ARTICLE IN PRESS

Journal of Multivariate Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

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 xxxx

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 $\mathbf{\Sigma}_{\mathbf{f}}$, 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 $\mathbf{V}\mathbf{\Sigma}_{\mathbf{f}}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p$ where $\mathbf{I}_p$ is the $p \times p$ identity matrix.

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016

The primary goal of the SupSVD model is to identify low rank structure in the observed data $\mathbf{X}$ using $\mathbf{Y}$. Namely, we want to estimate $\mathbf{YBV}^T + \mathbf{FV}^T$, where $\mathbf{YBV}^T$ is the deterministic part and $\mathbf{FV}^T$ is the random part. The deterministic signal is driven by $\mathbf{Y}$ and the random signal captures important structures from unknown sources. The two parts are related through the common loading matrix $\mathbf{V}$, and together they form the underlying low rank representation for $\mathbf{X}$. In practice, we substitute all model parameters by estimates obtained from the observed data, and replace the random matrix $\mathbf{F}$ by its best unbiased prediction.

The SupSVD model (2) is identifiable in terms of the coefficient matrix $\boldsymbol{\beta} = \mathbf{BV}^T$ and the covariance matrix $\boldsymbol{\Sigma} = \mathbf{FV}^T + \mathbf{E}$, but unidentifiable in terms of the specific parameters $\mathbf{B}$, $\mathbf{V}$, $\boldsymbol{\Sigma}_\mathbf{f}$, and $\sigma_\mathbf{e}^2$. To see this, let $\mathbf{B}^\star = \mathbf{BQ}$, $\mathbf{V}^\star = \mathbf{VQ}$, and $\boldsymbol{\Sigma}_\mathbf{f}^\star = \mathbf{Q}^T\boldsymbol{\Sigma}_\mathbf{f}\mathbf{Q}$ for any $r \times r$ orthogonal matrix $\mathbf{Q}$. It is easily seen that $\mathbf{BV}^T = \mathbf{B}^\star\mathbf{V}^{\star T}$ and $\mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T = \mathbf{V}^\star\boldsymbol{\Sigma}_\mathbf{f}^\star\mathbf{V}^{\star T}$. Namely, the two sets of parameters lead to the same Model (2). In particular, we define two sets of parameters to be *equivalent* when they give identical likelihood functions (see (6)).

For regression purpose knowing $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ is enough, but for dimension reduction purpose we need to obtain all specific parameters since each parameter has an important interpretation. For example, the columns of $\mathbf{V}$ can be interpreted as projection directions; the matrix $\boldsymbol{\Sigma}_\mathbf{f}$ gives the covariance structure of latent scores; each column of $\mathbf{B}$ indicates how the supervision matrix $\mathbf{Y}$ is related with the corresponding score vector. Therefore we impose the following constraints to identify the model.

(1) The $p \times r$ matrix $\mathbf{V}$ has orthonormal columns, i.e., $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$;
(2) The $r \times r$ matrix $\boldsymbol{\Sigma}_\mathbf{f}$ is diagonal with r distinct positive eigenvalues;
(3) The columns of $\mathbf{V}$ are sorted in the descending order in terms of column norms of $\mathbf{XV}$, and the first entry of each column is positive.

The first condition is commonly used in SVD analysis. Each loading vector corresponds with a projection direction. The orthonormality of loading vectors naturally leads to an orthogonal basis with unit lengths. The second condition implies that the latent variables in $\mathbf{U}$ are uncorrelated. We assume all diagonal entries to be positive and distinct to avoid indeterminacy of the loading vectors. In practice, this condition generally holds. The third condition rules out column and sign switches. In addition, we also assume that the supervision data matrix $\mathbf{Y}$ has linearly independent columns; in practice, one can discard linearly dependent columns in $\mathbf{Y}$. Under these conditions, the SupSVD model is identifiable. Hereafter, without special notice, we assume that the model satisfies all the aforementioned identifiability conditions. We comment that the identifiability conditions help us identify the unique representative in an equivalence class.

Proposition 1. *In Model (2), for any parameter set* $(\mathbf{B}, \mathbf{V}, \boldsymbol{\Sigma}_\mathbf{f}, \sigma_\mathbf{e}^2)$ *such that the largest* r *eigenvalues of* $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T + \sigma_\mathbf{e}^2\mathbf{I}$ *are distinct and greater than the remaining eigenvalues, there exists an unique parameter set that is equivalent with* $(\mathbf{B}, \mathbf{V}, \boldsymbol{\Sigma}_\mathbf{f}, \sigma_\mathbf{e}^2)$ *and satisfies the identifiability conditions.*

For cases in which two or more of the first r eigenvalues of $\boldsymbol{\Sigma}$ are equal, the above conditions are not sufficient for identifiability, and one may have to impose constraints on $\mathbf{B}$ as well. However, in real data examples, equal-eigenvalue cases rarely occur. Therefore, we can reasonably restrict our scope to models that satisfy the identifiability conditions.

2.2. Connections with existing models

The SupSVD model (2) has close connections with several existing models. On the one hand, when $\mathbf{B} = \mathbf{0}$, i.e., when the score matrix $\mathbf{U}$ equals to the random matrix $\mathbf{F}$, Model (2) reduces to

$$\mathbf{X} = \mathbf{FV}^T + \mathbf{E}. \tag{3}$$

In Model (3), each row of $\mathbf{X}$ is i.i.d. with mean zero and covariance matrix $\mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T + \sigma_\mathbf{e}^2\mathbf{I}_p$, which is exactly the r-component spiked covariance model for PCA, cf. [20,24,30]. In the model, the r columns of $\mathbf{V}$ are the first r principal component (PC) loadings, and the columns of $\mathbf{XV}$ are the corresponding PCs. Note that the PCA model is *unsupervised*, as the matrix $\mathbf{Y}$ does not appear in the model.

On the other hand, when the latent score matrix $\mathbf{U}$ is fully driven by $\mathbf{Y}$, i.e., $\boldsymbol{\Sigma}_\mathbf{f} = \mathbf{0}$, the SupSVD model reduces to

$$\mathbf{X} = \mathbf{YBV}^T + \mathbf{E}, \tag{4}$$

where for identifiability purposes we let $\mathbf{B}$ have orthogonal columns. We note that Model (4) is the reduced rank regression (RRR) model [19,25] with isotropic covariance structure (we will refer to isotropic RRR as RRR). The matrix $\mathbf{C} = \mathbf{BV}^T$ is the rank r coefficient matrix whose least square estimator is explicitly given in [25]. In this case, the true underlying structure of $\mathbf{X}$ is $\mathbf{YBV}^T$, whose column space is a subspace of the column space of $\mathbf{Y}$. In other words, the underlying structure is fully driven by the supervision information. We therefore refer to the RRR model as *fully supervised*.

