Covariate-driven factorization by thresholding for multi-block data

by

Xing Gao

B.S. in Economics, Ohio State University, 2012

M.A. in Applied Statistics, University of Pittsburgh, 2015

Submitted to the Graduate Faculty of
the Dietrich School of Arts and Sciences in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Xing Gao

It was defended on

March 27th 2020

and approved by

Satish Iyengar, Department of Statistics, University of Pittsburgh

Sungkyu Jung, Department of Statistics, Seoul National University

Gen Li, Department of Biostatistics, Columbia University

Yu Cheng, Department of Statistics, University of Pittsburgh

Zhao Ren, Department of Statistics, University of Pittsburgh

Dissertation Director: Satish Iyengar, Department of Statistics, University of Pittsburgh

Covariate-driven factorization by thresholding for multi-block data

Xing Gao, PhD

University of Pittsburgh, 2020

Multi-block data, where multiple groups of variables from different sources are observed for a common set of subjects, are routinely collected in many areas of science. Methods for joint factorization of such multi-block data are being developed to explore the potentially joint variation structure of the data. While most of the existing work focuses on delineating joint components, shared across all data blocks, from individual components, which is only relevant to a single data block, we propose to model and estimate partially-joint components across some, but not all, data blocks. If covariates, with potential multi-block structures, are available, then the components are further modeled to be driven by the covariate information. To estimate such a covariate-driven, block-structured factor model, we propose an iterative algorithm based on thresholding, by transforming the problem of signal segmentation into a grouped variable selection problem. The proposed factorization provides accurate estimation of individual and (partially) joint structures in multi-block data, as confirmed by simulation studies. In two real multi-block data sets from genomics and image analysis, we demonstrate that the estimated block structures facilitate easy interpretation of the major factors.

Keywords: Data integration; Factorization; Individual and joint variation extraction; Multi-block data decomposition; Principal component analysis; Supervised data decomposition.

Table of contents

List of tables

List of figures

1.0 Introduction

1.1 Background and relevant work

In modern data analysis, it becomes increasingly relevant to analyze multiple sets of data observed for a common set of samples. As a motivating example, the Cancer Genome Atlas (TCGA) project [34] collects multiple aspects of genomic information, such as gene expression, methylation and copy number variation, for a common set of patients. These heterogeneous datasets, or data blocks, may possess common signals. It has been of interest to segment the common signals in multiple blocks of genomic data as a form of low-rank approximation [4, 8, 18, 21, 37, 38].

Principal Component Analysis (PCA) is a well-known method in dimension reduction to learn about data variation pattern among the variables by using a small number of components. However, PCA only applies to one single data set and seeks for a linear combination of all variables. It is not able to effectively select important variables and further to separate joint and individual variations for an integrated multi-block data set. Sparse principal component analysis (SPCA) [39] incorporates elastic net with PCA to provide sparse loadings in its modified principal components. JIVE [21] extended from PCA and was proposed as a promising framework to study the individual and joint variation patterns across multiple data blocks. However, as pointed out by [4], there are cases that JIVE cannot separate such variation patterns as it defines for multi-block data. As an improvement, Angle-based Joint and Individual Variation Explained (AJIVE) [4] was proposed to capture the full-joint and individual variation patterns by measuring the principal angles among the score subspaces of the data blocks. Additionally, COBE [37] follows JIVE alike method that factorizes each data matrix in a multi-block data set into full-joint and individual structure. It provides an iterative algorithm to extract such structures from a quadratic optimization problem, which also provides a framework for better extraction of common components that reflect full-joint structures in multi-block data when these structures are actually weak [37]. GIPCA [38] is another method that uses an iterative algorithm to study the full-joint and individual relationship among variables in high-dimensional multi-block data, which allows for different types of data sources in multi-block data.

Other than these methods, as CCA [10] can work to address linear associations among variables from two data sets, some other methods extended from CCA that aim to learn variations among multiple data sets are also developed. Kettenring [15] developed an iterative procedure to find the canonical variables across multiple data sets by optimizing some function related to the correlation matrix. Such canonical variables helps to reveal the full-joint variation structure among these variables for multi-block data. An application and extension of this multi-set CCA [15] is discussed in [3] to address the variation patterns across multi-block FMRI and EEG data. Yoon et al. [36] introduced a truncated latent Gaussian copula model and uses a sequential LASSO [31] approach to find the canonical variables with sparsity for two sets of data without assumptions on the data types.

All methods above are unsupervised learning that do not consider any relevant supervision information that my possibly drive variation patterns in the primary data matrices. To incorporate available supervision information, Supervised Integrated Factor Analysis (SIFA) [18] is proposed for the decomposition of multi-block data by involving covariates to reveal the full-joint and individual variation patterns using a small number of factors. Supervised Singular Value Decomposition (SupSVD) [19] and reduced rank regression (RRR) [13] also consider the effect of supervision information in revealing the low rank structure of the primary data. However, SupSVD and RRR do not allow for multiple collections of data from different aspects for either the main data set or the supervision data set. Sparse reduced rank regression (SRRR) [2] improves RRR by considering predictor variable selection. But in terms of identifying group variation structures, [2, 13, 19] can only capture the full-joint variation structure for a concatenated multi-block data set. For the other methods mentioned above, while they are successful at the low-rank estimation of multi-block data when there are only full-joint and/or individual variation, none of these methods considers the partial-joint variation. Because of these major drawbacks, the interpretability of the variation patterns from these methods are still circumscribed.

While most of the previous work is focused on delineating the common signal, jointly related across *all* data blocks, from block-specific, *individual* signals, there has been a growing need to model a *partially-joint* common signal across some data blocks. SLIDE [8] is among the few attempts to model a partially joint structure without utilizing any supervision information; see also [22, 23] in the chemometrics literature. Group factor analysis (GFA) [16] that extends CCA to more than two sets can also identify the structural sparsity in its factor loadings, which can be used to reveal individual and partial joint variations. In line with these efforts, in the thesis, we propose a linear factorization of multiple data blocks with factors that are either fully joint, partially joint or individual. While [8, 22, 23] consider general block-wise structures in data blocks, we in addition incorporate additional information into the factorization. The auxiliary information, or covariates, is a potential driving factor, supervising the components [*cf.* [18, 19]]. This enables decomposing factor scores into the part relevant to the covariates and the part due to unknown sources, which in turn facilitates more detailed exploration of the association patterns among data blocks and identification of major drivers for common signals. The covariates can be clinical variables such as age, sex, ethnicity and pathologic stage of subjects. The covariates also consist of data blocks recording omics profiles such as single-nucleotide polymorphism (SNP) data. That is, the supervising covariate data set can also be block-wise structured and even high-dimensional.

1.2 Toy data example

Figure 1 shows a toy data example of the data structure we consider. Denote for $\mathbf{Y}_k$ the kth data block of size $n \times p_k$. By concatenating these data blocks $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3]$, a signal, or an additive linear component, is modeled as $\mathbf{u}\mathbf{v}^T$, where $\mathbf{v}$ is the $p \times 1$ loading vector ($p = \sum_{k=1}^{3} p_k$), $\mathbf{u}$ is the collection of n scores, further decomposed by $\mathbf{u} = \mathbf{X}\beta + \mathbf{f}$, where the covariate matrix $\mathbf{X}$ contains additional information. The distinction between the fully joint, partially joint and individual components is given by the zero patterns in $\mathbf{v}^T = (\mathbf{v}_{(1)}^T, \mathbf{v}_{(2)}^T, \mathbf{v}_{(3)}^T)$. For example, if only $\mathbf{v}_{(1)}$ is non-zero, then the score $\mathbf{u}$ is only relevant to the first data block, and we call such a signal the individual component. Likewise, if both $\mathbf{v}_{(1)}$ and $\mathbf{v}_{(2)}$ are non-zero, while $\mathbf{v}_{(3)} = \mathbf{0}$, then the component is partially joint across the first two blocks.

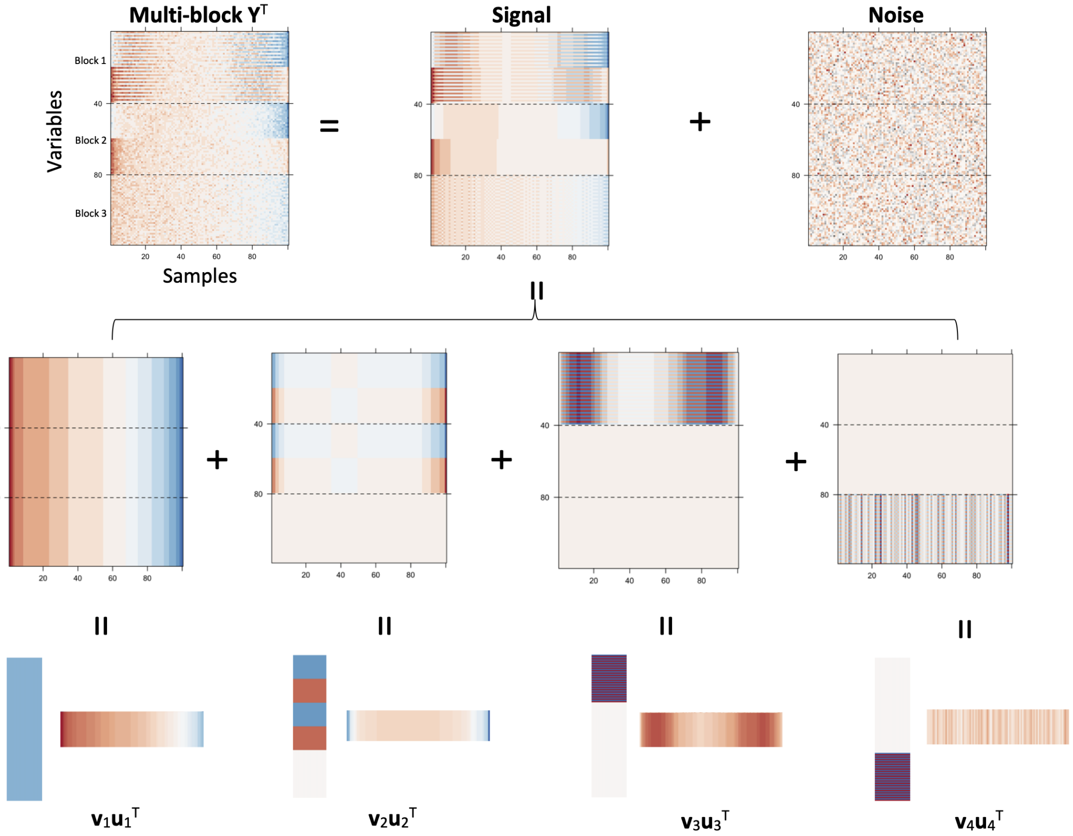


Figure 1: A toy data example. The concatenated three-block data matrix $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3]$ is decomposed into four components: $\mathbf{Y}^T = \sum_{j=1}^{4} \mathbf{v}_j \mathbf{u}_j^T + \mathbf{E}^T$, where $\mathbf{v}_1\mathbf{u}_1^T$ is a fully joint component, $\mathbf{v}_2\mathbf{u}_2^T$ is a partially joint component, $\mathbf{v}_3\mathbf{u}_3^T$ and $\mathbf{v}_4\mathbf{u}_4^T$ are two individual components.

Segmentation of common signals from potentially individual signals comes with a few challenges. First, without proper restrictions on the loading vectors and other parameters, a common signal can be exactly represented by a collection of individual signals, as previously observed in [18]. Second, introducing partially joint components sharply increases the model complexity. For K data blocks, there are $2^K - 1$ types of components, including K types of individuals, 1 fully joint and $2^K - K - 2$ partially joint components. We circumvent the model complexity issue by not specifically modeling the types of association patterns. Rather, we use a simple covariate-driven factor model and convert the problem of common signal segmentation into a (grouped) variable selection problem. The identifiability issue is then easily handled.

Note that for K data blocks, a loading vector $\mathbf{v}$ corresponding to a partially joint, or individual, component can be seen as block-wise *sparse*. If not all variables in a block is relevant, then the loadings $\mathbf{v}$ are not only block-wise sparse but also variable-wise sparse. With the above potential sparsity in mind, we devise a thresholding-based estimation scheme in the estimation of $\mathbf{v}$, while keeping $\|\mathbf{v}\| = 1$. We also allow the case where the covariate matrix $\mathbf{X}$ has several blocks of variables. Since the coefficient β in $\mathbf{u} = \mathbf{X}\beta + \mathbf{f}$ is equipped with a block structure, we regularize β to potentially have block-wise sparsity as well as variable-wise sparsity.

All in all, we propose a COvariate-driven, Block-wise Structured (COBS, for short) factorization of concatenated multi-block data, and an estimating algorithm. The inclusion of potentially-supervising covariate $\mathbf{X}$ makes the low-rank decomposition more interpretable. In particular, block-wise sparse β, together with block-wise sparse $\mathbf{v}$, identifies which covariate blocks are related to which primary data blocks. The proposed method facilitates exploration of multi-block data in revealing the major patterns of association among data blocks and identifying the main drivers of each signal.

In our numerical studies, the proposed method shows superior performances in correctly segmenting (partially) joint signals than some competitors [4, 8, 18, 19], and in the estimation of the factor loadings and coefficients. The proposed method is further utilized in exploring a TCGA breast cancer data set [12] and an image-feature data set [25]. In these real data examples, we demonstrate that the segmented components by the COBS factorization have insightful interpretation facilitated by the estimated block structure and the effect of covariates, and that the signals driven by covariates stand out in the factorization.

2.0 Model

In this section, we outline the data situation we have in mind. Suppose there are K blocks of data, each with p_k variables, all observed for a common sample of size n. These blocks are referred to as primary data blocks, and denoted by $\mathbf{Y}_k$, $k = 1, \ldots, K$. The concatenated data set of size $n \times p$ is $\mathbf{Y} = [\mathbf{Y}_1, \cdots, \mathbf{Y}_K]$, where $p = \sum_{k=1}^{K} p_k$. Throughout, we assume that $\mathbf{Y}$ are column-centered. A linear, structured factor model for the concatenated data is $\mathbf{Y} = \sum_{j=1}^{r} \mathbf{u}_j \mathbf{v}_j^T + \mathbf{E}$, where the n-vector of scores $\mathbf{u}_j$ is the unobserved factor scores, while the loading vector $\mathbf{v}_j^T = (\mathbf{v}_{j(1)}^T, \ldots, \mathbf{v}_{j(K)}^T)$ characterizes the effect of the jth factor to each data blocks. We call it *structured* as the interpretation of a component depends on the block structure of $\mathbf{v}_j$. This simple model encompasses many methodologies in the literature of multi-source data integration. As an instance, the method of Joint and Individual Variation (JIVE, for short) proposed by [21] and [4] assumes the linear factor model, but with a restriction that for each j, $\mathbf{v}_{j(k)}$s are either all non-zero, or at most one $\mathbf{v}_{j(k)}$ is non-zero, corresponding to joint or individual variations, respectively. We eliminate such a restriction, and allow $\mathbf{v}_j$ to have any pattern of block-wise zeros.

In particular, a $\mathbf{v}_j$ with only one non-zero $\mathbf{v}_{j(k)}$ is an individual component, whose factor only affects the kth data block. Any $\mathbf{v}_j$ with $\|\mathbf{v}_{j(k)}\| > 0$ for all k is interpreted as a full-joint component, whose factor jointly affects variations in all K blocks of data. All other block-wise zero patterns in $\mathbf{v}_j$ correspond to partial-joint components. For example, $\mathbf{v}_j = (\mathbf{v}_{j(1)}^T, \mathbf{v}_{j(2)}^T, 0 \ldots, 0)$ corresponds to the component, partially joint across the first two blocks.

When covariates are available, their relation to the primary data $\mathbf{Y}$ is modeled through the factor scores $\mathbf{u}_j$. Suppose there are G groups of variables in the covariate matrix $\mathbf{X}$. We write each by $\mathbf{X}_g$, $g = 1, \ldots, G$, so that $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_G]$ and let

$$\mathbf{u}_j = \sum_{g=1}^{G} \mathbf{X}_g \mathbf{b}_{j(g)} + \mathbf{f}_j = \mathbf{X}\mathbf{b}_j + \mathbf{f}_j. \tag{2.1}$$

Here, $\mathbf{b}_j^T = (\mathbf{b}_{j(1)}^T, \ldots, \mathbf{b}_{j(G)}^T)$ is the coefficient vector, and $\mathbf{f}_j$ is an unknown random source for the jth factor score. The decomposition (2.1) was used in [19] for the single covariate block case, i.e., $G = 1$.

We note that the roles of primary data blocks and covariate data blocks can be partially or entirely interchanged, depending on the specific goals of analysis. For example, in the analysis of TCGA breast cancer we present in Section 5.1, the three primary data blocks (GE, Meth, CNV) are factorized with supervision from the cancer subtypes. One may otherwise indicate the role of the CNV data block, taking only GE and Meth data blocks as primary and using CNV and subtypes as two blocks of covariate matrix. This enables the joint factorization of GE and Meth data blocks, potentially driven by the signals in CNV (and also by subtypes).

Putting it altogether, the primary data matrix is decomposed into

$$\begin{aligned}\mathbf{Y} = [\mathbf{Y}_1, \cdots, \mathbf{Y}_K] &= (\mathbf{X}\mathbf{B} + \mathbf{F})\mathbf{V}^T + \mathbf{E} \qquad (2.2)\\ &= \sum_{j=1}^{r} \mathbf{u}_j(\mathbf{v}_{j(1)}^T, \ldots, \mathbf{v}_{j(K)}^T) + \mathbf{E},\end{aligned}$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ is the $p \times r$ matrix of loading vectors, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_r)$ is the $q \times r$ matrix of coefficients, and $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_r)$ is the $n \times r$ matrix of random factors. We assume that the rows of $\mathbf{F}$ are independent with each other, and have mean zero and covariance matrix Σ_F. The noise matrix $\mathbf{E}$ consists of independent mean-zero random noises with variance σ_0^2, and $\mathbf{E}$ and $\mathbf{F}$ are independent.

For the special case of no covariate, or $\mathbf{B} = \mathbf{0}$, our model (2.2) includes that of [8], [21], and [37]. If partial-joint components are not allowed in (2.2), then it is the model used in [18].

We now put an identifiability condition for the parameters $\theta = (\mathbf{V}, \mathbf{B}, \Sigma_F, \sigma^2)$ of (2.2). Let $F_\theta(\mathbf{Y})$ be the distribution of $\mathbf{Y}$, parameterized by θ. We say the model parametrized by θ is identifiable if $F_{\theta_1}(\mathbf{Y}) = F_{\theta_2}(\mathbf{Y})$ implies $\theta_1 = \theta_2$. Note that the rows of $\mathbf{Y}$ are independent with each other but are not identically distributed, due to the covariates. Write $\mathbf{S}_X = \mathbf{X}^T\mathbf{X}/n$.

Proposition 1. *Suppose that the number of components r is fixed, and the covariate matrix $\mathbf{X}$ is treated fixed. Assume that for any $\theta = (\mathbf{V}, \mathbf{B}, \Sigma_F, \sigma^2)$ the following conditions are satisfied:*

(i) *$\mathbf{V}$ has orthogonal columns, i.e. $\mathbf{V}^T\mathbf{V} = \mathbb{I}_r$.*

(ii) *$\mathbf{B}^T\mathbf{S}_X\mathbf{B} + \Sigma_F$ is a diagonal matrix with r distinct and positive diagonal values, in descending order.*

(iii) *The $q \times q$ matrix $\mathbf{S}_X$ is of full rank.*

Then, the parameter θ of (2.2) is identifiable. If only conditions (i) and (ii) are satisfied, then $\theta_X = (\mathbf{V}, \mathbf{XB}, \Sigma_F, \sigma^2)$ is identifiable.

