Journal of Statistical Planning and Inference 175 (2016) 87–99

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Sparse principal component analysis with measurement errors

Jianhong Shi [a], Weixing Song [b,*]

[a] School of Mathematics and Computer Science, Shanxi Normal University, Linfen, Shanxi, 041000, China

[b] Department of Statistics, Kansas State University, Manhattan, KS 66503, United States

ARTICLE INFO

Article history:
Received 4 July 2015
Received in revised form 11 March 2016
Accepted 12 March 2016
Available online 18 March 2016

Keywords:
Lasso
Elastic net
Sparse principal component analysis
Measurement error
Bias correction

ABSTRACT

Traditional principal component analysis often produces non-zero loadings, which makes it hard to interpret the principal components. This drawback can be overcome by the sparse principal component analysis procedures developed in the past decade. However, similar work has not been done when the random variables or vectors are contaminated with measurement errors. Simply applying the existing sparse principal component analysis procedure to the error-contaminated data might lead to biased loadings. This paper tries to modify an existing sparse principal component procedure to accommodate the measurement error setup. Similar to error-free cases, we show that the sparse principal component for the latent variables can be formulated as a bias-corrected lasso (elastic net) regression problem based on the observed surrogates, efficient algorithms are also developed to implement the procedure. Numerical simulation studies are conducted to illustrate the finite sample performance of the proposed method.

1. Introduction

As an efficient dimensional reduction technique, principal component analysis (PCA) provides a sequence of orthogonal linear combinations of the random variables, called principal components (PCs), from a multi-dimensional random vector, which can sequentially capture the biggest variability among the data collected from the random vector. Through these PCs, one can extract the commonality contained in the vector, and hopefully, an informative explanation might follow. Many interesting applications of PCA can be found in the areas of engineering, biology, education and other social science, for example, the handwritten zip code classification in Hastie et al. (2001), the human face recognition in Hancock et al. (1996), and the gene expression data analysis in Alter et al. (2000), just name a few. However, the entries in the loading vectors usually are nonzero which makes the interpretation of the PCs difficult. To overcome this drawback, Henry (1958) proposed the famous rotation technique; Gorsuch (1983) recommended rotating with varimax to produce orthogonal PCs or promax to produce oblique PCs. For more information on determining the proper rotations, see Tabachnick and Fidell (2007); By restricting the loading values to be 0, 1, and −1 or other values, Vines (2000) proposed a simple principal component analysis; By imposing a L_1 constraint on the loading vectors directly, Jolliffe and Uddin (2003) proposed the SCoTLASS method. As noted in Zou et al. (2006), the SCoTLASS technique suffers from the high computational cost, and insufficient sparsity of loadings when a high percentage of explained variance is required. Being aware of that the PCA can be reformulated as a ridge regression problem, Zou et al. (2006) proposed a modified Sparse PCA (SPCA) by integrating

* Corresponding author.
E-mail address: weixing@ksu.edu (W. Song).

http://dx.doi.org/10.1016/j.jspi.2016.03.001
0378-3758/

the elastic net approach with the lasso procedure, to produce sparse loadings. In addition to its attractive regression type optimization idea, the popularity of the SPCA is enhanced by its very efficient algorithm. Lasso and elastic net are very popular variable selection procedures in high dimensional modeling, we will not introduce them here for the sake of brevity. More details on these methodology can be found in Tibshirani (1996), Efron et al. (2004), Zou and Hastie (2005), Zou and Trevor (2005) and the references therein.

