# Supervised Singular Value Decomposition and Its Asymptotic Properties

Gen Li
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Dan Yang
Department of Statistics and Biostatistics
Rutgers, The State University of New Jersey
Haipeng Shen
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Andrew B. Nobel
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

January 25, 2014

**Abstract**

We develop a supervised singular value decomposition (SupSVD) model for supervised dimension reduction. The research is motivated by applications 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 very general and includes the principal component analysis model and the reduced rank regression model as two extreme cases. We formulate the model in a hierarchical fashion using latent variables, and develop a modified expectation-maximization algorithm for parameter estimation, which is computationally efficient. The asymptotic properties for the estimated parameters are derived. We use comprehensive simulations and two real data examples to illustrate the advantages of the SupSVD model.

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

# 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. 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 capture the majority of the variation, can 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 is relevant. When multiple related data sets are available on the same set of samples, sharing information across the data sets may lead to recovery of a low rank structure that is more interpretable. Several approaches have been developed for analyzing such multiple data sets. See, for example, Bair et al. (2006) and Lock et al. (2013). In this paper, we propose a supervised SVD (SupSVD) model from a new perspective. We assume that the additional data set, which we refer to as the *supervision information*, is a potential driving factor for the low rank structure of the *primary* data of interest.

The assumption that the supervision information partially drives the low rank structure in the primary data is reasonable in many applications. For example, some genetics studies collect both gene expression and single-nucleotide polymorphism (SNP) data on the same group of subjects. One is interested in understanding intrinsic patterns of the gene expression data. Biologically, the expression levels of some genes are regulated by certain SNPs, which are known as the expression quantitative trait loci (eQTL). In other words, the SNP data indeed drive some underlying structure in the gene expression data, which can be better understood if we take advantage of the supervision (SNP) data. In this paper, Section 5.2.1 describes some gene expression data about 348 samples along with their subtype of breast cancer. One interesting question is to identify low-dimensional structures in the gene expression that are driven by the cancer subtypes.

We now introduce our 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 rows matched with $\mathbf{X}$ and $q$ columns. We

assume the intrinsic information in $\mathbf{X}$ is low dimensional with rank $r$ ($r \leq \min(n, p)$), and possibly driven by $\mathbf{Y}$. Our SupSVD model in matrix form 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 the $n \times r$ latent score matrix, $\mathbf{V}$ is the $p \times r$ full-rank loading matrix, and $\mathbf{B}$ is the $q \times r$ coefficient matrix, with $\mathbf{F}$ and $\mathbf{E}$ being $n \times r$ and $n \times p$ random error matrices respectively.

Overall, the SupSVD model reflects the fact that $\mathbf{X}$ has an intrinsic low rank structure and the structure is partially affected by $\mathbf{Y}$. In our earlier examples, $\mathbf{X}$ is the gene expression matrix, while $\mathbf{Y}$ contains the SNP information or the cancer subtype. The first equation in (1) is motivated by the additive-multiplicative low-rank approximation model for SVD, as in Dozier and Silverstein (2007) and Shabalin and Nobel (2013). It indicates that the observed data matrix $\mathbf{X}$ consists of the low rank structure $\mathbf{U}\mathbf{V}^T$ and the measurement errors $\mathbf{E}$. The second equation of (1) is a multivariate linear regression model for the latent score matrix $\mathbf{U}$. This regression setup implies that $\mathbf{U}$ is potentially driven by the supervision information in $\mathbf{Y}$. The matrix $\mathbf{F}$ captures information in $\mathbf{U}$ that cannot be explained by $\mathbf{Y}$.

Compared with SVD, the SupSVD method makes use of the additional supervision data set $\mathbf{Y}$. The potential advantages of SupSVD over SVD are two-fold. From an exploratory point of view, using additional information may help reveal interesting patterns which are otherwise undiscovered. In our first motivating example of Section 5.1.1, SupSVD successfully recovers the true subgroup structure in the data while SVD fails. From an interpretation point of view, the supervision information may enhance the interpretability of the extracted low rank structure. In the second example of Section 5.1.1, SupSVD loadings have clear meanings while SVD loadings do not. In summary, SupSVD outperforms SVD when the supervision information is indeed a driving factor of the low rank structure of the data of interest. When there is no need for supervision, for example in Case 2 of Section 5.1.2, SupSVD automatically adapts to the underlying model and performs as good as SVD.

There is a rich literature on dimension reduction of $\mathbf{X}$ in the presence of $\mathbf{Y}$, for example

sufficient dimension reduction by Cook and Ni (2005), supervised principal component by Bair et al. (2006), principal fitted component by Cook (2007), Cook and Forzani (2008). Moreover, the reduced rank regression (RRR) by Izenman (1975), Reinsel and Velu (1998) can also be viewed as a dimension reduction approach for $\mathbf{X}$ if we regress $\mathbf{X}$ on $\mathbf{Y}$. However, the focus of most methods is to find a dimension reduced version of $\mathbf{X}$ such that it keeps all information about $\mathbf{Y}$ in $\mathbf{X}$, which is very different from the scope of our current paper. Our primary goal is to obtain the comprehensive low rank structure of $\mathbf{X}$, whether or not related with $\mathbf{Y}$. The additional information in $\mathbf{Y}$ only offers potential guidance for dimension reduction of $\mathbf{X}$. In particular, the SVD method and the RRR method are two extreme cases for our method. As far as we know, little previous work directly studies 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 two real data examples. We conclude in Section 6. Proofs and technical details 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 the 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. We also assume that the random error matrices $\mathbf{E}$ and $\mathbf{F}$ are independent, and the rows of each matrix are independently identically distributed (i.i.d.) from multivariate normal $\mathcal{N}_p(\mathbf{0}, \sigma_{\mathbf{e}}^2\mathbf{I}_p)$ and $\mathcal{N}_r(\mathbf{0}, \mathbf{\Sigma_f})$ respectively, where $\mathbf{I}_p$ is the $p \times p$ identity matrix and $\mathbf{\Sigma_f}$ is a $r \times r$ positive definite matrix. The normality assumption allows us to derive maximum likelihood estimations in Section 3. The isotropic covariance assumption for $\mathbf{E}$ follows from the $r$-component spiked covariance model for principal component analysis (PCA), cf. Johnstone (2001) and Paul (2007), as well as the signal-plus-noise model for matrix reconstruction, cf. Shabalin and Nobel (2013).

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 are i.i.d. with covariance structure $\mathbf{\Sigma}$ equal to $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}$.

The primary goal of the SupSVD model is to extract the low rank structure from the observed data. Namely, we want to estimate $\mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{F}\mathbf{V}^T$, where $\mathbf{Y}\mathbf{B}\mathbf{V}^T$ is the deterministic part and $\mathbf{F}\mathbf{V}^T$ 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 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.

By default, the SupSVD model (2) is identifiable in terms of the coefficient matrix $\boldsymbol{\beta}$ and the covariance matrix $\mathbf{\Sigma}$, but unidentifiable in terms of the specific parameters $\mathbf{B}$, $\mathbf{V}$, $\mathbf{\Sigma_f}$, and $\sigma_{\mathbf{e}}^2$. To see this, let $\mathbf{B}^\star = \mathbf{B}\mathbf{Q}$, $\mathbf{V}^\star = \mathbf{V}\mathbf{Q}$, and $\mathbf{\Sigma}_{\mathbf{f}}^\star = \mathbf{Q}^T\mathbf{\Sigma_f}\mathbf{Q}$ for any $r \times r$ orthogonal matrix $\mathbf{Q}$. It is easily seen that $\mathbf{B}\mathbf{V}^T = \mathbf{B}^\star\mathbf{V}^{\star T}$ and $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T = \mathbf{V}^\star\mathbf{\Sigma}_{\mathbf{f}}^\star\mathbf{V}^{\star T}$. 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 (5).

For regression purpose knowing $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ is enough, but for dimension reduction purpose we need to obtain all specific parameters since each parameter carries important interpre-

tation. For example, the columns of $\mathbf{V}$ can be interpreted as projection directions; the matrix $\mathbf{\Sigma_f}$ 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 $\mathbf{\Sigma_f}$ 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. Besides, for identifiability we also assume that the supervision data matrix $\mathbf{Y}$ has linearly independent columns. One can always discard linearly dependent columns in $\mathbf{Y}$ to avoid redundant supervision information. As a result, the SupSVD model is rigorously identified. Hereafter, without special notice, we always 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. In particular, we have the following proposition.

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

For cases where some of the first $r$ eigenvalues of $\mathbf{\Sigma}$ are equal, the above conditions are not sufficient for identifiability. One may have to impose constraints on $\mathbf{B}$ as well. However, in real data examples, the equal-eigenvalue cases rarely happen. 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, Model (2) reduces to the following probabilistic model when $\mathbf{B} = \mathbf{0}$, i.e., when the score matrix $\mathbf{U}$ equals to the random matrix $\mathbf{F}$,

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

In Model (3), each row of $\mathbf{X}$ is i.i.d. from $\mathcal{N}_p(\mathbf{0}, \mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p)$, which is exactly the $r$-component spiked covariance model for principal component analysis (PCA), studied for example by Johnstone (2001), Paul (2007), and Shen et al. (2013). In the model, the columns of $\mathbf{V}$ are the principal component (PC) loadings, the columns of $\mathbf{XV}$ are the PCs, and the covariance matrix of the PCs is the diagonal matrix $\mathbf{\Sigma_f} + \sigma_{\mathbf{e}}^2\mathbf{I}_p$. 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., $\mathbf{\Sigma_f} = \mathbf{0}$, the SupSVD model reduces to the following model:

$$\mathbf{X} = \mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{E}, \tag{4}$$

where for identifiability purposes we let $\mathbf{B}$ have orthogonal columns . We note that Model (4) is the reduced rank regression (RRR) model with the isotropic covariance structure (in the context, we refer to the isotropic RRR as RRR). The matrix $\mathbf{C} = \mathbf{B}\mathbf{V}^T$ is the rank $r$ coefficient matrix whose least square estimator is explicitly given in Reinsel and Velu (1998). In this case, the true underlying structure of $\mathbf{X}$ is $\mathbf{Y}\mathbf{B}\mathbf{V}^T$, 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*.