A proof of Proposition 1 is deferred to Appendix.

The identifiability condition given in (i)—(iii) are not restictive. Conditions (i) and (iii) are commonly imposed respectively for linear dimension reduction and regression settings. In Condition (ii), we allow the covariate effect and the random effects be interwined in any form, and Σ_F is not required to be diagonal.

Our condition encourages the loadings to be driven by the covariates $\mathbf{XB}$, if $\mathbf{B}^T\mathbf{S}_X\mathbf{B}$ is much larger than Σ_F. In such a case, $\mathbf{Xb}_i$ is nearly orthogonal to $\mathbf{Xb}_j$ for $1 \le i < j \le r$. On the other hand, if Σ_F is larger than $\mathbf{B}^T\mathbf{S}_X\mathbf{B}$, then Σ_F should be close to a diagonal matrix. Note that one may impose the condition of diagonal Σ_F (with distinct diagonal values), in place of Condition (ii) of Proposition 1. Under such a situation, the corresponding model parameters are still identifiable [19].

In the next section, we assume diagonal Σ_F but without the restriction of distinct values, for interpretability and clarity of presentation. The estimation scheme under the general Σ_F is discussed in the appendix.

3.0 Estimation

While our model includes the joint and individual components, the distinction is not explicit at (2.2). Instead of explicitly modeling all types of components, we use a data-driven approach in segmenting the full joint, partially joint and individual components from the primary data blocks. Our algorithm is based on an iterative thresholding of the loading vectors, aiming at both block-wise sparsity and variable-wise sparsity.

3.1 General strategy

By writing (2.2) for each observation vector, we have

$$\begin{aligned}\mathbf{y}_i &= \mathbf{V}(\mathbf{B}^T\mathbf{x}_i + \mathbf{f}_i) + \mathbf{e}_i \\ &= \sum_{j=1}^{r} \mathbf{v}_j(\mathbf{b}_j^T\mathbf{x}_i + f_{ij}) + \mathbf{e}_i, \quad i = 1, \ldots, n, \end{aligned} \tag{3.1}$$

where $\mathbf{y}_i^T, \mathbf{f}_i^T, \mathbf{x}_i^T$, and $\mathbf{e}_i^T$ are respectively the ith row of $\mathbf{Y}$, $\mathbf{F}$, $\mathbf{X}$ and $\mathbf{E}$, and $\mathbf{b}_j$ is the jth column of $\mathbf{B}$. To facilitate the estimation and prediction of the random factor $\mathbf{f}_i$, we assume normality for $\mathbf{f}_i$ and $\mathbf{e}_i$. In particular, $\mathbf{e}_i \sim N_p(\mathbf{0}, \sigma_0^2\mathbb{I}_p)$ and $\mathbf{f}_i \sim N_r(\mathbf{0}, \Sigma_F)$, where $\Sigma_F = \text{diag}(\sigma_1^2, \ldots, \sigma_r^2)$, independently for all i.

We propose a sequential estimation of $(\mathbf{v}_j, \mathbf{b}_j)$ for $j = 1, \ldots, r$, for its computational efficiency. The choice of r can be from some existing rank selection methods in learning of PCA, such as using the elbow of cumulative scree plot by choosing the percentage of variance explained or using the sequential test for the skewness of squared residual scores that is proposed in [14]. With rank r selected, we will introduce a sequential approach to estimate the parameters layer by layer. For estimation of $(\mathbf{v}_1, \mathbf{b}_1)$, we merge the remaining $r-1$ components with the error term, and write $\mathbf{y}_i = \mathbf{v}_1(\mathbf{b}_1^T\mathbf{x}_i + f_{i1}) + \mathbf{e}_i'$ from which we obtain estimates $\hat{\mathbf{v}}_1$, $\hat{\mathbf{b}}_1$ and predictions $\hat{f}_{i1}$. Given the first $J-1$ predicted components $\hat{u}_{ij} = \hat{\mathbf{b}}_j^T\mathbf{x}_i + \hat{f}_{ij}$ and estimates $\hat{\mathbf{v}}_j$ (for $j = 1, \ldots, J-1$), the Jth component is estimated with the rank-1 model $\mathbf{y}_i^{[J]} := \mathbf{y}_i - \sum_{j=1}^{J-1} \hat{\mathbf{v}}_j \hat{u}_{ij} = \mathbf{v}(\mathbf{b}^T\mathbf{x}_i + f_i) + \mathbf{e}_i'$. The estimation of Σ_F and σ_0^2 is important not only for interpretation but also for the estimation of $\mathbf{v}$ and $\mathbf{b}$, and prediction of f_{ij}. We use the normal model in the estimation of Σ_F and σ_0^2 for each component. Our approach is an analog of the sequential definition of principal components (maximizing the variance of the projected), or the deflation algorithm [24] for sparse PCA. Since the columns of $\mathbf{V}$ are in the decreasing order of the "variances" $\mathbf{B}^T\mathbf{S}_X\mathbf{B} + \Sigma_F$, the estimated $\mathbf{V}$ also satisfy the identifiability condition.

In the next section, an iterative thresholding algorithm is presented for the estimation with the rank-1 model.

3.2 Estimation for rank-1 model

With the normality assumption and the constraint $\mathbf{v}^T\mathbf{v} = 1$, the -2 log-likelihood function of the rank-1 model $\mathbf{y}_i = \mathbf{v}(\mathbf{b}^T\mathbf{x}_i + f_i) + \mathbf{e}_i$, $i = 1, \ldots, n$, or $\mathbf{Y} = (\mathbf{X}\mathbf{b} + \mathbf{f})\mathbf{v}^T + \mathbf{E}$, is simplified to

$$\ell(\mathbf{b}, \mathbf{v}, \sigma_f^2, \sigma_0^2) = \frac{1}{\sigma_e^2}\left(\|\mathbf{Y} - \mathbf{X}\mathbf{b}\mathbf{v}^T\|_F^2 - \frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2\right) + c(\sigma_e^2, \sigma_f^2), \tag{3.2}$$

where $c(\sigma_e^2, \sigma_f^2) = n\log[2\pi(\sigma_e^2)^{p-1}(\sigma_f^2 + \sigma_e^2)]$. Our approach is to iteratively update each of $(\mathbf{v}, \mathbf{b}, \sigma_e^2, \sigma_f^2)$, while others fixed. The block-wise and variable-wise sparsity in $\mathbf{v}^T = (\mathbf{v}_{(1)}^T, \ldots, \mathbf{v}_{(K)}^T)$ and $\mathbf{b}^T = (\mathbf{b}_{(1)}^T, \ldots, \mathbf{b}_{(G)}^T)$ will be controlled by a set of tuning parameters $(\alpha_v, \lambda_v, \alpha_b, \lambda_b)$. The role of the tuning parameters will be discussed below.

Given $\mathbf{b}, \sigma_e^2, \sigma_f^2$, minimizing ℓ with respect to $\mathbf{v}$ is equivalent to

$$\max_{\mathbf{v}} \|\mathbf{Y}\mathbf{v} + \mathbf{X}\tilde{\mathbf{b}}\|_2^2 \text{ subject to } \mathbf{v}^T\mathbf{v} = 1, \tag{3.3}$$

where $\tilde{\mathbf{b}} = \sigma_e^2/\sigma_f^2\mathbf{b}$. To incorporate the block structure, one would attempt to regularize (3.3) by adding sparsity-inducing constraints or penalizing terms. However, the resulting optimization problems become unwieldy, because (3.3) is already unconventional. The objective function $F(\mathbf{v}) = \|\mathbf{Y}\mathbf{v} + \mathbf{X}\tilde{\mathbf{b}}\|_2^2$ is convex on the convex hull of the feasible region, but we wish to maximize rather than minimize. As a result, there are typically a few local maxima and many inflection points of $F(\mathbf{v})$ on the feasible region. Proposition 2 below transforms the unconventional multivariate optimization problem into a simple univariate optimization problem, which can be solved efficiently by, e.g., a bisection method. It provides the globally optimal solution $\tilde{\mathbf{v}}$ of (3.3).

Proposition 2. *Let $\mathbf{Y} = \mathbf{L}\mathbf{D}\mathbf{R}^T$ be the singular value decomposition of $\mathbf{Y}$, where $\mathbf{D}$ is the diagonal matrix with positive singular values $d_1, \ldots, d_N$, for some $N \leq \min(n, p)$. Let $\mathbf{z} = (z_1, \ldots, z_N)^T = \mathbf{L}^T\mathbf{X}\tilde{\mathbf{b}}$.*

(i) If $\mathbf{z} = 0$, then the solution of (3.3) is the first right singular vector of $\mathbf{Y}$.

(ii) If $z_i \neq 0$ for any $i = 1, \ldots, N$, then $\tilde{\mathbf{v}} = \mathbf{R}\mathbf{D}(\tilde{t}\mathbb{I}_N - \mathbf{D}^2)^{-1}\mathbf{z}$ is the solution of (3.3), where $\tilde{t}$ is the unique root of $\sum_{i=1}^N \frac{d_i^2 z_i^2}{t - d_i^2} = 1$, located on (t_1, t_2) for $t_1 = d_1^2 + d_1 z_1$ and $t_2 = d_1^2 + (\sum_{i=1}^N d_i^2 z_i^2)^{1/2}$.

In Proposition 2, we assumed that d_i's are distinct and z_i's are either all zero or all nonzero, which are true if $\mathbf{Y}$ is sampled from a continuous distribution.

To incorporate the block structure, we threshold the solution $\tilde{\mathbf{v}}^T = (\tilde{\mathbf{v}}_1^T, \ldots, \tilde{\mathbf{v}}_K^T)$ of (3.3) twice, for variable-wise sparsity and block-wise sparsity. The level of thresholding is controlled by the tuning parameters α_v, λ_v. The parameter $\alpha_v \in [0, 1]$ indicates whether we seek only block-wise sparse structure ($\alpha_v = 0$) or both block-wise and variable-wise sparse loadings ($\alpha_v > 0$). The parameter $\lambda_v \geq 0$ controls the overall degrees of sparsity in the estimate of $\mathbf{v}$. Larger values of λ_v lead to more sparse group-wise structure, which in turn entails more frequent identification of individual and partially-joint components. Given (α_v, λ_v), we set the thresholds by $\gamma_1 = \alpha_v\lambda_v$ and $\gamma_2 = (1 - \alpha_v)\lambda_v$ (for variable-wise and block-wise thresholding, respectively). For any $\gamma \geq 0$ and m, define the soft-thresholding function $s_\gamma : \mathbb{R}^m \to \mathbb{R}^m$by taking the ith element of $s_\gamma(\mathbf{c})$ as $\text{sign}(c_i)\max\{|c_i| - \gamma, 0\}$. Let

$$\mathbf{v}_{k,*} = \frac{\mathbf{c}_k}{\|\mathbf{c}_k\|} s_{\gamma_2}(\|\mathbf{c}_k\|), \ \mathbf{c}_k = s_{\gamma_1}(\tilde{\mathbf{v}}_k), \tag{3.4}$$

for $k = 1, \ldots, K$. The doubly thresholded $\mathbf{v}_*^T = (\mathbf{v}_{1,*}^T, \ldots, \mathbf{v}_{K,*}^T)$ is the solution of $\min_\mathbf{x} \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{v}}\|^2 + \gamma_1\|\mathbf{x}\|_1 + \gamma_2\|\mathbf{x}\|_{2,1}$, where $\|\mathbf{x}\|_{2,1} = \sum_{k=1}^K \|\mathbf{x}_k\|_2$ with the partition of $\mathbf{x}$ given by the prescribed block structure. The thresholding (3.4) will only keep sufficiently large $\tilde{\mathbf{v}}_k$s as non-zero, producing loading vectors for individual or partially joint components.

To ensure that the loading vector has the unit norm, we set $\hat{\mathbf{v}} = \mathbf{v}_*/\|\mathbf{v}_*\|$.

Given $\mathbf{v}$, minimizing ℓ with respect to $\mathbf{b}$ is equivalent to minimize $\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2$. Since it is now equivalent to a regression problem of regressing $\mathbf{Y}\mathbf{v}$ onto $\mathbf{X}$, our update for $\mathbf{b}$ is the solution of the sparse group lasso regression problem [29] of minimizing

$$\frac{1}{2n}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2 + \alpha_b\lambda_b\|\mathbf{b}\|_1 + (1-\alpha_b)\lambda_b\|\mathbf{b}\|_{2,1}. \tag{3.5}$$

The tuning parameter $\lambda_b \geq 0$ controls the overall degrees of sparsity in $\hat{\mathbf{b}}$. We set $\alpha_b = 0$ for strictly block-wise sparse coefficient, and $\alpha_b = 1$ when there is only one block in $\mathbf{X}$. Choosing $\alpha_b \in (0,1)$ results in both block-wise and variable-wise sparse coefficient estimates.

To obtain the solution $\hat{\mathbf{b}}$ of (3.5), we use the r package SGL [28] if $\alpha_b > 0$ or GLMNET [6] if $\alpha_b = 0$.

Given $(\mathbf{v}, \mathbf{b})$, we set $\hat{\sigma}_e^2$ and $\hat{\sigma}_f^2$ as the minimizers of ℓ, which are given by

$$\hat{\sigma}_e^2 = \frac{\|\mathbf{Y} - \mathbf{X}\mathbf{b}\mathbf{v}^T\|_F^2 - \|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2}{n(p-1)}, \ \hat{\sigma}_f^2 = \frac{\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2}{n} - \hat{\sigma}_e^2. \tag{3.6}$$

To summarize, to estimate the parameters $\theta = (\mathbf{v}, \mathbf{b}, \sigma_e^2, \sigma_f^2)$ of the rank-1 model, we begin with an initial value $\theta_{(0)}$ and iteratively update the elements $\mathbf{v}$, $\mathbf{b}$ and (σ_e^2, σ_f^2) as described above, until the changes in $\theta_{(t)}$ are small enough. The initial value $\theta_{(0)}$ is obtained for any given $\mathbf{v}_{(0)}$, as the updates for $\mathbf{b}, \sigma_e^2, \sigma_f^2$ only depend on a given $\mathbf{v}$. We set $\mathbf{v}_{(0)}$ as the first sample principal component direction of $\mathbf{Y}$.

Moreover, with the rank-one estimates of $\mathbf{b}$, $\mathbf{v}$, σ_f^2, σ_e^2 achieved and under the model assumptions described in Chapter 2, we will also be able to predict the random effect $\mathbf{f}$ given primary data $\mathbf{Y}$ and supervision information $\mathbf{X}$. More specifically, the elements in $\mathbf{f}$ are i.i.d. normal with mean 0 and variance σ_f^2 [denoted as $f_i \sim \mathcal{N}(0, \sigma_f^2)$, for $i = 1, \ldots, n$]. Each observation row of the primary data $\mathbf{Y}_i$ given $\mathbf{X}_i$ and f_i follows a multivariate normal distribution: $\mathbf{Y}_i|(\mathbf{X}_i, f_i) \sim \mathcal{N}_p(\mathbf{X}_i\mathbf{b}\mathbf{v}^T + f_i\mathbf{v}^T, \sigma_e^2\mathbb{I}_p)$, where $\mathbf{Y}_i$ and $\mathbf{X}_i$ are ith rows of $\mathbf{Y}$ and $\mathbf{X}$ respectively, f_i is a random factor for the ith observation, for $i = 1, \ldots, n$. Thus it can be shown that for $i = 1, \ldots, n$ the joint distribution of $(\mathbf{Y}_i, f_i)$ given $\mathbf{X}_i$ are i.i.d.

$$(\mathbf{Y}_i, f_i)|\mathbf{X}_i \sim \mathcal{N}_{p+1}\left[\begin{pmatrix} \mathbf{X}_i\mathbf{b}\mathbf{v}^T & 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2\mathbb{I}_p + \sigma_f^2\mathbf{v}\mathbf{v}^T & \sigma_f^2\mathbf{v} \\ \sigma_f^2\mathbf{v}^T & \sigma_f^2 \end{pmatrix}\right].$$

Then, it can also be found that

$$\mathbb{E}(f_i|\mathbf{Y}_i, \mathbf{X}_i) = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2}(\mathbf{Y}_i\mathbf{v} - \mathbf{X}_i\mathbf{b}). \tag{3.7}$$

As the n samples are i.i.d., then given the rank-one estimates $\hat{\mathbf{b}}$, $\hat{\mathbf{v}}$, $\hat{\sigma}_f^2$, $\hat{\sigma}_e^2$, the predictor of $\mathbf{f}$ is

$$\hat{\mathbf{f}} = \mathbb{E}(\mathbf{f}|\mathbf{Y}, \mathbf{X}) = \frac{\hat{\sigma}_f^2}{\hat{\sigma}_f^2 + \hat{\sigma}_e^2}(\mathbf{Y}\hat{\mathbf{v}} - \mathbf{X}\hat{\mathbf{b}}). \tag{3.8}$$

From this, the predictor of $\mathbf{u} = \mathbf{X}\mathbf{b} + \mathbf{f}$ is

$$\hat{\mathbf{u}} = \mathbf{X}\hat{\mathbf{b}} + \hat{\mathbf{f}} = (\hat{\sigma}_f^2 + \hat{\sigma}_e^2)^{-1}(\hat{\sigma}_f^2\mathbf{Y}\hat{\mathbf{v}} + \hat{\sigma}_e^2\mathbf{X}\hat{\mathbf{b}}). \tag{3.9}$$

Algorithm 1: One-Layer Iteration:

```
Input: Primary data $\mathbf{Y}$, and block index vector $\mathbf{G}_Y$; Supervision information $\mathbf{X}$,
       and group index vector $\mathbf{G}_X$; Tuning parameters $(\alpha_b, \alpha_v, \lambda_b, \lambda_v)$ .
Output: $\hat{\mathbf{b}}$, $\hat{\mathbf{v}}$, $\hat{\sigma}_e$, $\hat{\sigma}_f$, and $\hat{\mathbf{u}}$
Set initial parameters $\mathbf{v}_{[0]}$, $\mathbf{b}_{[0]}$, $\sigma^2_{e[0]}$, and $\sigma^2_{f[0]}$;
while $\|\mathbf{v}_{[i+1]} - \mathbf{v}_{[i-1]}\|_1 \geq$ tolerance and $\|\mathbf{b}_{[i+1]} - \mathbf{b}_{[i]}\|_1 \geq$ tolerance do
    Given ($\mathbf{b}_{[i+1]}$, $\mathbf{v}_{[i]}$, $\sigma^2_{e[i]}$, and $\sigma^2_{f1[i]}$), apply Proposition 2 and thresholding to get
      $\mathbf{v}_{[i+1]}$;
    Given ($\mathbf{b}_{[i]}$, $\mathbf{v}_{[i]}$, $\sigma^2_{e[i]}$, and $\sigma^2_{f1[i]}$), apply sparse group lasso [29] to equation 3.5 to
      get $\mathbf{b}_{[i+1]}$;
    Given ($\mathbf{b}_{[i+1]}$, $\mathbf{v}_{[i+1]}$, $\sigma^2_{e[i]}$, and $\sigma^2_{f1[i]}$), obtain $\sigma^2_{e[i+1]}$ from equation 3.6;
    Given ($\mathbf{b}_{[i+1]}$, $\mathbf{v}_{[i+1]}$, $\sigma^2_{e[i+1]}$, and $\sigma^2_{f1[i]}$), obtain $\sigma^2_{f1[i+1]}$ from equation 3.6;
    Set $i \leftarrow i + 1$.
end
Given final estimates $\hat{\mathbf{b}}$, $\hat{\mathbf{v}}$, $\hat{\sigma}_e$, $\hat{\sigma}_f$, obtain $\hat{\mathbf{u}}$ from equation 3.9
```

3.3 Multi-layer estimation

Suppose that we have estimates $(\hat{\mathbf{v}}_j, \hat{\mathbf{b}}_j)$ and the predictors $\hat{\mathbf{u}}_j$ for $j = 1, \ldots, J-1$ in the general model (3.1). For the estimation of the Jth component, we subtract the first $J-1$ components from $\mathbf{Y}$,

$$\mathbf{Y}^{[J]} = \mathbf{Y} - \sum_{j=1}^{J-1} \hat{\mathbf{u}}_j \hat{\mathbf{v}}_j^T,$$

and treat $\mathbf{Y}^{[J]}$ as the primary data block. The algorithm in Section 3.2 is then applied to the rank-1 model $\mathbf{Y}^{[J]} = (\mathbf{X}\mathbf{b} + \mathbf{f})\mathbf{v}^T + \mathbf{E}'$, from which we obtain $(\hat{\mathbf{v}}_J, \hat{\mathbf{b}}_J, \hat{\mathbf{u}}_J)$.

