Statistica Sinica **35** (2025), 1-22
doi:https://doi.org/10.5705/ss.202022.0319

AUTOMATIC SPARSE PCA FOR HIGH-DIMENSIONAL DATA

Kazuyoshi Yata* and Makoto Aoshima

University of Tsukuba

Abstract: Sparse principal component analysis (SPCA) methods have proven to efficiently analyze high-dimensional data. Among them, threshold-based SPCA (TSPCA) is computationally more cost-effective than regularized SPCA, based on L1 penalties. We herein present an investigation of the efficacy of TSPCA for high-dimensional data settings and illustrate that, for a suitable threshold value, TSPCA achieves satisfactory performance for high-dimensional data. Thus, the performance of the TSPCA depends heavily on the selected threshold value. To this end, we propose a novel thresholding estimator to obtain the principal component (PC) directions using a customized noise-reduction methodology. The proposed technique is consistent under mild conditions, unaffected by threshold values, and therefore yields more accurate results quickly at a lower computational cost. Furthermore, we explore the shrinkage PC directions and their application in clustering high-dimensional data. Finally, we evaluate the performance of the estimated shrinkage PC directions in actual data analyses.

Key words and phrases: Clustering, large p small n, PCA consistency, shrinkage PC directions, thresholding.

1. Introduction

High-dimensional, low-sample-size (HDLSS) data scenarios exist in many areas of modern science including genomics, medical imaging, text recognition, and finance. In recent years, substantial work has been conducted on HDLSS asymptotic theory, wherein the sample size n is fixed or $n/d \to 0$ is used as the data dimension $d \to \infty$. For principal component analysis (PCA), Jung and Marron (2009) and Yata and Aoshima (2009) investigated inconsistency properties for both the eigenvalues and principal component (PC) directions in a sample covariance matrix. Yata and Aoshima (2012) developed a new PCA method called the *noise-reduction methodology* and reported consistent estimators for both eigenvalues and PC directions in addition with the PC scores using this method. Sparse PCA (SPCA) methods have been investigated in several studies. For example, Zou and Hastie (2006), Shen and Huang (2008), and Lee, Huang and Hu (2010) considered a regularized SPCA (RSPCA) based on L1 penalties under high-dimensional settings. Johnstone and Lu (2009) proposed a

*Corresponding author. E-mail: yata@math.tsukuba.ac.jp