We note that our SupSVD model (2) has some overlap with the coordinate representation of the envelope model recently proposed by Cook et al. (2010). If we treat $\mathbf{Y}$ as an observed predictor, $\mathbf{X}$ as a multivariate response, $\mathbf{B}\mathbf{V}^T$ as a regression coefficient matrix, and $\mathbf{F}\mathbf{V}^T + \mathbf{E}$ as a structural error matrix, Model (2) happens to be equivalent with Model (3.2) in Cook et al. (2010). However, the two models arise in the analysis of very different problems, and have very different applications and interpretations. The goal of SupSVD

is to extract a low rank representation of the observed data matrix $\mathbf{X}$, while the envelope model aims at reducing the variation of coefficient estimation in regression. Due to the different aims, we impose identifiability conditions in our model, and estimate every single parameter, all of which possess particular interpretations in dimension reduction; Cook et al. (2010) focused on estimable subspaces spanned by parameters instead. The estimation procedures are also very different. Referring to the equivalent hierarchical model (1), we propose a modified EM algorithm that is computationally very efficient. The likelihood of the observed data usually converges to the maximum after a few iterations. For the envelope model, the authors directly maximize the likelihood function which involves the computationally costly optimization over a Grassmann manifold.

In summary, SupSVD can be viewed as a general model for supervised dimension reduction. It covers the unsupervised PCA and the fully supervised RRR as two extremes. It also connects with the envelope model if we view from a multivariate regression perspective.

# 3 Model Estimation

In this section, we describe the parameter estimation algorithm. Throughout the paper, we assume the rank of the underlying structure of $\mathbf{X}$ is known to be $r$. In practice, we can obtain rank estimation through the scree plot based on the estimated singular values.

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

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

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

$$\begin{aligned}\mathcal{L}(\mathbf{X}) = &-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det\left(\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p\right) \\ &-\frac{1}{2}\text{tr}\left((\mathbf{X}-\mathbf{Y}\mathbf{B}\mathbf{V}^T)(\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_{\mathbf{e}}^2\mathbf{I}_p)^{-1}(\mathbf{X}-\mathbf{Y}\mathbf{B}\mathbf{V}^T)^T\right)\end{aligned} \tag{5}$$

where the parameters satisfy the aforementioned identifiability conditions.

One way to estimate the parameters is to directly maximize the above likelihood function under the identifiability conditions. However, the direct constrained maximization is

challenging for two reasons: 1) $\mathbf{V}$ appears in both the mean and the variance of the normal distribution; 2) the constrained parameter space is not convex. As a remedy, below we propose a computationally efficient expectation-maximization-standardization (EMS) algorithm for parameter estimation.

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 as in (6): the conditional log likelihood of $\mathbf{X}$ given $\mathbf{U}$ (7), and the marginal log likelihood of $\mathbf{U}$ (8), as shown below,

$$\mathcal{L}(\mathbf{X}, \mathbf{U}) \;=\; \mathcal{L}(\mathbf{X}|\mathbf{U}) + \mathcal{L}(\mathbf{U}), \tag{6}$$

where

$$\mathrm{vec}(\mathbf{X}^T) \mid \mathbf{U} \;\sim\; \mathcal{N}_{np}\left(\mathrm{vec}(\mathbf{V}\mathbf{U}^T), \sigma_{\mathbf{e}}^2 \mathbf{I}_{np}\right), \tag{7}$$

$$\mathrm{vec}(\mathbf{U}^T) \;\sim\; \mathcal{N}_{nr}\left(\mathrm{vec}(\mathbf{B}^T\mathbf{Y}^T), \mathbf{I}_n \otimes \Sigma_{\mathbf{f}}\right). \tag{8}$$

The benefits of such separation are that the parameters $(\mathbf{B}, \mathbf{\Sigma_f})$ are isolated from $(\mathbf{V}, \sigma_{\mathbf{e}}^2)$, and each parameter only contributes to one part of the likelihood. More specifically, the joint log likelihood has the following explicit expression:

$$\begin{aligned}\mathcal{L}(\mathbf{X}, \mathbf{U}) \;\propto\; & -np\log\sigma_{\mathbf{e}}^2 - \sigma_{\mathbf{e}}^{-2}\mathrm{tr}\left((\mathbf{X}-\mathbf{U}\mathbf{V}^T)(\mathbf{X}-\mathbf{U}\mathbf{V}^T)^T\right) \\ & -n\log\det\mathbf{\Sigma_f} - \mathrm{tr}\left((\mathbf{U}-\mathbf{Y}\mathbf{B})\mathbf{\Sigma_f}^{-1}(\mathbf{U}-\mathbf{Y}\mathbf{B})^T\right).\end{aligned}$$

Below we describe the separate 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 $i$th iteration that 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., $\mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)})$. The conditional distribution of $\mathbf{U}$ given $\mathbf{X}$ and the previous parameter estimation $\theta^{(i)}$ is

$$\mathrm{vec}(\mathbf{U}^T) \mid \mathbf{X} \;\sim\; \mathcal{N}\left(\;\mathrm{vec}\left({\Theta_{\mathbf{U}|\mathbf{X}}^{(i)}}^T\right),\;\; \mathbf{I}_n \otimes \Omega_{\mathbf{U}|\mathbf{X}}^{(i)}\;\right), \tag{9}$$

where

$$\begin{aligned}\Theta_{\mathbf{U}|\mathbf{X}}^{(i)} \;&=\; \mathrm{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X}) = \left(\mathbf{Y}\mathbf{B}^{(i)}\left(\sigma_{\mathbf{e}}^{2(i)}\mathbf{\Sigma}_{\mathbf{f}}^{(i)^{-1}}\right) + \mathbf{X}\mathbf{V}^{(i)}\right)\left(\mathbf{I}_r + \sigma_{\mathbf{e}}^{2(i)}\mathbf{\Sigma}_{\mathbf{f}}^{(i)^{-1}}\right)^{-1}, \\ \Omega_{\mathbf{U}|\mathbf{X}}^{(i)} \;&=\; \left(\mathbf{\Sigma}_{\mathbf{f}}^{(i)^{-1}} + \sigma_{\mathbf{e}}^{-2(i)}\mathbf{I}_r\right)^{-1}.\end{aligned}$$

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 $\mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X},\mathbf{U})|\mathbf{X},\theta^{(i)})$ with respect to all the parameters under the identifiability constraints discussed in Section 2.1. The constrained optimization is challenging since it's not convex. Noticing that the joint distribution of $\mathbf{X}$ and $\mathbf{U}$ is identifiable even without the side conditions, we propose a modified EM algorithm to bypass the constrained optimization problem. More specifically, we first obtain the unconstrained optimizers of $\mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{X},\mathbf{U})|\mathbf{X},\theta^{(i)})$, and then find the unique set of parameters that is equivalent with the optimizers in terms of the SupSVD model, while satisfying the identifiability conditions.

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

$$\widehat{\mathbf{B}} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathrm{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X},\theta^{(i)}), \tag{10}$$

$$\widehat{\mathbf{V}} = \mathbf{X}^T\mathrm{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X},\theta^{(i)})\left[\mathrm{E}_{\mathbf{U}}(\mathbf{U}^T\mathbf{U}\ |\mathbf{X},\theta^{(i)})\right]^{-1}, \tag{11}$$

$$\widehat{\mathbf{\Sigma}_{\mathbf{f}}} = \frac{1}{n}\mathrm{E}_{\mathbf{U}}\left[(\mathbf{U}-\mathbf{Y}\widehat{\mathbf{B}})^T(\mathbf{U}-\mathbf{Y}\widehat{\mathbf{B}})\ |\mathbf{X},\theta^{(i)}\right], \tag{12}$$

$$\widehat{\sigma_{\mathbf{e}}^2} = \frac{1}{np}\mathrm{E}_{\mathbf{U}}\left[\mathrm{tr}((\mathbf{X}-\mathbf{U}\widehat{\mathbf{V}}^T)(\mathbf{X}-\mathbf{U}\widehat{\mathbf{V}}^T)^T)\ |\mathbf{X},\theta^{(i)}\right], \tag{13}$$

where the corresponding conditional expectations can be obtained from (9). Details can be found in the the Supplement, Section C.

**S Step** The obtained unconstrained optimizers $(\widehat{\mathbf{B}},\widehat{\mathbf{V}},\widehat{\mathbf{\Sigma}_{\mathbf{f}}},\widehat{\sigma_{\mathbf{e}}^2})$ almost always satisfy the condition of Proposition 1. Therefore, we can get the unique equivalent set of parameters that satisfy the identifiability conditions. More specifically, we first 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 resulting $\mathbf{V}^{(i+1)}$ and $\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}$ satisfy the identifiability conditions and are unique up to a column reordering. Then, we set $\mathbf{B}^{(i+1)}$ equal to $\widehat{\mathbf{B}}\widehat{\mathbf{V}}^T\mathbf{V}^{(i+1)}$. It's easy to see that:

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

Finally, we set $\sigma_{\mathbf{e}}^{2(i+1)} = \widehat{\sigma_{\mathbf{e}}^2}$. Furthermore, we reorder the columns of $\mathbf{V}^{(i+1)}$, and accordingly the columns of $\mathbf{B}^{(i+1)}$ and the rows/columns of $\mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}$, to have decreasing column norms of $\mathbf{XV}^{(i+1)}$. As a result, we get parameter estimates for the $(i+1)$th iteration as $\theta^{(i+1)} = (\mathbf{B}^{(i+1)}, \mathbf{V}^{(i+1)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}, \sigma_{\mathbf{e}}^{2(i+1)})$.