We estimate r layers of components, where r is pre-specified, or is determined to be $J-1$ if, at the estimation of Jth rank-1 model, the thresholding (3.4) results in $\hat{\mathbf{v}}_* = \mathbf{0}$.

While sequentially estimating the r layers, we have temporarily obtained $\hat{\sigma}_e^2$, which estimates the variance of e'_{ij} that varies from layer to layer. Thus, from the full model with r components, and with $(\hat{\mathbf{v}}_j, \hat{\mathbf{b}}_j)$ plugged in, we estimate σ_0^2 and $\Sigma_F = \text{diag}(\sigma_1^2, \ldots, \sigma_r^2)$ by the maximum likelihood estimates $\hat{\sigma}_0^2 = (n(p-r))^{-1}\|\mathbf{Y}\hat{\mathbf{V}}^{\perp}\|_F^2$, where $\hat{\mathbf{V}}^{\perp}$ is the $p \times (p-r)$ matrix formed by an orthonormal basis of the null space of $\hat{\mathbf{V}}$, and $\hat{\sigma}_j^2 = n^{-1}\|\mathbf{Y}\hat{\mathbf{v}}_j - \mathbf{X}\hat{\mathbf{b}}_j\|_2 - \hat{\sigma}_0^2$, for $j = 1, \ldots, r$.

3.4 Choice of tuning parameters

In our estimating algorithm, the parameters $(\alpha_b, \lambda_b, \alpha_v, \lambda_v)$ play important roles in the identification of the block structure and also in the variable selection. In particular, larger λ_v will lead to more frequent identification of individual and partially joint components. The algorithm with too small λ_v will not identify any individual components.

Searching the best tuning parameter $(\alpha_b, \lambda_b, \alpha_v, \lambda_v)$ in the four dimensional space is computationally infeasible. We resort to fix the parameters α_b and α_v as a pre-specified balance between block-wise sparsity and variable-wise sparsity. We allow λ_b and λ_v to be different for each rank-1 estimation (cf. Section 3.2). To choose the tuning parameters λ_b, λ_v, we utilize the Bayesian information criterion (BIC) [27], assuming the normal model.

As discussed in [40], utilizing BIC tends to choose the true sparse model if the true model is sparse. Since we implicitly assume that the true $\mathbf{V}$ is block-wise sparse, we thrive for selecting sparse models rather than maximizing a prediction accuracy.

Our definition of BIC for $\boldsymbol{\lambda} = (\lambda_b, \lambda_v)$, evaluated for each rank-1 model is as follows. Let $\hat{\theta}(\boldsymbol{\lambda}) = (\hat{\mathbf{b}}, \hat{\mathbf{v}}, \hat{\sigma}_f^2, \hat{\sigma}_e^2)$ be the estimated parameters of the rank-1 model, with $\boldsymbol{\lambda}$. Recall that $\ell(\theta)$ is the -2 log-likelihood evaluated at θ (3.2). We set

$$\mathrm{BIC}(\boldsymbol{\lambda}) := \frac{\ell\{\hat{\theta}(\boldsymbol{\lambda})\}}{np} + \frac{\omega_{(p,n)}}{np}\widehat{df}\{\hat{\theta}(\boldsymbol{\lambda})\},$$

where np is the number of observations in the concatenated data blocks, $\widehat{df}\{\hat{\theta}(\boldsymbol{\lambda})\}$ is the number of non-zero elements of $\hat{\mathbf{b}}$ and $\hat{\mathbf{v}}$, and $\omega(p,n) = \log(np)$. If the dimensionality of the primary data p is deemed high, we use $\omega(p,n) = 6(1+\delta)\log(np)$, as suggested by [7], for a small value of δ.

For each layer, in the estimation of the corresponding rank-1 model, the BIC is evaluated for a grid for $\boldsymbol{\lambda}$. We choose $\boldsymbol{\lambda}$ with the smallest $\mathrm{BIC}(\boldsymbol{\lambda})$.

4.0 Simulation studies

In this section, we numerically compare the performance of our proposal with competing methods. Our method will be referred to as COBS, a short name for COvariate-driven Block-wise Sparse factorization. We consider several competing methods from different literature aspects. In matrix decomposition for dimension reduction literature, we compare with the usual principal component analysis, given by the singular value decomposition (SVD) of the concatenated matrix $\mathbf{Y}$, the angle-based JIVE (AJIVE) estimation of [4], the structural learning and integrative decomposition (SLIDE) [8], the supervised SVD [19], and supervised integrated factor analysis (SIFA) of [18]. In multivariate linear regression, we compare with both reduced-rank regression (RRR) [13] and sparse reduced-rank regression (SRRR) [2] for their coverage in using supervision information for dimension reduction, where SRRR also allows for sparsity in estimating the coefficients for variable selection. We also compare with the group factor analysis [16] in terms of the high-dimension factor model. These methods will be utilized in the estimation of the factor loadings $\mathbf{V}$ and, if feasible, in the estimation of the coefficient matrix $\mathbf{B}$.

4.1 Simulation settings

For comprehensive comparisons, we not only consider the "COBS" models in Section 2, but also consider data situations following the models suggested in SIFA and JIVE. From the model (3.1), only allowing joint and individual components in $\mathbf{V}$ results in a SIFA model. If, in addition, $\mathbf{B} = \mathbf{0}$, then the model corresponds to a JIVE model.

Throughout, we assume that there are $K = 4$ equal-sized blocks of primary variables with both low-dimension and high-dimension application. For low-dimensional simulation, each block consists of $p_k = 25$ variables ($k = 1, \ldots, 4$) from a sample of size $n = 500$. For high-dimensional simulation, each block consists of $p_k = 100$ variables ($k = 1, \ldots, 4$) with sample size $n = 200$. The number of components is $r = 4$. The $n \times q$ covariate matrix $\mathbf{X}$

consists of randomly sampled standard normal variates, and we assume that there are $G = 4$ groups of equal size $q_g = 10$ $(g = 1, \ldots, 4)$. Furthermore, $\mathbf{e}_i \sim N_p(\mathbf{0}, \mathbb{I}_p)$ and $\mathbf{f}_i \sim N_4(\mathbf{0}, \Sigma_F)$ are independent. For each scenario below, we consider two values of Σ_F, relatively large or small. In both cases, we set Σ_F diagonal.

The three data situations we consider are dictated by $\mathbf{V}$ and $\mathbf{B}$.

(a) COBS. The loading matrix $\mathbf{V}_{(a)} = [\mathbf{v}_1, \ldots, \mathbf{v}_4]$ consists of the loadings producing two individuals ($\mathbf{v}_1$ and $\mathbf{v}_2$), one partial-joint ($\mathbf{v}_3$) and a full-joint component. : The $q \times r$ coefficient matrix $\mathbf{B}_{(a)}$ is also block-wise and variable-wise sparse and is the block-diagonal matrix $\mathbf{B}_{(a)} = \text{diag}(5\mathbf{b}, -4\mathbf{b}, -3\mathbf{b}, 2\mathbf{b})$ where $\mathbf{b} = (1, 1, 1, 0, \ldots, 0)^T$. Note that in this setting, the individual and partially joint components have greater variances than the full-joint component.

(b) SIFA. The loading matrix $\mathbf{V}_{(b)} = [\mathbf{v}_1, \ldots, \mathbf{v}_4]$ has a fully joint component $\mathbf{v}_1$ (corresponding to the largest variance), and three individual components. The coefficient matrix $\mathbf{B}_{(b)}$ is the same as $\mathbf{B}_{(a)}$ in the situation (a).

(c) JIVE. We set $\mathbf{V}_{(c)} = \mathbf{V}_{(b)}$ and $\mathbf{B}_{(c)} = \mathbf{0}$ so that neither a partially joint component nor a covariate effect is present. Even though there is no connection to the primary data, $\mathbf{X}$ is still available.

See the appendix for detailed data generating process.

In applying the proposed estimation algorithm, COBS, to the simulated data, we set the number of components $r = 4$ to be the true value. For the tuning parameters, we set $\alpha_b = \alpha_v = 0.2$, and choose λ_b and λ_v by the BIC. For situation (c), the model parameter $\mathbf{B}_{(c)} = \mathbf{0}$ is in principle not known, so $\mathbf{B}_{(c)}$ is estimated as well. Our algorithm does fit $\hat{\mathbf{B}} \approx \mathbf{0}$. Fitting a SIFA model requires pre-specifying the numbers of joint and each individual components. We supply the true numbers. For situation (a), the partial joint $\mathbf{v}_3$ is considered as a full-joint component for SIFA. To apply AJIVE, the number of "signal" components in each data block must be supplied. The true values are supplied for AJIVE as well. Note that the methods of SVD and SupSVD are simply applied to the concatenated data, assuming the true number of components $r = 4$ is known.

4.2 Evaluation criteria

The performances of the competing methods are evaluated by the following.

To compare the quality of subspace learning, we use the largest principal angle and the Grassmannian distance between the column spaces of $\mathbf{V}$ and $\hat{\mathbf{V}}$. We refer to [35] for a discussion on distances between linear subspaces. Let $0 \le d_1 \le \ldots \le d_r$ be the singular values of $\boldsymbol{\mathcal{V}}^T \hat{\boldsymbol{\mathcal{V}}}$, where $\boldsymbol{\mathcal{V}}$ (or $\hat{\boldsymbol{\mathcal{V}}}$) is any orthonormal basis of $\mathbf{V}$ (or $\hat{\mathbf{V}}$, respectively). The principal angles are $\theta_i = \arccos d_i$. The largest principal angle and the Grassmanian distance are then $\angle_P(\mathbf{V}, \hat{\mathbf{V}}) = \theta_1(180/\pi)$ and $d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}}) = \sqrt{\sum_{i=1}^r \theta_i^2}$.

To measure the quality of block identification, we introduce a permuted Hamming distance $l(\mathbf{V}, \hat{\mathbf{V}})$ with respect to the given block structure. Assuming both matrices are of size $p \times r$, let $S_{\mathbf{V}}$ be the $K \times r$ matrix whose (k, j)th element is 1 if $\mathbf{v}_{j(k)}$ is non-zero, and is 0 if $\mathbf{v}_{j(k)} = 0$. The Hamming distance between $S_{\mathbf{V}}$ and $S_{\hat{\mathbf{V}}}$ is the entrywise 1-norm of $S_{\mathbf{V}} - S_{\hat{\mathbf{V}}}$, which is the number of misidentified blocks in $\hat{\mathbf{V}}$, compared to $\mathbf{V}$. Since different methods have different identifiability conditions for $\mathbf{V}$, it is possible that an estimate $\hat{\mathbf{V}}$ is estimating a column-permuted $\mathbf{V}$. Thus, we define $l(\mathbf{V}, \hat{\mathbf{V}}) = \min_{\Pi} |S_{\mathbf{V}} - S_{\hat{\mathbf{V}}}\Pi|$, where the minimum is taken over all $r \times r$ permutation matrices Π, and $|\cdot|$ is the entrywise 1-norm.

We also use the squared Frobenius norm for $\mathbf{V} - \hat{\mathbf{V}}$ and $\mathbf{B} - \hat{\mathbf{B}}$. The use of the Frobenius norm requires that $r = r'$, where r is the dimension of $\mathbf{V}$ and r' is the dimension of $\hat{\mathbf{V}}$.

4.3 Simulation results

For each of situations (a)—(c) and for each choice of Σ_F (either large or small variance), we have simulated $n = 500$ independent sample for high-dimensional cases and $n = 200$ independent sample for low-dimensional cases. We applied all competing methods to estimate $\mathbf{V}$ and, if feasible, $\mathbf{B}$. This is repeated for 100 times, and the results are summarized in Tables 1 to 4. Overall, we confirm that our estimators $\hat{\mathbf{V}}$ and $\hat{\mathbf{B}}$ are among the closest to the true $\mathbf{V}$ and $\mathbf{B}$, compared with the competing estimators.

In particular, the accuracy of the subspace learning by COBS is best among all methods

		SVD	**AJIVE**	**SupSVD**	**SIFA**	**COBS**
(a) COBS + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	6.44 (0.46)	19.77 (21.63)	6.40 (0.45)	67.90 (35.50)	**2.43** (0.59)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.16 (0.01)	0.38 (0.40)	0.16 (0.01)	1.32 (0.69)	**0.05** (0.01)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	8.00 (0.00)	4.13 (1.37)	8.00 (0.00)	2.00 (0.00)	0.30 (0.58)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.08 (0.04)	8.08 (1.01)	0.33 (0.22)	5.55 (3.16)	**0.01** (0.01)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	-	-	16.88 (12.90)	205.14 (116.37)	**1.47** (0.65)
(b) SIFA + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	6.54 (0.43)	3.31 (0.39)	6.51 (0.43)	3.30 (0.38)	**1.70** (0.54)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.16 (0.01)	0.09 (0.01)	0.16 (0.01)	0.09 (0.01)	**0.04** (0.01)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	9.00 (0.00)	**0.00** (0.00)	9.00 (0.00)	**0.00** (0.00)	0.20 (0.65)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.08 (0.04)	8.60 (3.87)	0.33 (0.28)	**0.01** (0.00)	**0.01** (0.02)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	-	-	17.25 (15.62)	0.97 (0.28)	**1.66** (1.09)
(c) JIVE + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	14.27 (1.12)	8.50 (0.75)	14.30 (1.15)	8.59 (0.76)	**5.49** (1.99)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.38 (0.02)	0.23 (0.02)	0.38 (0.02)	0.23 (0.02)	**0.13** (0.03)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	9.00 (0.00)	**0.00** (0.00)	9.00 (0.00)	**0.00** (0.00)	0.44 (1.03)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.40 (0.25)	7.92 (4.07)	0.45 (0.28)	**0.06** (0.01)	0.11 (0.27)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	-	-	2.92 (0.39)	**0.01** (0.03)	0.05 (0.06)

		RRR	**SRRR**	**SPCA**	**GFA**	**SLIDE**
(a) COBS + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	7.14 (0.53)	7.14 (0.53)	5.43 (0.46)	11.11 (0.84)	14.00 (1.96)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.17 (0.01)	0.17 (0.01)	0.13 (0.01)	0.28 (0.02)	0.25 (0.03)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	8.00 (0.00)	8.00(0.00)	8.00(0.00)	3.02 (1.52)	**0.01** (0.10)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.05 (0.01)	0.05 (0.01)	0.07 (0.04)	7.72 (2.08)	6.89 (1.37)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	3.57 (0.87)	3.24 (0.85)	-	-	-
(b) SIFA + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	7.25 (0.56)	7.25 (0.56)	5.29 (0.43)	9.69 (0.66)	8.96 (0.40)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.17 (0.01)	0.17 (0.01)	0.12 (0.01)	0.27 (0.02)	0.17 (0.01)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	9.00 (0.00)	9.00 (0.00)	9.00 (0.00)	5.04(1.36)	**0.00** (0.00)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.05 (0.01)	0.05 (0.01)	0.07 (0.04)	8.06 (1.79)	6.85 (3.69)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	3.64 (0.77)	3.31 (0.75)	-	-	-
(c) JIVE + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	51.11 (9.48)	51.10 (9.46)	12.89 (1.10)	13.44 (1.10)	9.30 (2.33)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	1.31 (0.14)	1.31 (0.14)	0.34 (0.02)	0.37 (0.03)	0.24 (0.04)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	9.00 (0.00)	9.00 (0.00)	9.00 (0.00)	5.86 (1.32)	0.01 (0.10)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	3.84 (1.18)	3.84 (1.18)	0.37 (0.25)	8.04 (2.03)	4.62 (1.93)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	3.82 (0.38)	3.75 (0.37)	-	-	-

Table 1: High-dimensional results one. $n = 500, p = 100$. Large variances in unknown source. The means (standard deviations) of evaluation criteria, from 100 repetitions, are presented in the table. The bold number indicates the best result.