The SupSVD model (2) is also connected with the envelope model that was recently proposed by Cook et al. [8] and further developed in [32,7,11,10]. The envelope model is a parsimonious model for multivariate regression that is based on the assumption that variation in the response can be divided into two parts: a material part that is related to the predictor, and

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016

an immaterial part that is unrelated to the predictor. The envelope model achieves substantial efficiency gain in parameter estimation by focusing on the material part of the response. The coordinate version of the envelope model can be written as

$$\begin{aligned} \mathbf{y} &= \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{x} + \boldsymbol{\varepsilon} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T. \end{aligned} \tag{5}$$

Here $\mathbf{y}$ is a p-dimensional response, $\mathbf{x}$ is a q-dimensional predictor, $\boldsymbol{\Gamma}$ is $p \times r$ semi-orthogonal matrix and $\boldsymbol{\eta}$ is an $r \times q$ matrix. The product of $\boldsymbol{\Gamma}$ and $\boldsymbol{\eta}$ acts as a coefficient, while $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ are the intercept and the random error. The random error $\boldsymbol{\varepsilon}$ has covariance matrix $\boldsymbol{\Sigma}$ defined in the second equation, in which $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0)$ is orthogonal, and $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are positive definite.

If we regard the response as the primary data to be approximated, and the predictor as the supervision data, it can be shown that Model (5) coincides with Model (2). The covariance of (2) is slightly more specific than that of (5). However, we note that the two models arise in the analysis of different problems, and that they have different applications and interpretations. The SupSVD model attempts to extract a low rank representation of a primary data matrix, and is intended for dimension reduction problems in which auxiliary data is present. The goal of the envelope model is to reduce the variation of coefficient estimation in regression problems. Here we impose identifiability conditions on the model and estimate each parameter, as the parameters are directly interpretable in the context of dimension reduction. In [8] the authors focus on identifying estimable subspaces that are spanned by the parameters of their model; the parameters themselves are of less importance. In addition, fitting of the SupSVD and envelope models is carried out in fundamentally different ways. We describe a computationally efficient EM type algorithm to fit the model (1) for which the likelihood of the observed data usually converges to a local maximum after a few iterations. In order to fit the envelope model, the authors of [8] directly maximize the likelihood function, which involves optimization over a Grassmann manifold. We compared the computational speeds of both methods using various simulations, and in general the EM algorithm is faster.

SupSVD can be viewed as a general model for supervised dimension reduction. It encompasses unsupervised PCA and fully supervised RRR as two extremes. When the auxiliary information is irrelevant to low rank structure of the primary data, the SupSVD model reduces to the PCA model; when the underlying structure is totally driven by the auxiliary data, the SupSVD model reduces to the RRR model. It also connects with the envelope model from a multivariate regression point of view.

3. Model estimation

In this section, we describe the parameter estimation algorithm, incorporating the identifiability constraints discussed in Section 2.1. We assume multivariate normality for the random matrices $\mathbf{E}$ and $\mathbf{F}$ hereafter. To begin, we assume that the rank of the underlying structure of $\mathbf{X}$ is known to be r. Data-driven selection of the rank r is discussed at the end of this section.

Under the normality assumption for $\mathbf{E}$ and $\mathbf{F}$, we can obtain the distribution of the observed data $\mathbf{X}$ according to (2) as

$$\mathrm{vec}(\mathbf{X}^T) \sim \mathcal{N}_{np}\left(\mathrm{vec}(\mathbf{V}\mathbf{B}^T\mathbf{Y}^T),\ \mathbf{I}_n \otimes (\mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T + \sigma_\mathbf{e}^2\mathbf{I}_p)\right),$$

where $\mathrm{vec}(\cdot)$ is the column-stacking operator and $\otimes$ is the Kronecker product. Thus the log likelihood of $\mathbf{X}$ can be expressed explicitly as

$$\mathcal{L}(\mathbf{X}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T + \sigma_\mathbf{e}^2\mathbf{I}_p) - \frac{1}{2}\mathrm{tr}\left((\mathbf{X} - \mathbf{Y}\mathbf{B}\mathbf{V}^T)(\mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T + \sigma_\mathbf{e}^2\mathbf{I}_p)^{-1}(\mathbf{X} - \mathbf{Y}\mathbf{B}\mathbf{V}^T)^T\right), \tag{6}$$

where the parameters satisfy the identifiability conditions discussed above.

One way to estimate the parameters is to directly maximize the likelihood function (6) under the identifiability conditions. However, a direct constrained maximization is challenging for two reasons: (1) $\mathbf{V}$ appears in both the mean and the variance of the normal distribution; and (2) the constrained parameter space is not convex. As a remedy, we propose a modified expectation–maximization (EM) algorithm, namely an *expectation–maximization–standardization (EMS)* algorithm, to efficiently estimate the model parameters. The additional standardization step guarantees that the parameter estimates satisfy the identifiability conditions.

The latent matrix $\mathbf{U}$ in Model (1) naturally suggests the possibility of using the EM algorithm for parameter estimation. The joint log likelihood of $\mathbf{X}$ and $\mathbf{U}$, i.e., $\mathcal{L}(\mathbf{X}, \mathbf{U})$, can be separated into two parts: the conditional log likelihood of $\mathbf{X}$ given $\mathbf{U}$, and the marginal log likelihood of $\mathbf{U}$. In detail,

$$\mathcal{L}(\mathbf{X}, \mathbf{U}) = \mathcal{L}(\mathbf{X}|\mathbf{U}) + \mathcal{L}(\mathbf{U}), \tag{7}$$

where

$$\mathrm{vec}(\mathbf{X}^T) \mid \mathbf{U} \sim \mathcal{N}_{np}\left(\mathrm{vec}(\mathbf{V}\mathbf{U}^T), \sigma_\mathbf{e}^2\mathbf{I}_{np}\right), \quad \text{and} \tag{8}$$

$$\mathrm{vec}(\mathbf{U}^T) \sim \mathcal{N}_{nr}\left(\mathrm{vec}(\mathbf{B}^T\mathbf{Y}^T), \mathbf{I}_n \otimes \Sigma_\mathbf{f}\right). \tag{9}$$

The benefit of this separation is that the parameters $(\mathbf{B}, \boldsymbol{\Sigma}_\mathbf{f})$ are isolated from $(\mathbf{V}, \sigma_\mathbf{e}^2)$, and each parameter only contributes to one part of the likelihood. Using (7) the joint log likelihood has the following form:

$$\mathcal{L}(\mathbf{X}, \mathbf{U}) \propto -np\log\sigma_\mathbf{e}^2 - \sigma_\mathbf{e}^{-2}\mathrm{tr}\left((\mathbf{X} - \mathbf{U}\mathbf{V}^T)(\mathbf{X} - \mathbf{U}\mathbf{V}^T)^T\right) - n\log\det\boldsymbol{\Sigma}_\mathbf{f} - \mathrm{tr}\left((\mathbf{U} - \mathbf{Y}\mathbf{B})\boldsymbol{\Sigma}_\mathbf{f}^{-1}(\mathbf{U} - \mathbf{Y}\mathbf{B})^T\right).$$

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016

Below we describe the steps of the EMS algorithm, which is presented as Algorithm 1 at the end of this section. We use $\theta^{(i)} = (\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i)}, {\sigma_{\mathbf{e}}^2}^{(i)})$ to denote the parameter estimates obtained in the ith iteration, which satisfy the identifiability conditions.

