi(z)+d^2\sup_{z>0}\psi(z)\right)$$

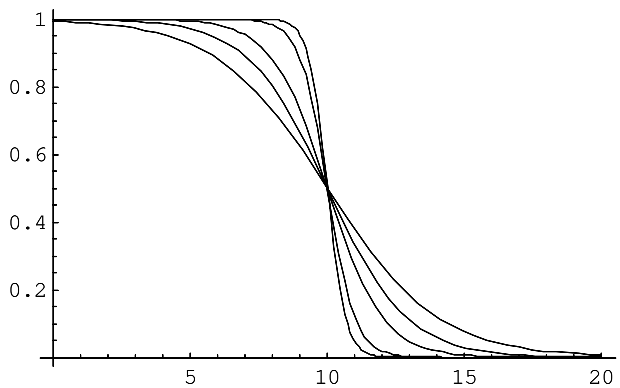
for any outlier ξ in R^p satisfying (18), where

$$C = \sum\left|\frac{\hat{\lambda}_j\|\tilde{\gamma}_I\tilde{\gamma}_j^{\mathrm{T}}+\tilde{\gamma}_j\tilde{\gamma}_I^{\mathrm{T}}\|}{\tilde{\lambda}_j(\hat{\lambda}_j-\hat{\lambda}_I)}\right|.$$

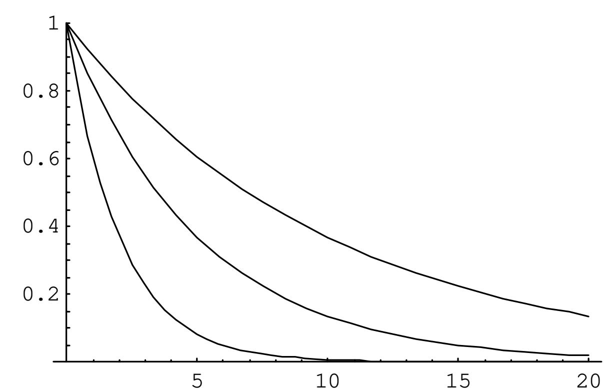
Therefore, the condition (19) for Ψ leads to the boundedness of the influence function IF_Ψ conditional on the condition (18).



(1) φ_1 with $\beta = 1$, $\eta = 7, 8, 9, 10, 11$.



(2) φ_1 with $\beta = 0.5, 0.7, 1, 2, 3$, $\eta = 10$.



(3) φ_2 with $\beta = 0.1, 0.2, 0.5$.

Figure 1: Graphs of ϕ for several tuning parameters.

4. Adaptive Selection for Tuning Parameters

Let α be a parameter in the generic function Ψ which defines our objective function (5). In this section, we focus on the role of the tuning parameter α. The performance of PCA by the proposed method generated by Ψ_α depends on the value of the tuning parameter α. As shown in (6) and (7), the generic function Ψ_1 involves tuning parameters β and η, and Ψ_2 involves β.

Thus, generic functions control the sensitivity to outliers by these tuning parameters. See Figure 1 for the graphs of the derivatives ψ_1 and ψ_2 of Ψ_1 and Ψ_2 for several tuning parameters.

The generic function Ψ_1 where $\eta \to \infty$ or $\beta \to 0$ yields the classical PCA. If we exactly assume Gaussian distribution, or $\varepsilon = 0$ in (1), then the classical PCA or $\psi(z) \equiv 1$, is recommended as the standard method. See Kamiya and Eguchi (2001) for a detailed discussion. This suggests that under the situation in which $\varepsilon \neq 0$, there exists an optimal selection for tuning parameters giving a generic function other than $\psi(z) \equiv 1$.

We propose herein a method of determining tuning parameters as in (6) and (7), based on K-fold cross-validation. See Subsection 7.10 in Hastie *et al.* (2001) for the detailed discussion. Throughout the section, we focus on batch data.

