DOI: 10.1111/biom.12698

Incorporating Covariates into Integrated Factor Analysis of Multi-View Data

Gen Li[1,*] and Sungkyu Jung[2]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York 10032, New York, U.S.A.
[2]Department of Statistics, University of Pittsburgh, Pittsburgh 15260, Pennsylvania, U.S.A.
[*]*email:* gl2521@cumc.columbia.edu

Summary. In modern biomedical research, it is ubiquitous to have multiple data sets measured on the same set of samples from different views (i.e., multi-view data). For example, in genetic studies, multiple genomic data sets at different molecular levels or from different cell types are measured for a common set of individuals to investigate genetic regulation. Integration and reduction of multi-view data have the potential to leverage information in different data sets, and to reduce the magnitude and complexity of data for further statistical analysis and interpretation. In this article, we develop a novel statistical model, called supervised integrated factor analysis (SIFA), for integrative dimension reduction of multi-view data while incorporating auxiliary covariates. The model decomposes data into joint and individual factors, capturing the joint variation across multiple data sets and the individual variation specific to each set, respectively. Moreover, both joint and individual factors are partially informed by auxiliary covariates via nonparametric models. We devise a computationally efficient Expectation–Maximization (EM) algorithm to fit the model under some identifiability conditions. We apply the method to the Genotype-Tissue Expression (GTEx) data, and provide new insights into the variation decomposition of gene expression in multiple tissues. Extensive simulation studies and an additional application to a pediatric growth study demonstrate the advantage of the proposed method over competing methods.

Key words: Data integration; Dimension reduction; Multi-source data; Principal component analysis; Supervision.

1. Introduction

In contemporary biomedical studies, researchers usually have access to multiple data sets for the same set of subjects from different views or heterogeneous sources. Such data are commonly referred to as multi-view data or multi-source data. For example, the Genotype-Tissue Expression (GTEx) project collects gene expression data from multiple human tissues for a common set of genotyped individuals to study genetic regulation (The GTEx Consortium, 2015). Different data sets may contain distinct but related information. It is important to understand the relations between variables in different sets, and leverage information across views for further statistical analysis such as inference, prediction and clustering. The process is often called data integration or data fusion.

Factor analysis is a popular tool for modeling dependence among multiple observed variables. It identifies a few latent factors that capture the majority of variation in data. The unknown factors and loadings in factor analysis are sometimes estimated via the principal component analysis (PCA). The obtained factors reduce the dimensionality of the original data and facilitate various statistical analyses. However, the conventional factor analysis only applies to a single data set. There is a pressing need for statistical methods that simultaneously identify the joint and individual structure in multiple data sets.

In addition to multiple primary data sets, auxiliary covariates are often collected on the same samples. In our motivating GTEx example, other than the gene expression data in multiple tissues, genotype data and experimental factors (e.g., batch effect) are also collected. These auxiliary data can be viewed as covariates, driving the underlying expression patterns in multiple tissues. Covariates are potential driving factors of the joint and individual structures in multi-view data. In other words, covariates provide *supervision* to the underlying patterns. Using covariates to inform the integration of multi-view data not only leads to accurate estimation of the underlying patterns but also provides highly interpretable results.

In this article, we develop a novel statistical framework called *Supervised Integrated Factor Analysis* (SIFA), for the integration and reduction of multi-view data informed by auxiliary covariates. SIFA decomposes multi-view data into low-rank joint structure and individual structure. It exploits a small number of joint factors to capture the shared patterns across all data sets, and separate individual factors to capture the specific patterns in each data set. Corresponding loading vectors identify the contribution of the variables to different factors. To allow auxiliary covariates to inform the latent structure, the model assumes each factor is potentially driven by the covariates and some random effects. We particularly consider regression models that flexibly accommodate parametric or nonparametric relations between factors and covariates. Through the regression models, the covariates exert supervision on the latent structure. We also extend the model to incorporate variable selection, in order to identify important covariates that drive different factors. Overall, SIFA provides a general framework for the covariate-driven factor analysis of multi-view data.

There is an extensive body of literature on the integrative analysis of multi-view data (Tseng, Ghosh and Zhou, 2015). Here, we particularly focus on data integration and reduction. Multiple factor analysis is an extension of the conventional factor analysis to multiple data sets (Abdi, Williams and Valentin, 2013). The idea is to merge multiple data with weights and perform the factor analysis on the combined data. However, the method does not distinguish joint and individual structure and may lead to misleading results. More recently, new methods have been developed to decompose the total variation of multiple data sets into shared and individual variation (Löfstedt, Hoffman and Trygg, 2013; Ray et al., 2014; Schouteden et al., 2014; Yang and Michailidis, 2016; Zhou et al., 2016). For example, Lock et al. (2013) adopts an iterative PCA approach to estimate the Joint and Individual Variation Explained (JIVE). However, a drawback of these methods is that they cannot take into account any auxiliary covariates in dimension reduction. When covariates are strongly associated with the latent structure of the multi-view data, incorporating the supervision effects from covariates promises to improve estimation accuracy and interpretability.

Recently, a couple of methods were proposed to allow covariates to inform factor analysis. Li et al. (2016) developed the Supervised Singular Value Decomposition (SupSVD) method that exploits linear models to accommodate covariates in dimension reduction of a primary data matrix. Later, Fan, Liao and Wang (2016) proposed the projected PCA that generalizes SupSVD by allowing nonparametric relations between covariates and factors. However, these methods are only suitable for a single data set, and cannot easily extend to multi-view data. To our best knowledge, there is no covariate-driven factor analysis method for multi-view data decomposition. Our proposed method will bridge the gap and provide a unified framework.

The rest of the article is organized as follows. In Section 2, we propose a semiparametric latent variable model for SIFA and develop an Expectation–Maximization (EM) algorithm to fit the model. In Section 3, extensive simulation studies are conducted to compare the proposed method with existing methods. In Section 4, we apply SIFA to the GTEx multi-tissue genetic data to offer novel insights into the decomposition of genetic variation in a gene set across multiple tissues. In Section 5, we discuss possible directions for future research. Technical details, additional simulation results, and an application to the decoupled growth amplitude and phase data from the Berkeley Growth Study can be found in the online supplementary material.

2. Integrated Factor Analysis Framework

In this section, we first introduce the latent variable model for SIFA and discuss its connection to existing methods. Then we elaborate two sets of identifiability conditions, and devise model fitting algorithms under respective conditions. Finally, we propose rank selection methods to determine the joint and individual ranks in the model.

2.1. *Model*

Let $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_K$ be K primary data matrices of size $n \times p_1, \cdots, n \times p_K$ for the same set of samples collected from K different sources. Each row corresponds to a sample and each column is a variable. Let $\boldsymbol{X}$ be an $n \times q$ data matrix containing covariates for the matched samples. The goal is to identify low-rank joint and individual patterns from the primary data matrices while accounting for the supervision effects from the covariates. Without loss of generality, we center each column of the primary data and the covariates to remove the mean effect of each variable.

We propose a latent variable model called SIFA for the integrative factor analysis of multiple data matrices. For $k = 1, \cdots, K$, the SIFA model is as follows (without special notice, the index k takes integer values from 1 to K):

$$\boldsymbol{Y}_k = \boldsymbol{J}_k + \boldsymbol{J}_k + \boldsymbol{E}_k, \tag{1}$$

$$\boldsymbol{J}_k = \boldsymbol{U}_0 \boldsymbol{V}_{0,k}^T, \tag{2}$$

$$\boldsymbol{A}_k = \boldsymbol{U}_k \boldsymbol{V}_k^T, \tag{3}$$

$$\boldsymbol{U}_0 = \boldsymbol{f}_0(\boldsymbol{X}) + \boldsymbol{F}_0, \tag{4}$$

$$\boldsymbol{U}_k = \boldsymbol{f}_k(\boldsymbol{X}) + \boldsymbol{F}_k. \tag{5}$$

In (1), we adopt a signal-plus-noise model to capture the important patterns in each data set. This type of model is commonly used in the dimension reduction literature (cf. Shabalin and Nobel, 2013). More specifically, the data matrix $\boldsymbol{Y}_k$ consists of signal $\boldsymbol{J}_k + \boldsymbol{A}_k$ and noise $\boldsymbol{E}_k$. The matrix $\boldsymbol{J}_k$ captures the joint structure shared across multiple sources, and the matrix $\boldsymbol{A}_k$ captures the individual structure specific to this data source. The noise matrix $\boldsymbol{E}_k$ is assumed to have independent and identically distributed (i.i.d.) entries from a normal distribution $\mathcal{N}(0, \sigma_k^2)$, where σ_k^2 is an unknown parameter.