E Step: We calculate the conditional expectation of $\mathcal{L}(\mathbf{X}, \mathbf{U})$ with respect to $\mathbf{U}$ given $\mathbf{X}$ and $\theta^{(i)}$, i.e., $\mathbb{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)})$. The conditional distribution of $\mathbf{U}$ given $\mathbf{X}$ and the previous parameter estimation $\theta^{(i)}$ is

$$\mathrm{vec}(\mathbf{U}^T) \mid \mathbf{X} \sim \mathcal{N}\left(\mathrm{vec}\left({\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}}^T\right), \mathbf{I}_n \otimes \Omega_{\mathbf{U}|\mathbf{X}}^{(i)}\right), \tag{10}$$

where

$$\Theta_{\mathbf{U}|\mathbf{X}}^{(i)} = \mathbb{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X}) = \left(\mathbf{Y}\mathbf{B}^{(i)}\left({\sigma_{\mathbf{e}}^2}^{(i)}{\mathbf{\Sigma}_{\mathbf{f}}^{(i)}}^{-1}\right) + \mathbf{X}\mathbf{V}^{(i)}\right)\left(\mathbf{I}_r + {\sigma_{\mathbf{e}}^2}^{(i)}{\mathbf{\Sigma}_{\mathbf{f}}^{(i)}}^{-1}\right)^{-1},$$

$$\Omega_{\mathbf{U}|\mathbf{X}}^{(i)} = \left({\mathbf{\Sigma}_{\mathbf{f}}^{(i)}}^{-1} + {\sigma_{\mathbf{e}}^{-2}}^{(i)}\mathbf{I}_r\right)^{-1}.$$

Note that the conditional expectation of $\mathbf{U}$ given $\mathbf{X}$ is a weighted average of $\mathbf{Y}\mathbf{B}^{(i)}$ and $\mathbf{X}\mathbf{V}^{(i)}$, where the weights are determined by ${\sigma_{\mathbf{e}}^2}^{(i)}$ and $\mathbf{\Sigma}_{\mathbf{f}}^{(i)}$.

M Step: We maximize $\mathbb{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)})$ with respect to all the parameters under the identifiability constraints in Section 2.1. The optimization is challenging since the constraint is not convex. As the joint distribution of $\mathbf{X}$ and $\mathbf{U}$ is identifiable even without the side conditions, we propose a modified EM algorithm that bypasses the constrained optimization problem. More specifically, we first obtain the unconstrained optimizers of $\mathbb{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)})$, and then find the unique set of parameters that is equivalent to the optimizers in terms of the SupSVD model, and that satisfies the identifiability conditions.

The unconstrained optimization problem can be solved analytically. Setting partial derivatives of $\mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)})$ with respective to each parameter to zero, we obtain

$$\widehat{\mathbf{B}} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbb{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X}, \theta^{(i)}), \tag{11}$$

$$\widehat{\mathbf{V}} = \mathbf{X}^T\mathbb{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X}, \theta^{(i)})\left[\mathbb{E}_{\mathbf{U}}(\mathbf{U}^T\mathbf{U}\,|\mathbf{X}, \theta^{(i)})\right]^{-1}, \tag{12}$$

$$\widehat{\mathbf{\Sigma}_{\mathbf{f}}} = \frac{1}{n}\mathbb{E}_{\mathbf{U}}\left[(\mathbf{U} - \mathbf{Y}\widehat{\mathbf{B}})^T(\mathbf{U} - \mathbf{Y}\widehat{\mathbf{B}})\,|\mathbf{X}, \theta^{(i)}\right], \tag{13}$$

$$\widehat{\sigma_{\mathbf{e}}^2} = \frac{1}{np}\mathbb{E}_{\mathbf{U}}\left[\mathrm{tr}((\mathbf{X} - \mathbf{U}\widehat{\mathbf{V}}^T)(\mathbf{X} - \mathbf{U}\widehat{\mathbf{V}}^T)^T)\,|\mathbf{X}, \theta^{(i)}\right], \tag{14}$$

where the corresponding conditional expectations can be obtained from (10). Details can be found in the Supplement, Section Appendix C.

S Step: The unconstrained optimizers $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_{\mathbf{f}}}, \widehat{\sigma_{\mathbf{e}}^2})$ in (11)–(14) typically satisfy the condition of Proposition 1. In this case, we can obtain the unique equivalent set of parameters that satisfy the identifiability conditions. In particular, we perform SVD on $\widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}_{\mathbf{f}}}\widehat{\mathbf{V}}^T$ to obtain the following eigen-decomposition:

$$\mathbf{V}^{(i+1)}\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}{\mathbf{V}^{(i+1)}}^T = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma}_{\mathbf{f}}}\widehat{\mathbf{V}}^T,$$

where the columns of $\mathbf{V}^{(i+1)}$ are the orthonormal eigenvectors and the diagonal entries of the diagonal matrix $\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}$ are the eigenvalues. In practice, the eigenvalues are almost always positive and distinct, so that the matrices $\mathbf{V}^{(i+1)}$ and $\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}$ satisfy the identifiability conditions and are unique up to a column reordering. Then, we set $\mathbf{B}^{(i+1)} = \widehat{\mathbf{B}}\widehat{\mathbf{V}}^T\mathbf{V}^{(i+1)}$ and ${\sigma_{\mathbf{e}}^2}^{(i+1)} = \widehat{\sigma_{\mathbf{e}}^2}$. It is easy to see that

$$\mathbf{B}^{(i+1)}{\mathbf{V}^{(i+1)}}^T = \widehat{\mathbf{B}}\widehat{\mathbf{V}}^T.$$

Lastly, we reorder the columns of $\mathbf{V}^{(i+1)}$, and accordingly the columns of $\mathbf{B}^{(i+1)}$ and the rows/columns of $\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}$, in order to ensure that the column norms of $\mathbf{X}\mathbf{V}^{(i+1)}$ are decreasing. As a result, we get parameter estimates $\theta^{(i+1)} = (\mathbf{B}^{(i+1)}, \mathbf{V}^{(i+1)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}, {\sigma_{\mathbf{e}}^2}^{(i+1)})$ for the $(i+1)$th iteration.

