Communications for Statistical Applications and Methods
2017, Vol. 24, No. 2, 143–154

https://doi.org/10.5351/CSAM.2017.24.2.143
Print ISSN 2287-7843 / Online ISSN 2383-4757

Probabilistic penalized principal component analysis

Chongsun Park[1,a], Morgan C. Wang[b], Eun Bi Mo[a]

[a]Department of Statistics, Sungkyunkwan University, Korea;
[b]Department of Statistics, University of Central Florida, USA

Abstract

A variable selection method based on probabilistic principal component analysis (PCA) using penalized likelihood method is proposed. The proposed method is a two-step variable reduction method. The first step is based on the probabilistic principal component idea to identify principle components. The penalty function is used to identify important variables in each component. We then build a model on the original data space instead of building on the rotated data space through latent variables (principal components) because the proposed method achieves the goal of dimension reduction through identifying important observed variables. Consequently, the proposed method is of more practical use. The proposed estimators perform as the oracle procedure and are root-*n* consistent with a proper choice of regularization parameters. The proposed method can be successfully applied to high-dimensional PCA problems with a relatively large portion of irrelevant variables included in the data set. It is straightforward to extend our likelihood method in handling problems with missing observations using EM algorithms. Further, it could be effectively applied in cases where some data vectors exhibit one or more missing values at random.

Keywords: probability model, variable selection, penalized likelihood, EM algorithm, non-convex penalty, oracle estimators

1. Introduction

Dimension reduction is an important topic in statistics and other fields such as image processing, data compression, pattern recognition, and statistical data mining. Dimension reduction is becoming more important due to the increase in the size of data and a larger number of variables to consider in the information age. Principal component analysis (PCA; Jolliffe, 2002) is a frequently used method in dimension reduction especially in multivariate statistical analysis. PCA often gives a relatively small number of linear combinations of variables that can effectively explain a large portion of the variance in a given data set. Each component still includes many non-zero coefficients on irrelevant variables (i.e., it poses an interpretation problem especially when the number of non-zero coefficients is large).

Several methods are available to aid the interpretation when each component has many non-zero coefficients. Jolliffe (1972, 1973) examined methods that discard irrelevant variables based on threshold values using multiple correlations, PCA itself, and clustering.These methods are very simple; however, this might be misleading as pointed out by Cadima and Jolliffe (1995). Other methods that aid in the interpretation of principal components include orthogonal rotation, similar to those used in factor analysis (Jolliffe, 1989, 1995), that restrict the coefficients of the components to a small set of possible values such as $-1, 0, 1$ (Hausman, 1982; Vines, 2000) and to introduce penalty functions to

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: cspark@skku.edu

Published 31 March 2017 / journal homepage: http://csam.or.kr

force the coefficients of irrelevant variables to zero (Jolliffe, 2002). Jolliffe (1995) pointed out the rotation method might have problems and the L_1 penalty function proposed by Jolliffe *et al.* (2003) might cause bias on coefficient estimates. Zou *et al.* (2006) exploit the regression/reconstruction error property of principal components in order to obtain sparse principal components. Witten *et al.* (2009) proposed a penalized matrix decomposition with L_1 penalty that results in a regularized version of singular value decomposition for sparse principal components and canonical correlation analysis. These are able to enhance computing efficiency applicability from principal components to other methods; however, still we need more advanced algorithms when we decide to use other non-convex penalties like hard penalties or smoothly clipped absolute deviation (SCAD) rather than least absolute shrinkage and selection operator (LASSO) type L_1 penalty function. Xie *et al.* (2010) generalize existing penalized model-based clustering approaches for mixture of factor analyzers with application to clustering.

In this article, we use several L_1 type and non-convex penalty functions that include the hard thresholding (HARD) penalty function proposed by Antoniadis (1997) and Fan (1997) along with probabilistic PCA (PPCA) proposed by Tipping and Bishop (1999a, 1999b) to solve the problem of non-zero coefficients on irrelevant variables. PPCA enables us to utilize the likelihood idea to prove that the consistency and the sparsity of coefficients estimates along with asymptotic normality could be established through the convex and non-convex penalty functions. We show that the method works as well as if the correct sub-model were known with or without missing observations at random.

In Section 2, the PPCA with latent variable model will be introduced. In Section 3, three penalty functions, HARD, SCAD, and LASSO (Jolliffe *et al.*, 2003) will be added to the PPCA established in Section 2. Procedures for finding the maximum likelihood estimates (MLE) through EM algorithm are given in Section 4. In Section 5, simulation studies and real examples are given. We conclude in Section 6.

2. Probabilistic PCA with latent variables model

PCA (Jolliffe, 2002) is a well-known technique for dimension reduction for multivariate data sets and can be expressed as a latent variable model (Anderson and Rubin, 1956; Lawley, 1953). Tipping and Bishop (1999b) has shown that PCA may be viewed as a maximum likelihood procedure based on a probability density model of the observed data.

Suppose that we have p-dimensional data vectors $\{\boldsymbol{x}_n\}$, $n \in \{1, \ldots, N\}$ and sample covariance matrix $\boldsymbol{S} = \sum_n (\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^T/N$, where $\bar{\boldsymbol{x}}$ is the data sample mean with N observations. Standard PCA is the same as the solving the eigenvalue problem

$$\boldsymbol{S}\boldsymbol{w}_j = \delta_j \boldsymbol{w}_j \quad \text{for } j = 1, \ldots, q.$$

Then the q principal components of the observed vector $\boldsymbol{x}_n$ are

$$\boldsymbol{c}_n = \boldsymbol{W}^T(\boldsymbol{x}_n - \bar{\boldsymbol{x}}) \quad \text{with } \boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_q)$$

such that q ($\leq p$) principal axes $\boldsymbol{w}_j$ are those orthonormal axes onto which the retained variance under projection is maximal. The components $\boldsymbol{c}_n$ are then uncorrelated such that the covariance matrix $\Sigma_n \boldsymbol{c}_n \boldsymbol{c}_n^T/N$ is diagonal with elements δ_js.