In (2) and (3), we assume that the joint and individual patterns for $\boldsymbol{Y}_k$ have low-rank decomposition. Let r_0 be the underlying rank of the joint structure and r_k be the rank of the individual structure in the kth data set. Correspondingly, $\boldsymbol{U}_0$ and $\boldsymbol{U}_k$ are $n \times r_0$ and $n \times r_k$ (latent) factor matrices, and $\boldsymbol{V}_{0,k}$ and $\boldsymbol{V}_k$ are $p_k \times r_0$ and $p_k \times r_k$ loading matrices. In particular, $\boldsymbol{U}_0$ contains r_0 joint factors shared across different data sets, and $\boldsymbol{V}_0 = (\boldsymbol{V}_{0,1}^T, \cdots, \boldsymbol{V}_{0,K}^T)^T$ contains r_0 corresponding joint loadings. The matrices $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ contain r_k individual factors and loadings, respectively. Following the convention of the factor analysis, we assume the factors are independent and the loadings are orthonormal within each matrix. Namely, $\boldsymbol{V}_k^T \boldsymbol{V}_k = \boldsymbol{I}_{r_k}$ for each $k = 0, 1, \cdots, K$, where $\boldsymbol{I}_{r_k}$ denotes the $r_k \times r_k$ identity matrix (we shall drop the subscript when it does not cause any confusion).

In order to capture the driving effects of covariates on the low-rank structure, we propose to regress the latent factors on the covariates via (4) and (5). The mapping functions $\boldsymbol{f}_k(\cdot) : \mathbb{R}^q \mapsto \mathbb{R}^{r_k}$ $(k = 0, 1, \cdots, K)$ are unknown parametric or nonparametric functions. With a slight abuse of notation, we use $\boldsymbol{f}_k(\boldsymbol{X})$ $(k = 0, 1, \cdots, K)$ to represent row-wise mappings. Namely, $\boldsymbol{f}_k(\boldsymbol{X})$ is an $n \times r_k$ matrix whose ith row corresponds to $\boldsymbol{f}_k(\boldsymbol{x}_{(i)})$, where $\boldsymbol{x}_{(i)}$ is the ith row of $\boldsymbol{X}$ $(i = 1, \cdots, n)$. The mapping functions capture flexible relations between covariates and the latent factors. In practice, users can determine whether to use nonparametric functions or parametric functions (e.g., linear functions). Any unknown variation in the factors is contained in the random matrices $\boldsymbol{F}_k$ $(k = 0, 1, \cdots, K)$. In particular, we assume each row of $\boldsymbol{F}_k$ follows an i.i.d. multivariate normal distribution with zero mean and an unknown diagonal covariance matrix $\boldsymbol{\Sigma}_k$ (with positive, distinct, and decreasing diagonal values). Moreover, we assume $\boldsymbol{F}_0$, $\boldsymbol{F}_k$'s, and $\boldsymbol{E}_k$'s are mutually independent.

The proposed SIFA model provides a general framework for the factor analysis of multi-view data. After rearranging the formulas, we get an equivalent form of the model as

$$\boldsymbol{Y}_k = \boldsymbol{f}_0(\boldsymbol{X})\boldsymbol{V}_{0,k}^T + \boldsymbol{f}_k(\boldsymbol{X})\boldsymbol{V}_k^T + \boldsymbol{F}_0\boldsymbol{V}_{0,k}^T + \boldsymbol{F}_k\boldsymbol{V}_k^T + \boldsymbol{E}_k. \qquad (6)$$

It is easy to see that the SIFA model decomposes the kth data matrix $\boldsymbol{Y}_k$ into five parts: 1) $\boldsymbol{f}_0(\boldsymbol{X})\boldsymbol{V}_{0,k}^T$ is the joint deterministic structure (because $\boldsymbol{f}_0(\boldsymbol{X})$ is shared across multiple data sources and non-random) driven by the covariates; 2) $\boldsymbol{f}_k(\boldsymbol{X})\boldsymbol{V}_k^T$ is the individual deterministic structure; 3) $\boldsymbol{F}_0\boldsymbol{V}_{0,k}^T$ is the joint random structure capturing any unknown variation; 4) $\boldsymbol{F}_k\boldsymbol{V}_k^T$ is the individual random structure; 5) $\boldsymbol{E}_k$ is the random noise. With proper identifiability conditions which we will discuss later, the SIFA model attributes the total variation to different parts. Different model components will facilitate different analyses. For example, the joint factors in $\boldsymbol{f}_0(\boldsymbol{X}) + \boldsymbol{F}_0$ can be potentially used for consensus clustering; the individual loadings in $\boldsymbol{V}_k$ can be used to investigate the dependence among variables in the kth data source.

We remark that the proposed SIFA model (6) subsumes many existing methods as special cases. When $K = 1$, that is, with only one primary data set $\boldsymbol{Y}$, there is no distinction between the joint structure and the individual structure. Consequently, the model degenerates to

$$\boldsymbol{Y} = (\boldsymbol{f}(\boldsymbol{X}) + \boldsymbol{F})\boldsymbol{V}^T + \boldsymbol{E},$$

which corresponds to the projected PCA model proposed by Fan, Liao and Wang (2016). In particular, if we let the function $\boldsymbol{f}(\cdot)$ be a linear mapping, that is, $\boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{B}$, where $\boldsymbol{B}$ is a $q \times r$ coefficient matrix, the above model further connects to the SupSVD model developed in Li et al. (2016). Furthermore, if we eliminate the covariate supervision by setting $\boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{0}$, the model degenerates to the conventional factor analysis model or the probabilistic PCA model (Tipping and Bishop, 1999). When $K \geq 2$, without accounting for the covariates (i.e., $\boldsymbol{f}_k(\boldsymbol{X}) = \boldsymbol{0}; k = 0, 1, \cdots, K$), the SIFA model reduces to

$$\boldsymbol{Y}_k = \boldsymbol{F}_0\boldsymbol{V}_{0,k}^T + \boldsymbol{F}_k\boldsymbol{V}_k^T + \boldsymbol{E}_k.$$

This coincides with the JIVE model (Fan, Liao and Wang, 2016) if we assume $\boldsymbol{F}_0$ and $\boldsymbol{F}_k$ are unknown score matrices for the joint and individual structures. The SIFA model unifies and generalizes the above models, and provides a general framework for the integration and reduction of multi-view data informed by covariates.

2.2. *Identifiability*

Suppose $\theta_0 = \{\boldsymbol{f}_0(\cdot), \boldsymbol{f}_k(\cdot), \boldsymbol{V}_0, \boldsymbol{V}_k, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_k, \sigma_k^2; k = 1, \ldots, K\}$ is a parameter set for Model (6), satisfying the basic conditions of $\boldsymbol{V}_k^T\boldsymbol{V}_k = \boldsymbol{I}$ and $\boldsymbol{\Sigma}_k$ being diagonal with distinct (decreasing) positive diagonal values for each $k = 0, 1, \cdots, K$. If there is only one primary data set (i.e., $K = 1$), the model is uniquely defined (Li et al., 2016). However, when there are multiple data sets, the above basic conditions are no longer sufficient for identifiability.

To be specific, let Θ be the collection of parameter sets θ satisfying the basic conditions and having equal likelihood $\mathcal{L}(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_K \mid \theta)$ (defined later in (7)) with θ_0 for any data $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_K$. Namely, any parameter set $\theta \in \Theta$ and θ_0 are *observationally equivalent* for Model (6), that is, Θ is the *equivalence class* of θ_0. We note that the collection Θ typically contains multiple elements (see the supplementary material for examples of some equivalent models). In other words, θ_0 is unidentifiable. This non-identifiability is mainly caused by the indistinguishable individual and joint structures. Different elements in Θ may have different sets of ranks, or the same set of ranks but different parameters. Additional regularity conditions are needed to enforce the identifiability of the SIFA model. For this, we propose two sets of sufficient conditions.