Each step of the EMS algorithm has an analytical expression and can be computed efficiently. Our numerical studies indicate that the algorithm is insensitive to initial values. In practice, we use the naive estimates from SVD as the initial values. The following proposition guarantees convergence of the EMS algorithm to a local optimum.

Proposition 2. *In each iteration of the EMS algorithm, the log likelihood of the observed data $\mathcal{L}(\mathbf{X})$ is monotonically nondecreasing. Therefore, the EMS algorithm always converges to some stationary point (maybe local maximum).*

The presentation in this section assumes that the rank r is known. In practice, the rank has to be determined from the data. In the numerical studies of Section 5.1, we adopt a popular practice within the PCA literature: using the scree plot of a primary data matrix to determine a proper rank. The rationale is that we assume the rank of the underlying signal of

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016

Algorithm 1 The EMS Algorithm for Parameter Estimation under the SupSVD Model

1: Set initial values for the parameters $(\mathbf{B}^{(0)}, \mathbf{V}^{(0)}, \Sigma_{\mathbf{f}}^{(0)}, \sigma_{\mathbf{e}}^{2(0)})$;
2: **while** $\mathcal{L}(\mathbf{X}|\theta^{(i+1)}) - \mathcal{L}(\mathbf{X}|\theta^{(i)}) >$ threshold **do**
3: **E Step:** Derive the conditional distribution (10) given $\theta^{(i)} = (\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \Sigma_{\mathbf{f}}^{(i)}, \sigma_{\mathbf{e}}^{2(i)})$;
4: **M Step:** Obtain the unconstrained optimizer $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\Sigma}_{\mathbf{f}}, \widehat{\sigma_{\mathbf{e}}^2})$ from (11)–(14);
5: **S Step:** Standardize $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\Sigma}_{\mathbf{f}}, \widehat{\sigma_{\mathbf{e}}^2})$ to get $\theta^{(i+1)} = (\mathbf{B}^{(i+1)}, \mathbf{V}^{(i+1)}, \Sigma_{\mathbf{f}}^{(i+1)}, \sigma_{\mathbf{e}}^{2(i+1)})$ that satisfy the identifiability conditions;
6: Set $i \leftarrow i + 1$.
7: **end while**

a primary data matrix is inherent. Auxiliary information is used to help recover the underlying low-rank structure more accurately, without altering the rank. Other rank selection methods that have been studied in the PCA literature, e.g., the permutation assessment method in [4] and the bi-cross-validation method in [23], are also appropriate in our framework. The likelihood ratio test approach of [8] could be used to select r in the SupSVD model as well.

4. Asymptotic analysis

In this section, we state the consistency and asymptotic normality of the SupSVD parameter estimates. Since the SupSVD model is overparameterized, i.e., unidentifiable without side conditions, standard asymptotics from the maximum likelihood framework do not apply directly. Instead, we refer to the asymptotic results in [28] for overparameterized structural models. A similar treatment can be found in [8].

Specifically, we first focus on the estimable functions $\boldsymbol{\beta} = \mathbf{B}\mathbf{V}^T$ and $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Sigma}_{\mathbf{f}}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}$, which uniquely define the likelihood function. In order to fit our analysis into the framework of [28], we rewrite the parameters as

$$\boldsymbol{\phi} = \begin{pmatrix} \mathrm{vec}(\mathbf{B}) \\ \mathrm{vec}(\mathbf{V}) \\ \mathrm{vech}(\boldsymbol{\Sigma}_{\mathbf{f}}) \\ \sigma_{\mathbf{e}}^2 \end{pmatrix} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix},$$

where the operator $\mathrm{vech}(\cdot)$ stacks the lower triangular part of a symmetric matrix into a vector. The estimable functions can then be expressed as

$$\mathbf{h}(\boldsymbol{\phi}) = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\beta}) \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \mathrm{vec}(\mathbf{B}\mathbf{V}^T) \\ \mathrm{vech}(\mathbf{V}\boldsymbol{\Sigma}_{\mathbf{f}}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}) \end{pmatrix} = \begin{pmatrix} h_1(\boldsymbol{\phi}) \\ h_2(\boldsymbol{\phi}) \end{pmatrix}. \tag{15}$$

For any $d \times d$ symmetric matrix $\boldsymbol{\Omega}$, we denote the $d(d+1)/2 \times d^2$ constant contraction matrix as $\mathbf{C}_d$, and the $d^2 \times d(d+1)/2$ constant expansion matrix as $\mathbf{E}_d$ to relate the operator $\mathrm{vech}(\cdot)$ and $\mathrm{vec}(\cdot)$, i.e., $\mathrm{vech}(\boldsymbol{\Omega}) = \mathbf{C}_d\mathrm{vec}(\boldsymbol{\Omega})$ and $\mathrm{vec}(\boldsymbol{\Omega}) = \mathbf{E}_d\mathrm{vech}(\boldsymbol{\Omega})$. Moreover, for any $l \times m$ matrix $\boldsymbol{\Gamma}$, we denote the $lm \times lm$ constant commutation matrix as $\mathbf{K}_{lm}$, i.e., $\mathrm{vec}(\boldsymbol{\Gamma}^T) = \mathbf{K}_{lm}\mathrm{vec}(\boldsymbol{\Gamma})$. We can obtain the following theorem, whose proof can be found in the Supplement, Section Appendix D.

Theorem 1. *Assume Model (2) and let* $\mathbf{h}(\cdot)$ *be as in (15). Denote* $\mathbf{H} = \partial\mathbf{h}(\boldsymbol{\phi})/\partial\boldsymbol{\phi}$, *and let* $\mathbf{J}$ *be the Fisher information of* $\mathbf{h}(\boldsymbol{\phi})$. *Let* $\widehat{\mathbf{h}}$ *be the maximum likelihood estimator of* $\mathbf{h}$. *Then,*

$$\sqrt{n}(\widehat{\mathbf{h}} - \mathbf{h}) \to_d \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{h}}), \tag{16}$$

where $\boldsymbol{\Sigma}_{\mathbf{h}} = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^{\dagger}\mathbf{H}^T$, *where* $\dagger$ *indicates the Moore–Penrose inverse. Specifically,*

$$\mathbf{H} = \begin{pmatrix} \mathbf{V} \otimes \mathbf{I}_q & (\mathbf{I}_p \otimes \mathbf{B})\mathbf{K}_{pr} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_p(\mathbf{V}\boldsymbol{\Sigma}_{\mathbf{f}} \otimes \mathbf{I}_p) & \mathbf{C}_p(\mathbf{V} \otimes \mathbf{V})\mathbf{E}_r & vech(\mathbf{I}_p) \end{pmatrix}$$

and

$$\mathbf{J} = \begin{pmatrix} \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{Y}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{E}_p^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_p \end{pmatrix}$$

where $\boldsymbol{\Sigma}_{\mathbf{Y}} = \lim_{n\to\infty} \mathbf{Y}\mathbf{Y}^T/n$.