		SVD	**AJIVE**	**SupSVD**	**SIFA**	**COBS**
(a) COBS + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	7.29 (0.55)	23.38 (21.97)	7.17 (0.51)	89.27 (4.18)	**2.66** (0.70)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.17 (0.01)	0.47 (0.43)	0.17 (0.01)	1.74 (0.10)	**0.06** (0.01)
	$l(\mathbf{V},\hat{\mathbf{V}})$	8.00 (0.00)	3.92 (1.45)	8.00 (0.00)	2.00 (0.00)	0.05 (0.22)
	$\|\mathbf{V}-\hat{\mathbf{V}}\|_F^2$	0.08 (0.04)	7.96 (0.92)	0.53 (0.32)	7.65 (0.42)	**0.00** (0.00)
	$\|\mathbf{B}-\hat{\mathbf{B}}\|_F^2$	-	-	22.80(16.76)	282.32 (11.41)	**0.64** (0.17)
(b) SIFA + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	7.30 (0.63)	3.71 (0.52)	7.21 (0.61)	3.68 (0.51)	**1.75** (0.52)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.17 (0.01)	0.10 (0.01)	0.17 (0.01)	0.10 (0.01)	**0.04** (0.01)
	$l(\mathbf{V},\hat{\mathbf{V}})$	9.00 (0.00)	**0.00** (0.00)	9.00 (0.00)	**0.00** (0.00)	0.16 (0.61)
	$\|\mathbf{V}-\hat{\mathbf{V}}\|_F^2$	0.08 (0.05)	8.04 (3.87)	0.59 (0.39)	**0.01** (0.00)	**0.01** (0.02)
	$\|\mathbf{B}-\hat{\mathbf{B}}\|_F^2$	-	-	25.63 (19.22)	**0.30** (0.09)	0.93 (1.07)
(c) JIVE + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	35.97 (4.25)	20.72 (1.81)	38.65 (8.18)	20.00 (1.75)	**16.78** (4.30)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.91 (0.06)	0.55 (0.04)	0.95 (0.12)	0.54 (0.04)	**0.42** (0.08)
	$l(\mathbf{V},\hat{\mathbf{V}})$	9.00 (0.00)	**0.00** (0.00)	9.00 (0.00)	**0.00** (0.00)	2.54 (1.63)
	$\|\mathbf{V}-\hat{\mathbf{V}}\|_F^2$	1.50 (0.72)	2.65 (0.78)	1.58 (0.68)	**0.30** (0.04)	0.75 (1.16)
	$\|\mathbf{B}-\hat{\mathbf{B}}\|_F^2$	-	-	1.23 (0.15)	**0.00** (0.01)	**0.00** (0.00)

		RRR	**SRRR**	**SPCA**	**GFA**	**SLIDE**
(a) COBS + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	7.30 (0.51)	7.30 (0.51)	6.25 (0.54)	11.85 (0.85)	14.89 (3.64)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.17 (0.01)	0.17 (0.01)	0.14 (0.01)	0.29 (0.02)	0.27 (0.06)
	$l(\mathbf{V},\hat{\mathbf{V}})$	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	3.55 (1.65)	**0.03** (0.22)
	$\|\mathbf{V}-\hat{\mathbf{V}}\|_F^2$	0.04 (0.01)	0.04 (0.01)	0.07 (0.04)	7.90 (2.13)	6.60 (1.55)
	$\|\mathbf{B}-\hat{\mathbf{B}}\|_F^2$	1.23 (0.29)	1.05 (0.29)	-	-	-
(b) SIFA + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	7.37 (0.62)	7.37 (0.62)	6.01(0.66)	10.03(0.70)	8.66 (0.36)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.17 (0.01)	0.17 (0.01)	0.14 (0.01)	0.27 (0.02)	0.17 (0.01)
	$l(\mathbf{V},\hat{\mathbf{V}})$	9.00 (0.00)	9.00 (0.00)	9.00 (0.00)	4.87 (1.14)	**0.00** (0.00)
	$\|\mathbf{V}-\hat{\mathbf{V}}\|_F^2$	0.04 (0.01)	0.04 (0.01)	0.07 (0.05)	7.89 (2.21)	5.89 (2.98)
	$\|\mathbf{B}-\hat{\mathbf{B}}\|_F^2$	1.23 (0.33)	1.05 (0.32)	-	-	-
(c) JIVE + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	82.73 (5.25)	82.75 (5.22)	34.59 (4.33)	-	21.65 (6.56)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	2.26 (0.13)	2.26 (0.13)	0.87 (0.06)	-	0.55 (0.11)
	$l(\mathbf{V},\hat{\mathbf{V}})$	9.00 (0.00)	9.00 (0.00)	9.00 (0.00)	-	0.12 (0.42)
	$\|\mathbf{V}-\hat{\mathbf{V}}\|_F^2$	6.11 (0.65)	6.11 (0.65)	1.40 (0.73)	-	4.24 (1.10)
	$\|\mathbf{B}-\hat{\mathbf{B}}\|_F^2$	2.23 (0.12)	2.20 (0.12)	-	-	-

Table 2: High-dimensional results two. $n = 500, p = 100$. Small variances in unknown source. The means (standard deviations) of evaluation criteria, from 100 repetitions, are presented in the table. The bold number indicates the best result.

		SVD	**AJIVE**	**SupSVD**	**SIFA**	**COBS**
(a) COBS + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	20.30 (1.30)	60.42 (12.59)	20.24 (1.30)	24.52 (15.74)	**11.04** (1.72)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.50 (0.02)	1.21 (0.23)	0.49 (0.02)	0.53 (0.29)	**0.23** (0.03)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	8.00 (0.00)	2.00(0.00)	8.00(0.00)	2.00 (0.00)	0.03 (0.22)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.39 (0.12)	8.02 (3.13)	1.12 (0.62)	0.86 (1.65)	**0.06** (0.02)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	-	-	48.2 (31.90)	23.94 (64.36)	**3.48** (1.69)
(b) SIFA + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	20.33 (1.20)	10.35 (0.87)	20.25 (1.19)	10.31 (0.85)	**5.96** (1.17)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.50 (0.02)	0.28 (0.01)	0.50 (0.02)	0.28 (0.01)	**0.15** (0.02)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	9.00 (0.00)	**0.00** (0.00)	9.00 (0.00)	**0.00** (0.00)	0.01 (0.10)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.38 (0.15)	8.16 (3.81)	0.99 (0.51)	0.08 (0.01)	**0.07** (0.39)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	-	-	41.44 (24.43)	**2.20** (0.57)	6.45 (23.60)
(c) JIVE + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	41.47 (3.31)	25.98 (1.50)	41.82 (3.42)	25.89 (1.48)	**23.82** (14.59)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	1.14 (0.05)	0.71 (0.03)	1.14 (0.05)	0.71 (0.03)	**0.59** (0.25)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	9.00 (0.00)	**0.00** (0.00)	9.00 (0.00)	**0.00** (0.00)	0.17 (0.71)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	1.94 (0.73)	8.66 (3.85)	2.17 (0.78)	**0.51** (0.05)	1.13 (1.72)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	-	-	11.13 (1.31)	**0.03** (0.05)	1.63 (1.27)

		RRR	**SRRR**	**SPCA**	**GFA**	**SLIDE**
(a) COBS + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	22.08 (1.57)	22.06 (1.58)	17.72 (1.81)	21.99 (1.20)	20.95 (1.29)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.53 (0.02)	0.53 (0.02)	0.43 (0.04)	0.50 (0.02)	0.42 (0.02)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	8.00 (0.00)	8.00(0.00)	8.00(0.00)	7.99 (0.10)	**0.00** (0.00)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.34 (0.05)	0.34 (0.05)	0.32 (0.13)	7.75 (1.83)	5.18 (1.42)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	11.23 (2.45)	9.12 (2.37)	-	-	-
(b) SIFA + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	22.09 (1.44)	22.07 (1.43)	15.29 (1.51)	14.41 (0.57)	10.58 (0.73)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	0.53 (0.02)	0.53 (0.02)	0.39 (0.03)	0.42 (0.02)	0.29 (0.01)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	9.00 (0.00)	9.00 (0.00)	9.00 (0.00)	8.88 (0.36)	**0.00** (0.00)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	0.34 (0.05)	0.34 (0.05)	0.27 (0.15)	8.21 (2.11)	5.49 (2.29)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	10.93 (2.38)	8.83 (2.29)	-	-	-
(c) JIVE + Large var	$\angle_P(\mathbf{V}, \hat{\mathbf{V}})$	76.30 (8.28)	76.34 (8.24)	36.41 (4.46)	27.75 (1.52)	25.91 (1.47)
	$d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$	2.09 (0.14)	2.09 (0.14)	0.97 (0.07)	0.68 (0.04)	0.71 (0.03)
	$l(\mathbf{V}, \hat{\mathbf{V}})$	9.00 (0.00)	9.00 (0.00)	9.00 (0.00)	9.87 (0.37)	**0.00** (0.00)
	$\Vert\mathbf{V} - \hat{\mathbf{V}}\Vert_F^2$	5.72 (0.81)	5.73 (0.82)	1.35 (1.03)	-	4.39 (0.92)
	$\Vert\mathbf{B} - \hat{\mathbf{B}}\Vert_F^2$	19.76 (1.30)	18.50 (1.20)	-	-	-

Table 3: Low-dimensional results one. $n = 200, p = 400$. Large variances in unknown source. The means (standard deviations) of evaluation criteria, from 100 repetitions, are presented in the table. The bold number indicates the best result.

		SVD	**AJIVE**	**SupSVD**	**SIFA**	**COBS**
(a) COBS + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	22.6 (1.62)	60.33 (11.43)	22.28 (1.57)	73.64 (26.90)	**13.16** (2.38)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.54 (0.02)	1.20 (0.20)	0.53 (0.20)	1.47 (0.53)	**0.27** (0.04)
	$l(\mathbf{V},\hat{\mathbf{V}})$	8.00 (0.00)	2.00(0.00)	8.00 (0.00)	2.00(0.00)	**0.00** (0.00)
	$\lVert\mathbf{V}-\hat{\mathbf{V}}\rVert_F^2$	0.41 (0.12)	8.39 (3.24)	1.67 (0.98)	5.99 (2.65)	**0.07** (0.02)
	$\lVert\mathbf{B}-\hat{\mathbf{B}}\rVert_F^2$	-	-	62.69 (43.21)	223.25 (104.49)	**1.20** (0.24)
(b) SIFA + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	22.35 (1.43)	11.43 (0.98)	22.00 (1.39)	11.23 (0.97)	**6.67** (1.15)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.54 (0.02)	0.31 (0.01)	0.53 (0.02)	0.30 (0.01)	**0.17** (0.02)
	$l(\mathbf{V},\hat{\mathbf{V}})$	9.00 (0.00)	**0.00** (0.00)	9.00 (0.00)	**0.00** (0.00)	**0.00** (0.00)
	$\lVert\mathbf{V}-\hat{\mathbf{V}}\rVert_F^2$	0.43 (0.12)	7.32 (4.17)	1.61 (0.80)	0.09 (0.01)	**0.04** (0.04)
	$\lVert\mathbf{B}-\hat{\mathbf{B}}\rVert_F^2$	-	-	59.33(33.69)	**0.63** (0.17)	2.05 (3.16)
(c) JIVE + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	83.48 (4.67)	70.65 (10.06)	84.77 (3.59)	60.37 (8.66)	83.48 (4.92)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	2.32 (0.10)	1.87 (0.20)	2.37 (0.10)	1.64 (0.14)	2.30 (0.11)
	$l(\mathbf{V},\hat{\mathbf{V}})$	9.00 (0.00)	1.50(1.97)	9.00 (0.00)	**0.00** (0.00)	8.09 (1.64)
	$\lVert\mathbf{V}-\hat{\mathbf{V}}\rVert_F^2$	5.86 (0.61)	7.50 (2.66)	6.00 (0.59)	**2.55** (0.40)	6.06 (0.76)
	$\lVert\mathbf{B}-\hat{\mathbf{B}}\rVert_F^2$	-	-	10.31 (1.40)	**0.03** (0.05)	0.53 (0.56)

		RRR	**SRRR**	**SPCA**	**GFA**	**SLIDE**
(a) COBS + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	22.58 (1.60)	22.53 (1.59)	19.85 (2.18)	23.96 (1.50)	23.17 (1.60)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.54 (0.02)	0.54 (0.02)	0.46 (0.04)	0.54 (0.02)	0.46 (0.03)
	$l(\mathbf{V},\hat{\mathbf{V}})$	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	8.00 (0.00)	**0.00** (0.00)
	$\lVert\mathbf{V}-\hat{\mathbf{V}}\rVert_F^2$	0.31 (0.03)	0.31 (0.03)	0.31 (0.11)	8.21 (1.77)	4.94 (1.24)
	$\lVert\mathbf{B}-\hat{\mathbf{B}}\rVert_F^2$	3.98 (0.86)	2.84 (0.79)	-	-	-
(b) SIFA + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	22.31 (1.42)	22.26 (1.41)	15.29 (1.51)	14.95 (0.65)	11.54 (0.95)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	0.54 (0.02)	0.54 (0.02)	0.39 (0.03)	0.44 (0.01)	0.31 (0.01)
	$l(\mathbf{V},\hat{\mathbf{V}})$	9.00 (0.00)	9.00 (0.00)	9.00 (0.00)	8.84(0.49)	**0.00** (0.00)
	$\lVert\mathbf{V}-\hat{\mathbf{V}}\rVert_F^2$	0.31 (0.03)	0.31 (0.02)	0.27 (0.15)	7.90 (2.14)	5.10 (2.20)
	$\lVert\mathbf{B}-\hat{\mathbf{B}}\rVert_F^2$	4.04 (0.81)	2.88 (0.77)	-	-	-
(c) JIVE + Small var	$\angle_P(\mathbf{V},\hat{\mathbf{V}})$	87.40 (1.96)	87.41 (1.99)	84.42 (3.84)	-	**34.81** (8.25)
	$d_{\mathcal{G}}(\mathbf{V},\hat{\mathbf{V}})$	2.74 (0.07)	2.74 (0.07)	2.40 (0.10)	-	**0.61** (0.16)
	$l(\mathbf{V},\hat{\mathbf{V}})$	9.00 (0.00)	9.00 (0.00)	9.00 (0.00)	-	11.94 (0.34)
	$\lVert\mathbf{V}-\hat{\mathbf{V}}\rVert_F^2$	7.24 (0.31)	7.24 (0.30)	6.05 (0.66)	-	-
	$\lVert\mathbf{B}-\hat{\mathbf{B}}\rVert_F^2$	16.45 (1.12)	15.28 (1.00)	-	-	-

Table 4: Low-dimensional results one. $n = 200, p = 400$. Small variances in unknown source. The means (standard deviations) of evaluation criteria, from 100 repetitions, are presented in the table. The bold number indicates the best result.

for most cases, as confirmed by smaller principal angles $\angle_P(\mathbf{V}, \hat{\mathbf{V}})$ (in degrees) and $d_{\mathcal{G}}(\mathbf{V}, \hat{\mathbf{V}})$. The only exception is under low-dimensional setting scenario (c) with small variances in unknown source. This situation has the smallest signal strength among all cases. Thus most of methods failed in this situation. The failure of AJIVE and SIFA in Scenario (a) is expected since the partially joint structure is not allowed in their estimating procedures. This results in a misidentification of noisy directions as a $\hat{\mathbf{v}}$, and the large values in the principal angles.

The permuted Hamming distance $l(\mathbf{V}, \hat{\mathbf{V}})$ measures the quality of block structure identification. Overall SLIDE provides the best group identification as its method mostly focuses on group structure retrieving. But our method is also accurate in identifying the structures most of the time. For Scenario (a), both high-dimentional and low-dimensional, only up to 6% of time, at most 1 block is misidentified and pretty perfect identification for the rest. The signal variances in (c) are much smaller than those in (a) and (b), and the block identification quality of COBS deteriorates. The (Sup)SVD, (Sparse) Reduced-Rank Regression, and Sparse PCA do not reflect any group structure (Sparse PCA works for variable selection but not group selection), and all components are treated as full-joint components. Note also that for SIFA and AJIVE we have supplied the exact number of joint and individual components, thus good performances of those methods in Scenarios (b) and (c) in terms of the block identification are expected. On the other hand, in Scenario (a), these methods completely fail in identifying the partially joint structure.

In terms of the squared loss, $\|\mathbf{V} - \hat{\mathbf{V}}\|_F^2$, COBS also provides the best estimates. Small squared loss indicates that not only the column spaces of $\mathbf{V}$ and $\hat{\mathbf{V}}$ are close but also the order of components in $\hat{\mathbf{V}}$ is well-identified according to the identifiability conditions we imposed in Proposition 1. Although SLIDE is perfect in group identification overall, but it tends to deviate from true order of the components. We also remark that the implicit models of SIFA and AJIVE assume larger variances in joint components than in individual components, which is not the case in Scenario (a). In (a), $\hat{\mathbf{v}}_1$ of SIFA and AJIVE are in fact estimators of either $\mathbf{v}_3$ or $\mathbf{v}_4$, resulting in large squared loss.

While COBS provides more accurate estimates of $\mathbf{B}$ than SupSVD and SIFA, we note that in Scenario (c) with large samples, even from the pure noise $\mathbf{X}$, independent with the primary data, COBS accurately provides $\mathbf{X}\hat{\mathbf{B}} \approx \mathbf{0}$.

5.0 Real data application

The proposed COBS factorization is applied in the exploration of multi-source genomic data and of multi-modal image-feature data.

5.1 TCGA breast cancer data

The breast cancer data set we use is a part of TCGA project [34], and contains multiple aspects of genetic information on a common set of subjects. In particular, the measurements for each subjects are grouped into the gene expression (GE), methylation (Me) and copy number variation (CNV). The goal of the analysis is to reveal the underlying variation patterns across these multiple sets of genetic data, potentially driven by the subtypes of breast cancer.

The data are preprocessed as done in [12]. This gives three blocks of variables (5125 genes (variables) for GE, 5036 for Me and 6115 for CNV) for a common sample of size $n = 770$. We further reduce dimensions by filtering in 200 genes with the largest standard deviations in each block, and form the primary multi-block data $\mathbf{Y} = (\mathbf{Y}_{\mathrm{GE}}, \mathbf{Y}_{\mathrm{Me}}, \mathbf{Y}_{\mathrm{CNV}})^T$ of size $n \times p$, where $p = 600$. The primary multi-block data set is then scaled to have zero mean and unit standard deviation. The cancer subtypes play the role of the covariate. There are five subtypes: Basal (133), Her2 (51), LumA (394), LumB (161), and Normal (31), where the class sizes are also given in the parenthesis. The rows of the covariate matrix $\mathbf{X}$ are binary; $(1, 0, 0, 0, 0)$ if the corresponding subject is labelled Basal, $(0, 1, 0, 0, 0)$ for Her2, and so on.

The proposed COBS algorithm is sequential in the sense that the rank-1 components are sequentially estimated. Specifying the overall rank of the signal, r in (2.2), is not necessary in probing the major structure of the multi-block data, which typically appears in the first few components with large variances.

We set $\alpha_v = 0$ (that is, we only seek block-wise sparsity) for demonstration. Other choices of α_v provide similar results for this data set, as detailed in the appendix. The tuning parameter λ_v determines the group structure, and is chosen by the BIC. Since there is one block of covariates, we fixed $\alpha_b = 1$, and λ_b is chosen by the BIC for the first component, which is then used for all subsequent layers. We highlight a few results from the COBS factorization on this dataset.

First, the identified block structure reveals that the first few components are either individual or partially joint. The block structure is reflected in the non-zero patterns of the loadings. For illustration, the estimated loadings of the first five components are

$$\hat{\mathbf{V}}_{1:5} = \begin{pmatrix} \hat{\mathbf{V}}_{1,\mathrm{GE}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \hat{\mathbf{V}}_{5,\mathrm{GE}} \\ \hat{\mathbf{V}}_{1,\mathrm{Me}} & \mathbf{0} & \hat{\mathbf{V}}_{3,\mathrm{Me}} & \mathbf{0} & \hat{\mathbf{V}}_{5,\mathrm{Me}} \\ \mathbf{0} & \hat{\mathbf{V}}_{2,\mathrm{CNV}} & \mathbf{0} & \hat{\mathbf{V}}_{4,\mathrm{CNV}} & \mathbf{0} \end{pmatrix}. \tag{5.1}$$

The first component represents a partially joint variation among GE and Me. The second component is an individual pattern in CNV only. Not shown in (5.1) is that the variance of the first factor is further decomposed into the covariate effect, explaining about 85%, and the unknown factor effect, explaining 15%, while the second factor is mostly about the unknown factor with 90% of the variance. This is further demonstrated in the scatterplot of the estimated scores of the first two components; see Fig. 2.

The scatterplot is appended by the b_{ij} estimates in $\widehat{\mathbf{B}}$ and its confidence interval, obtained by [32]. As expected, the subtypes are main drivers of the variation in the first component, with significantly different b_{i1}'s. In the second component, the subtypes are not significantly different.

Analyzing the pattern in (5.1) also reveals that GE and Me blocks are more related with each other than with the CNV block. The correlation network for the variables in respective blocks, contained in the appendix, confirms that variables in CNV have only weak associations to the variables in GE and Me.