First, we consider a set of *general conditions* for each $k = 1, \cdots, K$:

A1. Each submatrix $\boldsymbol{V}_{0,k}$ of the joint loading matrix $\boldsymbol{V}_0$ has full column rank;

A2. The columns in $\boldsymbol{V}_{0,k}$ and $\boldsymbol{V}_k$ are linearly independent, and $r_0 + r_k < p_k$.

Loosely speaking, Condition A1 guarantees that the joint loading matrix $\boldsymbol{V}_0$ indeed captures the joint structure, and does not contain any structure only pertaining to a subset of the K data sets. Condition A2 ensures that the joint and individual patterns are well separated, and are not interchangeable. With both conditions, Model (6) is identifiable as shown in the following proposition (the proof is postponed to the supplementary material).

PROPOSITION 1. *Let θ_0 be a parameter set satisfying Conditions A1 and A2. For any element θ in the equivalent class Θ of θ_0, if θ also satisfies Conditions A1 and A2, then θ is equal to θ_0 up to trivial sign changes. Moreover, by writing $r_0(\theta)$ as the rank of $\boldsymbol{V}_0$ in the parameter set θ, we have $r_0(\theta_0) \leq r_0(\theta)$ for all $\theta \in \Theta$.*

The proposition guarantees that the general conditions are sufficient for model identifiability. The identifiability is defined up to trivial column-wise sign changes in $\boldsymbol{V}_k$ and $\boldsymbol{U}_k$ $(k = 0, 1, \cdots, K)$. In practice, one could easily fix the signs by setting the first nonzero entry of each column of $\boldsymbol{V}_k$ to be positive. Correspondingly, the sign of each column of $\boldsymbol{U}_k$ is fixed.

Remark: Technically, the general conditions may slightly affect the generalizability of the model. Condition A1 rules out the possibility of any partially joint structure pertaining to multiple but not all data sets. Namely, the model cannot identify common patterns across a subset of data sets. The same issue exists for JIVE as well. This is a future research direction as discussed in Section 5. Nevertheless, *in practice, the general conditions are suitable for most applications.*

In some circumstances, it is desired to further restrict the model parameters for better interpretation and computation. We consider the following *orthogonal conditions*:

- B1. The columns of $\boldsymbol{V}_{0,k}$ are orthogonal with norm $1/\sqrt{K}$, that is, $\boldsymbol{V}_{0,k}^T\boldsymbol{V}_{0,k} = \frac{1}{K}\boldsymbol{I}$;
- B2. The columns in $\boldsymbol{V}_{0,k}$ and $\boldsymbol{V}_k$ are orthogonal ($\boldsymbol{V}_{0,k}^T\boldsymbol{V}_k = \boldsymbol{0}$), and $r_0 + r_k < p_k$.

Apparently, Conditions B1 and B2 are sufficient conditions for Conditions A1 and A2. Therefore, they are also sufficient conditions for the identifiability of the SIFA model. Condition B1 implies that different data sets contribute roughly equally to the joint factors $\boldsymbol{U}_0$ (i.e., columns in $\boldsymbol{V}_{0,k_1}$ and $\boldsymbol{V}_{0,k_2}$ have the same ℓ_2 norm, for $k_1 \neq k_2$). In many real applications (e.g., the GTEx example in Section 4), when the data are properly preprocessed, the equal contribution assumption can be well justified. Conditions B1 and B2 also imply that the combined loadings $(\boldsymbol{V}_{0,k}, \boldsymbol{V}_k)$ for the kth data set are mutually orthogonal. For high dimensional data, it is reasonable to assume that the orthogonality between different loadings holds (Ahn and Marron, 2010). When both assumptions are justified, it is beneficial to study the SIFA model under the orthogonal conditions. These conditions not only improve model interpretation, but also facilitate computation as discussed in the next subsection.

Remark: The SIFA model with the general conditions is equivariant under individual scaling of each data set. In other words, at the population level, the model is not affected by weighing multiple data sets differently. In practice, to avoid numerical instability, it is recommended to normalize different data sets to the same scale before estimation (e.g., set the Frobenius norm of every data set to be 1). The orthogonal conditions do not have the equivariant property under rescaling. Thus, if the orthogonal assumptions are justifiable, one should directly apply the method without scaling the data. See the supplementary material for more details. There, we also discuss the effect of imbalanced dimensions of different data sets.

2.3. *Algorithm*

To estimate the model parameters in $\theta_0 = \{\boldsymbol{f}_0(\cdot), \boldsymbol{f}_k(\cdot), \boldsymbol{V}_{0,k}, \boldsymbol{V}_k, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_k, \sigma_k^2; k = 1, \ldots, K\}$ for Model (6), we use a maximum likelihood approach. We assume all random variables are from normal distributions. For the ease of presentation, $\boldsymbol{V}_\star = \text{blkdiag}(\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K)$ denotes the combined individual loading matrix of size $\sum_{k=1}^K p_k \times \sum_{k=1}^K r_k$, which is a block-wise diagonal matrix with K diagonal blocks $\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K$. We also let $\boldsymbol{U}_\star = (\boldsymbol{U}_1, \cdots, \boldsymbol{U}_K) = (\boldsymbol{f}_1(\boldsymbol{X}) + \boldsymbol{F}_1, \cdots, \boldsymbol{f}_K(\boldsymbol{X}) + \boldsymbol{F}_K)$ denote the combined individual factor matrix. Let $\boldsymbol{Y}_\star = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_K)$ and $\boldsymbol{E}_\star = (\boldsymbol{E}_1, \cdots, \boldsymbol{E}_K)$ be the combined primary data matrix and noise matrix respectively. As a result, the SIFA model can be succinctly expressed as

$$\boldsymbol{Y}_\star = \boldsymbol{U}_0\boldsymbol{V}_0^T + \boldsymbol{U}_\star\boldsymbol{V}_\star^T + \boldsymbol{E}_\star.$$

The log likelihood function of the SIFA model is

$$\log \mathcal{L}(\boldsymbol{Y}_\star \mid \theta_0) = \sum_{i=1}^{n} \left[-\frac{\sum_{k=1}^K p_k}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_\star| - \frac{1}{2}(\boldsymbol{y}_{\star(i)} - \boldsymbol{\mu}_{\star(i)})^T \boldsymbol{\Sigma}_\star^{-1} (\boldsymbol{y}_{\star(i)} - \boldsymbol{\mu}_{\star(i)}) \right], \quad (7)$$

where $\boldsymbol{y}_{\star(i)}$ is a column vector corresponding to the ith row of $\boldsymbol{Y}_\star$, and $\boldsymbol{\mu}_{\star(i)}$ and $\boldsymbol{\Sigma}_\star$ are the marginal expectation and covariance matrix of $\boldsymbol{y}_{\star(i)}$, respectively. In particular,

$$\boldsymbol{\mu}_{\star(i)}^T = \boldsymbol{f}_0(\boldsymbol{x}_{(i)})\boldsymbol{V}_0^T + \left[\boldsymbol{f}_1(\boldsymbol{x}_{(i)})\boldsymbol{V}_1^T, \cdots, \boldsymbol{f}_K(\boldsymbol{x}_{(i)})\boldsymbol{V}_K^T\right],$$

where $\boldsymbol{x}_{(i)}$ is a column vector corresponding to the ith row of $\boldsymbol{X}$, and $\boldsymbol{f}_k(\boldsymbol{x}_{(i)})$ is a row vector of length r_k $(k = 0, 1, \cdots, K)$. The grand covariance matrix $\boldsymbol{\Sigma}_\star$ has the form

$$\boldsymbol{\Sigma}_\star = \boldsymbol{V}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^T + \boldsymbol{V}_\star\boldsymbol{\Sigma}_{\boldsymbol{F}}\boldsymbol{V}_\star^T + \boldsymbol{\Sigma}_{\boldsymbol{E}},$$

where$\boldsymbol{\Sigma}_{\boldsymbol{F}} = \text{blkdiag}(\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K)$ and $\boldsymbol{\Sigma}_{\boldsymbol{E}} = \text{blkdiag}(\sigma_1^2\boldsymbol{I}_{r_1}, \cdots, \sigma_k^2\boldsymbol{I}_{r_K})$. The optimization of the above log likelihood function under the identifiability conditions is computationally prohibitive because the likelihood function involves unknown nonparametric functions and the conditions are non-convex.