This PCA can be expressed as a latent variable model that relates p-dimensional random vector $\boldsymbol{x}$ to a corresponding q-dimensional vector of latent variables $\boldsymbol{c}$ as

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{c} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{2.1}$$

with the conventional assumption of $\boldsymbol{c} \sim N(\boldsymbol{0}, \boldsymbol{I})$. Using the isotropic noise model $N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ for $\boldsymbol{\epsilon}$ in conjunction with equation (2.1) implies that the $\boldsymbol{c}$-conditional probability distribution over $\boldsymbol{x}$-space is given by

$$\boldsymbol{x}|\boldsymbol{c} \sim N\left(\boldsymbol{W}\boldsymbol{c} + \boldsymbol{\mu}, \sigma^2\boldsymbol{I}\right).$$

The marginal distribution for $\boldsymbol{x}$ is obtained by integrating the latent variables and is also normal:

$$\boldsymbol{x} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Psi}\right),$$

with $\boldsymbol{\Psi} = \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$. If we replace $\sigma^2\boldsymbol{I}$ in $\boldsymbol{\Psi}$ with a diagonal matrix of positive but different values for each diagonal elements, then it becomes a well-known factor analysis. Finally, the log-likelihood for $\boldsymbol{W}$ and σ^2 becomes

$$l(\boldsymbol{W}, \sigma^2) = -\frac{N}{2}\left\{p\ln(2\pi) + \ln|\boldsymbol{\Psi}| + \mathrm{tr}\left(\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}\right)\right\},$$

where

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})(\boldsymbol{x}_n - \boldsymbol{\mu})^T.$$

Tipping and Bishop (1999b) suggested using the iterative EM algorithm to find estimates for $\boldsymbol{W}$ and σ^2 and mentioned that it could be effectively applied in cases where some of the data vectors exhibit one or more missing values at random and in other situations. We can find the conditional distribution of the latent variable $\boldsymbol{c}$ given the observed $\boldsymbol{x}$, that may be calculated using Bayes' rule,

$$\boldsymbol{c}|x \sim N\left(M^{-1}\boldsymbol{W}^T(\boldsymbol{x} - \boldsymbol{\mu}), \sigma^2\boldsymbol{M}^{-1}\right),$$

with $\boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I}$.

3. Penalized probabilistic PCA

We can consider the problem of extending the penalized likelihood idea to PPCA for variable selection in each principal component. A form of penalized likelihood becomes

$$l\left(\boldsymbol{W}, \sigma^2\right) - N\sum_{i=1}^{p}\sum_{j=1}^{q}p_\lambda(w_{ij}),$$

with w_{ij} as the element of $\boldsymbol{W}$ in its i^{th} row and j^{th} column and $p_\lambda(\cdot)$ as a penalty function. Fan and Li (2001) argued that unbiasedness, sparsity, and continuity are three properties possessed by a good penalty function, and suggested the SCAD penalty function as the best one for regression problems.

Several well-known penalty functions including SCAD penalty function are as follows.

1. L_p: $p_\lambda(w_{ij}) = \lambda|w_{ij}|^p$ and it becomes LASSO with $p = 1$ for least squares case.

2. HARD penalty:

$$p_\lambda(w_{ij}) = \lambda^2 - (|w_{ij}| - \lambda)^2 I(|w_{ij}| < \lambda).$$

3. SCAD penalty:

$$p_\lambda(w_{ij}) = \begin{cases} \lambda w_{ij}, & \text{if } w_{ij} < \lambda, \\ -\dfrac{w_{ij}^2 - 2a\lambda w_{ij} + \lambda^2}{2(a-1)}, & \text{if } \lambda \le w_{ij} < a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & \text{if } w_{ij} \ge a\lambda. \end{cases}$$

Unfortunately, none of three penalty functions simultaneously satisfy all of three properties mentioned above. L_p penalty function is biased and this causes serious problems especially when applied to PCA problems in which coefficients compete with each other due to orthonormality conditions. Fan and Li (2001) noted that consistency and oracle properties cannot be satisfied simultaneously for the L_1 penalty. The HARD penalty function is unbiased and sparse, but not continuous. SCAD behaves somewhat between L_1 and HARD and needs two dimensional burdensome generalized cross-validation (CV) or usual CV to find optimal values for two parameters, a and λ.

Overall, it appears reasonable to use HARD penalty for the PCA problem since it seems better at forcing coefficients of irrelevant variables to zero (Tables 1–3) and at the same time in preserving original directions after introducing the penalty function (Figure 1) in the maximum likelihood procedure.

The consistency and sparsity of the HARD and SCAD penalty function for our nonconcave penalized PPCA (PenPPCA) estimator can be established through similar procedures in Fan and Li (2001) or Fan and Peng (2004). There exists a penalized likelihood estimator that converges at the rate $O_p(n^{-1/2} + a_n)$, where $a_n = \max_i\{p'_{\lambda_n}(w_{ij}) : w_{ij} \neq 0\}$. This implies that for the HARD and SCAD penalty functions, the penalized likelihood estimator is root-n consistent if $\lambda_n \to 0$. Furthermore, it can be shown that such a root-n consistent estimator must satisfy $\hat{\boldsymbol{w}}_2 = \mathbf{0}$ and this implies that the penalized likelihood estimator performs as well as if $\boldsymbol{w}_{20} = \mathbf{0}$ were known under the assumption that the single selected component of $\boldsymbol{W}$ can be divided as

$$\boldsymbol{w}_0 = (w_{10}, \ldots, w_{p0}) = \left(\boldsymbol{w}_{10}^T, \boldsymbol{w}_{20}^T\right)^T$$

and $\boldsymbol{w}_{20} = \mathbf{0}$ without loss of generality.