We refer to the modified version of the EM algorithm as the *EMS* algorithm, where E is the classical E step, M is the unconstrained maximization step, and S stands for the additional standardization step. Since we have analytical expressions in each step of an iteration, the computation for EMS is very fast. 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 its convergence 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).*

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

1: Set initial values for the parameters $(\mathbf{B}^{(0)}, \mathbf{V}^{(0)}, \mathbf{\Sigma}_{\mathbf{f}}^{(0)}, \sigma_{\mathbf{e}}^{2(0)})$;
2: **while** $\mathcal{L}(\mathbf{X}|\theta^{(i+1)}) - \mathcal{L}(\mathbf{X}|\theta^{(i)}) >$threshold **do**
3: **E Step:** Derive the conditional distribution (9) given $\theta^{(i)} = (\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i)}, \sigma_{\mathbf{e}}^{2(i)})$;
4: **M Step:** Obtain the unconstrained optimizer $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_{\mathbf{f}}}, \widehat{\sigma_{\mathbf{e}}^2})$ from (10)-(13);
5: **S Step:** Standardize $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma}_{\mathbf{f}}}, \widehat{\sigma_{\mathbf{e}}^2})$ to get $\theta^{(i+1)} = (\mathbf{B}^{(i+1)}, \mathbf{V}^{(i+1)}, \mathbf{\Sigma}_{\mathbf{f}}^{(i+1)}, \sigma_{\mathbf{e}}^{2(i+1)})$ that satisfy the identifiability conditions;
6: Set $i \leftarrow i+1$.
7: **end while**

# 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 Shapiro (1986) for overparameterized structural models. A similar treatment can be found in Cook et al. (2010).

Specifically, we first focus on the estimable functions $\boldsymbol{\beta} = \mathbf{B}\mathbf{V}^T$ and $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T + \sigma^2_\mathbf{e}\mathbf{I}$, which uniquely define the likelihood function. In order to fit into Shapiro's framework, we rewrite the parameters in the following format:

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

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

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

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 vech$(\cdot)$ and vec$(\cdot)$, i.e., vech$(\boldsymbol{\Omega}) = \mathbf{C}_d$vec$(\boldsymbol{\Omega})$ and vec$(\boldsymbol{\Omega}) = \mathbf{E}_d$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., vec$(\boldsymbol{\Gamma}^T) = \mathbf{K}_{lm}$vec$(\boldsymbol{\Gamma})$. We can obtain the following theorem, whose proof can be found in the Supplement, Section D.

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

$$\sqrt{n}(\widehat{\mathbf{h}} - \mathbf{h}) \to_d \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{h}), \tag{14}$$

*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 $\sqrt{n}\text{vec}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ and $\sqrt{n}\text{vech}(\widehat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})$ are jointly asymptotically normally distributed with mean zero. Moreover, under the aforementioned identifiability conditions, we obtain the following asymptotic property for every single parameter in $\widehat{\boldsymbol{\phi}}$.

**Corollary 1** *Given (14), under the suitable 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. Moreover, the asymptotic covariance matrix of* $\sqrt{n}vec(\widehat{\mathbf{v}}_i-\mathbf{v}_i)$ *for* $i=1,\cdots,r$ *is given in the Supplement, Section E.*

# 5 Numerical Examples

We compare SupSVD with SVD and RRR using extensive simulations (Section 5.1) and two real data examples (Section 5.2). Section 5.1.1 presents the two motivating examples. Section 5.1.2 compares the three methods with data simulated from each of the models respectively to show the adaptivity of SupSVD. Section 5.1.3 illustrates the performances of the methods under a spectrum of settings ranging from PCA to RRR. Finally, in Section 5.2, we illustrate SupSVD using the breast cancer data from The Cancer Genome Atlas Network (2012) and the call center data from Shen and Huang (2008).

## 5.1 Simulation Studies

### 5.1.1 Two Motivating Examples

**Example 1:** Let $\mathbf{X}$ be an data matrix with 80 samples and 200 variables. The samples are divided into 4 equal-sized subgroups, which have different means in the first two dimension of $\mathbf{X}$. Specifically,

$$\mathbf{X}=\mathbf{u}_1\mathbf{v}_1^T+\mathbf{u}_2\mathbf{v}_2^T+\mathbf{E},$$

where $\mathbf{v}_1=(1,0,0,...,0)^T$, $\mathbf{v}_2=(0,1,0,...,0)^T$, $\mathbf{u}_1=(\text{rep}(16,20),\text{rep}(-16,20),\text{rep}(0,40))^T+\boldsymbol{\varepsilon}_1$, and $\mathbf{u}_2=(\text{rep}(0,40),\text{rep}(10,20),\text{rep}(-10,20))^T+\boldsymbol{\varepsilon}_2$. The notation $\text{rep}(a,b)$ denotes a vector of length $b$ whose entries are all equal to $a$. The random vectors $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are independently distributed from $\mathcal{N}(\mathbf{0},4\mathbf{I})$ and $\mathcal{N}(\mathbf{0},9\mathbf{I})$, respectively. The random matrix $\mathbf{E}$ has i.i.d. entries from $\mathcal{N}(0,16)$. The supervision information $\mathbf{Y}$ is the subgroup index.

This setting simulates the situation where the true underlying structure is partially driven by the supervision information, and partially affected by variations from unknown sources. Figure 1 shows the scatter plot of the true score vectors in the first two dimensions as well as the score vectors estimated by the different methods, with the subgroups indicated by different colors and symbols. Clearly, the results from SupSVD are the closest to the underlying truth. SupSVD not only explains a large portion of variation in the data, but also well separates the underlying subgroups. The SVD scores, although explaining slightly bigger variations, mix all the subgroups together. Namely, from an exploratory point of view, SVD fails to capture the subgroup structure in the data. The RRR scores, on the other hand, shrink the four subgroups into four points, and do not allow any sample-to-sample variation. This example shows that by incorporating the additional supervision information, SupSVD can better recover the true underlying structure.

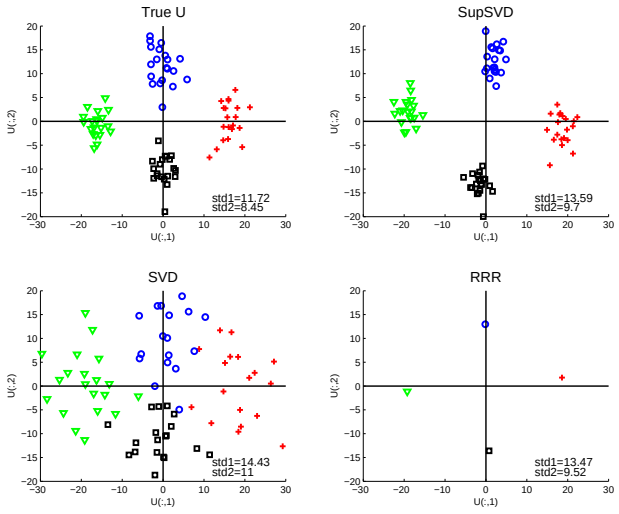


Figure 1: Example 1 - Scatter Plots of $\mathbf{U}_1$ and $\mathbf{U}_2$ from Different Methods. The standard deviations of two score vectors are given by std1 and std2.

**Example 2:** Let $\mathbf{X}$ be a $210 \times 100$ data matrix. The first two dimensions of $\mathbf{X}$ have 4 subgroups, each of which follows a bivariate normal distribution. Specifically, 105 samples are from $\mathcal{N}([-40, 30], [40, 0; 0, 1560])$; 35 samples are from $\mathcal{N}([10, -30], [55, 0; 0, 35])$; 35 samples are from $\mathcal{N}([40, -30], [120, 0; 0, 120])$; 35 samples are from $\mathcal{N}([70, 0], [60, 0; 0, 20])$. The other dimensions of $\mathbf{X}$ are i.i.d. $\mathcal{N}(0, 4)$. The supervision information $\mathbf{Y}$ is the subgroup

index.

Figure 2 contrasts the first two dimensions of $\mathbf{X}$ with the projected data onto the first two loading vectors obtained by each method. The projection from SupSVD is the most similar one with the truth, offering a good interpretation: the first direction separates the four subgroups; the second direction explains the variation within each subgroup. In comparison, both SVD and RRR have tilted loading directions. The variances explained by the first two components are similar among all three methods. This example indicates that SupSVD can improve interpretability by taking into account the supervision information.

