Two-stage Linked Component Analysis for Joint Decomposition of Multiple Biologically Related Data Sets

By Huan Chen

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, U.S.A.

hchen130@jhu.edu

Brian Caffo

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, U.S.A.

bcaffoweb@jhu.edu

Genevieve Stein-O'Brien

Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland, 21205, U.S.A.

gsteinobrien@jhmi.edu

Jinrui Liu

Department of Neurology, Johns Hopkins University, Baltimore, Maryland, 21205, U.S.A.

jliu179@jhmi.edu

Ben Langmead

Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, 21205, U.S.A.

langmea@cs.jhu.edu

Carlo Colantuoni*

Department of Neurology and Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland, 21205, U.S.A.

and

Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland, 21201, U.S.A.

ccolantu@jhmi.edu

and Luo Xiao*

Department of Statistics, North Carolina State University, Raleigh, North Carolina, 27695, U.S.A.

lxiao5@ncsu.edu

* Co-corresponding authors

Summary

Integrative analysis of multiple data sets has the potential of fully leveraging the vast amount of high throughput biological data being generated. In particular such analysis will be powerful in making inference from publicly available collections of genetic, transcriptomic and epigenetic data sets which are designed to study shared biological processes, but which vary in their target measurements, biological variation, unwanted noise, and batch variation. Thus, methods that enable the joint analysis of multiple data sets are needed to gain insights into shared biological processes that would otherwise be hidden by unwanted intra-data set variation. Here, we propose a method called two-stage linked component analysis (2s-LCA) to jointly decompose multiple biologically related experimental data sets with biological and technological relationships that can be structured into the decomposition. The consistency of the proposed method is established and its empirical performance is evaluated via simulation studies. We apply 2s-LCA to jointly analyze four data sets focused on human brain development and identify meaningful patterns of gene expression in human neurogenesis that have shared structure across these data sets. The code to conduct 2s-LCA has been complied into an R package "PJD", which is available at https://github.com/CHuanSite/PJD.

Some key words: Integrative methods, Joint decomposition, Low rank models, Multi-view data, Principal component analysis.

1. Introduction

1.1. Background

Heterogeneous data of various types are now being collected from scientific experiments and/or large scale health initiatives. Thus, it is of scientific interest to harness the collective discovery potential of such complex and growing data resources. For example, in biomedical research, high-dimensional gene expression and epigenetic data are produced to gain insight into cellular processes and disease mechanisms (Johnson et al., 2020; Micali et al., 2020; Clark et al., 2019; Stein-O'Brien et al., 2019). Another example is in the study of Alzheimer's disease where recent efforts have been focusing on combining brain imaging data, genetic data, as well as clinical outcomes in predicting disease (Nathoo et al., 2019). A third example is a large study profiling different states of wellness, where genetic, proteomic and metabolic data among other types of data are collected (Gao et al., 2020). Together, the different types of data provide a more comprehensive picture which has the potential of better characterizing and optimizing what it is to be healthy. In addition to these examples of studies with multimodal data collection, many public repositories are full of experimental data from single modalities that are related biologically. Therefore, in recent years, there has been growing interests as well as demands for understanding and utilizing these data in an integrative way (Lock et al., 2013; Yang & Michailidis, 2016; Li & Jung, 2017; Gaynanova & Li, 2017; Gao et al., 2020).

1.2. Motivating Data

Advances in RNA sequencing technologies have produced a large amount of data creating unprecedented opportunities for scientific discovery through data integration. We put together four experimental paradigms used to study brain development: *i*.van de Leemput et al. (2014); *ii*. Yao et al. (2017); *iii*. BrainSpan (2011); and *iv*. Nowakowski et al. (2017); see Fig. 1. Specifically, we use two *in vitro* data sets from cultured human

pluripotent stem cells subjected to neural differentiation, and two *in vivo* data sets of brain tissue across varying ages of human fetal development. Within the *in vitro* and *in vivo* studies two different RNA sequencing technologies were used: the two bulk studies extracted RNA from all the cells in tissue samples and expression levels were measured in aggregate for each tissue sample, while in the two single-cell studies, the RNA from each cell in a tissue sample was extracted and quantified individually. All four data sets focus on the creation of neurons from neural precursor cells across these times and hence are highly interrelated. However, there are sample specific details that introduce experiment specific variation beyond the expected technical batch effects. In the van de Leemput et al. (2014) bulk data set, data were collected at 9 time points across 77 days of neural differentiation with a total number of 24 *in vitro* samples. The Yao et al. (2017) data set was collected to model the production of human cortical neurons *in vitro*. There are about 2.7 thousand cells collected over across 54 days during *in vitro*. In the BrainSpan (2011) bulk data set, there are a total of 35 *in vivo* brain tissue samples. The Nowakowski et al. (2017) data set contains around 3.5 thousand single cells across different ages in the fetal brain. For visualization of these data sets, see the NeMO Analytics portal at https://nemoanalytics.org/p?l=ChenEtAl2021&g=DCX, where we have leveraged the gEAR platform to construct an integrated gene expression data viewer including all data sets used in the report (Orvis et al., 2020).