4. MLE with EM algorithm

MLE can be obtained via the EM algorithm (Green, 1990) as treating $\boldsymbol{c}_n$ as missing so completing the data set as $(\boldsymbol{x}_n, \boldsymbol{c}_n)$ (Tipping and Bishop, 1999b). The corresponding complete-data log-likelihood is:

$$l_{\boldsymbol{x},\boldsymbol{c}} = \sum_{n=1}^{N} \ln\{f(\boldsymbol{x}_n, \boldsymbol{c}_n)\},$$

with

$$f(\boldsymbol{x}_n, \boldsymbol{c}_n) = \left(2\pi\sigma^2\right)^{-\frac{p}{2}} \exp\left\{-\frac{\|\boldsymbol{x}_n - \boldsymbol{W}\boldsymbol{c}_n - \boldsymbol{\mu}\|^2}{2\sigma^2}\right\} (2\pi)^{-\frac{q}{2}} \exp\left\{\frac{\|\boldsymbol{c}_n\|^2}{2}\right\}.$$

Table 1: Average and standard deviation (in parenthesis) of the number of correct and incorrect 0 coefficients for $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ with L_1

		$\boldsymbol{W}_1$				$\boldsymbol{W}_2$			
N	λ	Correct		Incorrect		Correct		Incorrect	
		No missing	Missing	No missing	Missing	No missing	Missing	No missing	Missing
20	0	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-
	0.3	0.45 (0.61)	-	0.00 (0.00)	-	0.46 (0.58)	-	0.01 (0.10)	-
	0.7	1.64 (0.76)	-	0.02 (0.14)	-	1.67 (0.84)	-	0.11 (0.32)	-
	1.0	2.17 (0.73)	-	0.04 (0.20)	-	2.06 (0.68)	-	0.21 (0.41)	-
50	0	0.01 (0.10)	-	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-
	0.3	0.81 (0.62)	-	0.00 (0.00)	-	0.62 (0.60)	-	0.00 (0.00)	-
	0.7	2.17 (0.62)	-	0.00 (0.00)	-	2.17 (0.74)	-	0.03 (0.17)	-
	1.0	2.65 (0.52)	-	0.00 (0.00)	-	2.64 (0.58)	-	0.07 (0.26)	-
100	0	0.01 (0.10)	-	0.00 (0.00)	-	0.02 (0.14)	-	0.00 (0.00)	-
	0.3	0.99 (0.69)	-	0.00 (0.00)	-	1.01 (0.72)	-	0.00 (0.00)	-
	0.7	2.39 (0.60)	-	0.00 (0.00)	-	2.51 (0.60)	-	0.00 (0.00)	-
	1.0	2.76 (0.43)	-	0.00 (0.00)	-	2.72 (0.47)	-	0.00 (0.00)	-
300	0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.10)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.3	1.27 (0.70)	1.27 (0.67)	0.00 (0.00)	0.00 (0.00)	1.43 (0.74)	0.97 (0.66)	0.00 (0.00)	0.00 (0.00)
	0.7	2.78 (0.44)	2.66 (0.48)	0.00 (0.00)	0.00 (0.00)	2.67 (0.50)	2.58 (0.55)	0.00 (0.00)	0.00 (0.00)
	1.0	2.86 (0.38)	2.83 (0.38)	0.00 (0.00)	0.00 (0.00)	2.79 (0.43)	2.67 (0.47)	0.00 (0.00)	0.00 (0.00)
500	0	0.02 (0.14)	0.03 (0.17)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.3	1.48 (0.69)	1.32 (0.57)	0.00 (0.00)	0.00 (0.00)	1.25 (0.67)	1.27 (0.63)	0.00 (0.00)	0.00 (0.00)
	0.7	2.85 (0.36)	2.66 (0.50)	0.00 (0.00)	0.00 (0.00)	2.69 (0.49)	2.63 (0.51)	0.00 (0.00)	0.00 (0.00)
	1.0	2.89 (0.32)	2.78 (0.42)	0.00 (0.00)	0.00 (0.00)	2.76 (0.43)	2.69 (0.49)	0.00 (0.00)	0.00 (0.00)

Table 2: Average and standard deviation (in parenthesis) of the number of correct and incorrect 0 coefficients for $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ with hard thresholding