In the following discussion, we shall use the bold capital letter $\mathbf{X}$ to denote a $n \times p$ data matrix, its ith row is denoted by $\mathbf{x}_i$, $i = 1, 2, \ldots, n$, which will be viewed as a random sample from some population $\mathbf{x}$. The jth column of $\mathbf{X}$ will be denoted by X_j, $j = 1, 2, \ldots, p$. Often times, the quantity of interest $\mathbf{x}$ cannot be observed directly in practice, which is often called the latent variables or vector. Instead, a surrogate $\mathbf{z}$ can be observed, which is related to $\mathbf{x}$ in an additive way $\mathbf{z} = \mathbf{x} + \mathbf{u}$, where $\mathbf{u}$ is called the measurement error, $\mathbf{x}$ and $\mathbf{u}$ are independent. For some introduction on measurement error modeling, see Fuller (1987) and Carroll et al. (2006). Clearly, the PCA based on $\mathbf{x}$ is not feasible in this scenario. If we simply apply the existing SPCA procedure to the error-contaminated data, it might lead to biased loadings, just like the naive estimates in errors-in-variables regression models. Therefore, an interesting question is how to identify the loadings for $\mathbf{x}$ based on the sample from $\mathbf{z}$ and some additional information from $\mathbf{u}$. In the measurement error setup, the covariance matrix $\Sigma_{\mathbf{u}}$ of $\mathbf{u}$ is often assumed to be known. In the case of $\Sigma_{\mathbf{u}}$ being unknown, replicated observations on $\mathbf{x}$ are often used to obtain a consistent estimate of $\Sigma_{\mathbf{u}}$. More discussion on this case can be found in Section 3. In this paper, we will try to extend the SPCA of Zou et al. (2006) to the measurement error setup under the assumption of known $\Sigma_{\mathbf{u}}$. Throughout the paper, for any generic random vector $\mathbf{a}$, $\Sigma_{\mathbf{a}}$ denotes the population covariance of $\mathbf{a}$.

Note that the additive structure and the independence imply $\Sigma_{\mathbf{z}} = \Sigma_{\mathbf{x}} + \Sigma_{\mathbf{u}}$, so to find PCs for $\Sigma_{\mathbf{x}}$, one can directly work on $\Sigma_{\mathbf{z}} - \Sigma_{\mathbf{u}}$, provided the latter is known. However, this is rarely the case in practice. The PCA based on the sample covariance matrix is not as natural as the error-free situation simply because we do not know the sample covariance matrix of the latent vector $\mathbf{x}$. If we denote $S_{\mathbf{ab}} = n^{-1}\sum_{i=1}^{n}(\mathbf{a}_i - \bar{\mathbf{a}})^T(\mathbf{b}_i - \bar{\mathbf{b}})$, where $\bar{\mathbf{a}}$, $\bar{\mathbf{b}}$ are the mean vectors from the corresponding sequences, then simple algebra gives $S_{\mathbf{zz}} = S_{\mathbf{xx}} + S_{\mathbf{xu}} + S_{\mathbf{ux}} + S_{\mathbf{uu}}$. By subtracting $\Sigma_{\mathbf{u}}$ from both sides, we have $S_{\mathbf{zz}} - \Sigma_{\mathbf{u}} = S_{\mathbf{xx}} + S_{\mathbf{xu}} + S_{\mathbf{ux}} + S_{\mathbf{uu}} - \Sigma_{\mathbf{u}}$. Since $S_{\mathbf{xu}} + S_{\mathbf{ux}} + S_{\mathbf{uu}} - \Sigma_{\mathbf{u}}$ converges to 0 at the rate of $1/\sqrt{n}$ under quite general assumptions, so we can expect that the PCs based on $S_{\mathbf{xx}}$ could be well approximated by the PCs based on $S_{\mathbf{zz}} - \Sigma_{\mathbf{u}}$, but the effect of removing $S_{\mathbf{xu}} + S_{\mathbf{ux}} + S_{\mathbf{uu}} - \Sigma_{\mathbf{u}}$ from analysis on the resulting PCs should be investigated when the sample size is small. To adapt Zou et al. (2006)'s SPCA to our current setup, we have to find a way to transform the problem to a penalized linear regression problem.