Let $F(x)$ be a data distribution which is assumed to generate input data x. Then, we adopt a generalization error function for assessing the performance of a given $\hat{\mu}$ and $\hat{\Gamma}$ using

$$L(\hat{\theta}, F) = \int \cdots \int \Psi_0(z(x, \hat{\mu}, \hat{\Gamma}))dF(x)$$

where

$$\Psi_0(z) = \log \frac{1}{1 + \exp\{-\beta_0(z \cdot \eta_0)\}}$$

with $\alpha_0 = (\beta_0, \eta_0) = (50, 10)$. This choice of the error function is intended to achieve mild robustness to outliers. This is because we cannot obtain a sensible result if the error function itself is sensitive to outliers. The empirical error function is

$$L_{\text{emp}}(\hat{\theta}) = \frac{1}{n} \sum_{t=1}^{n} \Psi_0(z(x_t, \hat{\mu}, \hat{\Gamma}))$$

for given data $\{x_1, \ldots, x_n\}$ and would be unchanged by data contamination if $(\hat{\mu}, \hat{\Gamma})$ is a robust estimator. The choice of β_0 is universal, but that of η_0 should be adaptive. For example, the median of $\|x_I - \bar{\mu}\|$ with the usual centralizing vector $\bar{\mu}$ as a default value. Given a class of estimators $\hat{\theta}_\alpha$ by the generic function Ψ_α with the tuning parameter α, for example $\alpha = (\beta, \eta)$ in (6) or (7), we attempt to estimate the expected loss function, or the risk function associated with an estimator $\hat{\theta}_\alpha$, which is essentially

$$R(\hat{\theta}_\alpha, F) = E_F\{L(\hat{\theta}_\alpha, F)\},$$

where E_F denotes the mathematical expectation when input data $x_1, \cdots, x_n$ follow from the underlying distribution F.

Here we provide a method of selecting α^* which generates $\hat{\theta}_{\alpha^*}$ with good performance. We use K-fold validation to get a estimator of the generalization error $L(\hat{\theta}_\alpha, F)$. Here, we divide the data set $D = \{x_1, \ldots, x_n\}$ into K subsets $\{D_k = \{x_1^{(k)}, \ldots, x_{n_k}^{(k)}\} : k = 1, \cdots, K\}$ and so $D = \bigcup_{k=1}^{K} D_k$. Define

$$CV(\alpha) = \frac{1}{K} \sum_{k=1}^{K} L_{\text{emp}}^{(k)}(\hat{\theta}_\alpha^{(-k)}),$$

where $\hat{\theta}_\alpha^{(-k)}$ is the estimator based on the data set $\bigcup_{k' \neq k} D_{k'}$ and

$$L_{\text{emp}}^{(k)}(\theta) = \frac{1}{n_k} \sum_{I=1}^{n_k} L(x_I^{(k)}, \theta).$$

In this way, the estimator $\hat{\theta}_\alpha$ and D_k are statistically independent, which implies elimination of the bias in the empirical error $L_{\text{emp}}(\hat{\theta}_\alpha)$. In this formulation, we now define the optimal α^* by

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} CV(\alpha).$$

We will explore the performance of this selection method for synthetic data situations.

5. Simulation Study

We explore the robustness of our class of principal component subspaces in numerical experiments, focusing on the classical rule ($\psi(z) = 1$), the Xu-Yuille rule and the Gaussian kernel rule defined in (7). For our simulation study, we consider the following three types of outlier distributions H in the ε-contamination model defined by (1):

(i) Deterministic contamination: a sum of point-mass distributions at $x = \xi_j$ for $j = 1, \cdots, M$.

(ii) Structural contamination: the same Gaussian distribution $N(\mu_1, V_1)$, but with the structure in μ_1 and V_1 being quite different from with that in μ and V, that is to say, $\|\mu_1 - \mu\|$ or $\text{tr}(V_1 - V)^2$ is substantially large.

(iii) Distributional contamination: the same structure as in (μ, V) but the distribution is totally different from the Gaussian distribution $N(\mu, V)$.

First, we investigate the case in which ε is undetectably small, for example we take $\varepsilon = 0.03$. We have performed a numerical study for the behavior of our procedure for seven-dimensional data in the following setting: $\mu = (0, \cdots, 0)^{\mathrm{T}}, V = \text{diag}(5, 2, 3, 1, 0.5, 0.5, 0.5)$. As for (i), $\xi_1 = (0, 0, 0, 0, 0, 0, b)$, $\xi_2 = (0, 0, 0, 0, 0, b, 0)$ with a probability 0.5 for each.
In (ii) the distribution H is a Gaussian distribution with

$$\mu_1 = (1, \cdots, 1)^{\mathrm{T}}, V_1 = \text{diag}(0.5, 2, 3, 1, 0.5, 0.5, 0.5).$$

In (iii) the outlier, ξ has a distribution H of Cauchy-type of which the location-scatter structure is the same as (μ, V), that is, all the components of $V^{-\frac{1}{2}}(\xi - \mu)$ are independently and identically distributed according to a standard Cauchy distribution with density function $1/(\pi(1 + x^2))$.