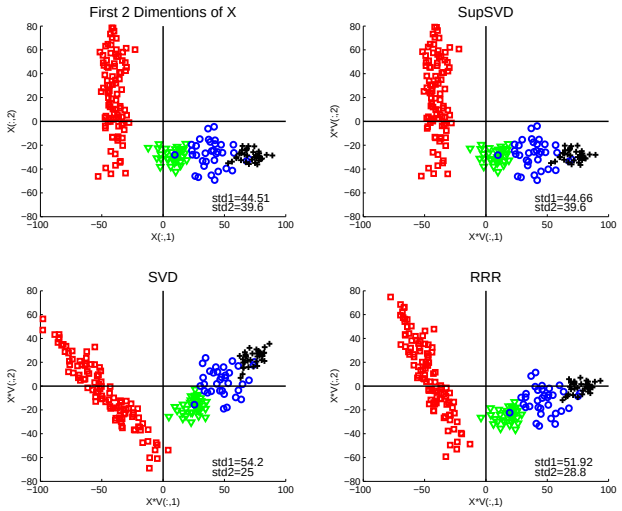


Figure 2: Example 2 - Scatter Plots of $\mathbf{XV}_1$ and $\mathbf{XV}_2$. The standard deviations of two projections are given by std1 and std2.

### 5.1.2 Adaptivity of SupSVD

We consider three simulation cases in this section to demonstrate the adaptivity of SupSVD, where the data are generated using one of the three models (PCA, RRR, SupSVD) respectively, and all three models are used to analyze the data. Most of the simulation parameters use the estimates obtained from the call center data of Section 5.2. The intention is to make the simulation as realistic as possible.

Specifically, we set the sample size $n = 210$, the dimension of $\mathbf{X}$ as $p = 68$, the dimension of the supervision data $\mathbf{Y}$ as $q = 4$, and the rank $r = 4$. We simulate the $210 \times 4$

supervision data matrix $\mathbf{Y}$ using i.i.d. $\mathcal{N}(0,1)$. The data matrix $\mathbf{X}$ is then generated from $\mathbf{X} = \mathbf{Y}\mathbf{B}\mathbf{V}^T + \mathbf{F}\mathbf{V}^T + \mathbf{E}$ where the parameters are defined below for the three cases.

(1) **Case 1:** Obtain the estimated parameters $\mathbf{B}$ and $\mathbf{V}$ from the call center data. Rescale the columns of $\mathbf{B}$ to have norm $(25, 16, 9, 1)$; set $\mathbf{\Sigma_f} = \text{diag}(45, 9, 6, 2)$ and $\sigma_\mathbf{e}^2 = 0.4$. Under this setting, the data are generated from the SupSVD model.

(2) **Case 2:** Set $\mathbf{V}$, $\mathbf{\Sigma_f}$ and $\sigma_\mathbf{e}^2$ to be the same as in **Case 1**. Set $\mathbf{B} = \mathbf{0}$. Under this setting, the data are generated from the PCA model.

(3) **Case 3:** Set $\mathbf{V}$ to be the same as in **Case 1**. Set $\mathbf{\Sigma_f} = \mathbf{0}$ and $\sigma_\mathbf{e}^2 = 16$. Orthogonalize the columns of $\mathbf{B}$ from **Case 1** and rescale the columns to have norm $(25, 16, 9, 1)$. Under this setting, the data are generated from the RRR model.

**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 by

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

where $\|\cdot\|_\mathbb{F}$ denotes the Frobenius norm, and $\mathbf{U}\mathbf{V}^T$ 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, $\widehat{\mathbf{U}} = \left(\mathbf{Y}\widehat{\mathbf{B}}(\widehat{\sigma_\mathbf{e}^2}\widehat{\mathbf{\Sigma_f}}^{-1}) + \mathbf{X}\widehat{\mathbf{V}}\right)\left(\mathbf{I}_r + \widehat{\sigma_\mathbf{e}^2}\widehat{\mathbf{\Sigma_f}}^{-1}\right)^{-1}$, where $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma_f}}, \widehat{\sigma_\mathbf{e}^2})$ is the parameter set estimated from the SupSVD approach. 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 (Golub and Van Loan (2012)) between the true subspace and the estimated subspace which is defined as

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

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

**Results** For each case, we repeat the simulation 100 times and present the median and the median absolute deviations (MAD) of each performance measurement for the three methods in Table 1. As one can see, SupSVD performs well under all three cases, which suggests that it is robust against model misspecification and adapts well to a wide range of practical situations. When the data are generated from the SupSVD model, SupSVD uniformly outperforms SVD and RRR in both parameter estimation and low rank recovery. When the data are generated from the PCA model, SupSVD and PCA have comparable performances, and both are significantly better than RRR; SupSVD is even better than SVD in terms of recovering the underlying structure. For the RRR setting, SupSVD is comparable with RRR and both are better than SVD for most measurements.

| | | SupSVD | SVD | RRR |
|---|---|---|---|---|
| Case 1 (SupSVD) | $MSE_{\mathbf{UV}^T}$ | **0.0324** (3.2E-3) | 0.0349 (3.2E-3) | 0.8946 (5.3E-2) |
| | $MSE_{\mathbf{V}}$ | **0.0002** (2.7E-4) | 0.0018 (5.9E-5) | 0.0036 (3.9E-4) |
| | $MSE_{\sigma_e^2}$ | **0.0001** (7.5E-5) | 0.0010 (2.3E-4) | 0.7635 (9.5E-2) |
| | $Angle_{\mathbf{V}}$ | **15.38** (1.25) | 15.42 (1.27) | 72.00 (9.95) |
| Case 2 (PCA) | $MSE_{\mathbf{UV}^T}$ | **0.0327** (3.5E-3) | 0.0354 (3.5E-3) | 0.8922 (4.6E-2) |
| | $MSE_{\mathbf{V}}$ | **0.0003** (1.5E-4) | **0.0003** (1.4E-4) | 0.0053 (8.5E-4) |
| | $MSE_{\sigma_e^2}$ | **0.0001** (8.9E-5) | 0.0011 (2.5E-4) | 0.7546 (8.2E-2) |
| | $Angle_{\mathbf{V}}$ | **15.50** (1.18) | **15.50** (1.19) | 81.74 (5.60) |
| Case 3 (RRR) | $MSE_{\mathbf{UV}^T}$ | 0.6765 (3.4E-2) | 2.4378 (7.2E-2) | **0.4672** (3.3E-2) |
| | $MSE_{\mathbf{V}}$ | 0.0036 (7.9E-4) | 0.0024 (1.5E-4) | **0.0018** (2.0E-4) |
| | $MSE_{\sigma_e^2}$ | 1.1836 (5.2E-1) | 6.2873 (1.1E0) | **0.4348** (3.0E-1) |
| | $Angle_{\mathbf{V}}$ | 79.56 (6.54) | 84.19 (4.20) | **63.90** (4.81) |

Table 1: Section 5.1.2 - Median(MAD) for Low Rank Structure Recovery Accuracy and Parameter Estimation Accuracy.

### 5.1.3 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 simplicity, we set $n = 210$, $p = 68$, $q = 1$,

and $r = 1$. Fill the $210 \times 1$ vector $\mathbf{Y}$ with standard normal random numbers. We simulate $\mathbf{X}$ from the SupSVD model, with $\mathbf{V}$ being the first column of $\mathbf{V}$ in Case 1, $\sigma_{\mathbf{e}}^2 = 16$, and $(\mathbf{B}, \mathbf{\Sigma_f}) \in \{(0, 36), (1, 25), (2, 16), (3, 9), (4, 0)\}$. Therefore, the SupSVD model ranges from the PCA model $\mathbf{X} = 6\mathbf{Z}\mathbf{V}^T + \mathbf{E}$ (where $\mathbf{Z}$ is a random vector with i.i.d. entries from standard normal distribution) to the RRR model $\mathbf{X} = 4\mathbf{Y}\mathbf{V}^T + \mathbf{E}$. 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 Figure 3. 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.

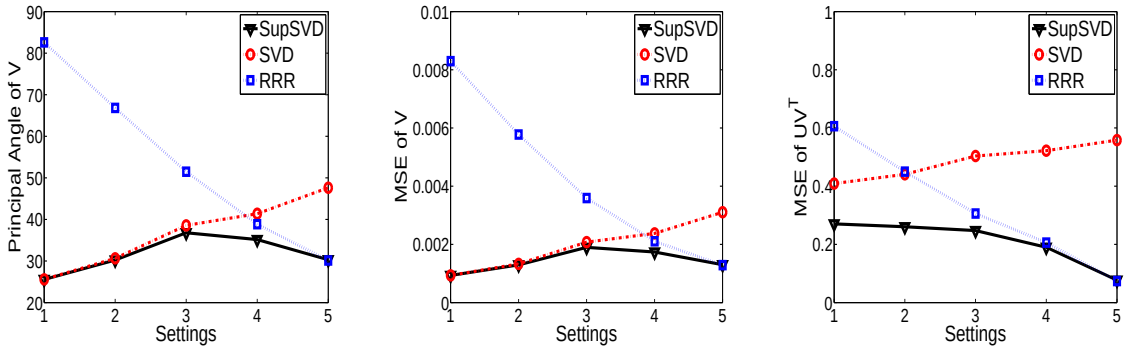


Figure 3: Section 5.1.3 - Median Curves for $Angle_{\mathbf{V}}$, $MSE_{\mathbf{V}}$, and $MSE_{\mathbf{UV}^T}$ Based on 100 Simulation Runs.

## 5.2 Real Data Examples

### 5.2.1 Breast Cancer Data

The first real data set is a gene expression data set for breast cancer patients, obtained from the The Cancer Genome Atlas (TCGA) project (The Cancer Genome Atlas Network, 2012). 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 for the cancer patients. In this case, we have additional information of

disease subtype for each sample. Conceptually, disease may modulate gene expressions, cf. the reactive model in Schadt et al. (2005). This means the cancer subtypes may partially drive the underlying structure of the gene expression data. Samples from the same subtype may share some common genetic variations. We use the subtype information as our supervision data and apply the SupSVD method.