As a result, we know that $\sqrt{n}\,\mathrm{vec}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{n}\,\mathrm{vech}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})$ are jointly asymptotically normally distributed with mean zero. Moreover, under the identifiability conditions, we obtain the following asymptotic property for each parameter in $\widehat{\boldsymbol{\phi}}$.

Corollary 1. *Given (16), under the identifiability conditions,* $\sqrt{n}\,vec(\widehat{\mathbf{B}} - \mathbf{B})$, $\sqrt{n}\,vec(\widehat{\mathbf{V}} - \mathbf{V})$, $\sqrt{n}\,diag(\widehat{\boldsymbol{\Sigma}}_{\mathbf{f}} - \boldsymbol{\Sigma}_{\mathbf{f}})$, *and* $\sqrt{n}\,(\widehat{\sigma_{\mathbf{e}}^2} - \sigma_{\mathbf{e}}^2)$ *are asymptotically jointly normal with mean zero. The asymptotic covariance matrix of* $\sqrt{n}\,(\widehat{\mathbf{v}}_i - \mathbf{v}_i)$, *where* $\widehat{\mathbf{v}}_i$ *and* $\mathbf{v}_i$ *are the ith columns of* $\widehat{\mathbf{V}}$ *and* $\mathbf{V}$ *respectively, is given in the Supplement, Section Appendix E.*

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016

5. Numerical examples

We compare SupSVD with SVD and RRR using extensive simulations (Section 5.1) and a real data example (Section 5.2). Section 5.1.1 compares the three methods with data simulated from each of the models respectively to show the adaptivity of SupSVD. Section 5.1.2 illustrates the performances of the methods under a spectrum of settings ranging from PCA to RRR. In Section 5.2, we illustrate SupSVD using the breast cancer data from [33]. Additional simulation and real data examples can be found in the Supplement, Section Appendix F, Appendix G and Appendix H..

5.1. Simulation studies

5.1.1. Adaptivity of SupSVD

We consider three simulation examples where the data are generated from each one of the three models (SupSVD, PCA, RRR) respectively. In particular, the PCA example illustrates a situation where the "supervision" data are actually not related to the primary data; the RRR example illustrates a situation where the underlying structure of primary data is fully driven by supervision. For each simulated example, we apply all three methods to analyze the simulated data, and demonstrate the adaptivity of SupSVD under different settings. We have tried a range of parameter settings in each case and the results are concordant across settings. Below we choose to only present representative results in each example.

In all three examples, we set the sample size $n = 100$, the dimension of $\mathbf{X}$ as $p = 68$, and the dimension of the supervision data $\mathbf{Y}$ as $q = 4$. The rank of the underlying structure is set to be $r = 2$. We fill in the supervision data matrix $\mathbf{Y}$ with numbers generated from a standard normal distribution. The loading vectors in $\mathbf{V}$ are set to be the first two orthogonal loadings with unit norms estimated from the call center data in the Supplement, Section Appendix H. The intention is to make the simulation setting as realistic as possible. In particular, the primary data matrix $\mathbf{X}$ is generated in the following ways for different examples.

(1) **Case 1 (SupSVD): X** is generated from the SupSVD model $\mathbf{X} = \mathbf{YBV}^T + \mathbf{FV}^T + \mathbf{E}$. The 4×2 fixed coefficient matrix $\mathbf{B}$ is standardized to have orthogonal columns with norm 3. The matrix $\mathbf{F}$ has i.i.d. rows from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma_f} = \mathrm{diag}(9, 4)$. The matrix $\mathbf{E}$ has i.i.d. entries from $\mathcal{N}(0, 3)$.
(2) **Case 2 (PCA): X** is generated from the PCA model $\mathbf{X} = \mathbf{FV}^T + \mathbf{E}$, where $\mathbf{F}$ is generated in the same way as in Case 1, and $\mathbf{E}$ has i.i.d. entries from a standard normal distribution.
(3) **Case 3 (RRR): X** is generated from the RRR model $\mathbf{X} = \mathbf{YBV}^T + \mathbf{E}$, where the 4×2 fixed coefficient matrix $\mathbf{B}$ is standardized to have orthogonal columns with norm 6 and 3 respectively. The error matrix $\mathbf{E}$ has i.i.d. entries from $\mathcal{N}(0, 3)$.

Performance measures. The three methods are compared in two aspects, *low rank structure recovery* and *parameter estimation*. The low rank recovery accuracy is measured by the mean square error (MSE) defined as

$$MSE_{\mathbf{UV}^T} = \frac{1}{np}\|\mathbf{UV}^T - \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T\|_{\mathbb{F}}^2,$$

where $\|\cdot\|_{\mathbb{F}}$ denotes the Frobenius norm, and $\mathbf{UV}^T$ and $\widehat{\mathbf{U}}\widehat{\mathbf{V}}^T$ are the true and estimated low rank structures respectively. For SVD, $\widehat{\mathbf{U}} = \mathbf{X}\widehat{\mathbf{V}}_{SVD}$; for RRR, $\widehat{\mathbf{U}} = \mathbf{Y}\widehat{\mathbf{B}}_{RRR}$; for SupSVD, $\widehat{\mathbf{U}} = \left(\mathbf{Y}\widehat{\mathbf{B}}(\widehat{\sigma_{\mathbf{e}}^2}\widehat{\mathbf{\Sigma}_{\mathbf{f}}}^{-1}) + \mathbf{X}\widehat{\mathbf{V}}\right)\left(\mathbf{I}_r + \widehat{\sigma_{\mathbf{e}}^2}\widehat{\mathbf{\Sigma}_{\mathbf{f}}}^{-1}\right)^{-1}$, where $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_{\mathbf{f}}}, \widehat{\sigma_{\mathbf{e}}^2})$ is the parameter set estimated from the SupSVD approach. We also considered other matrix norms such as 1-norm and 2-norm [15], and obtained similar results.

For parameter estimation, only the loading matrix $\mathbf{V}$ and the noise variance $\sigma_{\mathbf{e}}^2$ are common across the three methods. We use the following performance measures:

$$MSE_{\mathbf{V}} = \frac{1}{pr}\|\mathbf{V} - \widehat{\mathbf{V}}\|_{\mathbb{F}}^2, \qquad MSE_{\sigma_{\mathbf{e}}^2} = (\sigma_{\mathbf{e}}^2 - \widehat{\sigma_{\mathbf{e}}^2})^2.$$

Moreover, since the columns of a loading matrix form a basis for a projection subspace, we also measure the largest principal angle [15] between the true subspace and the estimated subspace which is defined as

$$Angle_{\mathbf{V}} = \frac{180}{\pi}\arccos(\min \mathrm{eig}(\mathbf{V}^T\widehat{\mathbf{V}})),$$

where $\min \mathrm{eig}(\cdot)$ denotes the minimal eigenvalue.

