Journal of Multivariate Analysis 146 (2016) 7–17

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Supervised singular value decomposition and its asymptotic properties

Gen Li [a,*], Dan Yang [b], Andrew B. Nobel [a], Haipeng Shen [a,c]

[a] Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, United States
[b] Department of Statistics and Biostatistics, Rutgers, The State University of New Jersey, United States
[c] School of Business, University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 21 February 2014
Available online 9 March 2015

AMS 2000 subject classifications:
62H12

Keywords:
Low rank approximation
Principal component analysis
Reduced rank regression
Supervised dimension reduction
SupSVD

ABSTRACT

A supervised singular value decomposition (SupSVD) model has been developed for supervised dimension reduction where the low rank structure of the data of interest is potentially driven by additional variables measured on the same set of samples. The SupSVD model can make use of the information in the additional variables to accurately extract underlying structures that are more interpretable. The model is general and includes the principal component analysis model and the reduced rank regression model as two extreme cases. The model is formulated in a hierarchical fashion using latent variables, and a modified expectation–maximization algorithm for parameter estimation is developed, which is computationally efficient. The asymptotic properties for the estimated parameters are derived. We use comprehensive simulations and a real data example to illustrate the advantages of the SupSVD model.

1. Introduction

As high dimensional data become increasingly common, dimension reduction becomes more and more important, since it is easier to visualize and analyze a low dimensional structure in high dimensional data. The singular value decomposition (SVD) is a fundamental tool used in multivariate analysis to decompose a high-dimensional data matrix into a sum of unit-rank layers ordered by importance. The first few layers, which often capture the majority of the variation, act as a low rank approximation or dimension reduction of the original data.

However, one drawback of SVD is that it only makes use of a single data set, and by default the resulting dimension reduction cannot incorporate any additional information that may be relevant. When multiple related data sets are available on the same set of samples, sharing information across data sets may lead to recovery of a low rank structure that is more interpretable. Several approaches have been developed for analyzing multiple data sets. For example, [22] develops an integrative approach to study joint and individual variations simultaneously; [2] develops a supervised principal component regression method to select predictors and do prediction. In this paper, we propose a supervised SVD (SupSVD) model to achieve dimension reduction that incorporates auxiliary information. We assume that the auxiliary data set, which we refer to as the *supervision*, is a potential driving factor for the low rank structure of the *primary* data of interest.

The assumption is reasonable in many applications. For example, some genetic studies collect both gene expression and single-nucleotide polymorphism (SNP) data on the same group of subjects. One interesting topic is to investigate

* Corresponding author.
E-mail addresses: ligen@live.unc.edu (G. Li), dyang@stat.rutgers.edu (D. Yang), nobel@email.unc.edu (A.B. Nobel), haipeng@email.unc.edu (H. Shen).

http://dx.doi.org/10.1016/j.jmva.2015.02.016
0047-259X/

intrinsic patterns of the expression data. Biologically, expression of some genes is regulated by SNPs known as expression quantitative trait loci (eQTL). In other words, SNPs indeed drive underlying structure in the gene expression data which one can potentially get a better understanding of if we take advantage of the supervision (SNP) data.

We now introduce the SupSVD model using matrix notation. Let $\mathbf{X}$ denote the data matrix of primary interest which has n rows (or samples) and p columns (or variables). Let $\mathbf{Y}$ denote the supervision data matrix which has n rows (matched with $\mathbf{X}$) and q columns. We assume that the intrinsic information in $\mathbf{X}$ is low dimensional with rank $r(r \leq \min(n, p))$, and is possibly driven by $\mathbf{Y}$, in a linear fashion. In matrix form, the SupSVD model can be expressed as follows:

$$\begin{cases} \mathbf{X} = \mathbf{U}\mathbf{V}^T + \mathbf{E}, \\ \mathbf{U} = \mathbf{Y}\mathbf{B} + \mathbf{F}, \end{cases} \tag{1}$$

where $\mathbf{U}$ is an $n \times r$ latent score matrix, $\mathbf{V}$ is a $p \times r$ full-rank loading matrix, and $\mathbf{B}$ is a $q \times r$ coefficient matrix, with $\mathbf{F}$ and $\mathbf{E}$ being $n \times r$ and $n \times p$ error matrices, respectively.