Specifically, the raw data set contains 17814 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 preprocess the data by imputing the missing values, removing genes with low expression level, and centering each gene. As a result, we obtain the final $\mathbf{X}$ as a column-centered data matrix with 348 samples and 645 genes. Based on the scree plot of the singular values of $\mathbf{X}$, we set the rank of the underlying structure to be 3.

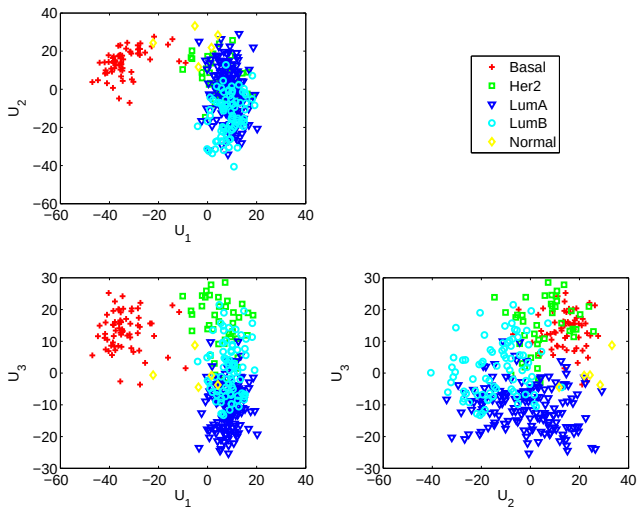


Figure 4: Breast Cancer Data - Scatter Plots of SupSVD Score Vectors. The 5 different subtypes are well separated by the first three score vectors.

Figure 4 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 mainly explains variations within each subtype. The third score vector roughly separates the Her2, LumA, and LumB subgroups. Figure 5 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 have several different patterns. The SupSVD method effectively captures important underlying patterns associated with the supervision information, while still allowing some within-group variation.

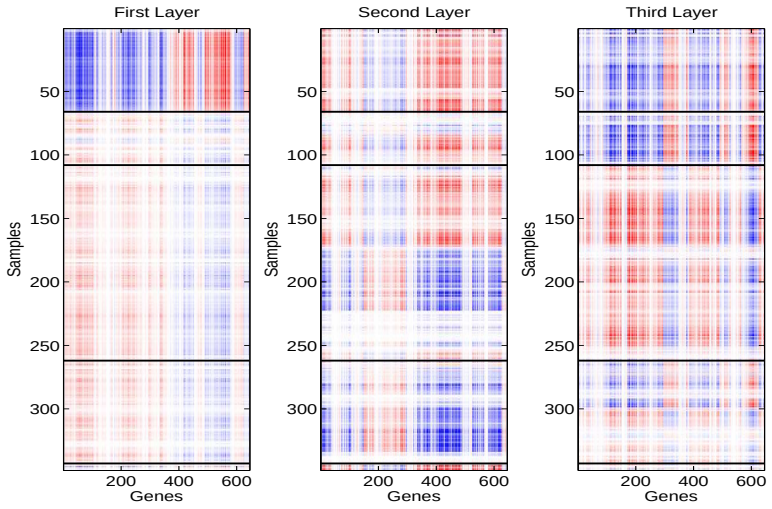


Figure 5: Breast Cancer Data - Heat Map of 1st Three Unit-rank SupSVD Structures of the Gene Expression Data. Blue is negative and red is positive. The samples are grouped in the order of Basal, Her2, LumA, LumB, Normal. The genes are reordered for better visualization.

### 5.2.2 Call Center Data

We then analyze the call center data previously studied by Shen and Huang (2008). The data record the number of agent-seeking calls to a banking call center during each 15-minute interval (from 7am to midnight) for 42 consecutive weeks. The goal is to understand the arrival pattern of calls and forecast future call volumes. We process the data set in the same way as in Shen and Huang (2008), and focus on the 210 weekdays since the weekends have very different patterns. After imputing missing data, replacing outliers, and applying the square root transformation $\sqrt{N + 1/4}$ where $N$ is the count data matrix, we get the

data matrix $\mathbf{X}$ with 210 rows (days) and 68 columns (15-minute intervals). Moreover, we center each column of $\mathbf{X}$ to have mean zero. The scree plot of the singular values of $\mathbf{X}$ suggests the rank to be 4. The supervision data matrix $\mathbf{Y}$ for this case contains the dummy variables for the day-of-week. Shen and Huang (2008) point out that the entries of $\mathbf{X}$ are approximatley normally distributed, and the weekday effect is a primary factor for the call volume patterns. Therefore, it makes sense to apply SupSVD in this case.

Due to the page limit, we omit the estimation results of SupSVD, and only present the follow-up forecasting results. We follow the forecasting procedure proposed by Shen and Huang (2008) (details can be found therein), but replacing their SVD with our SupSVD. We perform a rolling one-day-ahead forecasting scheme: use 150 days of data as the training set to predict the call volumes for the next day; then roll the forecasting window ahead for one day; repeat for 60 days. For each day, the forecasting accuracy is measured by the root mean squared error (RMSE) and the mean relative error (MRE) defined as

$$RMSE = \sqrt{\frac{1}{68}\sum_{i=1}^{68}(N_i - \widehat{N}_i)^2}, \qquad MRE = \frac{100}{68}\sum_{i=1}^{68}\frac{|N_i - \widehat{N}_i|}{N_i}$$

where $N_i$ and $\widehat{N}_i$ are the true and predicted call volumes in the $i$th interval of the next day.

Table 2 presents the comparison of forecasting performance between SVD and SupSVD. Clearly, SupSVD outperforms SVD. By using the additional weekday information to guide the dimension reduction, SupSVD captures more essential patterns in the call volumes and has a greater forecasting power.

| | RMSE | | | MRE | | |
|---|---|---|---|---|---|---|
| | Q1 | Median | Q3 | Q1 | Median | Q3 |
| SupSVD | **41.1769** | **50.1403** | **60.1383** | **4.4082** | **5.2686** | **6.5211** |
| SVD | 41.5895 | 50.6250 | 60.1436 | 4.4468 | 5.3321 | 6.6797 |

Table 2: Call Center Data - Comparison of Forecasting Accuracy between SVD and SupSVD. Results are based on one-day-ahead forecasting for 60 days.

# 6 Discussion

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

# Supplementary Materials

**Proofs and Technical Details:** Detailed proofs for all propositions, theorem and corollary, as well as details of the algorithm are provided in the online supplement to the article. (PDF file)

# Acknowledgements:

This research was partially supported by NSF Grants DMS-1127914, DMS-1106912, DMS-0907177 and DMS-1310002.

# References


Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Cook, R. D. (2007). Fisher Lecture: dimension reduction in regression. *Statistical Science*, 22(1):1–26.

Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20:927–1010.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association*, 100(470):410–428.

Dozier, B. P. and Silverstein, J. W. (2007). On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *Journal of Multivariate Analysis*, 98(4):678–694.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Annals of Applied Statistics*, 7(1):523–542.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, New York.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., et al. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717.

Shabalin, A. and Nobel, A. (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, 81(393):142.

Shen, D., Shen, H., and Marron, J. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *Journal of Multivariate Analysis*, 115:317–333.

Shen, H. and Huang, J. Z. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing & Service Operations Management*, 10(3):391–410.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

# Supplement: Proofs and Technical Details for "Supervised Singular Value Decomposition and Its Asymptotic Properties"

## A Proof of Proposition 1

PROOF. Let $(\mathbf{B}, \mathbf{V}, \mathbf{\Sigma_f}, \sigma_\mathbf{e}^2)$ be a parameter set such that $\mathbf{B}$ is a $q \times r$ matrix, $\mathbf{V}$ is a $p \times r$ matrix, $\mathbf{\Sigma_f}$ is a $r \times r$ positive definite matrix, and $\sigma_\mathbf{e}^2$ is a positive scalar. Moreover, let the largest $r$ eigenvalues of the $p \times p$ matrix $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T + \sigma_\mathbf{e}^2\mathbf{I}$ to be distinct and greater than the rest $p - r$ equal eigenvalues. It's equivalent to say that $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T$ has $r$ positive distinct eigenvalues. We have the eigen-decomposition of the $p \times p$ matrix $\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T$ as

$$\mathbf{V}\mathbf{\Sigma_f}\mathbf{V}^T = \widehat{\mathbf{V}}\widehat{\mathbf{\Sigma_f}}\widehat{\mathbf{V}}^T$$

where $\widehat{\mathbf{\Sigma_f}}$ is the $r \times r$ diagonal matrix containing the distinct eigenvalues, and $\widehat{\mathbf{V}}$ is the $p \times r$ orthonormal matrix containing the corresponding eigenvectors. Moreover, set $\widehat{\mathbf{B}} = \mathbf{B}\mathbf{V}^T\widehat{\mathbf{V}}$. Since $\mathbf{V}$ and $\widehat{\mathbf{V}}$ have the same column space, we know

$$\mathbf{B}\mathbf{V}^T = \widehat{\mathbf{B}}\widehat{\mathbf{V}}^T.$$

Therefore, the new parameter set $(\widehat{\mathbf{B}}, \widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma_f}}, \sigma_\mathbf{e}^2)$ is equivalent with the original parameter set in terms of Model (2), and satisfies the aforementioned identifiability conditions.

The uniqueness of the resulting parameter set is guaranteed by the uniqueness of the eigen-decomposition of the matrix with distinct eigenvalues.