Results. For each case, we repeat the simulation 100 times and present in Table 1 the median and the median absolute deviations (MAD) of each performance measurement for the three methods. The results clearly show that SupSVD performs favorably no matter which true model the data are generated from, while SVD and RRR only work well in their respective settings. This demonstrates that SupSVD, covering SVD and RRR as special cases, adapts to a wide range of practical situations. In practice, whenever additional information is available (whether it is truly supervision or not), SupSVD is always a good choice for dimension reduction. In these simulations, SupSVD always provides the best results, equivalent to (or better than) the method corresponding to the true data generative model.

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016

Table 1
Section 5.1.1—median(MAD) for low rank structure recovery accuracy and parameter estimation accuracy.

		SupSVD	SVD	RRR
Case 1 (SupSVD)	$MSE_{\mathbf{UV}^T}$	**0.1289** (0.0082)	0.1830 (0.0128)	0.2487 (0.0154)
	$MSE_{\mathbf{V}}$	**0.0025** (0.0005)	0.0036 (0.0013)	0.0080 (0.0048)
	$MSE_{\sigma_{\mathbf{e}}^2}$	**0.0104** (0.0066)	0.0357 (0.0127)	0.0075 (0.0060)
	$Angle_{\mathbf{V}}$	**23.1605** (1.3312)	**23.5571** (1.4463)	27.0765 (2.0600)
Case 2 (PCA)	$MSE_{\mathbf{UV}^T}$	**0.0497** (0.0035)	0.0606 (0.0036)	0.2066 (0.0138)
	$MSE_{\mathbf{V}}$	**0.0022** (0.0003)	**0.0022** (0.0003)	0.0199 (0.0027)
	$MSE_{\sigma_{\mathbf{e}}^2}$	**0.0009** (0.0007)	0.0035 (0.0014)	0.0231 (0.0056)
	$Angle_{\mathbf{V}}$	**25.0287** (2.3239)	**24.9046** (2.1729)	77.1232 (7.0102)
Case 3 (RRR)	$MSE_{\mathbf{UV}^T}$	**0.0659** (0.0051)	0.1845 (0.0097)	**0.0635** (0.0055)
	$MSE_{\mathbf{V}}$	0.0032 (0.0014)	0.0024 (0.0003)	**0.0018** (0.0002)
	$MSE_{\sigma_{\mathbf{e}}^2}$	**0.0082** (0.0064)	0.0329 (0.0130)	**0.0063** (0.0052)
	$Angle_{\mathbf{V}}$	**25.4285** (1.6554)	29.4099 (2.0742)	**25.2282** (1.5882)

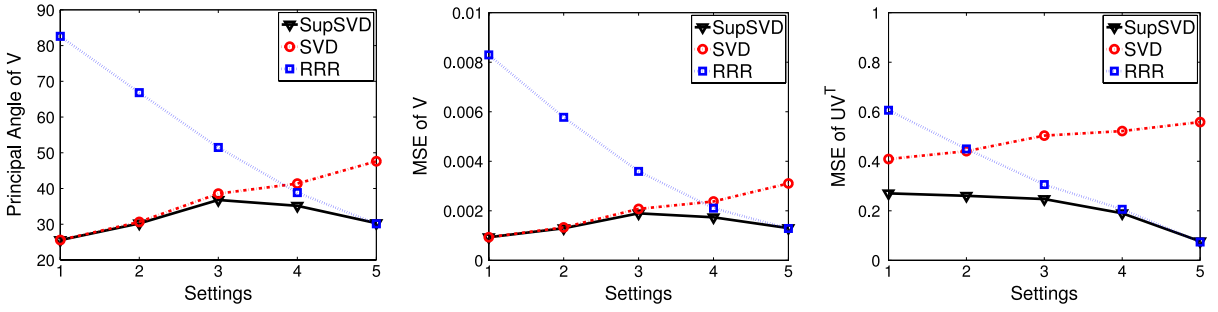

Fig. 1. Section 5.1.2—median curves for $Angle_{\mathbf{V}}$, $MSE_{\mathbf{V}}$, and $MSE_{\mathbf{UV}^T}$ based on 100 simulation runs.

Note that Table 1 also shows that the $MSE_{\mathbf{V}}$ of SupSVD is larger than the other two methods when the data are generated from the RRR model, i.e. in Case 3. We remark that this is due to the low identifiability of the SupSVD model when the true $\mathbf{\Sigma_f}$ is exactly zero. Numerically, SupSVD is still applicable but the estimated loading vectors are subject to an unstable orthogonal rotation. However, we comment that the estimated projection subspace of $\mathbf{V}$ (i.e., $Angle_{\mathbf{V}}$) and the low-rank recovery accuracy (i.e., $MSE_{\mathbf{UV}^T}$) are unaffected. Generally, it is very unlikely that the underlying structure of a primary data matrix is fully driven by supervision without any variations in practice. Therefore, we do not view this as a major drawback of SupSVD for practical use.

5.1.2. Comparison across a spectrum

We now compare SupSVD, SVD and RRR across a spectrum of simulation settings ranging from the PCA model to the RRR model. For easy presentation, we set $n = 210$, $p = 68$, $q = 1$, and $r = 1$. Fill the 210×1 vector $\mathbf{Y}$ with standard normal random numbers. We simulate $\mathbf{X}$ from the SupSVD model, with the loading vector being the first column of $\mathbf{V}$ in Case 1 above, $\sigma_{\mathbf{e}}^2 = 16$, and $(\mathbf{B}, \mathbf{\Sigma_f}) \in \{(0, 36), (1, 25), (2, 16), (3, 9), (4, 0)\}$, corresponding to Setting 1 to 5, respectively. Therefore, the SupSVD model ranges from the PCA model $\mathbf{X} = 6\mathbf{Z}\mathbf{V}^T + \mathbf{E}$ (Setting 1; $\mathbf{Z}$ is a random vector with i.i.d. entries from standard normal distribution) to the RRR model $\mathbf{X} = 4\mathbf{Y}\mathbf{V}^T + \mathbf{E}$ (Setting 5). Again, under each setting, we run 100 simulations and summarize the results.

To avoid redundancy, we only show the median curves of $MSE_{\mathbf{UV}^T}$, $MSE_{\mathbf{V}}$, and $Angle_{\mathbf{V}}$ for the methods in Fig. 1. We observe that SupSVD is uniformly the best over the spectrum of settings, with similar performance with SVD when the true underlying model is PCA, and similar performance with RRR when the true underlying model is RRR. Again, the results illustrate that SupSVD is a robust method that adapts well over a wide range of data-generating models.