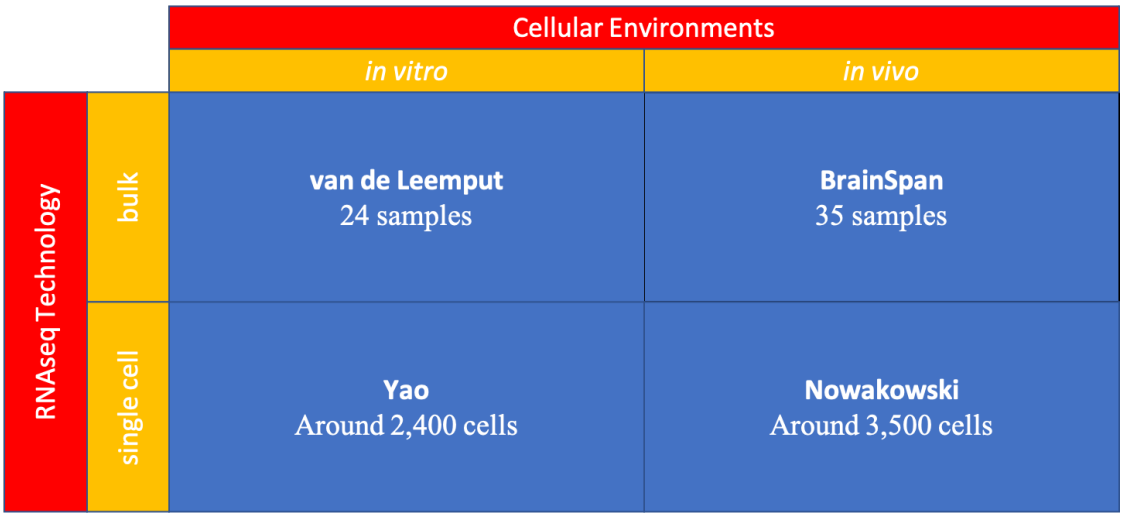

Fig. 1: Four gene expression data sets with a factorial design consisting of two factors: 1) cellular context: *in vivo* versus *in vitro*; 2) RNA sequencing technology: single cell versus bulk.

The primary goal of assembling these different data sets together is to define molecular dynamics that are common to all of them and are central in the creation of human neocortical neurons. In addition, we are interested in learning what cellular processes during *in vivo* development of the brain can be recapitulated in the *in vitro* systems and what processes are only present in the *in vivo* data. To achieve this goal, we need a new joint model that can properly model the two-way design exhibited by the four data sets and separate the different cellular processes.

1.3. Existing Literature

We distinguish two distinct types of data structure for multiple data sets: multi-view data and linked data. For multi-view data, e.g., multi-omics data, different sets of features are measured for each subject or unit with subject being the linking entity across data sets. For linked data, e.g., gene expression data for different types of cancer patients in The Cancer Genome Atlas (TCGA), the same set of features are measured for different groups of subjects. Therefore, multi-view data sets are correlated while linked data sets are independent. When data sets are organized such that each column corresponds to one sample, the former can be called vertically linked data as the data sets can be aligned vertically, while the latter can be called horizontally linked data as the data sets can be aligned horizontally (Richardson et al., 2016). For both types of data structures, a central goal of statistical analysis is to identify meaningful decomposition of variations across data sets. We discuss relevant methods below for both multi-view and linked data.

Canonical correlation analysis (CCA) is a useful method for extracting common variation across multi-view data. Various variants of CCA for high dimensional data sets have been proposed; see, e.g., Gao et al. (2020) and Min & Long (2020). However, CCA ignores potentially substantial variation present in individual data sets. To remedy this, the joint and individual variations explained (JIVE) method (Lock et al., 2013) and subsequent variants of JIVE, e.g., AJIVE (Feng et al., 2018), were developed to identify common variation shared across multiple data sets as well as individual variation specific to each data set. Going further, the structural learning and integrative decomposition (SLIDE) method (Gaynanova & Li, 2017) also accommodates variations that are shared by only a subset of data sets. Other relevant works include Li & Jung (2017) and Li et al. (2018a).