We observe in a series of simulations in the above setting that the classical procedure $(\psi(z) = 1)$ breaks down for a batch of data with most observations from $N(\mu, V)$ and a few outliers from H. The first principal component vector extracted only from the 98 simulated vectors is completely changed into the space of minor components after mixing with two outliers. In our experience, PCA has never been weakly perturbed by outlier contamination under the situation of setting (i) with a small b. Whether the classical PCA resists or completely breaks down against outliers is determined by b. If b ranges from 14 to 16, the breakdown occurs with about 50 percent proportion. If input data are simulated by setting (iii), then the classical PCA tends to break down with a higher frequency of

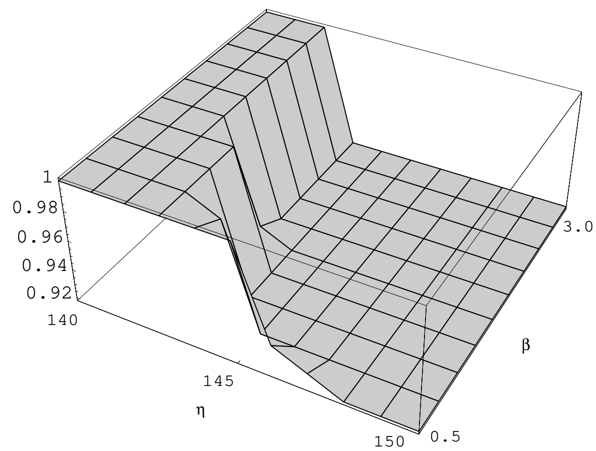


Figure 2: The plot of inner product of the PC vectors with/without outliers against β and η.

occurrence than for setting (ii). In almost all of the cases of setting (iii), the PCA breaks down. Thus, we observe that the distributional contamination is more severe than the structural contamination for the classical PCA.

We confine ourselves to a typical case of input vectors from the structural contamination. We obtained 270 input vectors of 200-dimension from $N(\mu, V)$, where

$$\mu = (0, \cdots, 0)^{\mathrm{T}}, \quad V = \mathrm{diag}(10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0.5, \cdots, 0.5),$$

and 30 outliers from $N(\mu_1, V_1)$, where

$$\mu_1 = (1, \cdots, 1)^{\mathrm{T}}, \quad V_1 = \mathrm{diag}(1, 9, 8, 7, 6, 5, 4, 3, 2, 1, 1, 1, \cdots, 1),$$

and observed that the inner product of the first principal component vectors based on the data of 270 vectors and on the data added to 30 outliers is 0.678 when using the classical PCA. On the other hand the inner product is 0.999 using procedure defined by (16) with $\beta = 0.5$ and $\eta = 130$. The RM algorithm is started with the initial vector $\gamma = (1/\sqrt{200}, \cdots, 1/\sqrt{200})^{\mathrm{T}}$ and $\mu = (0, \cdots, 0)$, which assigns less weight to the 30 outliers after 10 iterations.

In this procedure, we heuristically choose the tuning parameters β and η. We observe in Figure 2 that the performance is not so sensitive to the choice of β if $\eta < 145$. In the subsequent simulation, we will investigate the data-adaptive selection for tuning parameters using the 10-fold CV method in Section 4.

Secondly, we investigate the case in which ε is fairly large. Hence, we take $\varepsilon = 0.5$, which can be viewed as the worst case in our context. We focus on the case of structural contamination with the same setting as that with $\varepsilon = 0.03$. The situation is really an extreme case, and is beyond the usual context of outlier detection. Thus, we have several simulations involving this situation as the first step, as seen in Figure 3. A typical result gives 0.101 as the inner product between the first

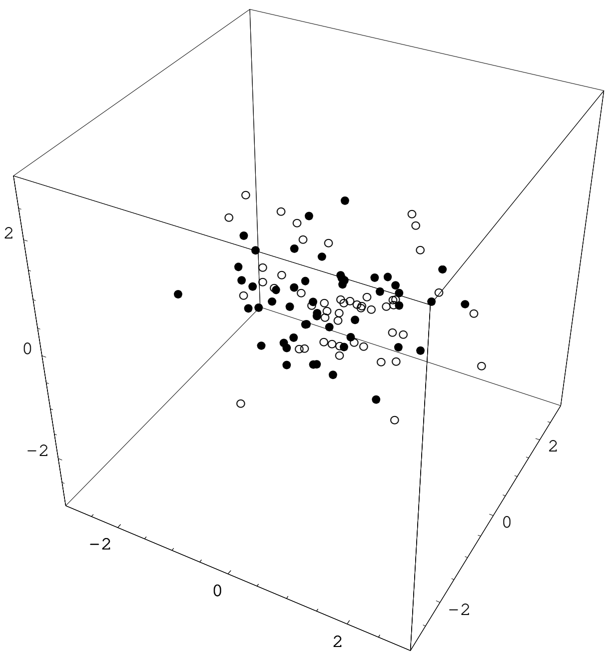


Figure 3: The plot of 50 observations and 50 outliers in the minor subspace.