5.2. Breast cancer data

We consider a real data set containing gene expression measurements from breast tumors, obtained from The Cancer Genome Atlas (TCGA) project [33]. A pointer to the publicly available data is at https://tcga-data.nci.nih.gov/docs/publications/brca_2012/. A primary goal is to understand underlying patterns of genetic variation among tumors. In this case, we have additional information of disease subtype for each tumor. We may regard cancer subtypes as a partial driver of the underlying structure of the gene expression data [26]. Samples from the same subtype will share common genetic variations. We use the subtype information as our supervision data and apply the SupSVD method.

The raw data set contains 17 814 genes and 348 samples. Out of the 348 samples, there are 5 subtypes of breast cancer with different number of samples in each subtype: Basal (66), Her2 (42), LumA (154), LumB (81), and Normal (5). We

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016

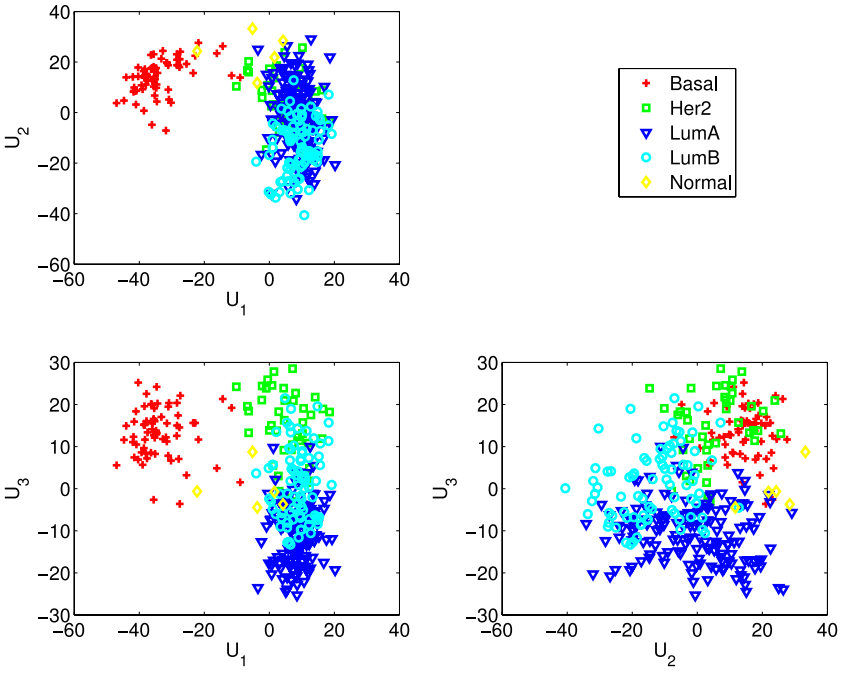

Fig. 2. Breast cancer data—scatter plots of SupSVD score vectors. The 5 different subtypes are well separated by the first three score vectors.

preprocessed the data in the same way as in [21]. We first imputed missing values with the k-nearest neighbors algorithm ($k = 10$), then removed genes with low variations across samples (standard deviation smaller than 1.5), and finally mean centered each gene. The result is a column-centered data matrix $\mathbf{X}$ with 348 samples and 645 genes. Based on the scree plot of the singular values of $\mathbf{X}$, we select the rank of the underlying structure to be 3.

Fig. 2 shows the scatter plots of the estimated SupSVD scores. The first score vector clearly separates the Basal subgroup from the rest. The second score vector captures variations within each subtype. The third score vector roughly separates the Her2, LumA, and LumB subgroups.

Fig. 3 presents the heat maps of the unit-rank structures from SupSVD. There are clear patterns driven by subtypes. For example, the first layer is dominated by the unique pattern in the Basal subgroup. The third layer shows patterns similar between Basal and Her2, but different among Her2, LumA and LumB. There are also within-group variations that are not driven by subtypes. For example, the LumA samples in the second layer clearly exhibit several different patterns. The SVD and RRR results are given in the Supplement, Section Appendix G. In comparison, SupSVD effectively captures important underlying patterns consisting of both between-group variations driven by the subtype information and within-group variations from unknown sources.

6. Discussion

In this paper, we propose a supervised dimension reduction model, SupSVD that takes advantage of auxiliary information to better recover the underlying low-rank structure in the primary data of interest. We focused on recovering comprehensive low-rank structures from the data with the potential guidance of the supervision information. The SupSVD model contains the PCA model and the RRR model as two extreme cases: when the supervision information is unrelated to the data of interest, SupSVD reduces to PCA; when the underlying structure is fully driven by the supervision information, SupSVD reduces to RRR. SupSVD automatically adjusts the amount of supervision used for dimension reduction without the use of tuning parameters. The proposed EMS algorithm for parameter estimation in SupSVD is computationally efficient. Asymptotic properties of SupSVD are derived for the resulting estimates. Simulation studies and real data applications clearly demonstrate the advantages and flexibility of the SupSVD method.

Dimension reduction is also useful in functional data analysis (FDA) to facilitate various subsequent analyses. For an overview of the FDA literature including recent advances, see [31,14,16,3]. We remark that our SupSVD method can be directly adapted to FDA through a basis approach. In particular, one can decompose discretized observations of functional data onto proper basis functions, obtain a coefficient matrix, and then apply SupSVD to the coefficient matrix. The low rank approximation obtained from SupSVD can then be converted back to the original functional space through the basis functions. Another approach is to first select important variables in discretized values of the function [1], and then apply SupSVD to the dimension-reduced vectors. Alternatively, our ongoing work attempts to extend the recent regularization formulation of functional principal component analysis [18,17] to incorporate supervision for FDA. We intend to impose