		$\boldsymbol{W}_1$				$\boldsymbol{W}_2$			
N	λ	Correct		Incorrect		Correct		Incorrect	
		No missing	missing	No missing	Missing	No missing	Missing	No missing	Missing
20	0	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-
	0.3	1.11 (0.78)	-	0.01 (0.10)	-	1.18 (0.76)	-	0.04 (0.20)	-
	0.7	1.94 (0.65)	-	0.04 (0.20)	-	1.86 (0.71)	-	0.17 (0.38)	-
	1.0	2.22 (0.71)	-	0.08 (0.27)	-	2.19 (0.66)	-	0.17 (0.38)	-
50	0	0.01 (0.10)	-	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-
	0.3	1.72 (0.64)	-	0.00 (0.00)	-	1.69 (0.80)	-	0.01 (0.10)	-
	0.7	2.07 (0.66)	-	0.00 (0.00)	-	2.23 (0.62)	-	0.04 (0.20)	-
	1.0	2.52 (0.69)	-	0.00 (0.00)	-	2.51 (0.60)	-	0.05 (0.22)	-
100	0	0.01 (0.10)	-	0.00 (0.00)	-	0.02 (0.14)	-	0.00 (0.00)	-
	0.3	1.97 (0.72)	-	0.00 (0.00)	-	1.96 (0.70)	-	0.00 (0.00)	-
	0.7	2.21 (0.69)	-	0.00 (0.00)	-	2.27 (0.62)	-	0.00 (0.00)	-
	1.0	2.66 (0.56)	-	0.00 (0.00)	-	2.63 (0.53)	-	0.01 (0.10)	-
300	0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.10)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.3	2.14 (0.65)	2.18 (0.66)	0.00 (0.00)	0.00 (0.00)	2.06 (0.63)	1.93 (0.64)	0.00 (0.00)	0.00 (0.00)
	0.7	2.37 (0.68)	2.42 (0.65)	0.00 (0.00)	0.00 (0.00)	2.32 (0.62)	2.18 (0.67)	0.00 (0.00)	0.00 (0.00)
	1.0	2.76 (0.43)	2.83 (0.38)	0.00 (0.00)	0.00 (0.00)	2.72 (0.47)	2.58 (0.52)	0.00 (0.00)	0.00 (0.00)
500	0	0.02 (0.14)	0.03 (0.17)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.3	2.24 (0.65)	2.17 (0.55)	0.00 (0.00)	0.00 (0.00)	2.04 (0.71)	2.06 (0.58)	0.00 (0.00)	0.00 (0.00)
	0.7	2.55 (0.67)	2.38 (0.57)	0.00 (0.00)	0.00 (0.00)	2.49 (0.60)	2.37 (0.61)	0.00 (0.00)	0.00 (0.00)
	1.0	2.76 (0.43)	2.72 (0.45)	0.00 (0.00)	0.00 (0.00)	2.78 (0.42)	2.64 (0.50)	0.00 (0.00)	0.00 (0.00)

Table 3: Average and standard deviation (in parenthesis) of the number of correct and incorrect 0 coefficients for $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ with smoothly clipped absolute deviation

N	λ	$\boldsymbol{W}_1$				$\boldsymbol{W}_2$			
		Correct		Incorrect		Correct		Incorrect	
		No missing	Missing	No missing	Missing	No missing	Missing	No missing	Missing
20	0	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-
	0.3	0.70 (0.70)	-	0.01 (0.10)	-	0.61 (0.63)	-	0.02 (0.14)	-
	0.7	1.57 (0.78)	-	0.01 (0.10)	-	1.40 (0.80)	-	0.107 (0.26)	-
	1.0	2.07 (0.73)	-	0.06 (0.24)	-	2.03 (0.76)	-	0.19 (0.39)	-
50	0	0.01 (0.10)	-	0.00 (0.00)	-	0.00 (0.00)	-	0.00 (0.00)	-
	0.3	1.19 (0.72)	-	0.00 (0.00)	-	1.10 (0.80)	-	0.01 (0.10)	-
	0.7	1.97 (0.73)	-	0.00 (0.00)	-	2.11 (0.67)	-	0.04 (0.20)	-
	1.0	2.60 (0.60)	-	0.00 (0.00)	-	2.58 (0.57)	-	0.04 (0.20)	-
100	0	0.01 (0.10)	-	0.00 (0.00)	-	0.02 (0.14)	-	0.00 (0.00)	-
	0.3	1.40 (0.71)	-	0.00 (0.00)	-	1.47 (0.80)	-	0.00 (0.00)	-
	0.7	2.25 (0.73)	-	0.00 (0.00)	-	2.33 (0.68)	-	0.00 (0.00)	-
	1.0	2.74 (0.44)	-	0.00 (0.00)	-	2.71 (0.46)	-	0.00 (0.00)	-
300	0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.10)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.3	1.99 (0.72)	1.87 (0.71)	0.00 (0.00)	0.00 (0.00)	1.94 (0.68)	1.54 (0.77)	0.00 (0.00)	0.00 (0.00)
	0.7	2.70 (0.48)	2.51 (0.60)	0.00 (0.00)	0.00 (0.00)	2.62 (0.58)	2.33 (0.60)	0.00 (0.00)	0.00 (0.00)
	1.0	2.86 (0.35)	2.82 (0.39)	0.00 (0.00)	0.00 (0.00)	2.80 (0.40)	2.65 (0.50)	0.00 (0.00)	0.00 (0.00)
500	0	0.02 (0.14)	0.03 (0.17)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.3	2.13 (0.63)	2.19 (0.71)	0.00 (0.00)	0.00 (0.00)	2.04 (0.60)	1.86 (0.57)	0.00 (0.00)	0.00 (0.00)
	0.7	2.84 (0.37)	2.58 (0.56)	0.00 (0.00)	0.00 (0.00)	2.62 (0.55)	2.50 (0.52)	0.00 (0.00)	0.00 (0.00)
	1.0	2.79 (0.43)	2.78 (0.42)	0.00 (0.00)	0.00 (0.00)	2.76 (0.43)	2.64 (0.43)	0.00 (0.00)	0.00 (0.00)

Now let $Q(\boldsymbol{W}', \sigma^{2'}|\boldsymbol{W}, \sigma^2) = E(\log\{f(\boldsymbol{c}|\boldsymbol{W}', \sigma^{2'})\}|\boldsymbol{x}, \boldsymbol{W}, \sigma^2)$. Then the EM algorithm is obtained by repeatedly replacing a trial estimate of $(\boldsymbol{W}, \sigma^2)$ by those $(\boldsymbol{W}', \sigma^{2'})$ maximizing

$$Q\left(\boldsymbol{W}', \sigma^{2'}\middle|\boldsymbol{W}, \sigma^2\right) - P_\lambda\left(\boldsymbol{W}'\right),$$

with $P_\lambda(\boldsymbol{W}') = N\sum_{i=1}^{p}\sum_{j=1}^{q} p_\lambda(w'_{ij})$.