## B Proof of Proposition 2

PROOF. Let $\boldsymbol{\theta}^{(i)} = (\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \mathbf{\Sigma}_\mathbf{f}^{(i)}, \sigma_\mathbf{e}^{2(i)})$ denote the EMS parameter estimation from the $i$th iteration. From the algorithm we know it satisfies the identifiability conditions. Let $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})$ denote the conditional expectation of the joint log likelihood. Namely,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}) &= \mathrm{E}_\mathbf{U}(\mathcal{L}(\mathbf{X}, \mathbf{U}|\boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(i)}) \\ &= \mathrm{E}_\mathbf{U}(\mathcal{L}(\mathbf{U}|\mathbf{X}, \boldsymbol{\theta})|\mathbf{X}, \boldsymbol{\theta}^{(i)}) + \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) \end{aligned}$$

Let $\widehat{\boldsymbol{\theta}}$ denote the unconstrained optimizer from the M step of EMS algorithm. Namely,

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)})$$

Referring to the information inequality that $\mathrm{E}_g(\log f) \leq \mathrm{E}_g(\log g)$ for any densities $f$ and $g$, we have

$$\mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{U}|\mathbf{X}, \widehat{\boldsymbol{\theta}})|\mathbf{X}, \boldsymbol{\theta}^{(i)}) \leq \mathrm{E}_{\mathbf{U}}(\mathcal{L}(\mathbf{U}|\mathbf{X}, \boldsymbol{\theta}^{(i)})|\mathbf{X}, \boldsymbol{\theta}^{(i)})$$

Combining with the fact $Q(\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}^{(i)}) \geq Q(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(i)})$, we know

$$\mathcal{L}(\mathbf{X}|\widehat{\boldsymbol{\theta}}) \geq \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}^{(i)})$$

Moreover, let $\boldsymbol{\theta}^{(i+1)}$ denote the equivalent parameter set that satisfies the identifiability conditions. We have

$$\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}^{(i+1)}) = \mathcal{L}(\mathbf{X}|\widehat{\boldsymbol{\theta}}) \geq \mathcal{L}(\mathbf{X}|\boldsymbol{\theta}^{(i)})$$

Therefore, the likelihood of the observed data $\mathbf{X}$ is monotonically nondecreasing with iterations. If we assume the maximum likelihood exists, the EMS algorithm can always converge.

# C Details of Algorithm 1

In the paper, we propose the EMS algorithm, which is a modified version of EM algorithm, to efficiently estimate the SupSVD model parameters. The detailed calculations for each step in each iteration are described below. We use $(\mathbf{B}^{(i)}, \mathbf{V}^{(i)}, \boldsymbol{\Sigma}_{\mathbf{f}}^{(i)}, \sigma_{\mathbf{e}}^{2(i)})$ to denote the estimations from the $i$th iteration.

**Initial estimation**: Our numerical studies indicate the algorithm is not sensitive to initial values. In practice, we apply SVD to the matrix $\mathbf{X}$ to get the initial estimation. More specifically, we first find the rank-$r$ approximation of $\mathbf{X}$ as

$$\mathbf{X} \approx \mathbf{U}\mathbf{V}^T \tag{15}$$

where $\mathbf{U}$ is the $n \times r$ semi-orthogonal matrix (i.e., the submatrix of the product of the left singular matrix and the diagonal singular value matrix), and $\mathbf{V}$ is the $p \times r$ matrix with

orthonormal columns (i.e., the submatrix of the right singular matrix). Here $\mathbf{V}$ is an initial estimation of $\mathbf{V}$ in our model. We treat $\mathbf{X} - \mathbf{U}\mathbf{V}^T$ as a random matrix with i.i.d. entries from $\mathcal{N}(0, \sigma_{\mathbf{e}}^2)$. Therefore we can get an initial estimation of $\sigma_{\mathbf{e}}^2$. Then we regress $\mathbf{U}$ on $\mathbf{Y}$ and assume that the multivariate residuals are i.i.d. with diagonal covariance structure. The regression coefficient matrix is an initial estimation of $\mathbf{B}$ and the diagonal covariance matrix is an initial estimation of $\mathbf{\Sigma_f}$.

**E step**: We have the conditional distribution (9) of $\mathbf{U}$ given $\mathbf{X}$ under the current parameter estimations. We can calculate the following quantities to be used in M step.

(1) First order conditional expectation:

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

(2) Second order conditional expectation:

$$\mathrm{E}_{\mathbf{U}}\left(\mathbf{U}^T\mathbf{U}|\mathbf{X}, \theta^{(i)}\right) = n\Omega^{(i)}_{\mathbf{U}|\mathbf{X}} + {\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}}^T\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}$$

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

(3) Conditional expectation of any quadratic form in $\mathbf{U}$:

$$\mathrm{E}_{\mathbf{U}}\left(\mathrm{tr}\left(\mathbf{U}\mathbf{\Delta}\mathbf{U}^T\right)|\mathbf{X}, \theta^{(i)}\right) = n\mathrm{tr}\left(\mathbf{\Delta}\Omega^{(i)}_{\mathbf{U}|\mathbf{X}}\right) + \mathrm{tr}\left(\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}\mathbf{\Delta}{\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}}^T\right)$$

where $\mathbf{\Delta}$ is any $r \times r$ symmetric matrix.

**M step**: We maximize the object function $\mathrm{E}_{\mathbf{U}}\left(\mathcal{L}(\mathbf{X}, \mathbf{U})|\mathbf{X}, \theta^{(i)}\right)$ without any constraints. Specifically, we set the partial derivatives of the conditional expectation with respect to all parameters to zero, and solve for the maximizer. Referring to the Leibniz's rule, we can exchange partial derivative with conditional expectation. We have

$$\begin{aligned}
\frac{\partial\mathcal{L}(\mathbf{X}, \mathbf{U})}{\partial\mathbf{B}} &= 2(\mathbf{Y}^T\mathbf{U} - \mathbf{Y}^T\mathbf{Y}\mathbf{B})\mathbf{\Sigma}_{\mathbf{f}}^{-1} \\
\frac{\partial\mathcal{L}(\mathbf{X}, \mathbf{U})}{\partial\mathbf{V}} &= 2\sigma_{\mathbf{e}}^{-2}(\mathbf{X}^T\mathbf{U} - \mathbf{V}\mathbf{U}^T\mathbf{U}) \\
\frac{\partial\mathcal{L}(\mathbf{X}, \mathbf{U})}{\partial\mathbf{\Sigma_f}} &= -n\mathbf{\Sigma}_{\mathbf{f}}^{-1} + \mathbf{\Sigma}_{\mathbf{f}}^{-1}(\mathbf{U} - \mathbf{Y}\mathbf{B})^T(\mathbf{U} - \mathbf{Y}\mathbf{B})\mathbf{\Sigma}_{\mathbf{f}}^{-1} \\
\frac{\partial\mathcal{L}(\mathbf{X}, \mathbf{U})}{\partial\sigma_{\mathbf{e}}^2} &= -np\sigma_{\mathbf{e}}^{-2} + \sigma_{\mathbf{e}}^{-4}\mathrm{tr}\left((\mathbf{X} - \mathbf{U}\mathbf{V}^T)(\mathbf{X} - \mathbf{U}\mathbf{V}^T)^T\right)
\end{aligned}$$

By setting the conditional expectations of the above items to zero, we have

$$\begin{aligned}
\widehat{\mathbf{B}} &= (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathrm{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X},\theta^{(i)}),\\
\widehat{\mathbf{V}} &= \mathbf{X}^T\mathrm{E}_{\mathbf{U}}(\mathbf{U}|\mathbf{X},\theta^{(i)})\left[\mathrm{E}_{\mathbf{U}}(\mathbf{U}^T\mathbf{U}\ |\mathbf{X},\theta^{(i)})\right]^{-1},\\
\widehat{\mathbf{\Sigma_f}} &= \frac{1}{n}\mathrm{E}_{\mathbf{U}}\left[(\mathbf{U}-\mathbf{Y}\widehat{\mathbf{B}})^T(\mathbf{U}-\mathbf{Y}\widehat{\mathbf{B}})\ |\mathbf{X},\theta^{(i)}\right],\\
\widehat{\sigma^2_{\mathbf{e}}} &= \frac{1}{np}\mathrm{E}_{\mathbf{U}}\left[\ \mathrm{tr}((\mathbf{X}-\mathbf{U}\widehat{\mathbf{V}}^T)(\mathbf{X}-\mathbf{U}\widehat{\mathbf{V}}^T)^T)\ |\mathbf{X},\theta^{(i)}\right].
\end{aligned}$$

By substituting the corresponding conditional expectations with the quantities obtained in E step, we have the following explicit expressions of all unconstrained maximizers

$$\begin{aligned}
\widehat{\mathbf{B}} &= (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}\\
\widehat{\mathbf{V}} &= \mathbf{X}^T\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}\left(n\Omega^{(i)}_{\mathbf{U}|\mathbf{X}}+\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}{}^T\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}\right)^{-1}\\
\widehat{\mathbf{\Sigma_f}} &= \frac{1}{n}\left(n\Omega^{(i)}_{\mathbf{U}|\mathbf{X}}+\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}{}^T\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}+\widehat{\mathbf{B}}^T\mathbf{Y}^T\mathbf{Y}\widehat{\mathbf{B}}-\widehat{\mathbf{B}}^T\mathbf{Y}^T\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}-\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}{}^T\mathbf{Y}\widehat{\mathbf{B}}\right)\\
\widehat{\sigma^2_{\mathbf{e}}} &= \frac{1}{np}\left(\mathrm{tr}(\mathbf{X}\mathbf{X}^T)-2\mathrm{tr}\left(\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}\widehat{\mathbf{V}}^T\mathbf{X}^T\right)+n\mathrm{tr}\left(\widehat{\mathbf{V}}^T\widehat{\mathbf{V}}\Omega^{(i)}_{\mathbf{U}|\mathbf{X}}\right)+\mathrm{tr}\left(\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}\widehat{\mathbf{V}}^T\widehat{\mathbf{V}}\Theta^{(i)}_{\mathbf{U}|\mathbf{X}}{}^T\right)\right)
\end{aligned}$$

where the parameters are estimated from previous iteration.