principal component vectors with and without 50 outliers by the classical procedure. Alternatively our procedure gives 0.833, so we see that the procedure can detect a more sensible direction vector than the classical procedure. The proposed procedure detects the heterogeneity of structure, as indicated in Figure 4. If we have more information on the contamination or outlying structure, then we can build a shaper model for the outlier distribution. For example, we might suggest a two component Gaussian mixture model and estimate the structure in a more complete situation via the EM algorithm. However, to expect exact information on the outliers is often unrealistic in the present situation.

We apply the RM algorithm to on-line input vectors for comparison with the gradient learning algorithm. For the simulation study, we assume a specific form of model (ii) with $\varepsilon = 0.1$ and Gaussian density with $\mu = 0, \mu_1 = (30,0,0,0,30)^{\mathrm{T}}$, $V = \mathrm{diag}(9,7,5,3,1)$, and $V_1 = \mathrm{diag}(1,1,3,3,5)$. Thus, the true principal component vector of V is $(1,0,0,0,0)$ in the simulation design. We observe that the RM algorithm is stable and attains rapid convergence from these on-line input vectors. However, the computational burden is much heavier than the on-line gradient algorithm, as shown in Figure 5.

We next investigate the numerical performance of the 10-fold CV method discussed in Section 4 under the model of the structural contamination as follows: 45 input vectors are simulated from a Gaussian density $N(0,\mathrm{diag}(9,7,5,3,1))$ and five outliers from $N((0,0,0,0,10),\mathrm{diag}(9,7,5,3,1))$. We observe that the 10-fold CV method has detected the optimal tuning parameter $\eta = 46$, as shown in Figures 6 (1) and (2), while the PCA is much less sensitive to β than to η, so we fix the optimal tuning parameter as $\beta = 1$.

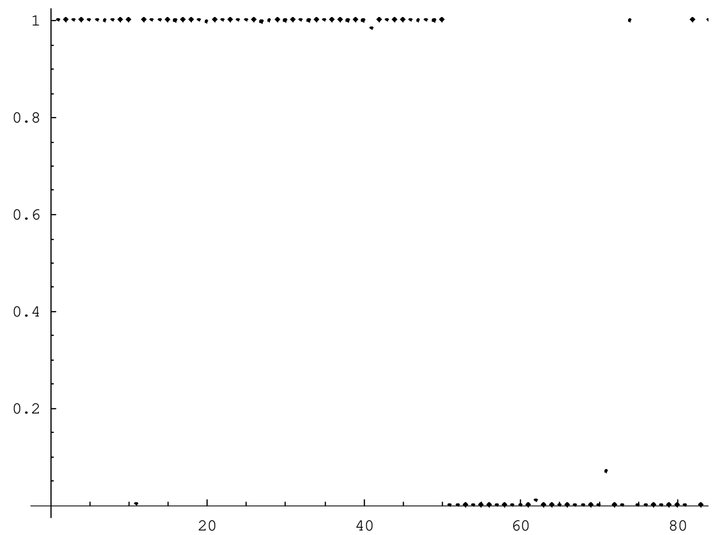


Figure 4: The plot of the weight function of ψ over 50 data with 50 outliers.

We propose a robust procedure for centralization of the data in (12). In the neural networks literature, such a variant for centralizing data has been ignored until now. Using the usual centralization is correct if all of the data are generated from Gaussian distribution. This is because the mean vector and covariance matrix are orthonormal as parameters, so that any influence on the principal vectors is independent of that on the mean vector. However, if the data in the mean vector are structured, this sometimes has a significant impact on the PCA. Here we consider a simple simulation study for investigating the difference between two procedures defined by adopting the weighted sample mean vector $\hat{\mu}$ in the RM algorithm and the sample mean vector $\bar{\mu}$ in classical PCA as the centralizer. We generate 180 observations from a Gaussian density $N(0,V)$ with $V = \mathrm{diag}(9,7,5,3,1)$ and 20 outliers from $N((0,0,20,0,0)^{\mathrm{T}},V)$. Figure 7 shows the two-dimensional score plot produced by the classical PCA based on only the 180 observations without any outliers, where the horizontal axis is taken as taken exactly as the first principal component vector. We observe that the first principal component vector by $\hat{\mu}$-centralization yields a proper direction.