Additionally, we decompose the total variation in the primary data into the variation due to the signal and noise. For this, we estimated $r = 62$ components, which covers about 80% of the total variance, and the analysis of variance is from all 62 components combined. The variation in the signal is decomposed into the effect of each subtype and the unknown factor effect. See Table 5. While the covariate explains less than 30% of total variation in

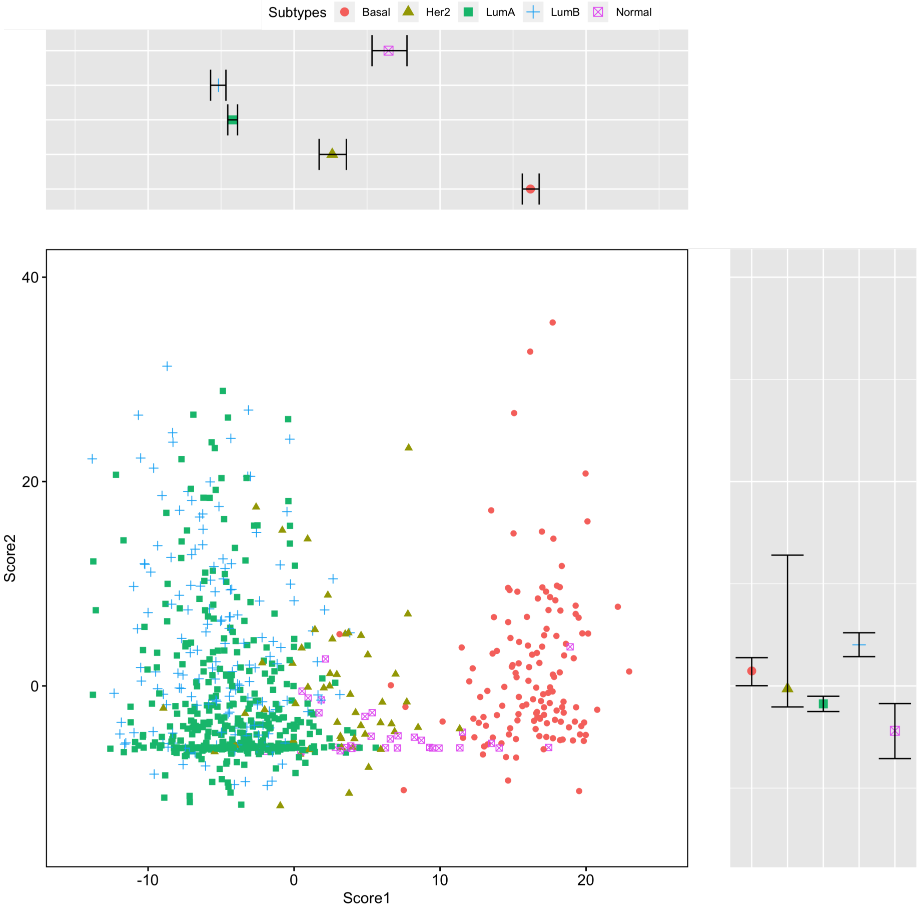


Figure 2: The first two component scores of the breast cancer data. Component 1 scores are mainly driven by the covariate effect, affecting the GE and Me blocks of variables. Component 2 scores are mostly from unknown factors, affecting the CNV block of variables. Also shown is the estimates **B** with 95% confidence intervals.

		Basal	**Her2**	**LumA**	**LumB**	**Normal**	**Total**
Gene Expression	**Supervision**	60.43%	7.78%	18.74%	9.80%	3.25%	26.84%
	Unknown	-	-	-	-	-	41.98%
	Noise	-	-	-	-	-	31.18%
Methylation	**Supervision**	55.60%	2.98%	10.66%	20.11%	10.65%	12.19%
	Unknown	-	-	-	-	-	46.67%
	Noise	-	-	-	-	-	41.14%
CNV	**Supervision**	17.29%	30.33%	16.84%	29.72%	5.81%	7.97%
	Unknown	-	-	-	-	-	79.30%
	Noise	-	-	-	-	-	12.73%
Overall:		**Supervision:** 15.67%		**Unknown:** 55.98%		**Noise:** 28.35%	

Table 5: TCGA breast cancer data: The proportion of variance explained proportion in terms of each component (i.e., supervision, unknown sources, and noise) in COBS model for each individual block and the overall multi-block data (Sum to 1 in terms of the "Total" for each block and "Overall" for the concatenated multi-block data sets). The variation in supervision part is further separated into different tumor subtype for each block.

each block, the mean effect of Basal explains most of the variation for the GE and Me block. In the CNV block, the mean effects of Her2 and LumB stand out, but they are dwarfed by the variation from unknown sources.

Next, we present more results by comparing COBS with other competing methods, such as SLIDE, AJIVE, GFA, SupSVD, and SIFA for the analysis of breast cancer data. We divide the data into half training and half testing. 50% of each cancer subtypes is sampled from the original samples. For subtypes with odd number samples, we round the number up. This results in $n_{\text{train}} = 387$ and $n_{\text{test}} = 383$. Using the training set, we apply each method to get the new variation directions $\hat{\mathbf{V}}_{\text{train}}$. Then the testing primary data is projected onto $\hat{\mathbf{V}}_{\text{train}}$, called the testing score (i.e., $\mathbf{Y}_{\text{test}}\hat{\mathbf{V}}_{\text{train}}$). All the comparisons are based on the testing score from each method.

We firstly sort the testing score components by their variances in descending order.

	COBS	SLIDE	AJIVE	GFA	SupSVD	SIFA
Comp. 1	72.73	76.21	65.20	64.07	71.95	69.75
Comp. 2	63.84	69.04	52.55	62.32	55.20	42.61
Comp. 3	42.63	44.71	42.80	53.73	46.16	37.01
Comp. 4	33.22	41.84	29.55	41.84	39.45	31.99
Comp. 5	28.31	41.47	28.64	35.07	31.40	31.69

Table 6: The variances of the testing score $\mathbf{Y}_{\text{text}}\hat{\mathbf{V}}_{\text{train}}$ for the top 5 components that have the largeset variances from each method.

One remark is that the group structures of the top 5 components from COBS follow the same pattern as the structure shown in 5.1. Table 6 summarizes the variances of the top 5 components from each method. For the first two components, SLIDE, AJIVE, GFA, SupSVD and SIFA all have larger variances from full-joint structures. With group structure identification considered, COBS still performs noticeably well.

Given this result, we perform hierarchical clustering by using the Euclidean distance and Ward's minimum variance method [26, 33], based on the first two components that have the largest variances from each method. To evaluate the performance of the clustering result, we use the Adjusted-Rand Index [11] to assess the cross-tabulation of the cluster groups (C) and the true group labels (L). The ARI is defined as:

$$\text{ARI} = \frac{\sum_{c=1}^{C}\sum_{l=1}^{L}\binom{N_{cl}}{2} - N_C N_L/\binom{N}{2}}{(N_C + N_L)/2 - N_C N_L/\binom{N}{2}}, \tag{5.2}$$

where C and L indicate cluster groups and true group labels accordingly, N_{cl} is the number of observations that is clustered in group c with true label l, $N_C = \sum_{c=1}^{C}\binom{N_{c\cdot}}{2}$, and $N_L = \sum_{l=1}^{L}\binom{N_{\cdot l}}{2}$. The closer the ARI to 1, the better the group cluster is. We use the mclust::adjustedRandIndex() [5] in R to calculate the ARI. The ARI from each method is, COBS 0.37, SLIDE 0.26, GFA 0.29, AJIVE 0.24, SupSVD 0.34, and SIFA 0.04. It turns out the ARI of all methods did not present good enough clusters. It is understandable here

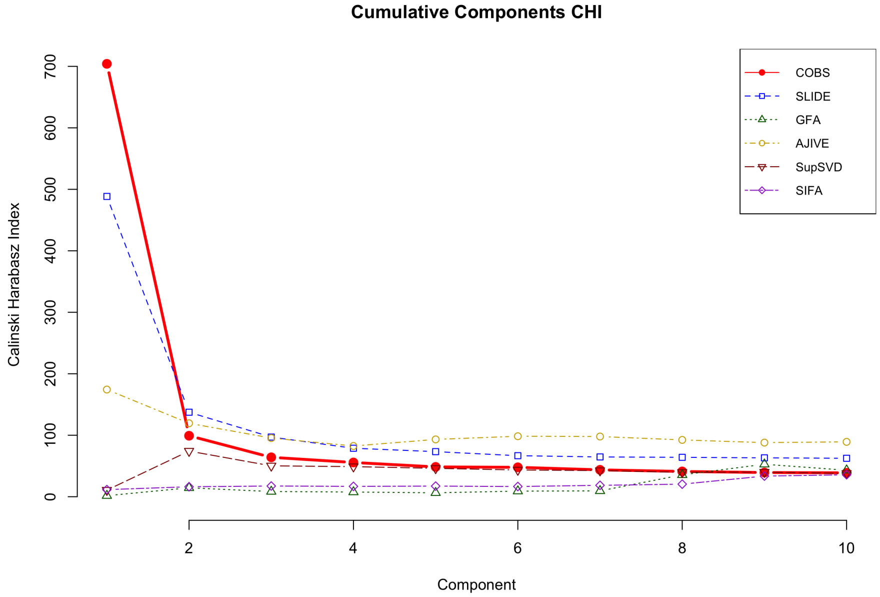


Figure 3: The Calinski-Harabasz Index of cumulative number of components from different methods. It presents that COBS factorization has the advantage for group separation with a smaller number of components. But SLIDE also performs comparably well in terms of CHI.

because of the large overlap between LumA and LumB subtypes, and also the small test samples in Normal and Her2 subtypes. But based on this, COBS still performs relatively better than the other methods in seperating LumA/LumB and Basal, which also agrees with the score scatters shown in Figure 2.

Moreover, for the illustration of the group separation of the testing score, we have extracted $r = 10$ components from each method, and computed the Calinski-Harabasz Index (CHI) [1], proportional to the ratio of sum of squares between groups (covariate effect) and the sum of squares of the residuals (unknown factor effects), cumulatively with the number of components. Figure 3 presents that the components from COBS factorization tends to have larger values of CHI, better group separation of the breast cancer subtypes, when the number of components is small.

5.2 Image-feature analysis of melanocytic lesions

In histopathology, digital image analysis of melanocytic lesions is often summarized by image features [25]. These image features are naturally grouped into color-related features and shape-related features, among others. Each skin slide results in multiple values of the features, and in this data example, we use the mean and standard deviation of each feature, computed from the image-feature data of size $n = 348$, as discussed in [25]. There are 18 features (variables) related to color, resulting in two blocks of variables called *Mean-Color* and *Sd-Color*; 13 shape-related features give additional two blocks, *Mean-Shape* and *Sd-Shape*. Each observation is labeled either nevi or melanoma, and we use this information as a covariate.

The data set has four blocks of variables, with block sizes 18, 18, 13 and 13. Each variable is transformed so that their distributions are approximately standard normal. The COBS factorization is applied with $\lambda_b = 0$, i.e., the ordinary least square for the estimation of $\mathbf{B}$, as there is only one binary variable for the covariate. Aiming at both the identification of the block structure and variable selection (within each block), we set $\alpha_v = 0.5$. The tuning parameter λ_v is chosen by the BIC.

The result of analysis is summarized in Table 7, for the first six estimated components. The proposed method not only identifies mostly partial-joint components but also selects a few variables in each block. Note that since Mean-Color and Sd-Color (or Mean-Shape and Sd-Shape) are computed from the same features, it is expected that they are often coupled.

We have further compared the COBS factorization with some of the competing methods, as discussed in Chapter 4 and in breast cancer data analysis. We also sort the score components by their variances in descending order from each method. Table 8 presents the five components with the largest variances. Through investigation of the results, we find that SLIDE, AJIVE, GFA, SupSVD and SIFA all provide largest variance from fully-joint components, however COBS has largest variance in its original first component that is partially-joint between Mean-Shape and Sd-Shape.

Additionally, the degrees of separation of the two groups (nevi and melanoma) are compared, since this binary information is used as a covariate in fitting COBS, SIFA, and

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5	Comp. 6
Mean-Color	0	5	12	11	6	6
Sd-Color	0	5	8	10	11	3
Mean-Shape	11	0	2	4	0	5
Sd-Shape	5	0	0	0	0	2

Table 7: The numbers of non-zero loadings, counted for each block and each component, from the analysis of melanoma data are shown to reveal the block structure identification and variable selection result. Mean-Color and Sd-Color blocks have each 18 variables and the other two blocks have 13 variables each.

COBS	SLIDE	AJIVE	GFA	SupSVD	SIFA
8.22	8.89	7.34	9.30	9.77	10.77
6.08	7.86	6.00	6.95	6.09	4.60
5.52	7.20	4.37	6.69	5.88	4.52
4.29	6.47	4.37	5.63	4.10	4.44
3.90	5.21	4.06	4.52	3.99	3.97

Table 8: Variances in descending order of five score components for each method.

SupSVD. For this, we have extracted $r = 13$ components (for 80% of total variation), and computed the CH Index from each component and also from the cumulative components output from COBS. The CHI is also computed in such two ways from component scores estimated from the other methods. Figure 4 illustrates the sorted CHI in descending order for each method and Figure 5 is the CHI for the cumulative components output from each method (AJIVE and SIFA output full-joint components in the beginning). In Figure 4, SLIDE and GFA are better in terms of the first component that has the largest CHI but all

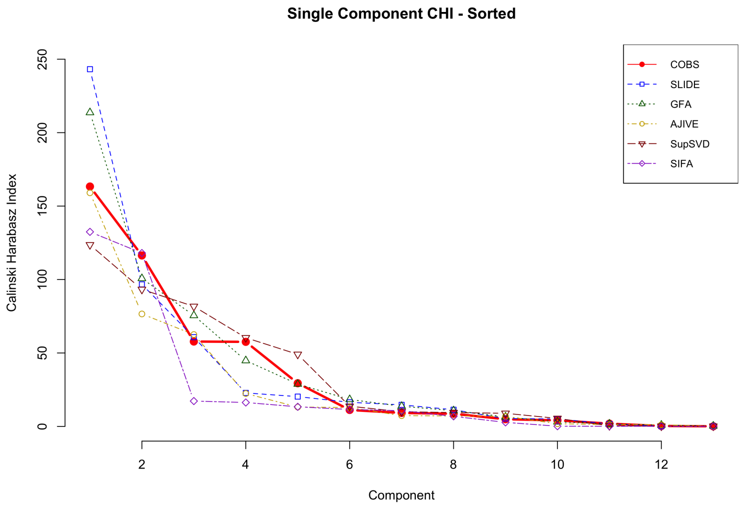


Figure 4: The CH-Index of each single component. SLIDE and GFA are better with the top component. All the other methods perform relevant similar to each other.

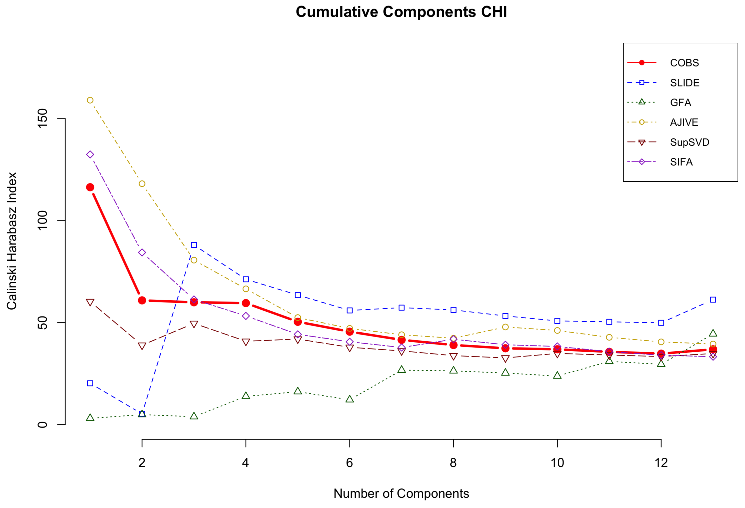


Figure 5: The CH-Index of cumulative components. AJIVE seems to be overall better in distinguishing melanoma and nevi in this situation.

the other methods are pretty similar to each other. But contrarily, GFA does not perform well in Figure 5. Similar situation happens for AJIVE and SIFA, where AJIVE and SIFA are better compared with the other methods in Figure 5 but they do not stand out in Figure 4. In both figures, COBS and SupSVD maintain relevant stable performance in distingushing Melanoma and Nevi, with COBS performs slightly better.

We conclude from Table 8, Figure 4 and Figure 5 together, that COBS performs overall well in identifying the group variation structures. At the meantime, COBS explains larger variances in the primary multi-block data and distinguishes the skin melanocytic lesions through structural score components.

6.0 Restricted maximum likelihood estimation in COBS

In this chapter, we will discuss another method, restricted maximum likelihood (ReML) [9], to estimate the variances (i.e., σ_e^2 and Σ_F) in COBS model 2.2 and then compare with the estimates in equation 3.6 that are calculated based on MLE. ReML estimates are commonly used in linear models to produce unbiased estimates for the variance terms. In the first section we will introduce how to apply ReML in COBS model to estimate σ_e^2 and the elements in Σ_F. Then we will then show that ReML estimates for COBS are unbiased, however the MLE elements in Σ_F as equation 3.6 are biased.

6.1 ReML estimates in COBS

The idea in ReML is to maximize a modified likelihood function that is free from the mean in the original distribution so that the estimates will not be biased due to the degree of freedom in the estimator. So we will apply this idea into COBS model. Without loss of generality, the ReML estimates are derived from the full model 2.2 with rank r, since the rank-one model will just be a special case by setting $r = 1$.

Consider the full model 2.2 and the assumptions regarding to the unknonwn source $\mathbf{F}$ and error $\mathbf{E}$ disucussed in Chapter 2. We know that each row of $\mathbf{Y}$ given $\mathbf{X}$ is multivariate normal with mean $\mathbb{E}(\mathbf{Y}_i|\mathbf{X}_i) = \mathbf{X}_i\mathbf{B}\mathbf{V}^T$ and covariance $\Sigma = \sigma_e^2\mathbb{I}_p + \mathbf{V}\Sigma_F\mathbf{V}^T$, for $i = 1, \ldots, n$. To make it easier to derive ReML next, we rewrite the distribution of $\mathbf{Y}$ given $\mathbf{X}$ in the matrix normal distribution format as

$$\mathbf{Y} \sim \mathcal{MVN}\Big(\mathbf{XBV}^T, \mathbb{I}_n, \Sigma\Big) \equiv \text{vec}(\mathbf{Y}) \sim \mathcal{N}_{np}\Big(\text{vec}(\mathbf{XBV}^T), \mathbb{I}_n \otimes \Sigma\Big),$$

where vec($\mathbf{Y}$) means the vectorization of the rows in $\mathbf{Y}$ given supervision data $\mathbf{X}$, and $\otimes$ indicates the Kronecker product. The original log-likelihood function of $\mathbf{Y}$ depends on the mean part $\mathbf{XBV}^T$. ReML is to maximize the log-likelihood function of the error contrast term $\mathbf{SY}$ for $\mathbf{S}$ such that $\mathbf{SX} = \mathbf{0}$. We define $\mathbf{S}$ as a full rank matrix with dimension $k \times n$.

Then based on the distribution of $\mathbf{Y}$, we can write the matrix normal distribution of $\mathbf{SY}$ as $\mathbf{SY} \sim \mathcal{MVN}\Big(\mathbf{0}, \mathbf{SS}^T, \Sigma\Big)$, so that the -2 log-likelihood function of $\mathbf{SY}$ is

$$\ell\big(\theta(\Sigma)\big) = k\log(|\Sigma|) + \text{trace}[\Sigma^{-1}(\mathbf{SY})^T(\mathbf{SS}^T)^{-1}(\mathbf{SY})] + c(\mathbf{S}),$$

where $c(\mathbf{S}) = np\log(2\pi) + p\log(|\mathbf{SS}^T|)$.