**S step**: As in the Supplement, Section A, we standardize the parameter set $(\widehat{\mathbf{B}},\widehat{\mathbf{V}},\widehat{\mathbf{\Sigma_f}},\widehat{\sigma^2_{\mathbf{e}}})$ by first eigen-decomposing $\widehat{\mathbf{V}}\widehat{\mathbf{\Sigma_f}}\widehat{\mathbf{V}}^T$ as $\mathbf{V}^{(i+1)}\mathbf{\Sigma}^{(i+1)}_{\mathbf{f}}\mathbf{V}^{(i+1)^T}$, and then set $\mathbf{B}^{(i+1)}=\widehat{\mathbf{B}}\widehat{\mathbf{V}}^T\mathbf{V}^{(i+1)}$ and $\sigma^{2(i+1)}_{\mathbf{e}}=\widehat{\sigma^2_{\mathbf{e}}}$. As a result, $(\mathbf{B}^{(i+1)},\mathbf{V}^{(i+1)},\mathbf{\Sigma}^{(i+1)}_{\mathbf{f}},\sigma^{2(i+1)}_{\mathbf{e}})$ is the set of parameter estimations from the current iteration that satisfies the identifiability conditions.

**Stopping rule**: As shown in the Supplement, Section B, the log likelihoods of the observed data are monotonically nondecreasing with iterations. We evaluate the log likelihood at each iteration and terminate the algorithm when the increase between two iterations is below $10^{-5}$.

# D Proof of Theorem 1

Proof. The proof is similar with the proof of Theorem 5.1 in Cook et al. (2010).

The SupSVD model (2) can be written as a simple multivariate linear model

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

where the coefficient matrix $\boldsymbol{\beta}$ and the noise matrix $\boldsymbol{\varepsilon}$ are equal to $\mathbf{B}\mathbf{V}^T$ and $\mathbf{F}\mathbf{V}+\mathbf{E}$ separately. Rows of the residual matrix is i.i.d. from $\mathcal{N}_p(\mathbf{0},\boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}=\mathbf{V}\boldsymbol{\Sigma}_\mathbf{f}\mathbf{V}^T+\sigma^2_\mathbf{e}\mathbf{I}$. Let $\mathbf{h}=\begin{pmatrix}\text{vec}(\boldsymbol{\beta})\\ \text{vech}(\boldsymbol{\Sigma})\end{pmatrix}$ denote the true parameters that satisfy the overparameterized structural constraints, and let $\widehat{\mathbf{h}}_{full}$ denote the unconstrained maximum likelihood estimation of the multivariate regression model. From classic asymptotic theories for maximum likelihood we know $\sqrt{n}(\widehat{\mathbf{h}}_{full}-\mathbf{h})$ is asymptotically normally distributed with mean equal to zero and covariance equal to $\mathbf{J}^{-1}$, i.e., the inverse of the Fisher information matrix of $\mathbf{h}$.

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

where $\boldsymbol{\Sigma}_\mathbf{Y}=\lim_{n\to\infty}\mathbf{Y}\mathbf{Y}^T/n$ and $\mathbf{E}_p$ is the expansion matrix relating vech() with vec().

In order to apply Shapiro's theorem, we define a discrepancy function $F(\cdot,\cdot)$ as follows. It is proportional to the log likelihood difference between $\widehat{\mathbf{h}}_{full}$ and any parameter set $\mathbf{h}$ that satisfies the overparameterized structural constraints.

$$\begin{aligned}F(\widehat{\mathbf{h}}_{full},\mathbf{h}) &= \text{tr}((\mathbf{X}-\mathbf{Y}\boldsymbol{\beta})^T(\mathbf{X}-\mathbf{Y}\boldsymbol{\beta})\boldsymbol{\Sigma})+n\log|\boldsymbol{\Sigma}|\\ &\quad -\text{tr}((\mathbf{X}-\mathbf{Y}\widehat{\boldsymbol{\beta}}_{full})^T(\mathbf{X}-\mathbf{Y}\widehat{\boldsymbol{\beta}}_{full})\widehat{\boldsymbol{\Sigma}}_{full})-n\log|\widehat{\boldsymbol{\Sigma}}_{full}|\end{aligned}$$

It's straightforward to see that $F(\widehat{\mathbf{h}}_{full},\mathbf{h})$ is nonnegative, equal to 0 if and only if $\mathbf{h}=\widehat{\mathbf{h}}_{full}$. Moreover, $F(\widehat{\mathbf{h}}_{full},\mathbf{h})$ is twice continuously differentiable in terms of $\mathbf{h}$ and $\widehat{\mathbf{h}}_{full}$. Besides, from the regularity of the normal likelihood we know, there is no neighborhood of $\widehat{\mathbf{h}}_{full}$ such that $F(\widehat{\mathbf{h}}_{full},\mathbf{h})$ is zero for all $\mathbf{h}$ in it. Therefore, from Shapiro's theorem, we know the minimizer of $F(\widehat{\mathbf{h}}_{full},\cdot)$, or equivalently, the maximizer of the log likelihood function under the structural constraints, has the asymptotic normality. More specifically,

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

and $\boldsymbol{\Sigma}_\mathbf{h}=\mathbf{P}\boldsymbol{\Gamma}\mathbf{P}^T$, where $\mathbf{P}=\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T\mathbf{J}$ is the projection matrix and $\boldsymbol{\Gamma}$ is the asymptotic covariance matrix for $\widehat{\mathbf{h}}_{full}$. Here the matrix $\mathbf{J}$ is the Fisher Information of $\mathbf{h}$ as $n$ goes to infinity, and the matrix $\mathbf{H}$ is the Jacobian matrix of $\mathbf{h}$ with respect to the overparameterized model parameters. The symbol † denotes the Moore-Penrose inverse. Particularly, under normality we know that $\boldsymbol{\Gamma}=\mathbf{J}^{-1}$, so that the asymptotic covariance

matrix $\mathbf{\Sigma_h}$ can be simplified as $\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$. The derivation of the Jacobian matrix $\mathbf{H}$ follows from basic matrix calculus, which can also be found in Cook et al. (2010). Specifically,

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

where $\mathbf{C}_p$ is the $p(p+1)/2 \times p^2$ constant contraction matrix; $\mathbf{E}_r$ is the $r^2 \times r(r+1)/2$ constant expansion matrix; $\mathbf{K}_{pr}$ is the $pr \times pr$ constant commutation matrix.

# E Proof of Corollary 1

Proof. We follow the procedure in Anderson (1963) to prove all parameter estimations are jointly asymptotically normal, and derive the asymptotic covariance for the estimated loading vectors $\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i)$, where $i = 1, \cdots, r$.

First, we introduce some notations. We know from Theorem 1 that $\sqrt{n}(\mathrm{vech}(\widehat{\mathbf{\Sigma}}) - \mathrm{vech}(\mathbf{\Sigma})) \to_d \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_0$ is the $p(p+1)/2 \times p(p+1)/2$ lower corner submatrix of $\mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$. We can decompose $\widehat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ as

$$\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Delta}}\widehat{\mathbf{\Gamma}}^T, \quad \mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Delta}\mathbf{\Gamma}$$

where $\widehat{\mathbf{\Gamma}} = (\widehat{\mathbf{V}}, \widehat{\mathbf{V}}_\perp)$, $\widehat{\mathbf{\Delta}} = \begin{pmatrix} \widehat{\mathbf{\Sigma_f}} + \widehat{\sigma^2_\mathbf{e}}\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \widehat{\sigma^2_\mathbf{e}}\mathbf{I}_{p-r} \end{pmatrix}$, $\mathbf{\Gamma} = (\mathbf{V}, \mathbf{V}_\perp)$, $\mathbf{\Delta} = \begin{pmatrix} \mathbf{\Sigma_f} + \sigma^2_\mathbf{e}\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \sigma^2_\mathbf{e}\mathbf{I}_{p-r} \end{pmatrix}$. For notation purpose, we write $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_r)$, and $\mathrm{diag}(\mathbf{\Sigma_f}) = (\sigma^2_{\mathbf{f},1}, \cdots, \sigma^2_{\mathbf{f},r})$. The parameters $(\mathbf{V}, \mathbf{\Sigma_f}, \sigma^2_\mathbf{e})$ and $(\widehat{\mathbf{V}}, \widehat{\mathbf{\Sigma_f}}, \widehat{\sigma^2_\mathbf{e}})$ satisfy the identifiability conditions. Following the idea in Anderson (1963), we denote

$$\mathbf{M} \triangleq \sqrt{n}(\mathbf{\Gamma}^T\widehat{\mathbf{\Sigma}}\mathbf{\Gamma} - \mathbf{\Delta}) = \sqrt{n}(\mathbf{\Gamma}^T\widehat{\mathbf{\Gamma}}\widehat{\mathbf{\Delta}}\widehat{\mathbf{\Gamma}}^T\mathbf{\Gamma} - \mathbf{\Delta}).$$

It's easily seen that $\mathbf{M}$ is asymptotically normally distributed with asymptotic mean