4.1. E-step

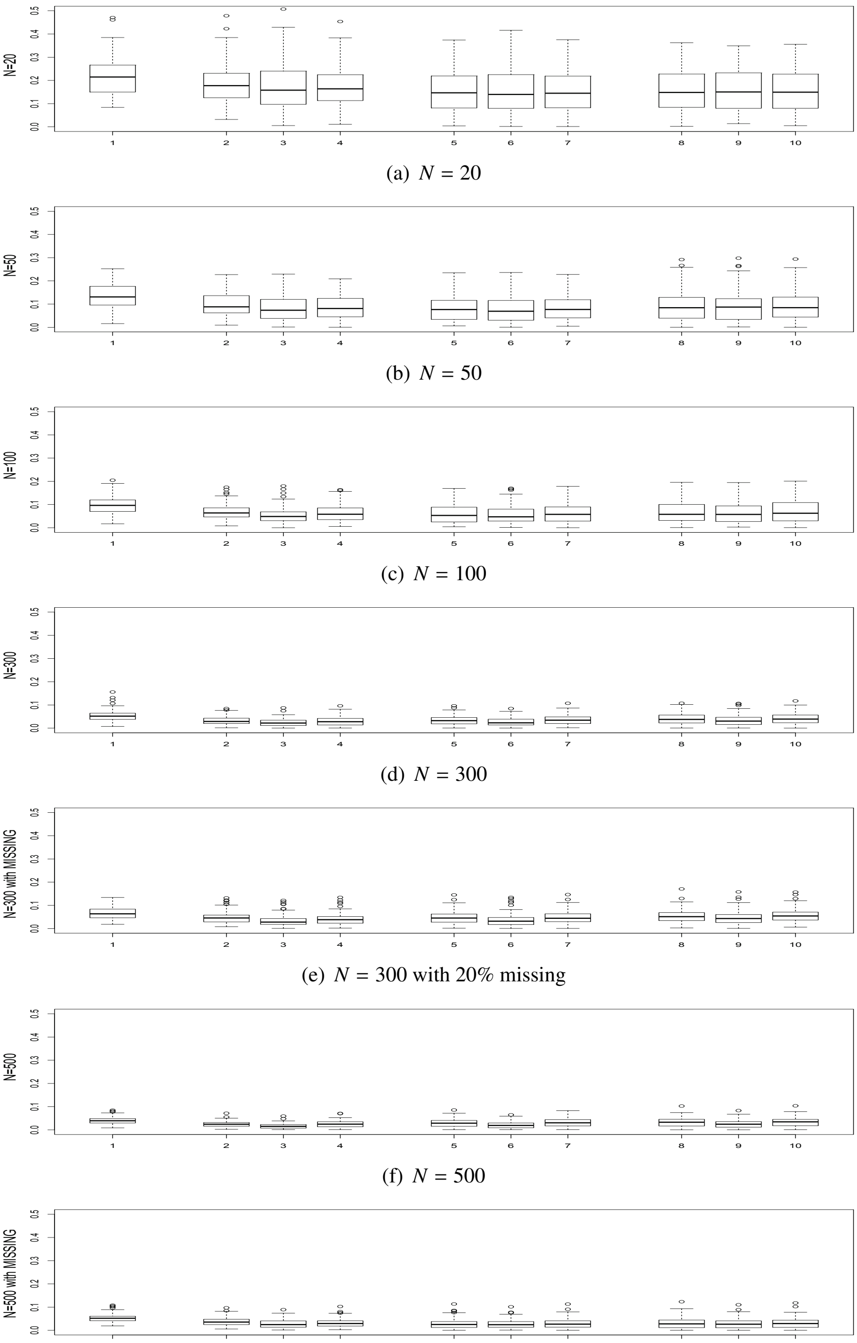
E-step takes the expectation of $l_{\boldsymbol{x},\boldsymbol{c}}$ with respect to the distribution $f(\boldsymbol{c}_n|\boldsymbol{x}_n, \boldsymbol{W}, \sigma^2)$:

$$\begin{aligned}\langle l_{\boldsymbol{x},\boldsymbol{c}}\rangle = -\sum_{n=1}^{N}\Bigg\{&\frac{p}{2}\ln\sigma^2 + \frac{1}{2}\mathrm{tr}\left(\left\langle \boldsymbol{c}_n\boldsymbol{c}_n^T\right\rangle\right) + \frac{1}{2\sigma^2}(\boldsymbol{x}_n - \boldsymbol{\mu})^T(\boldsymbol{x}_n - \boldsymbol{\mu})\\ &-\frac{1}{\sigma^2}\langle \boldsymbol{c}_n\rangle^T\boldsymbol{W}^T(\boldsymbol{x}_n - \boldsymbol{\mu}) + \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{W}^T\boldsymbol{W}\left\langle \boldsymbol{c}_n\boldsymbol{c}_n^T\right\rangle\right)\Bigg\},\end{aligned}$$

where we have omitted terms independent of the model parameters and

$$\begin{aligned}\langle \boldsymbol{c}_n\rangle &= \boldsymbol{M}^{-1}\boldsymbol{W}^T(\boldsymbol{x}_n - \boldsymbol{\mu}),\\ \left\langle \boldsymbol{c}_n\boldsymbol{c}_n^T\right\rangle &= \sigma^2\boldsymbol{M}^{-1} + \langle \boldsymbol{c}_n\rangle\langle \boldsymbol{c}_n\rangle^T,\end{aligned}$$

in which $\boldsymbol{M} = \boldsymbol{W}^T\boldsymbol{W} + \sigma^2\boldsymbol{I}$ as before.