We derive $\mathbf{S}$ from the eigen-decomposition of $\mathbb{I}_n - \mathbf{H}$, where $\mathbf{H}$ is the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. With full-rank $\mathbf{X}$, $\mathbb{I}_n - \mathbf{H}$ is idempotent with rank $n-q$. The eigenvalues of $\mathbb{I}_n - \mathbf{H}$ will be either 1 or 0. We define $\mathbf{S}^T$ as the first $k = n-q$ eigen-vectors where the corresponding eigenvalues are 1. Then it follows that $\mathbf{S}^T\mathbf{S} = \mathbb{I}_n - \mathbf{H}$ and $\mathbf{SS}^T = \mathbb{I}_k$. And we have shown in chapter 3 that $|\Sigma| = (\sigma_e^2)^p\prod_{j=1}^r(1+\frac{\sigma_j^2}{\sigma_e^2})$ and $\Sigma^{-1} = \frac{1}{\sigma_e^2}\mathbb{I}_p - \sum_{j=1}^r \frac{\sigma_j^2}{\sigma_e^2(\sigma_e^2+\sigma_j^2)}\mathbf{v}_j\mathbf{v}_j^T$. Then the -2 log-likelihood function of $\mathbf{SY}$ can be further derived as

$$\begin{aligned}\ell\big(\theta(\Sigma)\big) = & \frac{\text{trace}[\mathbf{Y}^T(\mathbb{I}_n - \mathbf{H})\mathbf{Y}]}{\sigma_e^2} - \sum_{j=1}^r \text{trace}[\frac{\sigma_j^2}{\sigma_e^2(\sigma_e^2+\sigma_j^2)}(\mathbf{Yv}_j)^T(\mathbb{I}_n - \mathbf{H})(\mathbf{Yv}_j)] \\ & + k(p-r)\log(\sigma_e^2) + k\sum_{j=1}^r \log(\sigma_e^2 + \sigma_j^2) + c, \end{aligned} \tag{6.1}$$

where $c = np\log(2\pi) + p\log(k)$, $k = n-q$. Taking partial derivatives w.r.t σ_e^2 and σ_j^2 for $j = 1, \ldots, r$ from equation 6.1, the ReML estimates are

$$\begin{aligned}\hat{\sigma}^2_{e[\text{ReML}]} &= \frac{\text{trace}[\mathbf{Y}^T(\mathbb{I}_n - \mathbf{H})\mathbf{Y}] - \sum_{j=1}^r \text{trace}[(\mathbf{Yv}_j)^T(\mathbb{I}_n - \mathbf{H})(\mathbf{Yv}_j)]}{k(p-r)} \\ \hat{\sigma}^2_{j[\text{ReML}]} &= \frac{\text{trace}[(\mathbf{Yv}_j)^T(\mathbb{I}_n - \mathbf{H})(\mathbf{Yv}_j)]}{k} - \hat{\sigma}^2_{e[\text{ReML}]} \end{aligned} \tag{6.2}$$

To apply ReML estimates above in COBS for the rank-one model, we substitute equation 3.6 with equation 6.2 setting $r = 1$ at each layer in the rank-one model as described in section 3.2. In terms of the full model, we can directly update σ_e^2 and elements in Σ_F using equation 6.2 after we have all layers updated.

6.2 ReML vs. MLE in COBS

In this section, we will show that ReML estimates for the variances in COBS are unbiased. Before we start the proof, we need to clarify the distribution of $\mathbf{Y}\mathbf{v}_j, j = 1, \ldots, r$ first, that $\mathbf{Y}\mathbf{v}_j$ is normally distributed with mean $\mathbf{XB}$ and variance $\sigma_e^2 + \sigma_j^2$. It follows that:

$$\begin{aligned}
\mathbb{E}(\text{trace}[\mathbf{Y}^T(\mathbb{I}_n - \mathbf{H})\mathbf{Y}]) &= \text{trace}[\mathbb{E}(\mathbf{Y}^T(\mathbb{I}_n - \mathbf{H})\mathbf{Y})] \\
&= \text{trace}[\mathbb{E}(\mathbf{Y})^T(\mathbb{I}_n - \mathbf{H})\mathbb{E}(\mathbf{Y})] + \text{trace}[(\mathbb{I}_n - \mathbf{H}) \otimes \Sigma] \\
&= 0 + k \text{ trace}(\Sigma) \\
&= kp\ \sigma_e^2 + k\sum_{j=1}^{r} \sigma_j^2 \qquad (6.3)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(\text{trace}[(\mathbf{Y}\mathbf{v}_j)^T(\mathbb{I}_n - \mathbf{H})(\mathbf{Y}\mathbf{v}_j)]) &= \text{trace}\Big[\mathbb{E}[(\mathbf{Y}\mathbf{v}_j)^T(\mathbb{I}_n - \mathbf{H})(\mathbf{Y}\mathbf{v}_j)]\Big] \\
&= 0 + \text{trace}[(\mathbb{I}_n - \mathbf{H})(\sigma_e^2 + \sigma_j^2)] \\
&= k(\sigma_e^2 + \sigma_j^2). \qquad (6.4)
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}(\hat{\sigma}^2_{e[\text{ReML}]}) &= \frac{kp\sigma_e^2 + k\sum_{j=1}^{r}\sigma_j^2 - \sum_{j=1}^{r} k(\sigma_e^2 + \sigma_j^2)}{k(p-r)} = \sigma_e^2 \qquad (6.5) \\
\mathbb{E}(\hat{\sigma}^2_{j[\text{ReML}]}) &= \frac{k(\sigma_e^2 + \sigma_j^2)}{k} - \sigma_e^2 = \sigma_j^2,\ j = 1, \ldots, r. \qquad (6.6)
\end{aligned}$$

All ReML Estimates for COBS are clearly unbiased for either rank-one model or the full model.

In Section 3.2 equation 3.6, the estimation of the variances follows the MLE scheme at each layer. Since there is no closed form solution for $\mathbf{b}$ at each layer, we will discuss a special case where $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{v}$ that is estimated from maximizing the unpenalized likelihood function treating $\mathbf{v}$ as given. Under this special case, $\hat{\sigma}_f^2$ shown in equation 3.6 is biased but $\hat{\sigma}_e^2$ is unbiased. To clarify the notation, here $\mathbf{v}$ indicates the rank-one model parameter, which is equivalent as $\mathbf{v}_j$ in the full model. We will show the biasedness of $\hat{\sigma}_f^2$ for the rank-one model here.

Plugging $\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{v}$ into equation 3.6, we have

$$\hat{\sigma}_e^2 = \frac{\|\mathbf{Y} - \mathbf{H}\mathbf{Y}\mathbf{v}\mathbf{v}^T\|_F^2 - \|\mathbf{Y}\mathbf{v} - \mathbf{H}\mathbf{Y}\mathbf{v}\|_2^2}{n(p-1)}, \ \hat{\sigma}_f^2 = \frac{\|\mathbf{Y}\mathbf{v} - \mathbf{H}\mathbf{Y}\mathbf{v}\|_2^2}{n} - \hat{\sigma}_e^2. \tag{6.7}$$

We still find the expection of the numerators first.

$$\begin{aligned}
\mathbb{E}(\|\mathbf{Y} - \mathbf{H}\mathbf{Y}\mathbf{v}\mathbf{v}^T\|_F^2) &= \mathbb{E}[\text{trace}(\mathbf{Y} - \mathbf{H}\mathbf{Y}\mathbf{v}\mathbf{v}^T)^T(\mathbf{Y} - \mathbf{H}\mathbf{Y}\mathbf{v}\mathbf{v}^T)] \\
&= \mathbb{E}[\text{trace}(\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{H}\mathbf{Y}\mathbf{v}\mathbf{v}^T)] \\
&= \mathbb{E}[\text{trace}(\mathbf{Y}^T\mathbf{Y})] - \mathbb{E}[\text{trace}\big((\mathbf{Y}\mathbf{v})^T\mathbf{H}(\mathbf{Y}\mathbf{v})\big)] \\
&= \text{trace}((\mathbf{X}\mathbf{b}\mathbf{v}^T)^T(\mathbf{X}\mathbf{b}\mathbf{v}^T)) + \text{trace}(\mathbb{I}_n \otimes \Sigma) \\
&\quad - \text{trace}((\mathbf{X}\mathbf{b})^T(\mathbf{X}\mathbf{b})) - \text{trace}(\mathbf{H}(\sigma_e^2 + \sigma_j^2)) \\
&= n(p\sigma_e^2 + \sigma_f^2) - q(\sigma_e^2 + \sigma_f^2)
\end{aligned} \tag{6.8}$$

and

$$\begin{aligned}
\mathbb{E}(\|\mathbf{Y}\mathbf{v} - \mathbf{H}\mathbf{Y}\mathbf{v}\|_2^2) &= \mathbb{E}[(\mathbf{Y}\mathbf{v} - \mathbf{H}\mathbf{Y}\mathbf{v})^T(\mathbf{Y}\mathbf{v} - \mathbf{H}\mathbf{Y}\mathbf{v})] \\
&= \mathbb{E}[(\mathbf{Y}\mathbf{v})^T(\mathbf{Y}\mathbf{v}) - (\mathbf{Y}\mathbf{v})^T\mathbf{H}(\mathbf{Y}\mathbf{v})] \\
&= \mathbb{E}\big(\text{trace}[(\mathbf{Y}\mathbf{v})^T(\mathbb{I}_n - \mathbf{H})(\mathbf{Y}\mathbf{v})]\big) \\
&= (n-q)(\sigma_e^2 + \sigma_f^2)
\end{aligned} \tag{6.9}$$

Then it follows that

$$\mathbb{E}(\hat{\sigma}_e^2) = \frac{n(p\sigma_e^2 + \sigma_f^2) - q(\sigma_e^2 + \sigma_f^2) - (n-q)(\sigma_e^2 + \sigma_f^2)}{n(p-1)} = \sigma_e^2 \tag{6.10}$$

$$\mathbb{E}(\hat{\sigma}_f^2) = \frac{(n-q)(\sigma_e^2 + \sigma_f^2)}{n} - \sigma_e^2 = \sigma_f^2 - \frac{q}{n}(\sigma_e^2 + \sigma_f^2). \tag{6.11}$$

So in this special case, the MLE of σ_f^2 for rank-one model is biased and the MLE of σ_e^2 remains unbiased. This special case can also be generalized to full-rank model estimation as the equation to estimate the elements $\sigma_j^2, j = 1, \ldots, r$ in Σ_F given $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{V}$ follows same format as what we have shown above. Refer to section 3.3 for the details of multi-layer estimation for COBS model.

Overall, we conclude that ReML can be an alternative method for COBS variances estimation. Without relying on the mean part, the estimates from the error contrast part in COBS model are all unbiased, which is also an advantage over the use of MLE of the variances.

7.0 Summary

We have introduced a covariate-driven, block-wise structured factorization for integrative analysis of multi-block data. Unlike most other low-rank factorization method in literature, our conceived model is flexible to incorporate any form of block structure, especially the partially joint components. The data-driven distinction between full-joint, partial-joint and individual structures by thresholding is seen to work well in simulation studies and real data analysis. We in addition incorporate covariate effects which potentially drive the segmented factors. The proposed algorithm is computationally efficient for moderately large dataset. This leaves the development of a more efficient algorithm suitable for ultra-high dimensional data blocks as a future topic of investigation. Finally, an important question we have not discussed is whether the identified block structure is really there. Inference on the COBS factorization involves two important sub-questions of 1) selection consistency of the block-wise structure and 2) post-selection inference of $\mathbf{B}$ and $\mathbf{V}$. These types of questions are primarily being answered in the context of regression [30], but have not been addressed in the integrative dimension reduction context, as observed by [20]. Theoretical tools developed for sparse PCA [17] and the usual lasso estimator may be used to provide the consistency, post-selection test procedures and intervals for the coefficients and loadings.

Appendix A Technical details

A.1 Proofs of propositions 1 and 2

Proof of Proposition 1.

Take any two parameter-set $\theta_1 = (\mathbf{V}_1, \mathbf{B}_1, \Sigma_1, \tau_1^2)$ and $\theta_1 = (\mathbf{V}_2, \mathbf{B}_2, \Sigma_2, \tau_2^2)$ and assume that $F_{\theta_1}(\mathbf{Y}) = F_{\theta_1}(\mathbf{Y})$. We will verify $\theta_1 = \theta_2$ if Conditions (i)—(iii) hold. The equality of the distributions implies that (a) $\mathbb{E}_{\theta_1}(\mathbf{Y}) = \mathbb{E}_{\theta_2}(\mathbf{Y})$ and (b) $\mathbb{E}_{\theta_1}\text{trace}(\mathbf{Y}\mathbf{Y}^T) = \mathbb{E}_{\theta_2}\text{trace}(\mathbf{Y}\mathbf{Y}^T)$ among others. These in turn leads that, from (a),

$$\mathbf{X}\mathbf{B}_1\mathbf{V}_1^T = \mathbf{X}\mathbf{B}_2\mathbf{V}_2^T, \tag{A.1}$$

and from (b)

$$\mathbf{V}_1 \left[\mathbf{B}_1^T\mathbf{S}_X\mathbf{B}_1 + \Sigma_1\right] \mathbf{V}_1^T + \tau_1^2\mathbb{I}_p = \mathbf{V}_2 \left[\mathbf{B}_2^T\mathbf{S}_X\mathbf{B}_2 + \Sigma_2\right] \mathbf{V}_2^T + \tau_2^2\mathbb{I}_p. \tag{A.2}$$

The uniqueness of eigenvalue decomposition of symmetric matrices, together with Condition (i) leads that $\tau_1^2 = \tau_2^2$ and $\text{span}(\mathbf{V}_1) = \text{span}(\mathbf{V}_2)$ from (A.2). Moreover, the sets of eigenvalues of $\mathbf{M}_1 = \mathbf{B}_1^T\mathbf{S}_X\mathbf{B}_1 + \Sigma_1$ and $\mathbf{M}_2 = \mathbf{B}_2^T\mathbf{S}_X\mathbf{B}_2 + \Sigma_2$ must be equal. By Condition (ii), both the eigenvalues of $\mathbf{M}_1$ and $\mathbf{M}_2$ are exactly their diagonal values. Since these eigenvalues are distinct and in a particular order, we have $\mathbf{V}_1 = \mathbf{V}_2$. Then, (A.1) becomes $\mathbf{X}\mathbf{B}_1 = \mathbf{X}\mathbf{B}_2$. (Thus, without assuming Condition (iii), $\mathbf{X}\mathbf{B}$ is identifiable.) Replacing $\mathbf{X}\mathbf{B}_1$ by $\mathbf{X}\mathbf{B}_2$ in $\mathbf{M}_1 = \mathbf{M}_2$ gives $\Sigma_1 = \Sigma_2$. Finally, by Condition (iii), $\mathbf{X}^T\mathbf{X}$ is invertible. Thus $\mathbf{X}^T\mathbf{X}\mathbf{B}_1 = \mathbf{X}^T\mathbf{X}\mathbf{B}_2$ leads that $\mathbf{B}_1 = \mathbf{B}_2$. □

Proof of Proposition 2.

Write $\mathbf{Y} = \mathbf{L}\mathbf{D}\mathbf{R}^T$ for the singular value decomposition of $\mathbf{Y}$, Here, $\mathbf{L}$ is the $n \times N$ matrix consisting of orthogonal columns, where $N \leq \min(n, p)$ is the number of positive singular values, and $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_N)$ with $d_i > d_{i+1}$ assumed. Then,

$$\begin{aligned}\max_{\mathbf{v}} \|\mathbf{Y}\mathbf{v} + \mathbf{X}\tilde{\mathbf{b}}\|_2^2 &= \max_{\mathbf{v}} \|\mathbf{D}\mathbf{R}^T\mathbf{v} + \mathbf{L}^T\mathbf{X}\tilde{\mathbf{b}}\|_2^2 + \|(\mathbb{I}_n - \mathbf{L}\mathbf{L}^T)\mathbf{X}\tilde{\mathbf{b}}\|_2^2 \\ &= \max_{\mathbf{w}} \|\mathbf{w} + \mathbf{z}\|_2^2 + c \\ &= \max_{\mathbf{w}} \sum_{i=1}^{N} (w_i + z_i)^2 + c \qquad \text{(A.3)}\end{aligned}$$

where $\mathbf{z} = \mathbf{L}^T\mathbf{X}\tilde{\mathbf{b}}$, $\mathbf{w} = \mathbf{D}\mathbf{R}^T\mathbf{v} \in \mathbb{R}^N$ satisfying $\mathbf{w}^T\mathbf{D}^{-2}\mathbf{w} = \sum_{i=1}^{N} w_i^2/d_i^2 \leq 1$ and c is a constant.

Note that the feasible region for $\mathbf{w}$ is an ellipsoid whose principal axes are along the coordinate axes. If $\mathbf{z} = 0$, then $\tilde{\mathbf{w}} = (d_1, 0, \ldots, 0)^T$ is the maximizer. We now assume $z_i > 0$ for all $i = 1, \ldots, N$ without loss of generality. An inspection of (A.3) leads that the maximizer $\hat{\mathbf{w}}$ should satisfy $\hat{w}_i \geq 0$ for all i and $\mathbf{w}^T\mathbf{D}^{-2}\mathbf{w} = 1$. Introducing the Lagrange multiplier t, the maximizer $\tilde{\mathbf{w}}$ of (A.3) with the constraint $\mathbf{w}^T\mathbf{D}^{-2}\mathbf{w} = 1$ is a stationary point of $A(\mathbf{w}, \phi) = \sum_{i=1}^{N}(w_i + z_i)^2 - t(\sum_{i=1}^{N} w_i^2/d_i^2 - 1)$. The first order condition gives

$$z_i = \left(\frac{t}{d_i^2} - 1\right) w_i,$$

for all i. The maximizer w_i is then $w_i = d_i^2 z_i/(t - d_i^2)$, for a t. Since both z_i and w_i have the same sign, we have $\frac{t}{d_i^2} - 1 > 0$ for all i, which in turn leads to $t > d_1^2$. The desired t is the root of

$$h(t) = \sum_{i=1}^{N} \frac{w_i^2}{d_i^2} - 1 = \sum_{i=1}^{N} \frac{d_i^2 z_i^2}{t - d_i^2} - 1,$$

on $t > d_i^2$. Since $h'(t) < 0$, h is strictly decreasing, and for $t_1 = d_1^2 + d_1 z_1$ and $t_2 = d_1^2 + (\sum_{i=1}^{N} d_i^2 z_i^2)^{1/2}$, $h(t_1) > 0 > h(t_2)$. Thus the unique root $\tilde{t}$ is located on (t_1, t_2). The maximizer $\tilde{\mathbf{w}} = (\tilde{w}_i)$ is $\tilde{w}_i = d_i^2 z_i/(\tilde{t} - d_i^2)$.

The maximizer $\tilde{\mathbf{v}}$ is then $\tilde{\mathbf{v}} = \mathbf{R}\mathbf{D}^{-1}\tilde{\mathbf{w}}$. □

A.2 Estimating algorithm for general Σ_F

In the main article, we have only discussed the estimation of COBS factorization under the special case $\Sigma_F = \text{diag}(\sigma_1^2, \ldots, \sigma_r^2)$. We discuss the modification of the algorithm to the general Σ_F case. To obtain the general Σ_F estimate, we only need to replace σ_j^2 estimation in Section 3.3.