$\mathrm{E}(m_{ij}) = 0$ and asymptotic covariance

$$\begin{aligned}
\mathrm{E}(m_{ij}m_{gh}) &= \mathrm{E}(\mathbf{v}_i^T\sqrt{n}(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{v}_j\mathbf{v}_g^T\sqrt{n}(\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma})\mathbf{v}_h) \\
&= \mathrm{E}(\mathbf{v}_j^T\otimes\mathbf{v}_i^T\sqrt{n}(\mathrm{vec}(\widehat{\mathbf{\Sigma}})-\mathrm{vec}(\mathbf{\Sigma}))\mathbf{v}_h^T\otimes\mathbf{v}_g^T\sqrt{n}(\mathrm{vec}(\widehat{\mathbf{\Sigma}})-\mathrm{vec}(\mathbf{\Sigma}))) \\
&= \mathbf{v}_j^T\otimes\mathbf{v}_i^T\mathrm{E}(\sqrt{n}(\mathrm{vec}(\widehat{\mathbf{\Sigma}})-\mathrm{vec}(\mathbf{\Sigma}))\sqrt{n}(\mathrm{vec}(\widehat{\mathbf{\Sigma}})-\mathrm{vec}(\mathbf{\Sigma}))^T)\mathbf{v}_h\otimes\mathbf{v}_g \\
&= \mathbf{v}_j^T\otimes\mathbf{v}_i^T\mathbf{E}_p\mathrm{E}(\sqrt{n}(\mathrm{vech}(\widehat{\mathbf{\Sigma}})-\mathrm{vech}(\mathbf{\Sigma}))\sqrt{n}(\mathrm{vech}(\widehat{\mathbf{\Sigma}})-\mathrm{vech}(\mathbf{\Sigma}))^T)\mathbf{E}_p^T\mathbf{v}_h\otimes\mathbf{v}_g \\
&= \mathbf{v}_j^T\otimes\mathbf{v}_i^T\mathbf{E}_p\mathbf{\Sigma}_0\mathbf{E}_p^T\mathbf{v}_h\otimes\mathbf{v}_g
\end{aligned}$$

Moreover, we denote $\mathbf{T} \triangleq \mathbf{\Gamma}^T\widehat{\mathbf{\Gamma}}$ and partition it as

$$\mathbf{T} = \begin{pmatrix} \mathbf{v}_1^T\widehat{\mathbf{v}}_1 & \cdots & \mathbf{v}_1^T\widehat{\mathbf{v}}_r & \mathbf{v}_1^T\widehat{\mathbf{V}}_\perp \\ \vdots & & \vdots & \vdots \\ \mathbf{v}_r^T\widehat{\mathbf{v}}_1 & \cdots & \mathbf{v}_r^T\widehat{\mathbf{v}}_r & \mathbf{v}_r^T\widehat{\mathbf{V}}_\perp \\ \mathbf{V}_\perp^T\widehat{\mathbf{v}}_1 & \cdots & \mathbf{V}_\perp^T\widehat{\mathbf{v}}_r & \mathbf{V}_\perp^T\widehat{\mathbf{V}}_\perp \end{pmatrix} = \begin{pmatrix} t_{11} & \cdots & t_{1r} & T_{1\perp} \\ \vdots & & \vdots & \vdots \\ t_{r1} & \cdots & t_{rr} & T_{r\perp} \\ T_{\perp 1} & \cdots & T_{\perp r} & T_{\perp\perp} \end{pmatrix}.$$

Accordingly, we partition $\mathbf{M}$ as

$$\mathbf{M} = \begin{pmatrix} m_{11} & \cdots & m_{1r} & M_{1\perp} \\ \vdots & & \vdots & \vdots \\ m_{r1} & \cdots & m_{rr} & M_{r\perp} \\ M_{\perp 1} & \cdots & M_{\perp r} & M_{\perp\perp} \end{pmatrix}.$$

Following the proof verbatim in Section 2 of Anderson (1963), we know the diagonal values of $\sqrt{n}(\widehat{\mathbf{\Delta}}-\mathbf{\Delta})$ are asymptotically normally distributed. In other words, $\sqrt{n}\mathrm{diag}(\widehat{\mathbf{\Sigma}_\mathbf{f}}-\mathbf{\Sigma}_\mathbf{f})$ and $\sqrt{n}(\widehat{\sigma_\mathbf{e}^2}-\sigma_\mathbf{e}^2)$ are jointly asymptotically normal. The diagonal blocks of $\mathbf{T}$ have the limiting distribution $\sqrt{n}(t_{ii}^2-1) \to_d 0$ $(i = 1, \cdots, r)$ and $\sqrt{n}(T_{\perp\perp}T_{\perp\perp}^T - \mathbf{I}_{p-r}) \to_d \mathbf{0}$. For the off diagonal blocks of $\mathbf{T}$, $\sqrt{n}t_{ij}$ $(i, j = 1, \cdots, r; i \neq j)$ has the same limiting distribution as $m_{ij}/(\sigma_{\mathbf{f},i}^2 - \sigma_{\mathbf{f},j}^2)$; $\sqrt{n}T_{i\perp}$ $(i = 1, \cdots, r)$ has the same limiting distribution as $M_{i\perp}/\sigma_{\mathbf{f},i}^2$; and $\sqrt{n}T_{\perp j}$ $(j = 1, \cdots, r)$ has the same limiting distribution as $M_{\perp j}/\sigma_{\mathbf{f},j}^2$.

In order to get the limiting distribution of $\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i)$ $(i = 1, \cdots, r)$, we notice that

$$\begin{aligned}
\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i) &= \sqrt{n}(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\widehat{\mathbf{v}}_i - \mathbf{v}_i) \\
&= \sqrt{n}(\sum_{j=1}^{r} \mathbf{v}_j\mathbf{v}_j^T\widehat{\mathbf{v}}_i + \mathbf{V}_\perp\mathbf{V}_\perp^T\widehat{\mathbf{v}}_i - \mathbf{v}_i) \\
&= \sqrt{n}(\mathbf{v}_i\mathbf{v}_i^T\widehat{\mathbf{v}}_i - \mathbf{v}_i) + \sqrt{n}\sum_{j\le r, j\neq i} \mathbf{v}_j\mathbf{v}_j^T\widehat{\mathbf{v}}_i + \sqrt{n}\mathbf{V}_\perp\mathbf{V}_\perp^T\widehat{\mathbf{v}}_i \\
&= \sqrt{n}\mathbf{v}_i(t_{ii} - 1) + \sqrt{n}(\mathbf{v}_1 \cdots \mathbf{v}_{i-1}, \mathbf{v}_{i+1} \cdots \mathbf{v}_r, \mathbf{V}_\perp)(t_{1i} \cdots t_{(i-1)i}, t_{(i+1)i} \cdots t_{ri}, T_{\perp i}^T)^T \\
&= \sqrt{n}\mathbf{v}_i(t_{ii} - 1) + \sqrt{n}\boldsymbol{\Gamma}_{-i}\mathbf{t}_{(-i)i}
\end{aligned}$$

where $\boldsymbol{\Gamma}_{-i}$ is the submatrix of $\boldsymbol{\Gamma}$ without the $i$th column and $\mathbf{t}_{(-i)i}$ is the $i$th column of $\mathbf{T}$ without the $i$th entry. The limiting distribution of the first term is 0. The limiting distribution of the second term can be substituted by the limiting distribution of corresponding $\mathbf{M}$ components. Therefore, we have the following two have the same limiting distribution.

$$\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i) =_d \boldsymbol{\Gamma}_{-i}\boldsymbol{\Delta}_i\mathbf{m}_{(-i)i}$$

where $\boldsymbol{\Delta}_i$ is the $(p-1) \times (p-1)$ submatrix of $\boldsymbol{\Delta} - (\sigma_{\mathbf{f},i}^2 + \sigma_{\mathbf{e}}^2)\mathbf{I}_p$ without the $i$th row and $i$th column, and $\mathbf{m}_{(-i)i}$ is the $i$th column of $\mathbf{M}$ without the $i$th entry. From previous derivation, we know the limiting distribution of $\mathbf{m}_{(-i)i}$ is multivariate normal with mean $\mathbf{0}$ and covariance $(\mathbf{v}_i^T \otimes \boldsymbol{\Gamma}_{-i}^T)\mathbf{E}_p\boldsymbol{\Sigma}_0\mathbf{E}_p^T(\mathbf{v}_i \otimes \boldsymbol{\Gamma}_{-i})$. Therefore, we have

$$\sqrt{n}(\widehat{\mathbf{v}}_i - \mathbf{v}_i) \to_d \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{v}_i})$$

where $\boldsymbol{\Sigma}_{\mathbf{v}_i} = \boldsymbol{\Gamma}_{-i}\boldsymbol{\Delta}_i(\mathbf{v}_i^T \otimes \boldsymbol{\Gamma}_{-i}^T)\mathbf{E}_p\boldsymbol{\Sigma}_0\mathbf{E}_p^T(\mathbf{v}_i \otimes \boldsymbol{\Gamma}_{-i})\boldsymbol{\Delta}_i\boldsymbol{\Gamma}_{-i}^T$, for $i = 1, \cdots, r$.

Lastly, since $\mathbf{B} = \mathbf{B}\mathbf{V}^T\mathbf{V} = \boldsymbol{\beta}\mathbf{V}$ and $\mathbf{V}$ can be expressed as a function of $\boldsymbol{\Sigma}$, $\mathbf{B}$ can be expressed as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. According to the joint asymptotic normality of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, it's obvious that $\sqrt{n}\text{vec}(\widehat{\mathbf{B}} - \mathbf{B})$ is also asymptotically normal.