Overall, the SupSVD model captures situations in which $\mathbf{X}$ has an intrinsic low rank structure and the structure is partially affected by $\mathbf{Y}$. The first equation in (1) is motivated by the additive-multiplicative low-rank approximation model for SVD, as in [12,27]. It indicates that the observed data matrix $\mathbf{X}$ consists of the low rank structure $\mathbf{U}\mathbf{V}^T$ plus measurement errors $\mathbf{E}$. We use a multivariate linear regression model to capture the potential supervising effect of $\mathbf{Y}$ on the score matrix $\mathbf{U}$. In particular, the matrix $\mathbf{F}$ captures information in $\mathbf{U}$ that cannot be explained by $\mathbf{Y}$. We note that very recently Fan et al. [13] proposed a projected principal component analysis (PCA) method that generalizes the second equation of (1) to a semi-parametric model.

Compared with the SVD, the SupSVD model incorporates the auxiliary information in $\mathbf{Y}$. The potential advantages of SupSVD over SVD are two-fold. First, using additional information may help reveal interesting patterns that might otherwise be undiscovered. Second, the low rank structure recovered by the SupSVD model might have superior interpretability. Evidence can be found in the simulated examples in the Supplement, Section Appendix F. Overall we find that SupSVD performs favorably when the supervision information is indeed a driving factor of low rank data. When auxiliary data are irrelevant, for example in Case 2 of Section 5.1.1, SupSVD automatically adapts to the situation and performs as well as SVD.

There is a rich literature on dimension reduction of a data matrix $\mathbf{X}$ in the presence of auxiliary information $\mathbf{Y}$, for example sufficient dimension reduction [9], supervised principal components [2], and principal fitted components [5,6]. Moreover, reduced rank regression (RRR) [19,25] can also be viewed as a dimension reduction approach for $\mathbf{X}$ if we regress $\mathbf{X}$ on $\mathbf{Y}$. The focus of most existing methods is to find a dimension reduced version of $\mathbf{X}$ that keeps all the information about $\mathbf{Y}$. This is different from the scope of the current paper. Here our primary goal is to identify low rank structure of $\mathbf{X}$, whether or not the structure is related to the auxiliary information $\mathbf{Y}$. The auxiliary information $\mathbf{Y}$ offers guidance for the dimension reduction of $\mathbf{X}$. To the best of our knowledge, our work is the first to address this topic.

The rest of the paper is organized as follows. In Section 2, we give more details of the SupSVD model, and explain its connections with existing models. In Section 3, we propose a modified version of the expectation–maximization (EM) algorithm for parameter estimation. The asymptotic properties of the estimates are discussed in Section 4. In Section 5, we compare different methods using extensive simulations and apply SupSVD to a real data example. We conclude in Section 6, with a brief discussion of potential extensions of our framework to functional data analysis. Proofs, technical details, and additional numerical examples can be found in supplemental materials.

2. The SupSVD model

In this section, we describe the SupSVD method in detail. Section 2.1 gives an equivalent formulation of the model, and discusses identifiability conditions. Section 2.2 establishes connections of the proposed model with some existing methods.

2.1. An equivalent form of the model

In Model (1), if we substitute the latent matrix $\mathbf{U}$ in the first equation with the second equation, we get an equivalent form for the SupSVD model as:

$$\mathbf{X} = \mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{F}\mathbf{V}^T + \mathbf{E}. \tag{2}$$

Without loss of generality, we assume that both $\mathbf{X}$ and $\mathbf{Y}$ are column-centered; hence, the model does not have intercepts. The random matrices $\mathbf{E}$ and $\mathbf{F}$ are assumed independent. Each entry of the error matrix $\mathbf{E}$ is independently identically distributed (i.i.d.) with mean zero and variance $\sigma_{\mathbf{e}}^2$. This follows the signal-plus-noise model for matrix reconstruction, cf. [27], as well as the r-component spiked covariance model for PCA, cf. [20,24]. Each row of $\mathbf{F}$ is i.i.d. with mean zero and covariance matrix $\mathbf{\Sigma_f}$, which is an unknown $r \times r$ positive definite matrix.

Furthermore, Model (2) can be viewed as a special setup of a multivariate linear regression model

$$\mathbf{X} = \mathbf{Y}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where the coefficient matrix $\boldsymbol{\beta}$ is $\mathbf{B}\mathbf{V}^T$ of rank $\min(r, q)$, and the random noise matrix $\boldsymbol{\varepsilon}$ is $\mathbf{F}\mathbf{V}^T + \mathbf{E}$. The rows of the noise matrix $\boldsymbol{\varepsilon}$ are i.i.d. with covariance $\mathbf{\Sigma}$ equal to $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p$ where $\mathbf{I}_p$ is the $p \times p$ identity matrix.