Recall that with the condition $\Sigma_F = \text{diag}(\sigma_1^2, \ldots, \sigma_r^2)$, from the full model with r components, and with $(\hat{\mathbf{v}}_j, \hat{\mathbf{b}}_j)$ plugged in, we estimate σ_0^2 and $\Sigma_F = \text{diag}(\sigma_1^2, \ldots, \sigma_r^2)$ by the maximum likelihood estimates. These are given by $\hat{\sigma}_0^2 = (n(p-r))^{-1}\|\mathbf{Y}\hat{\mathbf{V}}^{\perp}\|_F^2$, where $\hat{\mathbf{V}}^{\perp}$ is the $p \times (p-r)$ matrix formed by an orthonormal basis of the null space of $\hat{\mathbf{V}}$, and $\hat{\sigma}_j^2 = n^{-1}\|\mathbf{Y}\hat{\mathbf{v}}_j - \mathbf{X}\mathbf{b}_j\|_2 - \hat{\sigma}_0^2$, for $j = 1, \ldots, r$.

For the general Σ_f case, the negative log-likelihood function with $(\hat{\mathbf{v}}_j, \hat{\mathbf{b}}_j)$ plugged in is proportional to

$$\ell(\Sigma_f, \sigma_0^2) = n\log(2\pi|\Sigma|) + \sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i\hat{\mathbf{B}}\hat{\mathbf{V}}^T)\Sigma^{-1}(\mathbf{y}_i - \mathbf{x}_i\hat{\mathbf{B}}\hat{\mathbf{V}}^T)^T,$$

where $\Sigma = \sigma_0^2\mathbf{I}_p + \hat{\mathbf{V}}\Sigma_f\hat{\mathbf{V}}^T$. The minimizer of ℓ is then

$$\begin{aligned}\hat{\sigma}_0^2 =& (n(p-r))^{-1}\|\mathbf{Y}\hat{\mathbf{V}}^{\perp}\|_F^2,\\ \hat{\Sigma}_f =& \hat{\mathbf{V}}^T(\mathbf{S}_0 - \hat{\sigma}_0^2\mathbf{I}_p)\hat{\mathbf{V}},\end{aligned}$$

where $\mathbf{S}_0 = n^{-1}\mathbf{Y}^T\mathbf{Y}$.

A.3 Derivation of major equations

A.3.1 Derivation of equation 3.2: the -2 log-likelihood function

Under COBS model assumptions, the log-likelihood function is

$$\begin{aligned}\mathcal{L}[\mathbf{Y}(\mathbf{b}, \mathbf{v}, \sigma_e^2, \sigma_e^2)|\mathbf{X}]) &= \log P(\mathbf{Y}_1, \mathbf{Y}_2, \ldots \mathbf{Y}_n|\mathbf{b}_{\mathbf{1}}, \mathbf{v}, \sigma_e^2, \sigma_f^2)\\ &= -\frac{n}{2}\log(2\pi|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}\mathbf{v}^T)\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}\mathbf{v}^T)^T, \quad \text{(A.4)}\end{aligned}$$

where $\mathbf{\Sigma} = \sigma_e^2\mathbf{I}_p + \sigma_f^2\mathbf{v}\mathbf{v}^T$.

Under the constraint $\|\mathbf{v}\|_2^2 = 1$,

$$\begin{aligned}
|\mathbf{\Sigma}| &= \det(\sigma_e^2\mathbf{I}_p + \sigma_f^2\mathbf{v}\mathbf{v}^T) = (\sigma_e^2)^p \det(\mathbf{I}_p + \frac{\sigma_f^2}{\sigma_e^2}\mathbf{v}\mathbf{v}^T) \\
&= (\sigma_e^2)^p \det(1 + \frac{\sigma_f^2}{\sigma_e^2}\mathbf{v}^T\mathbf{v}) = (\sigma_e^2)^{p-1}(\sigma_f^2 + \sigma_e^2),
\end{aligned}$$

and

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_e^2}\mathbf{I}_p - \frac{\sigma_{f1}^2}{\sigma_e^2(\sigma_{f1}^2 + \sigma_e^2)}\mathbf{v}\mathbf{v}^T.$$

Therefore, it can be found that

$$\begin{aligned}
&(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}\mathbf{v}^T)\mathbf{\Sigma}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}\mathbf{v}^T)^T \\
&\quad = \frac{1}{\sigma_e^2}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}\mathbf{v}^T)(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}\mathbf{v}^T)^T - \frac{\sigma_f^2}{\sigma_e^2(\sigma_f^2 + \sigma_e^2)}(\mathbf{Y}_i\mathbf{v} - \mathbf{X}_i\mathbf{b})(\mathbf{Y}_i\mathbf{v} - \mathbf{X}_i\mathbf{b})^T
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}\mathbf{v}^T)\mathbf{\Sigma}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\mathbf{b}\mathbf{v}^T)^T \\
&= \frac{1}{\sigma_e^2}\|\mathbf{Y} - \mathbf{X}\mathbf{b}\mathbf{v}^T\|_F^2 - \frac{\sigma_f^2}{\sigma_e^2(\sigma_f^2 + \sigma_e^2)}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2.
\end{aligned}$$

Thus, the log-likelihood function A.4 is simplified as

$$\begin{aligned}
&\mathcal{L}[\mathbf{Y}(\mathbf{b}, \mathbf{v}, \sigma_e^2, \sigma_e^2)|\mathbf{X}] \\
&\quad = -\frac{n}{2}\log[2\pi(\sigma_e^2)^{p-1}(\sigma_f^2 + \sigma_e^2)] - \frac{1}{2\sigma_e^2}\|\mathbf{Y} - \mathbf{X}\mathbf{b}\mathbf{v}^T\|_F^2 + \frac{\sigma_f^2}{2\sigma_e^2(\sigma_f^2 + \sigma_e^2)}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2.
\end{aligned} \tag{A.5}$$

The -2 log-likelihood function is then expressed as

$$\ell(\mathbf{b}, \mathbf{v}, \sigma_f^2, \sigma_0^2) = \frac{1}{\sigma_e^2}\left(\|\mathbf{Y} - \mathbf{X}\mathbf{b}\mathbf{v}^T\|_F^2 - \frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2\right) + c(\sigma_e^2, \sigma_f^2), \tag{A.6}$$

where $c(\sigma_e^2, \sigma_f^2) = n\log[2\pi(\sigma_e^2)^{p-1}(\sigma_f^2 + \sigma_e^2)]$.

A.3.2 Derivation of equation 3.3: minimize ℓ w.r.t. $\mathbf{v}$

To minimize the -2 log-likelihood function w.r.t. $\mathbf{v}$ given the other parameters fixed, we minimize the following objective function under the constraint $\|\mathbf{v}\|_2^2 = 1$:

$$\begin{aligned}
F(\mathbf{v}) &= \|\mathbf{Y} - \mathbf{X}\mathbf{b}\mathbf{v}^T\|_F^2 - \frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2 \\
&= -\frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2}\sum_{i=1}^{n}\|\mathbf{Y}_{i\cdot}\mathbf{v}\|_2^2 - 2(1 - \frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2})\sum_{i=1}^{n}[(\mathbf{Y}_{i\cdot}\mathbf{v})(\mathbf{X}_{i\cdot}\mathbf{b})] \\
&+ (\|\mathbf{v}\|_2^2 - \frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2})\sum_{i=1}^{n}(\mathbf{X}_{i\cdot}\mathbf{b})^2\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\|\mathbf{Y}_{i\cdot}\|_2^2 \\
&= -\frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2}\|\mathbf{Y}\mathbf{v} + \frac{\sigma_e^2}{\sigma_f^2}\mathbf{X}\mathbf{b}\|_2^2 + \|\mathbf{X}\mathbf{b}\|_2^2\|\mathbf{v}\|_2^2 + C_0(\mathbf{Y}, \mathbf{X}, \mathbf{b}, \sigma_e^2, \sigma_f^2) \\
&\overset{\|\mathbf{v}\|_2^2=1}{=\!=} -\frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2}\|\mathbf{Y}\mathbf{v} + \frac{\sigma_e^2}{\sigma_f^2}\mathbf{X}\mathbf{b}\|_2^2 + C_1(\mathbf{Y}, \mathbf{X}, \mathbf{b}, \sigma_e^2, \sigma_f^2), \qquad (A.7)
\end{aligned}$$

where C_0 and C_1 are functions free of $\mathbf{v}$ and can be ignored. Thus, minimizing $F(\mathbf{v})$ w.r.t. $\mathbf{v}$ is equivalent to the problem

$$\max_{\mathbf{v}} \|\mathbf{Y}\mathbf{v} + \mathbf{X}\tilde{\mathbf{b}}\|_2^2 \text{ subject to } \mathbf{v}^T\mathbf{v} = 1,$$

where $\tilde{\mathbf{b}} = \sigma_e^2/\sigma_f^2\mathbf{b}$.

A.3.3 Derivation of the minimization problem of ℓ w.r.t. $\mathbf{b}$

To minimize the -2 log-likelihood function w.r.t. $\mathbf{b}$ given the other parameters fixed, we begin with the objective function $\|\mathbf{Y} - \mathbf{X}\mathbf{b}\mathbf{v}^T\|_F^2 - \frac{\sigma_f^2}{\sigma_f^2+\sigma_e^2}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2$. By the properties of Frobenius Norm: $\|\mathbf{A}\|_F^2 = tr(\mathbf{A}\mathbf{A}^T)$, for any orthogonal matrix $\mathbf{W}$ such that $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{I}$, we have $\|\mathbf{A}\mathbf{W}\|_F^2 = tr(\mathbf{A}\mathbf{W}\mathbf{W}^T\mathbf{A}^T) = \|\mathbf{A}\|_F^2$.
Then, define

$$\mathbf{W} = \begin{pmatrix}\mathbf{v} & \mathbf{v}_2^{\perp} & \mathbf{v}_3^{\perp} & ... & \mathbf{v}_p^{\perp}\end{pmatrix}, \text{ such that } \mathbf{v}^T\mathbf{W} = \begin{pmatrix}1 & 0 & 0 & ... & 0\end{pmatrix}.$$

This implies

$$\begin{aligned}
\|\mathbf{Y} - \mathbf{X}\mathbf{b}\mathbf{v}^T\|_F^2 &= \|\mathbf{Y}\mathbf{W} - \mathbf{X}\mathbf{b}\mathbf{v}^T\mathbf{W}\|_F^2 \\
&= \|\begin{pmatrix}\mathbf{Y}\mathbf{v} & \mathbf{Y}\mathbf{v}_2^{\perp} & ... & \mathbf{Y}\mathbf{v}_p^{\perp}\end{pmatrix} - \begin{pmatrix}\mathbf{X}\mathbf{b}_1 & 0 & ... & 0\end{pmatrix}\|_F^2 \\
&= \|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2 + \|\mathbf{X}\begin{pmatrix}\mathbf{v}_2^{\perp} & \mathbf{v}_3^{\perp} & ... & \mathbf{v}_p^{\perp}\end{pmatrix}\|_F^2. \qquad \text{(A.8)}
\end{aligned}$$

Therefore, the objective function is equivalent to the form

$$\begin{aligned}
&[1 - \frac{\sigma_f^2}{\sigma_f^2 + \sigma_e^2}]\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2 + \|\mathbf{X}\begin{pmatrix}\mathbf{v}_2^{\perp} & ... & \mathbf{v}_p^{\perp}\end{pmatrix}\|_F^2 \\
&= \frac{\sigma_e^2}{\sigma_f^2 + \sigma_e^2}\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2 + \|\mathbf{X}\begin{pmatrix}\mathbf{v}_2^{\perp} & ... & \mathbf{v}_p^{\perp}\end{pmatrix}\|_F^2. \qquad \text{(A.9)}
\end{aligned}$$

Terms free of $\mathbf{b}$ can be ignored, and the minimization of the -2 log-likihood function ℓ w.r.t $\mathbf{b}$ is equivalent to the minimization of $\|\mathbf{Y}\mathbf{v} - \mathbf{X}\mathbf{b}\|_2^2$, which then can be considered as a linear regression problem of regressing $\mathbf{Y}\mathbf{v}$ onto $\mathbf{X}$.

Appendix B Additional details of simulation studies

In this section, we provide more details about data generating process with respect to the three simulation settings discussed in Section 4.1.

(a) **COBS**

Multi-block data $\mathbf{Y}$ is generated from the COBS model $\mathbf{Y} = (\mathbf{XB} + \mathbf{F})\mathbf{V}^T + \mathbf{E}$. The coefficient matrix $\mathbf{B}$ and the loading matrix $\mathbf{V}$ follows the setting shown in Section 4.1. In particular, we set

$$\mathbf{V}_{(a)} = \begin{bmatrix} \mathbf{v}_{1(1)} & \mathbf{0} & \mathbf{0} & \mathbf{v}_{4(1)} \\ \mathbf{0} & \mathbf{v}_{2(2)} & \mathbf{0} & \mathbf{v}_{4(2)} \\ \mathbf{0} & \mathbf{0} & \mathbf{v}_{3(3)} & \mathbf{v}_{4(3)} \\ \mathbf{0} & \mathbf{0} & \mathbf{v}_{3(4)} & \mathbf{v}_{4(4)} \end{bmatrix}.$$

In this setting, the first block of the supervision information is related to the variation direction in the first block. Similarly, the second block of $\mathbf{X}$ picks the individual variation in the second block of $\mathbf{Y}$. The third loading vector indicates the partial-joint variation across the third and fourth blocks in $\mathbf{Y}$, which is picked by the third group of $\mathbf{X}$. The last group of $\mathbf{X}$ is associated the full-joint variation across all data blcoks in $\mathbf{Y}$. Under COBS setting, we consider two different scenarios by setting Σ_F relatively large and small to discuss how the variances from unknown sources can affect the estimates. For larger variances, we set $\Sigma_F = diag(10, 8, 6, 4)$ corresponding to the result (a) in table 1 and 3. For smaller variances, we set $\Sigma_F = diag(2.5, 2, 1.5, 1)$ that corresponds to result (a) in table 2 and 4.

To apply SVD and SupSVD with data from COBS model, we supply the true intrinsic rank $r = 4$. To apply AJIVE, we set the intrinsic rank in each block for its required initial setting as $r = (2, 2, 2, 2)$. To apply SIFA, the rank of the joint component is $r_0 = 2$ and the individual ranks are $r_1 = 1, r_2 = 1, r_3 = r_4 = 0$.

(b) **SIFA**

The loading matrix $\mathbf{V}_{(b)} = [\mathbf{v}_1, \ldots, \mathbf{v}_4]$ has a fully joint component $\mathbf{v}_1$ (corresponding to the largest variance), and three individual components:

$$\mathbf{V}_{(b)} = \begin{bmatrix} \mathbf{v}_{1(1)} & \mathbf{v}_{2(1)} & \mathbf{0} & \mathbf{0} \\ \mathbf{v}_{1(2)} & \mathbf{0} & \mathbf{v}_{3(2)} & \mathbf{0} \\ \mathbf{v}_{1(3)} & \mathbf{0} & \mathbf{0} & \mathbf{v}_{4(3)} \\ \mathbf{v}_{1(4)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Note that the fourth data block has no individual component. The coefficient matrix $\mathbf{B}_{(b)}$ is the same as $\mathbf{B}_{(a)}$ in the situation (a).

Under SIFA model, each data set in the multi-block data $\mathbf{Y}$ is generated from $\mathbf{Y}_k = \mathbf{J}_k+\mathbf{A}_k+\mathbf{E}_k = \mathbf{U}_0\mathbf{V}_{0,k}^T+\mathbf{U}_k\mathbf{V}_k^T+\mathbf{E}_k$, where the joint and individual factors are from the linear model $\mathbf{U}_0 = \mathbf{X}\mathbf{B}_0+\mathbf{F}_0$ and $\mathbf{U}_k = \mathbf{X}\mathbf{B}_k+\mathbf{F}_k$. Here we set $k = 1, \ldots, 4$, which indicates the four blocks in data $\mathbf{Y}$. Then the coefficient matrix $\mathbf{B}$ is formed as $(\mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_4)$ and loading matrix $\mathbf{V}$ is formed as $(\mathbf{V}_0, \mathbf{V}_1, \ldots, \mathbf{V}_4)$. Each row of $\mathbf{F}_0$ and $\mathbf{F}_k$ will follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ_{F_0} and Σ_{F_k} respectively.

To make sure that data sets can be compatible with the application of COBS and AJIVE, we will set $\mathbf{B}_4 = \mathbf{0}$, $\mathbf{V}_4 = \mathbf{0}$, and $\Sigma_{F_4} = \sigma_{f_4}^2 = 0$, which means that there is no individual pattern existing in the fourth data block. For the other columns in $\mathbf{B}$ and $\mathbf{V}$, we will follow the setting in COBS as described above and also in section 4.1.

We also adopt two different scenarios of the variances from unknown sources by setting larger variances scenario (correspoding to (b) in table 1 and 3) as $\Sigma_{F_0} = \sigma_{f_0}^2 = 10$, $\Sigma_{F_1} = \sigma_{f_1}^2 = 8$, $\Sigma_{F_2} = \sigma_{f_2}^2 = 6$, $\Sigma_{F_3} = \sigma_{f_3}^2 = 4$, $\Sigma_{F_4} = \sigma_{f_4}^2 = 0$, and smaller variances scenario (correspoding to (b) in table 2 and 4) as $\Sigma_{F_0} = \sigma_{f_0}^2 = 2.5$, $\Sigma_{F_1} = \sigma_{f_1}^2 = 2$, $\Sigma_{F_2} = \sigma_{f_2}^2 = 1.5$, $\Sigma_{F_3} = \sigma_{f_3}^2 = 1$, $\Sigma_{F_4} = \sigma_{f_4}^2 = 0$.

To fit SVD, SupSVD, and COBS with data genrated from SIFA, we set the intrinsic rank for $\mathbf{Y}$ as $r = 4$. To fit AJIVE, the intrinsic rank in each block for initial setting is $r = (2, 2, 2, 1)$. For SIFA, the joint signal has rank $r_0 = 1$ and the individual signals have ranks $r_1 = 1, r_2 = 1, r_3 = 1, r_4 = 0$.

(c) **JIVE**

We notice that SIFA is a generalized model from JIVE by incorporating covariates. In this sense, we generate multi-block data $\mathbf{Y}$ following most of the SIFA setting above but only need to change $\mathbf{B} = \mathbf{0}$ so that to remove the supervision information. This will also gurantee that the simulated data will be compatible with the other methods. Similar to the setting in SIFA, we set $\mathbf{V}_4 = \mathbf{0}$, and $\Sigma_{F_4} = \sigma^2_{f_4} = 0$ to rule out individual pattern in the fourth data block. Without the auxiliary covariates, the variation pattern in $\mathbf{Y}$ will all come from the random factor $\mathbf{F}$. We will also adopt same scenarios of Σ_F's as in SIFA above to compare how the variance in this random factor will affect the results from each methods. Result (c) in table 1 and 3 is from larger Σ_F's setting and (c) in table 2 and 4 is from smaller Σ_F's.

Under JIVE setting, the rank selections for each method are also the same as in SIFA above. That is, we use rank $r = 4$ to apply SVD, SupSVD, and COBS. The intrinsic rank in each block for initial setting in AJIVE is $r = (2, 2, 2, 1)$. For SIFA, the rank for joint variation pattern is $r_0 = 1$ and the individual ranks are $r_1 = 1, r_2 = 1, r_3 = 1, r_4 = 0$.

