Integrative sparse principal component analysis

Kuangnan Fang, Xinyan Fan, Qingzhao Zhang, Shuangge Ma

PII: S0047-259X(17)30526-2
DOI: https://doi.org/10.1016/j.jmva.2018.02.002
Reference: YJMVA 4324

To appear in: *Journal of Multivariate Analysis*

Received date : 2 September 2017

Please cite this article as: K. Fang, X. Fan, Q. Zhang, S. Ma, Integrative sparse principal component analysis, *Journal of Multivariate Analysis* (2018), https://doi.org/10.1016/j.jmva.2018.02.002

Integrative sparse principal component analysis

Kuangnan Fang[a,b], Xinyan Fan[a], Qingzhao Zhang[a,b,c], Shuangge Ma[d,*]

[a] *Department of Statistics, School of Economics, Xiamen University*
[b] *Fujian Key Laboratory of Statistical Science, Xiamen University*
[c] *The Wang Yanan Institute for Studies in Economics, Xiamen University*
[d] *Department of Biostatistics, Yale School of Public Health*

Abstract

In the analysis of data with high-dimensional covariates and small sample sizes, dimension reduction techniques have been extensively employed. Principal component analysis (PCA) is perhaps the most popular dimension reduction technique. To remove noise effectively and generate more interpretable results, the sparse PCA (SPCA) technique has been developed. In high dimension, the analysis of a single dataset often generates unsatisfactory results. In a series of studies under the "regression analysis + variable selection" setting, it has been shown that integrative analysis provides an effective way of pooling information from multiple independent datasets and outperforms single-dataset analysis and many alternative multi-datasets analyses, especially including the classic meta-analysis. In this study, with multiple independent datasets, we propose conducting dimension reduction using a novel iSPCA (integrative SPCA) approach. Penalization is adopted for regularized estimation and selection of important loadings. Advancing from the existing integrative analysis studies, we further impose contrasted penalties, which may generate more accurate estimation/selection. Multiple settings on the similarity across datasets are comprehensively considered. Consistency properties of the proposed approach are established, and effective computational algorithms are developed. A wide spectrum of simulations demonstrate competitive performance of iSPCA over the alternatives. Two sets of data analysis further establish its practical applicability.

Keywords: Contrasted penalization, Integrative analysis, Sparse PCA

1. Introduction

Data with high-dimensional covariates and small-to-moderate sample sizes abound. Extensive methodological and theoretical studies have been conducted. In particular, dimension reduction techniques such as principal component analysis (PCA), partial least squares (PLS), independent component analysis (ICA), and others, have been proposed. PCA is arguably the most popular of all. It can assist in understanding underlying data structures, clustering analysis, regression analysis, and many other tasks. We refer to [7, 12, 15, 17] and other publications for methodological, theoretical, and numerical studies on PCA in high-dimensional settings. In many practical studies, it has been suggested that only a small subset of variables are relevant, while others are "noise". To identify relevant variables and generate more interpretable results, the sparse PCA (SPCA) technique has been developed, which applies regularized estimation to generate sparse loadings. In the literature, methodological studies on SPCA include [6, 14, 18, 24, 27], theoretical studies include [5, 16], and numerical studies include [2], among others.

Despite many promising successes, it is still often observed that results generated from analyzing a single dataset are unsatisfactory. This can be partly seen from our numerical study. Although there may be multiple contributing factors, the most important is perhaps the small sample size. For many scientific problems, there are multiple independent studies with comparable settings, e.g., with the same set of variables measured on subjects with similar characteristics [3]. In a series of studies under the "regression analysis + variable selection" settings [19, 26], it has

*Corresponding author
Email address: shuangge.ma@yale.edu (Shuangge Ma)