The paper is organized as follows. Section 2 discusses the PCA based on population covariance matrices, a simple argument and some numerical examples are presented to show that PCA based on the covariance matrix of the surrogates often leads to biased PCs, but in a particular case, the PCs obtained from the matrices of the surrogates and the latent variables are the same! The direct bias-corrected SPCA approximation is introduced in Section 3, followed by the efficient algorithm developed for the proposed procedure, as well as some remarks on the adjusted total variances, the computational complexity of the algorithm and how to apply the proposed method when $\Sigma_{\mathbf{u}}$ is unknown but replicated observations are available. Numerical studies are conducted in Section 4, and all the theoretical derivations are postponed to Appendix.

2. PCA based on population covariance matrices

Before we work on the sample covariance matrices, it might be more illuminating to investigate the effect of measurement errors on the PCA based on the population covariance matrices. Note that $\Sigma_{\mathbf{z}} = \Sigma_{\mathbf{x}} + \Sigma_{\mathbf{u}}$, the PC directions of $\mathbf{x}$ might not be the same as those of $\mathbf{z}$ because of the perturbation of the measurement error. However, if $\Sigma_{\mathbf{u}}$ is diagonal and all the diagonal entries are equal, then the PC directions of $\mathbf{x}$ and $\mathbf{z}$ are indeed the same. To see this point, assume that $\Sigma_{\mathbf{u}} = \sigma^2 I$, and the spectral decomposition of $\Sigma_{\mathbf{x}}$ is $\mathbf{QMQ}^T$, where $\mathbf{M} = \mathrm{diag}(m_k^2)$. Then we must have

$$\mathbf{Q}^T\Sigma_{\mathbf{z}}\mathbf{Q} = \mathbf{Q}^T\Sigma_{\mathbf{x}}\mathbf{Q} + \sigma^2\mathbf{Q}^T\mathbf{Q} = \mathrm{diag}(m_j^2 + \sigma^2).$$

While maintaining the direction of principal components, the above result also implies the magnitudes along the principal components are inflated by an additive factor σ^2. For the general case, the eigenvalue–eigenvector relationship between $\Sigma_{\mathbf{z}}$ and $\Sigma_{\mathbf{x}}$ becomes complicated, however, $\Sigma_{\mathbf{z}} - \Sigma_{\mathbf{u}} = \Sigma_{\mathbf{x}}$ suggests us to study the PCA of $\Sigma_{\mathbf{x}}$, one can study the PCA of $\Sigma_{\mathbf{z}} - \Sigma_{\mathbf{u}}$. For illustration purpose, we choose

$$\Sigma_{\mathbf{x}} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \qquad \Sigma_{\mathbf{u}}^{(1)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \Sigma_{\mathbf{u}}^{(2)} = \begin{pmatrix} 1 & 0 \\ 0 & 0.5 \end{pmatrix}, \qquad \Sigma_{\mathbf{u}}^{(3)} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}.$$

Accordingly, let $\Sigma_{\mathbf{z}}$ be $\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{u}}^{(j)}$ with $j = 1, 2, 3$, that is, the latent vector $\mathbf{x}$ is contaminated with three different measurement errors. Fig. 1 shows the principal components of $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{z}}$ with $\Sigma_{\mathbf{u}}$ defined above. It is easy to see that the principal components of $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{z}}$ are the same for $\Sigma_{\mathbf{u}}^1$, and different for the latter two cases.

Knowing the covariance matrix $\Sigma_{\mathbf{z}}$ is an ideal case. More realistically, the observations for $\mathbf{z}$ are available, therefore the principal component analysis should be based on $S_{\mathbf{zz}} - \Sigma_{\mathbf{u}}$, where $S_{\mathbf{zz}}$ is the sample covariance matrix of $\mathbf{z}_i$, $i = 1, 2, \ldots, n$. Note that the bias-corrected statistic $S_{\mathbf{zz}} - \Sigma_{\mathbf{u}}$ is a consistent estimator of $\Sigma_{\mathbf{x}}$, so when the sample size is small, the performance of the principal component analysis based on $S_{\mathbf{zz}} - \Sigma_{\mathbf{u}}$ may not be very satisfying and the results should be cautiously interpreted. In particular, in the finite sample cases or if the dimension p is larger than the sample size n, $S_{\mathbf{zz}} - \Sigma_{\mathbf{u}}$