Appendix C Additional results from real data application

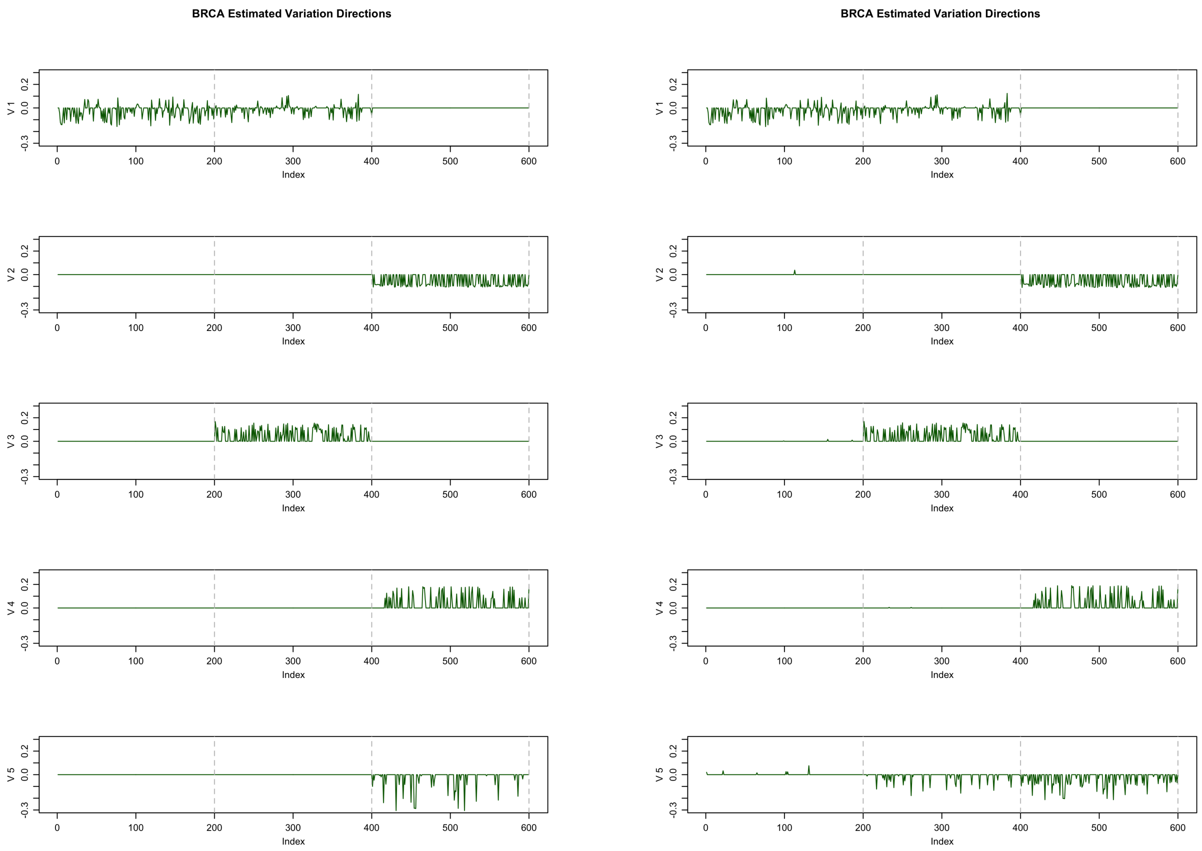
In this chapter, we provide more results from the application of COBS on breast cancer data and the image feature analysis.

C.1 TCGA breast cancer data

In section 5.1 we discussed the results by fixing $\alpha_v = 0$, which means we only consider group identification in estimating $\mathbf{V}$. The group structures are shown in equation 5.1. Here we compare the results with two other different settings as $\alpha_v = 0.5$ and $\alpha_v = 1$. The tuning parameter λ_v is chosen by BIC. We still fix $(\alpha_b, \lambda_b) = (1, 0.0025)$.

Figure 6 shows the top five layers of $\hat{\mathbf{V}}$ under two new different settings of α_v. The group identifications are the same as in equation 5.1 (with $\alpha_v = 0$), except the fifth component. But they all present stronger association between gene expression and methylation, than the association with CNV. With $\alpha_v = 1$, we no longer seek block-wise sparsity. This results in capturing negligible amount of signals in, e.g., one small loading in GE for $\mathbf{v}_2$. Ignoring such loadings of small magnitude, the estimated loading matrix is indeed quite close to that under $\alpha_v = 0$ setting. Figure 7 and figure 8 are the scatterplots of the estimated scores of the first two components with $\alpha_v = 0.5$ and $\alpha_v = 1$ correspondingly. They both clearly show similar pattern as in figure 2 in the main manuscript but just up to sign changes in the second component. Similar patterns in these results indeed prove that component 1 is majorly driven by the covariates that affects the gene expression and methylation variables, while the component 2 is mainly from unknown scores that affect copy number variation variables.

Additionally, as discussed in previous sections, it has been seen that genes in gene expression and methylation have a stronger connection compared with the connection with copy number variation. Thus, by constructing the correlation networks among the genes, we can check if there is any biological relationship among the three data blocks. We select



(a) Top five layers of $\hat{\mathbf{V}}$ under setting $\alpha_v = 0.5$.

(b) Top five layers of $\hat{\mathbf{V}}$ under setting $\alpha_v = 1$.

Figure 6: Variable and group realizations in $\hat{\mathbf{V}}$ for the top five layers. Variable indices 1-200 correspond to gene expression (GE), 201-400 to methylation (Me), and 401-600 to copy number variation (CNV).

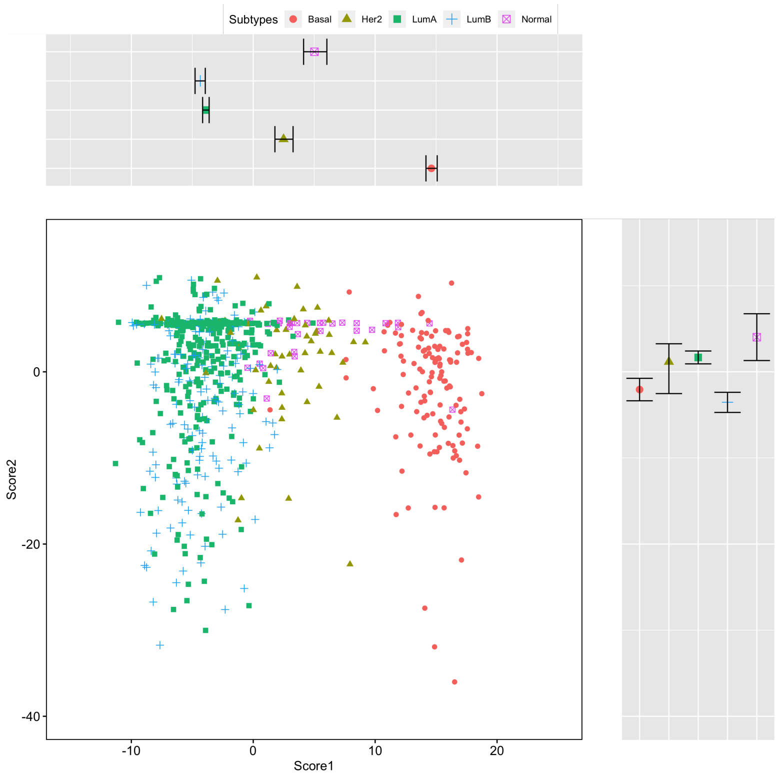


Figure 7: The first two component scores of the breast cancer data under the setting $\alpha_v = 0.5$.

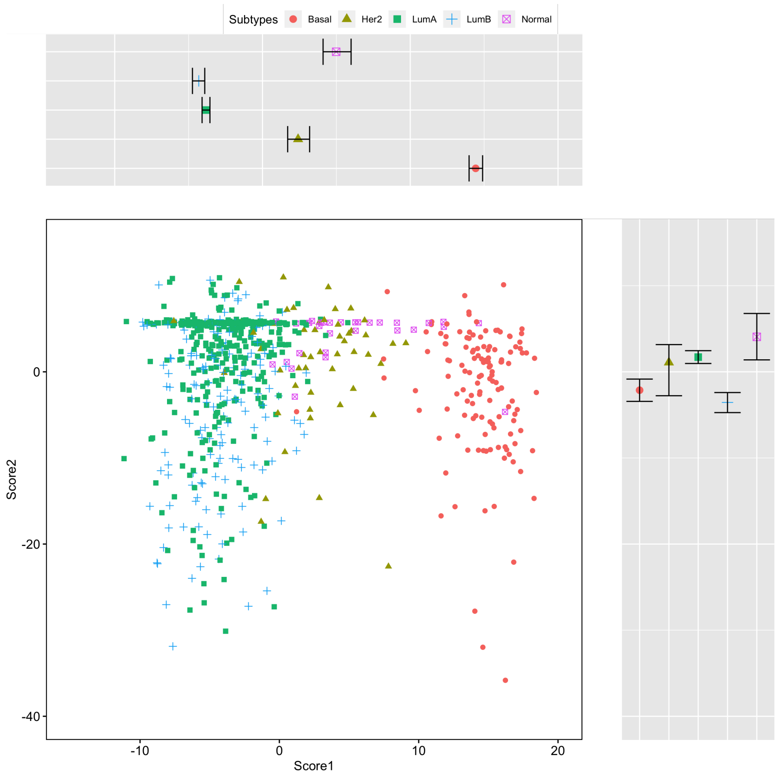
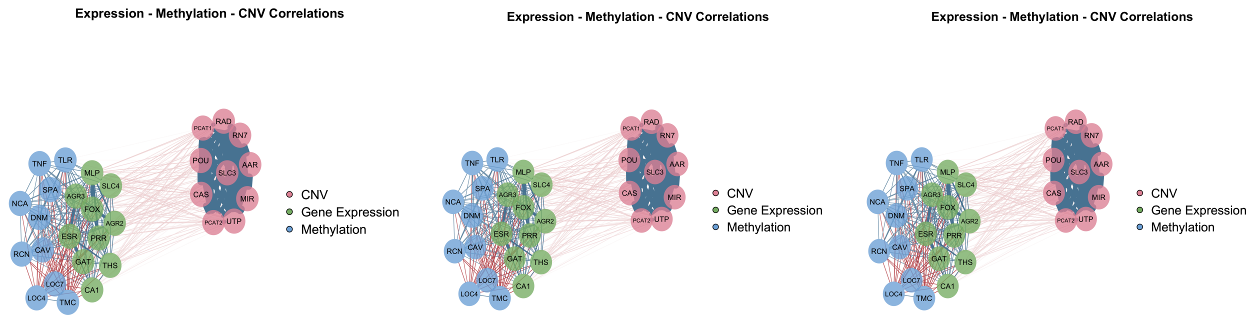


Figure 8: The first two component scores of the breast cancer data under the setting $\alpha_v = 1$.

10 genes in each data block that have the largest loadings from the first component (corresponding to gene expression and methylation) and second component (corresponding to copy number variation) to construct the correlation networks with the multi-block data $\mathbf{Y}$ under each setting (i.e., $\alpha_v = 0$, $\alpha_v = 0.5$, and $\alpha_v = 1$) as in Figure 9. The correlation networks clearly show that genes in gene expression and methylation are correlated with each other more significantly compared with CNV. However, the genes in CNV are more correlated with each other compared with the correlation in gene Expression or methylation. The three correlation networks present same selected genes and correlation significance, which again prove the connections among these genes in associated with the variation directions across gene expression and methylation are stronger than the connections with CNV.

We also decompose the total variation in the multi-block data $\mathbf{Y}$ into variation into different effect in COBS model under both setting $\alpha_v = 0.5$ and $\alpha_v = 1$ as shown in table 9 and 10 respectively. Similar conclusion can also be addressed here that Basal tumor explains most of the variation in gene expression and methylation data block. But Her2 and LumB explain more variation in CNV. With all three choices of α_v, a large portion of the variance is still hidden in the unknown sources which may be of interest in future. And the noise takes larger portion under $\alpha_v = 1$ compared with the other two choices due to more sparse estimates in $\hat{\mathbf{V}}$ under this setting. Overall we can see different choices of α_v can produce similar interpretation of the results for the breast cancer data analysis.



(a) Genes are selected undersetting $\alpha_v = 0$.

(b) Genes are selected undersetting $\alpha_v = 0.5$.

(c) Genes are selected undersetting $\alpha_v = 1$.

Figure 9: Correlation networks among gene expression (genes selected from component 1), methylation (genes selected from component 1), and CNV genes (genes selected from component 2). Blue indicates positive correlation. Red indicates negative correlation. Thickness indicates the magnitude of the correlation and color depth indicates the significance of the correlation.

		Basal	**Her2**	**LumA**	**LumB**	**Normal**	**Total**
Gene Expression	**Supervision**	62.34%	6.45%	17.90%	10.06%	3.25%	25.63%
	Unknown	-	-	-	-	-	21.44%
	Noise	-	-	-	-	-	52.93%
Methylation	**Supervision**	57.26%	3.58%	11.57%	18.28%	9.31%	12.86%
	Unknown	-	-	-	-	-	34.19%
	Noise	-	-	-	-	-	52.95%
CNV	**Supervision**	20.28%	27.95%	16.76%	28.77%	6.24%	7.72%
	Unknown	-	-	-	-	-	74.03%
	Noise	-	-	-	-	-	18.25%
Overall:		**Supervision:** 15.41%		**Unknown:** 43.22%		**Noise:** 41.37%	

Table 9: TCGA breast cancer data with $\alpha_v = 0.5$: The proportion of variance explained proportion in terms of each component (i.e., supervision, unknown sources, and noise) in COBS model for each individual block and the overall multi-block data (Sum to 1 in terms of the "Total" for each block and "Overall" for the concatenated multi-block data sets). The variation in supervision part is further separated into different tumor subtype for each block.

		Basal	**Her2**	**LumA**	**LumB**	**Normal**	**Total**
Gene Expression	**Supervision**	61.87%	6.35%	18.10%	10.21%	3.47%	26.71%
	Unknown	-	-	-	-	-	20.08%
	Noise	-	-	-	-	-	53.21%
Methylation	**Supervision**	58.12%	3.50%	11.37%	17.74%	9.27%	13.14%
	Unknown	-	-	-	-	-	33.08%
	Noise	-	-	-	-	-	53.78%
CNV	**Supervision**	17.49%	25.51%	17.86%	32.13%	7.01%	9.30%
	Unknown	-	-	-	-	-	70.30%
	Noise	-	-	-	-	-	20.40%
Overall:		**Supervision:** 16.38%		**Unknown:** 41.15%		**Noise:** 42.47%	

Table 10: TCGA breast cancer data with $\alpha_v = 1$: The proportion of variance explained proportion in terms of each component (i.e., supervision, unknown sources, and noise) in COBS model for each individual block and the overall multi-block data (Sum to 1 in terms of the "Total" for each block and "Overall" for the concatenated multi-block data sets). The variation in supervision part is further separated into different tumor subtype for each block.

Bibliography

[1] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[2] Lisha Chen and Jianhua Z Huang. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545, 2012.

[3] Nicolle M Correa, Tom Eichele, Tülay Adalı, Yi-Ou Li, and Vince D Calhoun. Multi-set canonical correlation analysis for the fusion of concurrent single trial erp and functional mri. *Neuroimage*, 50(4):1438–1445, 2010.

[4] Qing Feng, Meilei Jiang, Jan Hannig, and JS Marron. Angle-based joint and individual variation explained. *Journal of Multivariate Analysis*, 166:241–265, 2018.

[5] Chris Fraley, Adrian Raftery, Luca Scrucca, Thomas Murphy, and Michael Fop. *'mclust' R Package*, 2019. `https://cran.r-project.org/package=mclust` [Accessed: Mar 2020].

[6] Jerome Friedman, Trevor Hastie, Rob Tibshirani, Balasubramanian Narasimhan, Noah Simon, and Junyang Qian. *'GLMNET' R Package*, 2018. `https://cran.r-project.org/package=glmnet` [Accessed: Nov 2019].

[7] Xin Gao and Raymond J Carroll. Data integration with high dimensionality. *Biometrika*, 104(2):251–272, 2017.

[8] Irina Gaynanova and Gen Li. Structural learning and integrative decomposition of multi-view data. *Biometrics*, 75(4):1121–1132, 2019.

[9] David A Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383–385, 1974.

[10] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[11] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[12] Zhiguang Huo and George Tseng. Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *The Annals of Applied Statistics*, 11(2):1011–1039, 2017.

[13] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

[14] Sungkyu Jung, Myung Hee Lee, and Jeongyoun Ahn. On the number of principal components in high dimensions. *Biometrika*, 105(2):389–402, 2018.

[15] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 1971.

[16] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. Group factor analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 26(9):2136–2147, Sep. 2015.

[17] Jing Lei, Vincent Q Vu, et al. Sparsistency and agnostic inference in sparse pca. *The Annals of Statistics*, 43(1):299–322, 2015.

[18] Gen Li and Sungkyu Jung. Incorporating covariates into integrated factor analysis of multi-view data. *Biometrics*, 73(4):1433–1442, 2017.

[19] Gen Li, Dan Yang, Andrew B Nobel, and Haipeng Shen. Supervised singular value decomposition and its asymptotic properties. *Journal of Multivariate Analysis*, 146:7–17, 2016.

[20] Quefeng Li and Lexin Li. Integrative factor regression and its inference for multimodal data analysis. *arXiv preprint arXiv:1911.04056*, 2019.

[21] Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, 2013.

[22] Tommy Löfstedt, Daniel Hoffman, and Johan Trygg. Global, local and unique decompositions in onpls for multiblock data analysis. *Analytica Chimica Acta*, 791:13–24, 2013.

[23] Tommy Löfstedt and Johan Trygg. Onpls—a novel multiblock method for the modelling of predictive and orthogonal variation. *Journal of Chemometrics*, 25(8):441–455, 2011.

[24] Lester W. Mackey. Deflation methods for sparse pca. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1017–1024. Curran Associates, Inc., 2009.

[25] Jayson Miedema, James Stephen Marron, Marc Niethammer, David Borland, John Woosley, Jason Coposky, Susan Wei, Howard Reisner, and Nancy E Thomas. Image and statistical analysis of melanocytic histology. *Histopathology*, 61(3):436–444, 2012.

[26] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295, 2014.

[27] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[28] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. *'SGL' R Package*, 2018. `https://cran.r-project.org/package=SGL` [Accessed: Nov 2019].

[29] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[30] Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

[31] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[32] Ryan Tibshirani, Rob Tibshirani, Jonathan Taylor, Joshua Loftus, Stephen Reid, and Jelena Markovi. *'selectiveInference' R Package*, 2019. `https://cran.r-project.org/package=selectiveInference` [Accessed: Nov 2019].

[33] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[34] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113, 2013.

[35] Ke Ye and Lek-Heng Lim. Schubert varieties and distances between subspaces of different dimensions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1176–1197, 2016.

[36] Grace Yoon, Raymond J Carroll, and Irina Gaynanova. Sparse semiparametric canonical correlation analysis for data of mixed types. *arXiv preprint arXiv:1807.05274*, 2018.

[37] Guoxu Zhou, Andrzej Cichocki, Yu Zhang, and Danilo P Mandic. Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 27(11):2426–2439, 2016.

[38] Huichen Zhu, Gen Li, and Eric F Lock. Generalized integrative principal component analysis for multi-type data with block-wise missing structure. *(To appear in) Biostatistics*, 09 2018.

[39] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

[40] Hui Zou, Trevor Hastie, Robert Tibshirani, et al. On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.