On General Adaptive Sparse Principal Component Analysis

Chenlei Leng and Hansheng Wang

The method of *sparse principal component analysis* (S-PCA) proposed by Zou et al. (2006) is an attractive approach to obtain sparse loadings in principal component analysis (PCA). S-PCA was motivated by reformulating PCA as a least squares problem so that a lasso penalty on the loading coefficients can be applied. In this article, we propose new estimates to improve S-PCA on the following two aspects. Firstly, we propose a method of *simple adaptive sparse principal component analysis* (SAS-PCA), which uses the adaptive lasso penalty (Zou, 2006; Wang et al., 2007) instead of the lasso penalty in S-PCA. Secondly, we replace the least squares objective function in S-PCA by a general least squares objective function. This formulation allows us to study many related sparse PCA estimators under one unified theoretical framework and leads to the method of *general adaptive sparse principal component analysis* (GAS-PCA). Compared with SAS-PCA, GAS-PCA enjoys much improved finite sample performance. In addition, we show that when a BIC-type criterion is used for selecting the tuning parameters, the resulting estimates are consistent in variable selection. Numerical studies are conducted to compare the finite sample performance of various competing methods.

Key Words: Adaptive Lasso; BIC; GAS-PCA; LARS; Lasso; S-PCA; SAS-PCA

Supplementary Materials

1. Data Sets

`teaching.txt` This is the dataset describing teaching evaluation scores of 251 courses in Guanghua School of Management, Peking University during the

Chenlei Leng is Assistant Professor at Department of Statistics and Applied Probability, National University of Singapore, Singapore, (stalc@nus.edu.sg). Hansheng Wang is Associate Professor at Guanghua School of Management, Peking University, Beijing, P. R. China, 100871 (hansheng@gsm.pku.edu.cn)

This document describes supplementary material to an article published in the Journal of Computational and Graphical Statistics.

period from 2002 to 2004. Each row corresponds to one course and records the average scores. Each column corresponds to one of the following nine questions:

(Q1) I think this is a good course;

(Q2) The course improves my knowledge;

(Q3) The schedule is reasonable;

(Q4) The course is difficult;

(Q5) The course pace is too fast;

(Q6) The course load is very heavy;

(Q7) The text book is good;

(Q8) The reference book is helpful;

(Q9) Open this course is necessary.

2. Computer Code

`GAS.r` The code used in the paper.

`ex.r` A simple example on how to use GAS.r.