(a) $N = 20$

(b) $N = 50$

(c) $N = 100$

(d) $N = 300$

(e) $N = 300$ with 20% missing

(f) $N = 500$

(g) $N = 500$ with 20% missing

Figure 1: *Angles between true and estimated 1^{st} component. From the left to the right in each boxplot: $\lambda = 0$ with no penalty (1), $\lambda = 0.3$ for L_1 (2) · HARD (3) · SCAD (4), $\lambda = 0.7$ for L_1 (5) · HARD (6) · SCAD (7), $\lambda = 1.0$ for L_1 (8) · HARD (9) · SCAD (10) [Plot numbers are in (·).]. HARD = hard thresholding; SCAD = smoothly clipped absolute deviation.*

4.2. M-step

In M-step $\langle l_{\boldsymbol{x},\boldsymbol{c}}\rangle - P_\lambda(\boldsymbol{W})$ is maximized with respect to $\boldsymbol{W}$ and σ^2. First, the optimal estimate $\widetilde{\boldsymbol{W}}$ for $\boldsymbol{W}$ can be obtained by solving

$$\frac{\partial Q\left(\boldsymbol{W}', \sigma^{2'} \middle| \boldsymbol{W}, \sigma^2\right)}{\partial \boldsymbol{W}'} - \frac{\partial P_\lambda\left(\boldsymbol{W}'\right)}{\partial \boldsymbol{W}'} = 0$$

and it becomes

$$\begin{aligned}
&\frac{\partial Q\left(\boldsymbol{W}', \sigma^{2'} \middle| \boldsymbol{W}, \sigma^2\right)}{\partial \boldsymbol{W}'} - \frac{\partial P_\lambda\left(\boldsymbol{W}'\right)}{\partial \boldsymbol{W}'} \\
&= \sum_{n=1}^{N} \left\{ \frac{1}{\sigma^{2'}} (\boldsymbol{x}_n - \boldsymbol{\mu}) \langle \boldsymbol{c}_n \rangle^T - \frac{1}{\sigma^{2'}} \boldsymbol{W} \left\langle \boldsymbol{c}_n \boldsymbol{c}_n^T \right\rangle \right\} - \frac{\partial P_\lambda\left(\boldsymbol{W}'\right)}{\partial \boldsymbol{W}'} \\
&= 0
\end{aligned} \tag{4.1}$$

We need to solve equation (4.1) to estimate optimal solution for the $\widetilde{\boldsymbol{W}}$. As an example, for a L_1 penalty function, it would be

$$\begin{aligned}
\widehat{\boldsymbol{W}} &= \left\{ \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}) \langle \boldsymbol{c}_n \rangle^T - \sigma^{2'} \lambda N 1_{p\times q} \left(\operatorname{sgn}\left(w'_{ij} \right) \right) \right\} \left[\sum_{n=1}^{N} \left\langle \boldsymbol{c}_n \boldsymbol{c}_n^T \right\rangle \right]^{-1} \\
&= \left[\boldsymbol{S}\boldsymbol{W}\boldsymbol{M}^{-T} - \sigma^{2'} \lambda 1_{p\times q} \left(\operatorname{sgn}\left(w'_{ij} \right) \right) \right] \left[\sigma^{2'} \boldsymbol{M}^{-1} + \boldsymbol{M}^{-1} \boldsymbol{W}^T \boldsymbol{S}\boldsymbol{W}\boldsymbol{M}^{-T} \right]^{-1}.
\end{aligned}$$

However, for HARD and SCAD penalty functions that include some or all terms of w'_{ij} from $\boldsymbol{W}'$, we need to use an iterative algorithm to find an optimal $\boldsymbol{W}'$. Since it is not practical to find the information matrix with respect to $\boldsymbol{W}$, we should use an optimization method that only uses the first derivatives. A steepest descent algorithm using line search would be one possible choice from several well-known algorithms. We used “nlminb” function in R for our line search procedure.

Second, the estimate for σ^2 can be calculated as

$$\begin{aligned}
\widehat{\sigma^2} &= \frac{1}{Np} \sum \left\{ \|\boldsymbol{x}_n - \boldsymbol{\mu}\|^2 - 2\langle \boldsymbol{c}_n \rangle^T \widetilde{\boldsymbol{W}}^T (\boldsymbol{x}_n - \boldsymbol{\mu}) + \operatorname{tr}\left(\left\langle \boldsymbol{c}_n \boldsymbol{c}_n^T \right\rangle \widetilde{\boldsymbol{W}}^T \widetilde{\boldsymbol{W}} \right) \right\} \\
&= \frac{1}{p} \operatorname{tr}\left(\boldsymbol{S} - \boldsymbol{S}\boldsymbol{W}\boldsymbol{M}^{-1} \widetilde{\boldsymbol{W}}^T \right).
\end{aligned}$$

Finally, E- and M-steps are iterated in sequence until the algorithm is judged to have converged.

We conclude the proposed algorithm for PenPPCA by mentioning appropriate initial values for $\boldsymbol{W}$ and ways to find an optimal λ.

- Initial $\boldsymbol{W}$: Coefficient estimates from PCA would be reasonable initial values for $\boldsymbol{W}$’s and for each subsequent component in the PenPPCA.

- Optimal λ: For an optimal λ generalized or usual CV techniques may be used as suggested by Breiman (1995), Fu (1998), and Tibshirani (1996).