6. Discussion

We have discussed a class of procedures for PCA based on the generic functions Ψ. The derivative ψ gives a weight expressing the degree of confidence for each input vector being an outlier. The robust procedure for the PCA gives less weights to input vectors having long residual vectors when projected onto the principal component subspace. We emphasis that the μ-portion is defined to be a weighted mean with the same weights as in the Γ-portion while the usual PCA employs the naive centralization with constant weights.

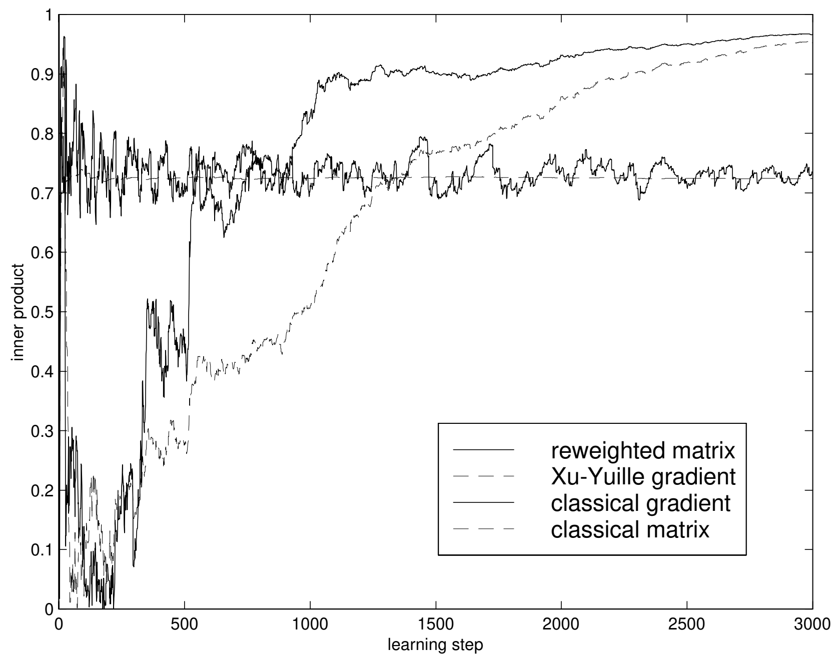


Figure 5: The inner products of the true vector $(1,0,0,0)$ and the PC vectors by RM, Xu-Yuille gradient, the classical gradient and classical matrix algorithms.

Our major point is the adaptive selection of a set of tuning parameters which control the degree of robustification. In empirical studies, we observe that the robustness performance is sensitive to the selection of tuning parameters. K-fold cross validation properly gives the adaptive selection for tuning parameters in accordance with data. However the selection method is done only for batch data but it cannot be applied to on-line data, which we must post as a future research. The RM algorithm needs the evaluation of eigenvalues and eigenvectors of the full matrix. In this respect it requires heavy computational burdens, whereas the convergence is stable and rapid relative to the gradient algorithm. The RM algorithm must be improved when the dimension of the input vector is considerably high. There is room for improvement in solving the k-dominant eigenvectors from a computational point of view.

Another interesting issue would be the breakdown point of the method proposed in the present paper as a global measure of robustness. In the previous literature the breakdown point has been considered as estimation of covariance (scatter) matrix other than estimation of principal component vector. However our method does not directly fit the theory since the method is not only a function of covariance matrix. We will need more discussion for this problem to be challenged as a future problem.

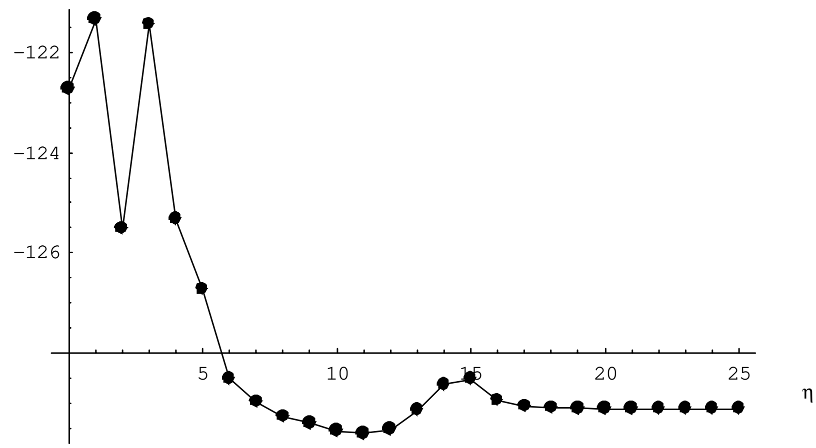


Figure 6: The plot of 10-fold CV for the Xu-Yuille rule for $\beta = 1$ fixed.