To circumvent the computational issue, we resort to the hierarchical form of the SIFA model in (1)–(5) and treat $\boldsymbol{U}_0$ and $\boldsymbol{U}_\star$ as latent variables, and derive an *Expectation–Maximization* (EM) algorithm. Specifically, in the E step, we calculate the conditional distribution of the latent variables $(\boldsymbol{U}_0, \boldsymbol{U}_\star)$ given the data $\boldsymbol{Y}_\star$ and the previously estimated model parameters. In the M step, we maximize the conditional expectation of the joint log likelihood of the latent variables and the data. The joint log likelihood is partitioned into the log likelihood of $(\boldsymbol{U}_0, \boldsymbol{U}_\star)$ and the conditional log likelihood of $\boldsymbol{Y}_\star$ given $(\boldsymbol{U}_0, \boldsymbol{U}_\star)$. Furthermore, since the latent variables $\boldsymbol{U}_0, \boldsymbol{U}_1, \cdots, \boldsymbol{U}_K$ are mutually independent, the log likelihood of $(\boldsymbol{U}_0, \boldsymbol{U}_\star)$ is further partitioned. Consequently, the M step is to solve the following problems under the respective identifiability conditions:

$$\max_{\boldsymbol{f}_k(\cdot), \boldsymbol{\Sigma}_k} \quad \mathbb{E}_{\boldsymbol{U}_k|\boldsymbol{Y}_\star} \mathcal{L}(\boldsymbol{U}_k), \quad k = 0, 1, \cdots, K, \quad (8)$$

$$\max_{\boldsymbol{V}_0, \boldsymbol{V}_\star, \sigma_1^2, \cdots, \sigma_K^2} \quad \mathbb{E}_{\boldsymbol{U}_0, \boldsymbol{U}_\star|\boldsymbol{Y}_\star} \mathcal{L}(\boldsymbol{Y}_\star|\boldsymbol{U}_0, \boldsymbol{U}_\star), \quad (9)$$

where $\mathbb{E}_{\boldsymbol{U}_0, \boldsymbol{U}_\star|\boldsymbol{Y}_\star}(\cdot)$ represents the conditional expectation with respect to $(\boldsymbol{U}_0, \boldsymbol{U}_\star)$. For simplicity, hereafter we will use $\mathbb{E}(\cdot)$ to denote the conditional expectations. Below we shall outline the key steps of the M step. More details can be found in the supplementary material.

It can be shown that in (8) each entry of the vector-valued function $\boldsymbol{f}_k(\cdot) = (f_{k,1}(\cdot), \cdots, f_{k,r_k}(\cdot))$ can be separately estimated via solving a least square problem

$$\widehat{f_{k,j}(\cdot)} = \arg\min_{f_{k,j}(\cdot)} \|\mathbb{E}(\boldsymbol{u}_{k,j}) - f_{k,j}(\boldsymbol{X})\|_{\mathbb{F}}^2, \quad j = 1, \cdots, r_k; \; k = 0, 1, \cdots, K, \quad (10)$$

where $\boldsymbol{u}_{k,j}$ is the jth column of $\boldsymbol{U}_k$, and $\|\cdot\|_{\mathbb{F}}$ denotes the Frobenius norm. If $f_{k,j}(\cdot)$ is a parametric function, the above problem can be solved via a Newton–Raphson method. In particular, if linear, it is explicitly solved via the ordinary least squares. If $f_{k,j}(\cdot)$ is nonparametric, the problem becomes nonparametric regression. Standard kernel methods and spline-based methods can be readily applied here (cf. Fan and Gijbels, 1996; Hollander, Wolfe and Chicken, 2013). When the dimension of the covariates is high, we can assume $f_{k,j}(\cdot)$ to be an additive model and easily incorporate variable selection through penalized methods (Tibshirani, 1996; Ravikumar et al., 2009). To sum up, regardless of the forms of the functions, $\{\boldsymbol{f}_k(\cdot)\}_{k=0,\cdots,K}$ can be easily estimated using existing methods.

Subsequently, it is easy to obtain the closed-form optimizer of (8) with respect to $\boldsymbol{\Sigma}_k$ as:

$$\widehat{\boldsymbol{\Sigma}_k} = \frac{1}{n}\mathrm{diag}\left\{\mathbb{E}\left[\left(\boldsymbol{U}_k - \widehat{\boldsymbol{f}}_k(\boldsymbol{X})\right)^T\left(\boldsymbol{U}_k - \widehat{\boldsymbol{f}}_k(\boldsymbol{X})\right)\right]\right\},$$
$$k = 0, 1, \cdots, K,$$

where $\mathrm{diag}(\boldsymbol{S})$ is the diagonal matrix consisting of the diagonal values of $\boldsymbol{S}$, and $\widehat{\boldsymbol{f}}_k(\cdot)$'s are the estimated covariate functions.

From (9), we obtain the estimates of the loading matrices and the noise variances under different identifiability conditions.

Under the general conditions A1 and A2, there are no explicit solutions of (9) for $\boldsymbol{V}_0$ and $\boldsymbol{V}_\star$. We propose to iteratively update the estimates of the loading matrices in a block-wise coordinate descent fashion. In particular, we cycle through the following steps: given $\boldsymbol{V}_0$, update $\boldsymbol{V}_k$'s in parallel via solving

$$\min_{\boldsymbol{V}_k:\boldsymbol{V}_k^T\boldsymbol{V}_k=\boldsymbol{I}} \quad \mathbb{E}\|\boldsymbol{Y}_k - \boldsymbol{U}_0\boldsymbol{V}_{0,k}^T - \boldsymbol{U}_k\boldsymbol{V}_k^T\|_{\mathbb{F}}^2; \tag{11}$$

and given $\boldsymbol{V}_\star$, update $\boldsymbol{V}_0$ by solving

$$\min_{\boldsymbol{V}_0:\boldsymbol{V}_0^T\boldsymbol{V}_0=\boldsymbol{I}} \quad \sum_{k=1}^{K}\sigma_k^{-2}\mathbb{E}\|\boldsymbol{Y}_k - \boldsymbol{U}_k\boldsymbol{V}_k^T - \boldsymbol{U}_0\boldsymbol{V}_{0,k}^T\|_{\mathbb{F}}^2. \tag{12}$$

It can be shown that the optimization problem (11) has a closed-form solution $\widehat{\boldsymbol{V}_k} = \boldsymbol{L}\boldsymbol{R}^T$, where $\boldsymbol{L}$ and $\boldsymbol{R}$ contain the left and right singular vectors of $\boldsymbol{Y}_k^T\mathbb{E}(\boldsymbol{U}_k) - \boldsymbol{V}_{0,k}\mathbb{E}(\boldsymbol{U}_0^T\boldsymbol{U}_k)$. The optimization (12) does not have an analytical solution due to the possibly different σ_k^2's. As a remedy, we relax the orthogonality constraint in (12) temporarily, and obtain an intermediate estimator of $\boldsymbol{V}_{0,k}$ as

$$\widetilde{\boldsymbol{V}_{0,k}} = \left[\boldsymbol{Y}_k^T\mathbb{E}(\boldsymbol{U}_0) - \boldsymbol{V}_k\mathbb{E}(\boldsymbol{U}_k^T\boldsymbol{U}_0)\right]\left[\mathbb{E}(\boldsymbol{U}_0^T\boldsymbol{U}_0)\right]^{-1}.$$

To impose the orthogonality constraint, the final estimator of $\boldsymbol{V}_0$ is the eigenvectors of $\widetilde{\boldsymbol{V}_0}\widehat{\boldsymbol{\Sigma}_0}\widetilde{\boldsymbol{V}_0}^T$. Correspondingly, we update the diagonal values of $\widehat{\boldsymbol{\Sigma}_0}$ to be the eigenvalues of $\widetilde{\boldsymbol{V}_0}\widehat{\boldsymbol{\Sigma}_0}\widetilde{\boldsymbol{V}_0}^T$. This additional standardization step ensures that $\boldsymbol{\Sigma}_\star$ in the likelihood function (7) remains unchanged. A similar approach was used in Li et al. (2016). As a result, the loading matrices are estimated under the general conditions. We remark that in practice, a one-step update in each EM iteration is usually accurate enough and there is no need to iterate.