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016

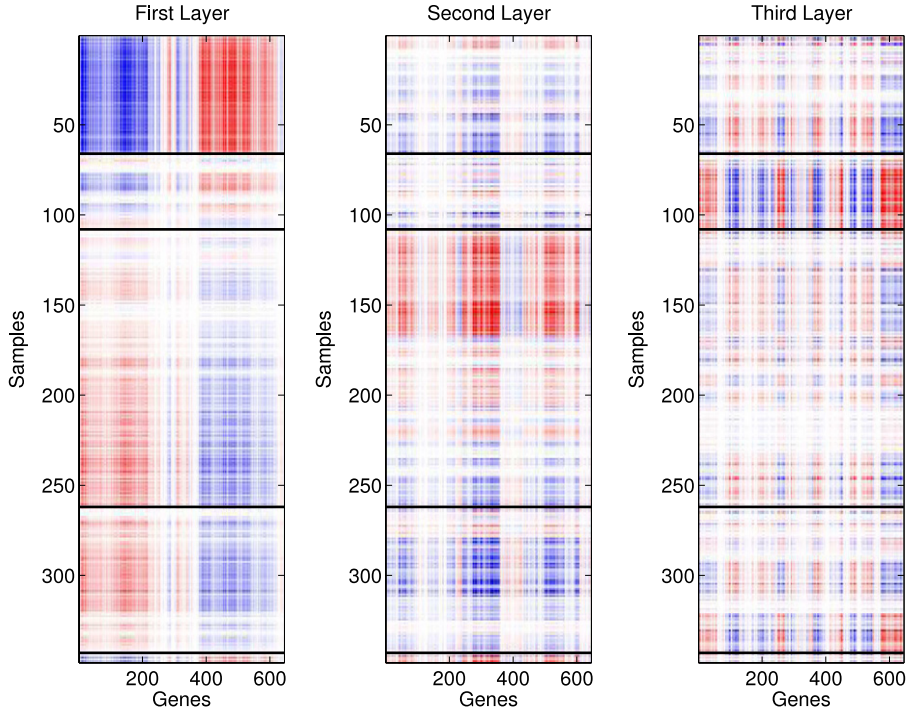

Fig. 3. Breast cancer data—heat map of first three unit-rank SupSVD structures of the gene expression data. Blue is negative and red is positive. The samples are grouped in the order of Basal, Her2, LumA, LumB, Normal. The genes are reordered for better visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

both sparsity [29] and roughness regularization to incorporate both high-dimensional multivariate data as well as infinite-dimensional functional data.

Acknowledgments

We would like to thank the editor, and two referees for their thorough and insightful reviews of the manuscript, which led to significant improvements in both content and presentation. This research was partially supported by National Science Foundation Grants DMS-0907177 (ABN, GL), DMS-1106912 (HS, GL), DMS-1127914 (DY), DMS-1310002 (ABN, GL), and DMS-1407655 (HS, GL).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2015.02.016.

References

[1] G. Aneiros, P. Vieu, Variable selection in infinite-dimensional problems, Statist. Probab. Lett. 94 (0) (2014) 12–20.
[2] E. Bair, T. Hastie, D. Paul, R. Tibshirani, Prediction by supervised principal components, J. Amer. Statist. Assoc. 101 (473) (2006) 119–137.
[3] E.G. Bongiorno, E. Salinelli, A. Goia, P. Vieu, Contributions in Infinite-dimensional Statistics and Related Topics, Società Editrice Esculapio, 2014.
[4] A. Buja, N. Eyuboglu, Remarks on parallel analysis, Multivariate Behav. Res. 27 (4) (1992) 509–540.
[5] R.D. Cook, Fisher Lecture: dimension reduction in regression, Statist. Sci. 22 (1) (2007) 1–26.
[6] R.D. Cook, L. Forzani, Principal fitted components for dimension reduction in regression, Statist. Sci. 23 (4) (2008) 485–501.
[7] R. Cook, I. Helland, Z. Su, Envelopes and partial least squares regression, J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (5) (2013) 851–877.
[8] R.D. Cook, B. Li, F. Chiaromonte, Envelope models for parsimonious and efficient multivariate linear regression, Statist. Sinica 20 (2010) 927–1010.
[9] R.D. Cook, L. Ni, Sufficient dimension reduction via inverse regression, J. Amer. Statist. Assoc. 100 (470) (2005) 410–428.
[10] R.D. Cook, Z. Su, Scaled envelopes: scale-invariant and efficient estimation in multivariate linear regression, Biometrika 100 (4) (2013) 939–954.
[11] R.D. Cook, X. Zhang, Simultaneous envelopes for multivariate linear regression, Technometrics 57 (1) (2015) 11–25.
[12] B.P. Dozier, J.W. Silverstein, On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices, J. Multivariate Anal. 98 (4) (2007) 678–694.
[13] J. Fan, Y. Liao, W. Wang, 2014. Projected principal component analysis in factor models. arXiv:1406.3836.
[14] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer, 2006.
[15] G.H. Golub, C.F. Van Loan, Matrix Computations, Vol. 3, JHU Press, 2012.
[16] L. Horváth, P. Kokoszka, Inference for Functional Data with Applications, Vol. 200, Springer, 2012.
[17] J.Z. Huang, H. Shen, A. Buja, The analysis of two-way functional data using two-way regularized singular value decompositions, J. Amer. Statist. Assoc. 104 (488) (2009).
[18] J.Z. Huang, H. Shen, A. Buja, et al., Functional principal components analysis via penalized rank one approximation, Electron. J. Stat. 2 (2008) 678–695.

[19] A.J. Izenman, Reduced-rank regression for the multivariate linear model, J. Multivariate Anal. 5 (1975) 248–264.
[20] I.M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, Ann. Statist. 29 (2) (2001) 295–327.
[21] E.F. Lock, D.B. Dunson, Bayesian consensus clustering, Bioinformatics (2013) btt425.
[22] E.F. Lock, K.A. Hoadley, J.S. Marron, A.B. Nobel, Joint and individual variation explained (JIVE) for integrated analysis of multiple data types, Ann. Appl. Stat. 7 (1) (2013) 523–542.
[23] A.B. Owen, P.O. Perry, Bi-cross-validation of the svd and the nonnegative matrix factorization, Ann. Appl. Stat. (2009) 564–594.
[24] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statist. Sinica 17 (2007) 1617–1642.
[25] G.C. Reinsel, R.P. Velu, Multivariate Reduced-Rank Regression: Theory and Applications, Springer, New York, 1998.
[26] E.E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S.K. Sieberts, S. Monks, M. Reitman, C. Zhang, et al., An integrative genomics approach to infer causal associations between gene expression and disease, Nature Genet. 37 (7) (2005) 710–717.
[27] A. Shabalin, A. Nobel, Reconstruction of a low-rank matrix in the presence of Gaussian noise, J. Multivariate Anal. 118 (2013) 67–76.
[28] A. Shapiro, Asymptotic theory of overparameterized structural models, J. Amer. Statist. Assoc. 81 (393) (1986) 142.
[29] H. Shen, J.Z. Huang, Sparse principal component analysis via regularized low rank matrix approximation, J. Multivariate Anal. 99 (6) (2008) 1015–1034.
[30] D. Shen, H. Shen, J. Marron, Consistency of sparse PCA in high dimension, low sample size contexts, J. Multivariate Anal. 115 (2013) 317–333.
[31] B. Silverman, J. Ramsay, Functional Data Analysis, Springer, 2005.
[32] Z. Su, R.D. Cook, Partial envelopes for efficient estimation in multivariate linear regression, Biometrika 98 (1) (2011) 133–146.
[33] The Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours, Nature 490 (7418) (2012) 61–70.

Please cite this article in press as: G. Li, et al., Supervised singular value decomposition and its asymptotic properties, Journal of Multivariate Analysis (2015), http://dx.doi.org/10.1016/j.jmva.2015.02.016