Sparse Principal Component Analysis

Hui ZOU, Trevor HASTIE, and Robert TIBSHIRANI

Principal component analysis (PCA) is widely used in data processing and dimensionality reduction. However, PCA suffers from the fact that each principal component is a linear combination of all the original variables, thus it is often difficult to interpret the results. We introduce a new method called sparse principal component analysis (SPCA) using the *lasso* (*elastic net*) to produce modified principal components with sparse loadings. We first show that PCA can be formulated as a regression-type optimization problem; sparse loadings are then obtained by imposing the lasso (elastic net) constraint on the regression coefficients. Efficient algorithms are proposed to fit our SPCA models for both regular multivariate data and gene expression arrays. We also give a new formula to compute the total variance of modified principal components. As illustrations, SPCA is applied to real and simulated data with encouraging results.

Key Words: Arrays; Gene expression; Lasso/elastic net; Multivariate analysis; Singular value decomposition; Thresholding.

1. INTRODUCTION

Principal component analysis (PCA) (Jolliffe 1986) is a popular data-processing and dimension-reduction technique, with numerous applications in engineering, biology, and social science. Some interesting examples include handwritten zip code classification (Hastie, Tibshirani, and Friedman 2001) and human face recognition (Hancock, Burton, and Bruce 1996). Recently PCA has been used in gene expression data analysis (Alter, Brown, and Botstein 2000). Hastie et al. (2000) proposed the so-called *gene shaving* techniques using PCA to cluster highly variable and coherent genes in microarray datasets.

PCA seeks the linear combinations of the original variables such that the derived variables capture maximal variance. PCA can be computed via the singular value decomposition (SVD) of the data matrix. In detail, let the data $\mathbf{X}$ be a $n \times p$ matrix, where n and p are the

Hui Zou is Assistant Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: *hzou@stat.umn.edu*). Trevor Hastie is Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: *hastie@stat.stanford.edu*). Robert Tibshirani is Professor, Department of Health Research Policy, Stanford University, Stanford, CA 94305 (E-mail: *tibs@stat.stanford.edu*).

Journal of Computational and Graphical Statistics, Volume 15, Number 2, Pages 265–286
DOI: 10.1198/106186006X113430