5. Numerical comparisons

This section tests the accuracy of the proposed approach and compare the performance of our method with existing ones. In the first small simulation study, we reported the number of true and false zeroes and angles between true and estimated component for 100 replicated simulated data sets with 20% and without missing values. Numerical comparison of our newly proposed method with simplified component technique-LASSO (ScoTLASS) (Jolliffe *et al.*, 2003) is included in the second real data example. All numerical comparisons are conducted using R codes.

5.1. Small simulation study

For any given vector of positive real numbers and an orthogonal matrix, we can find a covariance matrix whose eigenvalues are the elements of a given vector, and whose eigenvectors are the columns of a given matrix. The data sets for the study are simulated as follows. This is based on the observation that $\boldsymbol{x}$ is marginally distributed as normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Psi}= \boldsymbol{W}\boldsymbol{W}^T + \sigma^2\boldsymbol{I}$. Further we can set $\boldsymbol{\mu}$ to zero without loss of generality. The following sets of data are generated 100 times for each combination.

- N: Number of observations is considered are $20, 50, 100, 300, 500$.
- p: Number of variables is fixed at 5.
- $\boldsymbol{\mu}$: Mean vector of $\boldsymbol{x}$ is set to $\boldsymbol{0}$ without loss of generality.
- Eigenvectors of

$$\boldsymbol{\Psi} = \begin{pmatrix} 0.95 & 0 & -0.27 & 0.15 & -0.07 \\ 0 & 0.95 & 0.16 & 0.24 & -0.10 \\ 0.32 & 0 & 0.81 & -0.44 & 0.22 \\ 0 & -0.32 & 0.49 & 0.73 & -0.36 \\ 0 & 0 & 0 & 0.45 & 0.90 \end{pmatrix}$$

(Eigenvectors are columns of the matrix and the last three vectors are selected with no special purpose.).

- Eigenvalues of $\boldsymbol{\Psi}$ are $(2.5, 1.5, 0.4, 0.3, 0.3)^T$.

We compare the first two components associated with the largest and the second largest eigenvalues for assessing behavior of the proposed PenPPCA. We look at estimated directions and their standardized values for comparison with true W and σ^2. The average number of zero estimates for true zero (correct) coefficients for each case, and zero estimates for non-zero coefficients (incorrect) are of particular concern. We included the results of only four preset λs of 0 (no penalty), 0.3, 0.7 and 1.0 in Table 1 through Table 3 with L_1, HARD, and SCAD penalty, respectively. In addition, boxplots of angle between true and estimated 1^{st} component were also reported.

The results show that the number of true zeros increases as n increases and is bigger when the λ is 1 than other three cases. Most of the estimates for true non-zero coefficients are also different from zero when the number of observation is larger than 100. We can therefore say that, with a relatively large number of observations and an appropriate λ our method effectively forces estimates of true zero to zero and at the same time seldom gives zero estimates for true non-zero coefficients regardless of

Table 4: Descriptions for 13 variables measured

Variable	Description
x_1	The top diameter in inches
x_2	The length in inches
x_3	The moisture content, % of dry weight
x_4	The specific gravity at the time of the test
x_5	The oven-dry specific gravity
x_6	The number of annual rings at the top
x_7	The number of annual rings at the base
x_8	The maximum bow in inches
x_9	The distance of the point of maximum bow from the top in inches
x_{10}	The number of knot whorls
x_{11}	The length of clear prop from the top in inches
x_{12}	The average number of knots per whorl
x_{13}	The average diameter of the knots in inches

penalty functions used. For each fixed n, the number of correct 0 coefficients tends to converge to 3 for a range of λ values until one variables dominates with coefficient estimate (after standardizing) 1.

From boxplots in Figure 1, HARD penalty function have smallest bias in most of the cases and standard deviations are also relatively smaller than other penalty functions. However, there seems to have no big difference between cases with and without missing observations.

5.2. A real example

We considered the pitprop data set, introduced by Jeffers (1967), in which a PCA was conducted on the correlation matrix of 13 measurements made on a sample size of 180 pitprops cut from Corsican pine timber from East Anglia. Descriptions for variables measured are included in Table 4.

In this example, we numerically compare the proposed method with LASSO approach in PCA (ScoTLASS) for several values of tuning parameters. We tried three values of 0.5, 1, and 1.5 for tuning parameter λ and found that there seems to be too many 0 estimates with λ of 1.5 compared to the results of ScoTLASS, in which they reported results of $t = 1.75$ and $t = 2.25$. Variables with 0 coefficients for λs of 0.5 and 1.0 for our method together with the results of Jolliffe *et al.* (2003) are reported in Table 5.

Overall, variables with 0 estimates are quite different between PenPPCA and ScoTLASS in 5 components except in the 1^{st} one. In the first component, variables x_1, x_2, x_7, x_8, x_9, and x_{10} are relevant in both methods. Variables of 0 estimates vary with tuning parameter t in ScoTLASS. Conversely, the proposed method is consistent in that variables with 0 estimates for tuning parameter $\lambda = 0.5$ become 0 for larger λ values with new variables added with 0 estimates.

It is interesting to note that ScoTLASS looks like giving fewer 0 estimates for components as its variance becomes smaller. In the 6^{th} component only one estimate is 0 for variable x_{10} in all values of t.