Under the orthogonal conditions B1 and B2, the computation can be greatly simplified. The loading matrices $\boldsymbol{V}_{0,k}$ and $\boldsymbol{V}_k$ can be estimated simultaneously and explicitly. By writing $\boldsymbol{W}_k = (\sqrt{K}\boldsymbol{V}_{0,k}, \boldsymbol{V}_k)$, the optimization (9) is equivalent to

$$\min_{\boldsymbol{W}_k:\boldsymbol{W}_k^T\boldsymbol{W}_k=\boldsymbol{I}} \left\|\boldsymbol{Y}_k - \left(\frac{1}{\sqrt{K}}\mathbb{E}(\boldsymbol{U}_0), \mathbb{E}(\boldsymbol{U}_k)\right)\boldsymbol{W}_k^T\right\|_{\mathbb{F}}^2,$$

which is exactly an orthogonal Procrustes problem (Gower and Dijksterhuis, 2004). The optimizer has the explicit expression as $\widehat{\boldsymbol{W}_k} = (\sqrt{K}\widehat{\boldsymbol{V}_{0,k}}, \widehat{\boldsymbol{V}_k}) = \boldsymbol{L}\boldsymbol{R}^T$, where $\boldsymbol{L}$ and $\boldsymbol{R}$ contain the left and right singular vectors of $\boldsymbol{Y}_k^T\left(1/\sqrt{K}\mathbb{E}(\boldsymbol{U}_0), \mathbb{E}(\boldsymbol{U}_k)\right)$. Subsequently, it is easy to decouple $\widehat{\boldsymbol{V}_{0,k}}$ and $\widehat{\boldsymbol{V}_k}$, and obtain closed-form estimators for different loading matrices.

Once the loading matrices are estimated, solving (9) with respect to σ_k^2's, we obtain the closed-form optimizers as:

$$\widehat{\sigma_k^2} = \frac{1}{np_k}\mathbb{E}\|\boldsymbol{Y}_k - \boldsymbol{U}_k\widehat{\boldsymbol{V}_k}^T - \boldsymbol{U}_0\widehat{\boldsymbol{V}_{0,k}}^T\|_{\mathbb{F}}^2, \; k = 1, \cdots, K.$$

A step-by-step description of the algorithm can be found in the supplementary material.

2.4. *Rank Selection*

Up to now, we assume the ranks for the joint and individual structures in the SIFA model are known. In practice, we often need to estimate the ranks from data. The choice of the ranks is crucial for parameter estimation and model interpretation. In total, there are $K+1$ ranks to be determined. Here, we propose a two-step procedure to get a crude estimate of the ranks, and an optional likelihood cross validation (LCV) method for refining the estimate.

Since Model (6) can be viewed as a special form of a signal-plus-noise model, a natural first step is to estimate the rank of the underlying signal of each data set $\boldsymbol{Y}_k$ (denoted as $r_k^\star$) and the rank of the underlying signal of the combined data set $\boldsymbol{Y}_\star$ (denoted as $r_{\text{total}}^\star$), respectively. There are many existing methods to this purpose, such as the scree plot, the total variance explained criterion, hypothesis testing methods. Users can choose their favorite methods. Once estimated, we use $r_k^\star$ and $r_{\text{total}}^\star$ to calculate the ranks for different structures in Model (6). Under either set of identifiability conditions, the following equations hold:

$$r_{\text{total}}^\star = r_0 + \sum_{k=1}^{K} r_k, \quad r_k^\star = r_0 + r_k,$$

for $k = 1, \cdots, K$, where $r_0, r_1, \cdots, r_K$ are the joint and individual ranks for the SIFA model. Solving the equation system, we get

$$r_0 = \frac{\sum_{k=1}^{K} r_k^\star - r_{\text{total}}^\star}{K-1}, \quad r_k = r_k^\star - r_0,$$

which serve as good initial estimators of the ranks. Numerically, the estimate of r_0 may be non-integer or even negative when $K > 2$. In that case, we suggest rounding the estimate to the nearest non-negative integer. Then we plug it into the second equation to get an estimate of r_k. If the estimate of r_k is negative, it can be set to 0.

The above two-step procedure typically provides a good initial estimate of the ranks. If it is desired to further refine the rank estimation, one may exploit a more computationally intensive N-fold LCV approach. The idea is to randomly split the samples into N groups across different data sets. In each run, we withhold one group as the testing set and use the remaining $N - 1$ groups as the training set to fit Model (6) with different sets of ranks. For each set of ranks, the corresponding LCV score is the value of negative log likelihood, evaluated using (7) on the testing data. We repeat the procedure N times, and choose the set of ranks corresponding to the smallest average LCV score. A more detailed description can be found in the supplementary material.

3. Simulation Studies

In this section, we conduct comprehensive simulation studies to demonstrate the advantage of the proposed methods. We compare SIFA (under both sets of identifiability conditions) with JIVE (the original version and a covariate-adjusted version, denoted by cov-JIVE), SupSVD, and PCA. For cov-JIVE, we first regress different data sets on the covariates, and then apply JIVE to the residuals.

3.1. *Simulation Settings*

We consider two primary data sets $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ (i.e., $K = 2$) on the same set of samples with sample size $n = 500$, and dimension $p_1 = p_2 = 200$. The data are simulated from Model (6) with different parameters. We first consider three settings where, loosely speaking, the generative models are JIVE, SIFA under the general conditions (denoted as SIFA-A), and SIFA under the orthogonal conditions (denoted as SIFA-B). In particular, the SIFA-A and SIFA-B models employ linear models between covariates (with dimension $q = 10$) and latent factors. The true ranks of the joint and individual patterns are $r_0 = 2$, $r_1 = r_2 = 3$. Some important features of these settings are described below.

- **Setting 1** (JIVE Model): For $k = 0, 1, 2$, the factors in $\boldsymbol{U}_k$ are randomly generated and mutually independent (with $\boldsymbol{f}_k(\cdot) = \boldsymbol{0}$ in (6)); the loadings in $\boldsymbol{V}_k$ and the covariance $\boldsymbol{\Sigma}_k$ satisfy the basic conditions. The measurement errors in $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are i.i.d. with different variances. To test the robustness of the proposed method, we randomly generate 10 covariates unrelated with the factors, and incorporate them in the SIFA estimation.
- **Setting 2** (SIFA-A Model): The joint and individual factors are generated from the linear model $\boldsymbol{U}_k = \boldsymbol{X}\boldsymbol{B}_k + \boldsymbol{F}_k$ for $k = 0, 1, 2$. The loadings in $\boldsymbol{V}_0, \boldsymbol{V}_1$, and $\boldsymbol{V}_2$ are filled with random numbers and standardized to satisfy the general conditions. To make them further deviate from the orthogonal conditions, we intentionally choose $\boldsymbol{V}_{0,k}$ not orthogonal to $\boldsymbol{V}_k$ $(k = 1, 2)$, and artificially vary the norm of each column in $\boldsymbol{V}_{0,1}$ and $\boldsymbol{V}_{0,2}$.
- **Setting 3** (SIFA-B Model): The factors are generated in the same way as in Setting 2. The true loadings are specifically normalized to satisfy the orthogonal conditions. We note that the SIFA-B model is a special case of the SIFA-A model.