References

Amari, S. -I. Neural theory of association and concept formation. *Biological Cybernetics,* **26,** 175–185, 1977.

Campbell, N. A. Robust procedures in multivariate analysis 1: Robust covariance estimation. *Applied Statistics* **29**, 231-237, 1980.

Caussinus, H. and Ruiz, A. Interesting projections of multidimensional data by means of generalized principal component analysis. *COMPSTAT 90,* 121–126, 1990.

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

De la Torre, F. and Black, M. Robust principal component analysis for computer vision. *International Conference on Computer Vision,* 2001.

Hampel, F. R. The influence curve and its role in robust estimation. *Journal of the American Statistical Association,* **69,** 383–393, 1974.

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

Hastie, T. J., Tibshirani, R. J. and Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, 2001.

Haykin, S. *Neural Networks.* Prentice Hall, 1999.

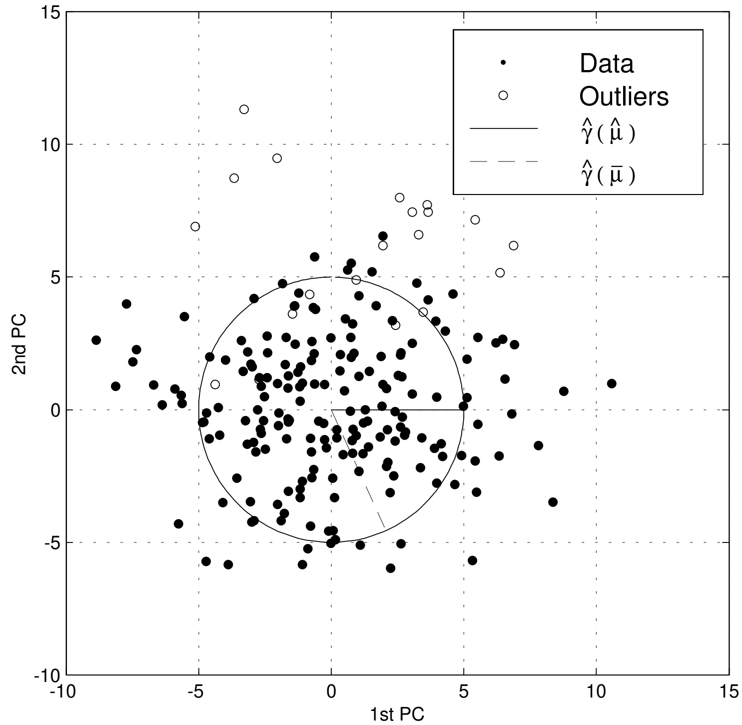


Figure 7: The effect of centralization ways.

Higuchi, I. and Eguchi, S. The influence function of principal component analysis by self-organizing rule. *Neural Computation,* **10,** 1435–1444, 1998.

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology,* **24,** 417–441, 1933.

Huber, P. J. *Robust Statistics.* Wiley, 1981.

Hyvarinen, A. and Oja, E. A fast fixed-point algorithm for independent component analysis. *Neural Computation,* **9,** 1483–1492, 1997

Kamiya, H. and Eguchi, S. A class of robust principal component vectors. *Journal of Multivariate Analysis,* **76**, 239–269, 2001.

McLachlan, G. J. and Krishnan, T. *The EM Algorithm and Extensions.* Wiley, 1997.

Oja, E. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology,* **15,** 267–273, 1982.

Oja, E. Neural networks, principal components and subspaces. *International Journal of Neural Systems,* **1,** 61–68, 1989.

Tanaka, Y. Sensitivity analysis in principal component analysis: influence on the subspace spanned by principal components. *Communications in Statistics -Theory and Methods,* **17,** 3157–3175, 1988.

Wu, C. F. J. On the convergence properties of the EM algorithm. *Annals of Statistics,* **11,** 95–103, 1983.

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