6. Discussions

We proposed a variable selection method in PCA via penalized likelihood approaches. From the family of penalty functions HARD seems to be the best in preserving original direction of coefficients for each principal component with root-n consistency and sparsity. The orthonormality of estimates for components are not guaranteed; however, we could apply the Gram-Schmidt method to obtain orthonormal components as in usual PCA if required. However, we found that orthonormalized direc-

Table 5: Coefficients of 0 estimates for several methods

Component	Method	Variable												
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
1	ScoTLASS (t = 1.75)			√	√	√	√					√	√	√
	ScoTLASS (t = 2.25)			√	√	√						√	√	√
	PenPPCA (λ = 0.50)						√							
	PenPPCA (λ = 1.00)				√		√					√		√
2	ScoTLASS (t = 1.75)	√				√			√	√	√	√		√
	ScoTLASS (t = 2.25)					√		√		√				
	PenPPCA (λ = 0.50)			√	√	√							√	
	PenPPCA (λ = 1.00)		√	√	√	√							√	√
3	ScoTLASS (t = 1.75)	√	√			√			√	√		√	√	
	ScoTLASS (t = 2.25)				√				√	√	√	√		
	PenPPCA (λ = 0.50)	√	√			√	√		√	√				√
	PenPPCA (λ = 1.00)	√	√			√	√	√	√	√			√	√
4	ScoTLASS (t = 1.75)			√	√	√	√	√		√		√		
	ScoTLASS (t = 2.25)					√		√				√		
	PenPPCA (λ = 0.50)							√		√	√			
	PenPPCA (λ = 1.00)	√	√			√	√	√		√	√			
5	ScoTLASS (t = 1.75)				√					√		√		
	ScoTLASS (t = 2.25)													
	PenPPCA (λ = 0.50)			√						√	√	√		
	PenPPCA (λ = 1.00)	√	√	√	√		√	√	√	√	√	√		
6	ScoTLASS (t = 1.75)													
	ScoTLASS (t = 2.25)										√			
	PenPPCA (λ = 0.50)	√	√	√					√					
	PenPPCA (λ = 1.00)	√	√	√	√	√			√			√	√	

ScoTLASS = simplified component technique-least absolute shrinkage and selection operator; PenPPCA = penalized probabilistic principal component analysis.

tions seem to be very close to those from the usual PCA algorithm in most cases.

The number of relevant components could be decided from estimates of $\sigma^2(p-q)$. Those values are quite close to the sum of eigenvalues for components not included in the first q components. Theoretical asymptotic tests for deciding the number of relevant components are under consideration. Clearly estimates of $\sigma^2(p-q)$ can be used to get an idea on the amount of variance explained by the first p components. The further asymptotic normality of estimates is straightforward; however, they are not included, since they are not frequently mentioned in standard PCA.

The proposed method can be successfully applied to high-dimensional PCA problems with a relatively large portion of irrelevant variables included in the data set. It is straightforward to extend our likelihood method in handling problems with missing observations using EM algorithms.

References

Anderson TW and Rubin H (1956). Statistical inference in factor analysis. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, 111–150.

Antoniadis A (1997). Wavelets in statistics: a review, *Journal of the Italian Statistical Society*, **6**, 97–144.

Breiman L (1995). Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373–384.

Cadima J and Jolliffe IT (1995). Loadings and correlations in the interpretation of principal components, *Journal of Applied Statistics*, **22**, 203–214.

Fan J (1997). Comments on ‘wavelets in statistics: a review’ by A. Antoniadis, *Journal of the Italian Statistical Society*, **6**, 131–138.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan J and Peng H (2004). Nonconcave penalized likelihood with a diverging number of parameters, *The Annals of Statistics*, **32**, 928–961.

Fu WJ (1998). Penalized regressions: the bridge versus the LASSO, *Journal of Computational and Graphical Statistics*, **7**, 397–416.

Green PJ (1990). On use of the EM for penalized likelihood estimation, *Journal of the Royal Statistical Society Series B (Methodological)*, **52**, 443–452.

Hausman RE (1982). Constrained multivariate analysis. In SH Zanckis and JS Rustagi (Eds), *Optimisation in Statistics: With a View Towards Applications in Management Science and Operations Research* (pp. 137–151), North-Holland, Amsterdam.

Jeffers JNR (1967). Two case studies in the application of principal component analysis, *Applied Statistics*, **16**, 225–236.

Jolliffe IT (1972). Discarding variables in a principal component analysis. I: artificial data, *Applied Statistics*, **21**, 160–173.

Jolliffe IT (1973). Discarding variables in a principal component analysis. II: real data, *Applied Statistics*, **22**, 21–31.

Jolliffe IT (1989). Rotation of ill-defined principal components, *Applied Statistics*, **38**, 139–147.

Jolliffe IT (1995). Rotation of principal components: choice of normalization constraints, *Journal of Applied Statistics*, **22**, 29–35.

Jolliffe IT (2002). *Principal Component Analysis*, Springer-Verlag, New York.

Jolliffe IT, Trendafilov NT, and Uddin M (2003). A modified principal component technique based on the LASSO, *Journal of Computational and Graphical Statistics*, **12**, 531–547.

Lawley DN (1953). A modified method of estimation in factor analysis and some large sample results. In *Uppsala Symposium on Psychological Factor Analysis, Number 3 in Nordisk Psykologi’s Monograph Series* (pp. 35–42), Almqvist and Wiksell, Uppsala.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B (Methodological)*, **58**, 267–288.

Tipping ME and Bishop CM (1999a). Mixtures of probabilistic principal component analyzers, *Neural computation*, **11**, 443–482.

Tipping ME and Bishop CM (1999b). Probabilistic principal component analysis, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **61**, 611–622.

Vines SK (2000). Simple principal components, *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **49**, 441–451.

Witten DM, Tibshirani R, and Hastie T (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics*, **10**, 515–534.

Xie B, Pan W, and Shen X (2010). Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data, *Bioinformatics*, **26**, 501–508.

Zou H, Hastie T, and Tibshirani R (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265–286.

Received November 21, 2016; Revised February 6, 2017; Accepted February 25, 2017