For each simulation setting, we fit JIVE, cov-JIVE, SIFA-A, and SIFA-B to the multiple simulated data sets, and fit PCA and SupSVD to the concatenated data $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$. We incorporate covariates for cov-JIVE, SIFA-A, SIFA-B, and SupSVD. In particular, when fitting the SIFA models, we set the functions in (10) to be linear, and use the ordinary least squares to estimate the coefficients. To avoid ambiguity, these model models are fitted with the true ranks. We set the rank for PCA and SupSVD to be $r_0 + r_1 + r_2$. We assess the performance of the LCV for rank selection in the next section.

To compare the loading estimation in JIVE, cov-JIVE, SIFA-A, and SIFA-B, we use the Grassmannian metric (Mattila, 1999) between the true loadings in $\boldsymbol{V}_k$ and the estimated loadings in $\widehat{\boldsymbol{V}}_k$ for each $k = 0, 1, 2$. The metric is defined as $d_{\mathcal{G}}(\boldsymbol{V}_k, \widehat{\boldsymbol{V}}_k) = \sqrt{\sum_{i=1}^{r_k} \mathrm{acos}(\delta_i)^2}$, where δ_i is the ith singular value of $\boldsymbol{V}_k^T \widehat{\boldsymbol{V}}_k$. We also evaluate the maximal principal angle $\measuredangle(\boldsymbol{V}, \widehat{\boldsymbol{V}})$ (Björck and Golub, 1973) between the subspaces in $\mathbb{R}^{p_1+p_2}$ spanned by the true loading vectors in $\boldsymbol{V} = (\boldsymbol{V}_0, \mathrm{blkdiag}(\boldsymbol{V}_1, \boldsymbol{V}_2))$ and the estimated ones, across all methods. To evaluate the accuracy of the estimated low-rank structure, we use $\|\boldsymbol{U}\boldsymbol{V}^T - \widehat{\boldsymbol{U}}\widehat{\boldsymbol{V}}^T\|_{\mathbb{F}}$ where $\boldsymbol{U} = (\boldsymbol{U}_0, \boldsymbol{U}_1, \boldsymbol{U}_2)$. The matrix $\widehat{\boldsymbol{U}}$ represents the estimated score matrix for PCA and JIVE, or the conditional expectation of the latent factor matrix for SupSVD, SIFA-A, and SIFA-B.

We also conduct comprehensive simulation studies to investigate: 1) the goodness of fit when the relations between covariates and latent factors are nonlinear; 2) the overfitting issue when nonparametric functions are used in the presence of linear relations; 3) the rank misspecification effect on the performance; 4) the violation of the Gaussian assumption; 5) the effect of rescaling different data sets; 6) the scalability of SIFA-A and SIFA-B in high dimension. The simulation settings and results are contained in the supplementary material.

3.2. *Rank Estimation by LCV*

We briefly demonstrate the efficacy of the LCV method using a simulated example. Data are generated under Setting 3, with the chosen true ranks to be $(r_0, r_1, r_2) = (2, 3, 3)$. Additional examples under Settings 1 and 2 are provided in the supplementary material. We particularly consider nine candidate rank sets in the neighborhood of the true rank set: $(r_0, r_1, r_2) \in \{(1, 2, 2), (2, 2, 2), (3, 2, 2), (1, 3, 3), (2, 3, 3), (3, 3, 3), (3, 4, 3), (3, 4, 4), (4, 4, 4)\}$. We conduct a 10-fold LCV. The evaluated LCV scores (i.e., the negative log likelihood values of test samples) for different candidate sets in each cross validation run are shown in Figure 1. The average score reaches the minimum at the true rank set. Namely, the LCV method correctly selects the true ranks.

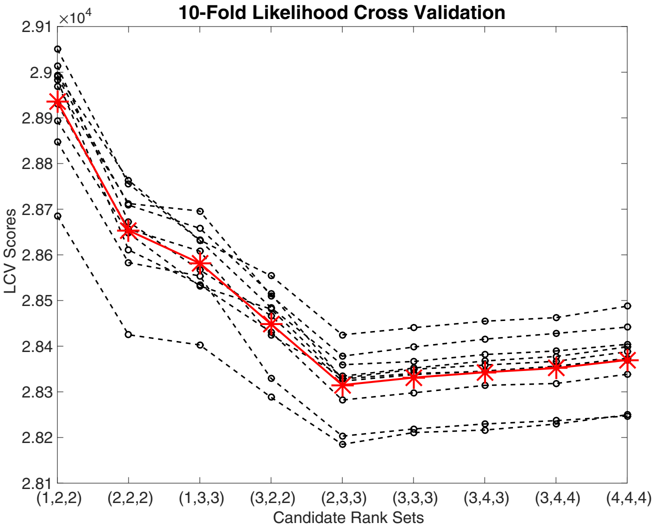

Figure 1. The LCV scores for 10-fold cross validation on nine candidate rank sets. Each dashed line with circles contains corresponds to the LCV scores (negative log likelihood values) in one cross validation run. The solid line with stars contains the average LCV scores for different rank sets.

3.3. *Simulation Results*

For each setting, we repeat the simulation 100 times and summarize the results. The results are summarized in Table 1. In Setting 1, both SIFA-A and SIFA-B perform similarly to JIVE in terms of the loading estimation, even if the generative model is JIVE (i.e., the covariates are unrelated to the factors). Remarkably, the SIFA methods provide the best low-rank structure recovery accuracy among all. The reason is similar to the argument in Li et al. (2016): the shrinkage effect of the conditional expectation of the factors reduces estimation variance. In Setting 2, SIFA-A provides the best performance in all criteria. SIFA-B is suboptimal because the orthogonal conditions are severely violated. In Setting 3, the SIFA-B method performs the best, followed closely by SIFA-A. Both are significantly better than the competing methods. In practice, when the orthogonal conditions are well justified, SIFA-B is favorable due to the ultra-fast computation and accurate estimation. Otherwise, we recommend the use of the SIFA-A method.

4. GTEx Data Analysis

In this section, we apply the proposed method to the multi-tissue genetic data from the GTEx project. We use the phs000424.v6 data which are available at http://www.gtexportal.org/ (registration required for data access). Technical details of data preprocessing and rank estimation can be found in the supplementary material.

The GTEx project collects gene expression data from multiple tissues (e.g., muscle, blood, skin) from the same set of subjects. We use the SIFA method to identify cross-tissue and tissue-specific gene expression patterns, and quantify the heritability of phenotypes representing expressions of a group of genes. Addressing the questions is integral to the fulfillment of the GTEx goal (The GTEx Consortium, 2015).

We particularly focus on the p53 signaling pathway in three tissues, that is, muscle, blood, and skin, for the illustration purpose. The analysis can be easily generalized to other gene sets or tissues. After proper preprocessing and normalization, we obtain 191 genes on 204 common samples in each tissue, denoted by $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3$. Each gene expression is standardized. In addition, we have the auxiliary data of sex, genotyping platform index, and genetic variants for each sample as covariates. To reduce the dimensionality of the genetic variants, we obtain the top 30 principal components, which capture the

Table 1

Simulation results under Setting 1, 2, and 3 (each with 100 simulation runs). The mean and standard deviation of each criterion for each method are shown in the table. The best results are highlighted in bold.

		SIFA-A	SIFA-B	JIVE	cov-JIVE	SupSVD	PCA
Setting 1	$d_G(\boldsymbol{V}_0, \widehat{\boldsymbol{V}}_0)$	0.61(0.03)	**0.60**(0.03)	0.69(0.07)	0.68(0.06)		
(JIVE)	$d_G(\boldsymbol{V}_1, \widehat{\boldsymbol{V}}_1)$	0.82(0.06)	**0.81**(0.06)	0.91(0.17)	0.89(0.16)		
	$d_G(\boldsymbol{V}_2, \widehat{\boldsymbol{V}}_2)$	1.32(0.17)	1.33(0.17)	1.31(0.18)	**1.30**(0.17)		
	$\measuredangle(\boldsymbol{V}, \widehat{\boldsymbol{V}})$	64.73(10.98)	65.33(10.96)	65.16(11.31)	**64.35**(10.98)	87.02(2.74)	86.76(2.77)
	$\Vert\boldsymbol{U}\boldsymbol{V}^T - \widehat{\boldsymbol{U}}\widehat{\boldsymbol{V}}^T\Vert_{\mathbb{F}}$	193.28(2.85)	**193.26**(2.73)	240.49(4.53)	287.86(3.74)	239.21(2.82)	279.95(3.05)
Setting 2	$d_G(\boldsymbol{V}_0, \widehat{\boldsymbol{V}}_0)$	**0.37**(0.02)	1.01(0.05)	0.76(0.02)	1.40(0.14)		
(SIFA-A)	$d_G(\boldsymbol{V}_1, \widehat{\boldsymbol{V}}_1)$	**0.27**(0.01)	1.06(0.08)	0.28(0.01)	1.41(0.03)		
	$d_G(\boldsymbol{V}_2, \widehat{\boldsymbol{V}}_2)$	**0.52**(0.02)	0.67(0.03)	0.62(0.04)	1.80(0.09)		
	$\measuredangle(\boldsymbol{V}, \widehat{\boldsymbol{V}})$	**27.67**(1.42)	44.46(2.42)	33.34(2.27)	88.30(1.43)	39.40(2.30)	46.97(4.28)
	$\Vert\boldsymbol{U}\boldsymbol{V}^T - \widehat{\boldsymbol{U}}\widehat{\boldsymbol{V}}^T\Vert_{\mathbb{F}}$	**169.21**(1.73)	207.74(2.03)	207.22(2.33)	296.14(2.11)	200.77(2.30)	235.44(3.00)
Setting 3	$d_G(\boldsymbol{V}_0, \widehat{\boldsymbol{V}}_0)$	0.38(0.03)	**0.30**(0.01)	0.61(0.02)	0.65(0.06)		
(SIFA-B)	$d_G(\boldsymbol{V}_1, \widehat{\boldsymbol{V}}_1)$	0.25(0.01)	**0.24**(0.01)	0.25(0.01)	1.16(0.21)		
	$d_G(\boldsymbol{V}_2, \widehat{\boldsymbol{V}}_2)$	0.35(0.01)	**0.34**(0.01)	0.36(0.01)	1.77(0.07)		
	$\measuredangle(\boldsymbol{V}, \widehat{\boldsymbol{V}})$	15.03(0.56)	**14.97**(0.57)	15.86(0.73)	85.93(2.83)	26.21(1.09)	27.71(1.30)
	$\Vert\boldsymbol{U}\boldsymbol{V}^T - \widehat{\boldsymbol{U}}\widehat{\boldsymbol{V}}^T\Vert_{\mathbb{F}}$	171.99(1.65)	**171.51**(1.66)	204.64(1.97)	290.82(3.33)	200.77(1.81)	230.80(2.05)

majority of variation in the genotype data. The covariates are denoted by $\boldsymbol{X}$.

We first estimate the ranks for the joint and individual patterns. We use the two-step procedure described in Section 2.4, and exploit a variance explained criterion in the first step (with a preset 90% threshold). The joint and individual ranks are estimated to be $r_0 = 26$, $r_1 = 24$, $r_2 = 5$, and $r_3 = 20$. Note that the individual rank for blood ($r_2 = 5$) is much smaller than that for muscle or skin. From the viewpoint of the expression pattern richness, blood is very different from the other two tissues. This is generally concordant with the previous discoveries (The GTEx Consortium, 2015).

We fit a SIFA-B model to the data with linear relations between the covariates and the latent factors. For comparison, we also fit a JIVE model with the same ranks. The estimated joint and individual patterns are shown in Figure 2 (for the SIFA-B model) and Figure 3 (for the JIVE model). By taking into account the auxiliary covariates, the patterns obtained by the SIFA-B model are more discernable than those from the JIVE model. The joint structure in Figure 2 clearly captures the shared patterns among samples across tissues, while the individual structure distinguishes different tissues. We also quantify the variation explained by different parts in both methods (see Figure 4). The SIFA-B decomposition attributes more variation to the individual structure than the JIVE method, which is consistent with the domain knowledge that the p53 gene expressions are highly tissue specific (Ribeiro et al., 2001; Tendler et al., 1999). The tissue-specific expression patterns may be used to investigate tissue identity and functions.

To quantify the heritability of the derived phenotypes (i.e., joint and individual scores) representing the p53 gene expressions, we calculate the variation explained by different components of the SIFA model. The results are summarized in Table 2. Within the joint structure (common across all tissues), the genetic variants explain about 17% of the variation, which is concordant with the general belief in the literature (Brown et al., 2015). The sex and the platform information take up 2 and 2.5% of the variation, respectively. The vast majority of the variation remains unexplained, which provokes further investigation. The individual structure for each tissue has a similar decomposition to the joint structure. An interesting finding is that sex is not a major contributor to the individual gene expression patterns in blood. The derived pathway expression phenotypes could also potentially be used to discover associations with clinical outcome and environmental factors. Due to the lack of such information in the GTEx data, we do not further pursue it here.

5. Discussion

In this article, we develop a supervised integrated factor analysis framework for reduction and integration of multi-view data. It decomposes multiple related data sets into

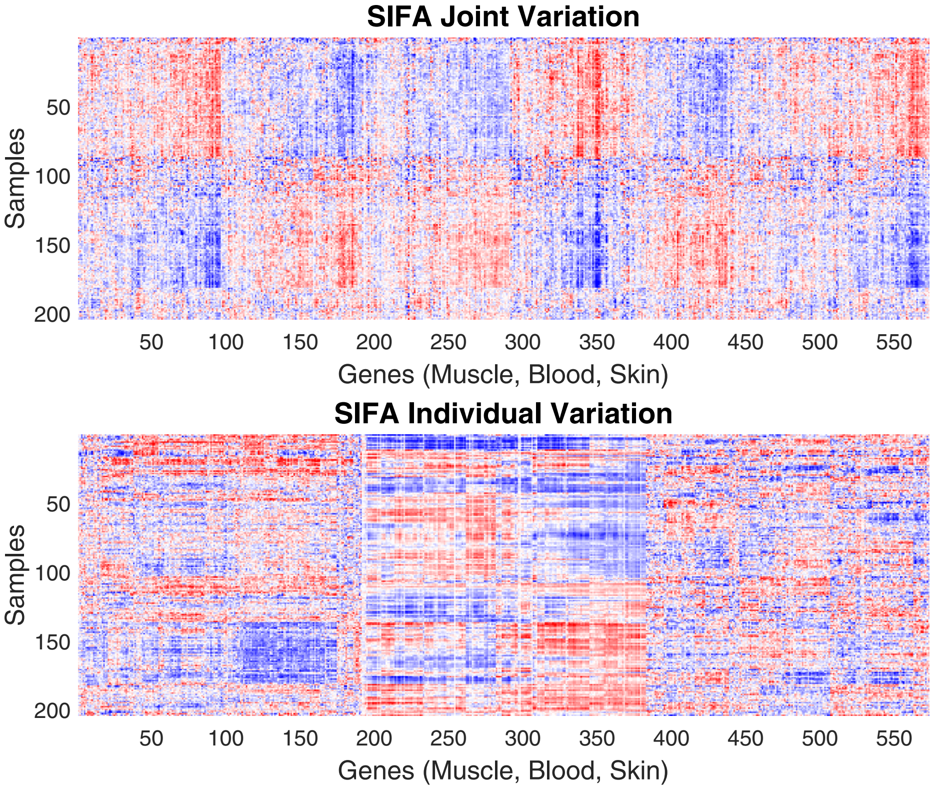

Figure 2. GTEx example: the heat maps of the joint and individual gene expression patterns for the p53 signaling pathway identified by the SIFA-B model. For visualization purpose, we reorder samples across three tissues and genes in each tissue. Top panel: the joint structure in three tissues; Bottom panel: the individual structures in three tissues. In each panel, the samples are matched across tissues.

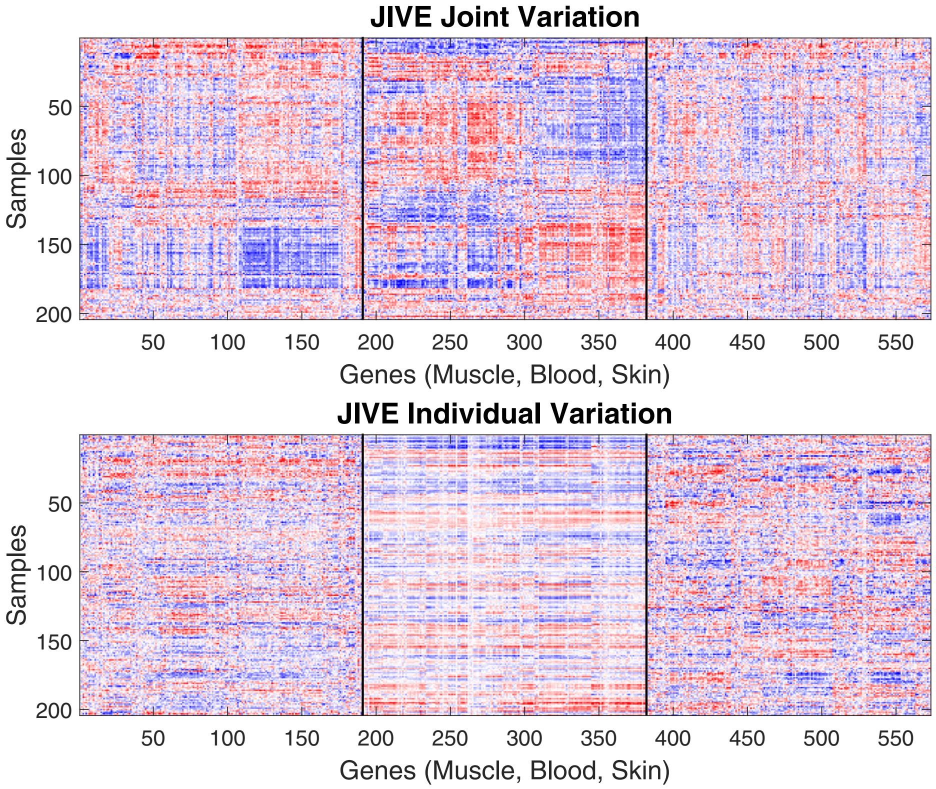

Figure 3. GTEx example: the heat maps of the joint and individual gene expression patterns for the p53 signaling pathway identified by the JIVE model. The rows and columns are ordered in the same way as in Figure 2.

joint and individual structures, while incorporating covariate supervision through parametric or nonparametric models. We investigate the identifiability of the model under two sets of conditions, the general conditions and the orthogonal conditions, each being useful in separate situations. An efficient EM algorithm with some variants is devised to fit the model. In particular, it is very easy to capture nonlinear relations between covariates and latent factors, and to incorporate variable selection of covariates. The comprehensive simulation studies demonstrate the efficacy of the proposed methods. With application to the GTEx data, we provide new insights into the genetic variation of a gene set across multiple tissues.

There are several directions for future research. First of all, it is of potential interest to generalize the current framework to accommodate non-normal data. Second, the model may

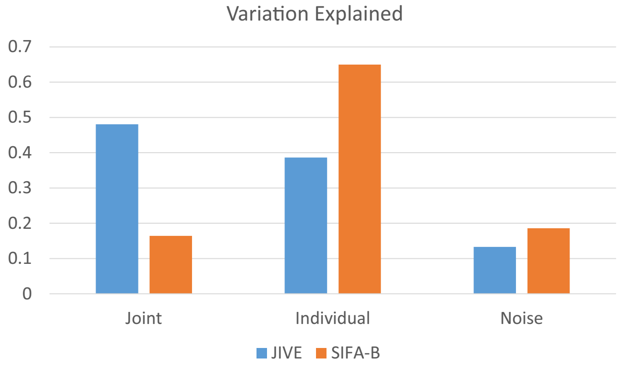

Figure 4. GTEx example: the variation explained by different components in the JIVE model and the SIFA-B model, respectively.

Table 2

GTEx example: the genetic variation explained by different factors in different tissues. For each tissue, the last column gives the percentage explained by the joint and individual structure, and the noise (add up to 1). The variation in the joint (individual) structure is further attributed to the genotype, sex, platform, and other unknown sources (add up to 1).

	Genotype	Sex	Platform	Unknown	Total
Muscle					
Joint	17.09%	2.03%	2.56%	78.32%	16.44%
Individual	15.55%	2.66%	1.74%	80.05%	65.29%
Noise					18.27%
Blood					
Joint	17.09%	2.03%	2.56%	78.32%	16.44%
Individual	14.05%	0.65%	0.90%	84.39%	63.56%
Noise					20.00%
Skin					
Joint	17.09%	2.03%	2.56%	78.32%	16.44%
Individual	16.55%	1.52%	1.04%	80.89%	66.06%
Noise					17.50%

be specially modified to capture partially joint structure pertaining to multiple but not all data sets. This is especially relevant when multiple data sets are naturally grouped at the source level. Third, customized rank estimation methods need further investigation.

6. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2, 3, and 4 are available with this article at the *Biometrics* website on Wiley Online Library. Matlab code implementing the proposed methods is available at https://github.com/reagan0323/SIFA.

Acknowledgment

We would like to thank the Editor, the Associate Editor, and two reviewers for the insightful comments. Gen Li was partially supported by the Calderone Junior Faculty Award by the Mailman School of Public Health at Columbia University. Sungkyu Jung was partially supported by the National Science Foundation grant DMS-1307178.

References

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: Principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics* **5**, 149–179.

Ahn, J. and Marron, J. (2010). The maximal data piling direction for discrimination. *Biometrika* **97**, 254–259.

Björck, K. and Golub, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of Computation* **27**, 579–594.

Brown, A., Ding, Z., Viñuela, A., Glass, D., Parts, L., Spector, T., et al. (2015). Pathway based factor analysis of gene expression data produces highly heritable phenotypes that associate with age. *G3: Genes— Genomes— Genetics* g3–114.

Fan, J. and Gijbels, I. (1996). Local polynomial modelling and its applications: Monographs on statistics and applied probability, **66**. Boca Raton, Florida: CRC Press.

Fan, J., Liao, Y., and Wang, W. (2016). Projected principal component analysis in factor models. *Annals of Statistics* **44**, 219–254.

Gower, J. C. and Dijksterhuis, G. B. (2004). Procrustes problems, **3**. Oxford, UK: Oxford University Press.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). Nonparametric statistical methods. Hoboken, NJ: John Wiley & Sons.

Li, G., Yang, D., Nobel, A. B., and Shen, H. (2016). Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis* **146**, 7–17.

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**, 523–542.

Löfstedt, T., Hoffman, D., and Trygg, J. (2013). Global, local and unique decompositions in onpls for multiblock data analysis. *Analytica Chimica Acta* **791**, 13–24.

Mattila, P. (1999). Geometry of sets and measures in Euclidean spaces: fractals and rectifiability, **44**. Cambridge, UK: Cambridge University Press.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B* **71**, 1009–1030.

Ray, P., Zheng, L., Lucas, J., and Carin, L. (2014). Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics* **30**, 1370–1376.

Ribeiro, R. C., Sandrini, F., Figueiredo, B., Zambetti, G. P., Edson Michalkiewicz, E., Lafferty, A. R., et al. (2001). An inherited p53 mutation that contributes in a tissue-specific manner to pediatric adrenal cortical carcinoma. *Proceedings of the National Academy of Sciences* **98**, 9330–9335.

Schouteden, M., Van Deun, K., Wilderjans, T. F., and Van Mechelen, I. (2014). Performing disco-sca to search for distinctive and common information in linked data. *Behavior Research Methods* **46**, 576–587.

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.

Tendler, Y., Weisinger, G., Coleman, R., Diamond, E., Lischinsky, S., Kerner, H., et al. (1999). Tissue-specific p53 expression in the nervous system. *Molecular Brain Research* **72**, 40–46.

The GTEx Consortium (2015). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58**, 267–288.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B* **61**, 611–622.

Tseng, G. C., Ghosh, D., and Zhou, X. J. (2015). Integrating Omics Data. New York, NY: Cambridge University Press.

Yang, Z. and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **32**, 1–8.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. P. (2016). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems* **27**, 2426–2439.

Received May 2016. *Revised February* 2017.
Accepted March 2017.