A pioneering method for linked data is common principal component analysis (CPCA, Flury (1984)), which can identify the same set of orthogonal eigenvectors shared across data sets. CPCA was extended to extract shared signal subspace present across data sets (Flury, 1987) and to determine the number of shared components (Wang et al., 2019). Another extension based on matrix decomposition, population value decomposition (Crainiceanu et al., 2011) extends CPCA to matrix-valued data, e.g., neuroimaging and electrophysiology data.

We briefly mention a few other related works. Park & Lock (2020) jointly analyzed multiple data sets for heterogeneous groups of objects with heterogeneous feature sets. Gao et al. (2020) and Wang & Allen (2019) considered clustering problems for multi-view data. Li et al. (2018b) proposed a regression model with multi-view data as covariates.

1.4. Our Contribution

We propose a joint decomposition model for linked multiple data sets with a general design, e.g., the 2 by 2 factorial design in our motivating data, thus extending the aforementioned CPCA method which essentially assumes the linked multiple data sets have a one-way design. The goal of the joint decomposition is to identify signal subspaces that are either shared or not shared across data sets. In particular, similar in spirit to the SLIDE method for multi-view data, our model allows for common signal subspace that are shared by all data sets, partially-shared signal subspaces that are shared by only subsets of all data sets, and individual signal subspaces specific to each data set. It is important to note that, for our method, it is fixed signal subspaces rather than latent random scores that are shared across data sets. Another significant difference is that the existence of partially-shared signal subspaces for the proposed model is determined by

design, which is scientifically meaningful for our motivating data and somehow similar to Park & Lock (2020) for bi-dimensional data. To the best of our knowledge, joint decomposition of linked data sets with a general design has not been considered before in the literature.

Our model estimation consists of a simple and intuitive two-stage procedure and is built on existing literature in principal component analysis. From a computational perspective, it is straightforward and avoids complex and computationally challenging optimization methods, e.g., biconvex optimization methods as in JIVE and SLIDE. More importantly, the theoretic properties of the proposed method for a general class of decomposition models are studied. Specifically, we establish the consistency of the proposed method for estimating the signal subspaces. In particular, we prove the consistency of the proposed method for estimating the ranks of the signal subspaces. Note that most existing methods in the literature lack theoretic justification for the choice of rank.

The rest of the paper is organized as follows. In Section 2, we first describe our model for linked data sets with a 2 by 2 design, as motivated by our data example, and then extend the model for data sets with a general design. A two-step model estimation method is given in the Section. In Section 3, we study the theoretic properties of the proposed model estimation method. In Section 4, we carry out a simulation study to evaluate the empirical performance of our proposed method. In Section 5, we apply the proposed method to the motivating data and also carry out a validation study using additional data sets. Finally, in Section 6, we discuss how our work can be extended to the context of high-dimensional data.

2. Models

2.1. Model Formulation

First, consider a 2 by 2 factorial design, as in our motivating data. Let i be the index for the cells' system, with $i = 1$ for *in vivo* and $i = 2$ for *in vitro*. Similarly, let j be the index for the RNA sequencing technology, with $j = 1$ for single cell and $j = 2$ for bulk. Let $\mathbf{Y}_{ij} \in \mathbb{R}^{n_{ij} \times p}$ be the (i, j)th data set consisting of n_{ij} tissue samples or cells of dimension p. For simplicity, hereafter we shall use sample to refer to either tissue sample or cell. Denote by $y_{ijk} \in \mathbb{R}^p$ the kth sample for the (i, j)th data set so that $\mathbf{Y}_{ij} = [y_{ij1}, \ldots, y_{ijn_{ij}}]$. Thus, the samples are of the same dimension for all data sets and they are assumed independent between the data sets and from each other.

Consider a two-way latent factor model,

$$y_{ijk} = \mathbf{A} f_{ijk} + \mathbf{B}_i g_{ijk} + \mathbf{C}_j h_{ijk} + \mathbf{D}_{ij} \ell_{ijk} + \epsilon_{ijk}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{p \times r_A}$, $\mathbf{B}_i \in \mathbb{R}^{p \times r_{B_i}}$, $\mathbf{C}_j \in \mathbb{R}^{p \times r_{C_j}}$ and $\mathbf{D}_{ij} \in \mathbb{R}^{p \times r_{D_{ij}}}$ are fixed orthonormal matrices of components with associated random scores $f_{ijk} \in \mathbb{R}^{r_A}$, $g_{ijk} \in \mathbb{R}^{r_{B_i}}$, $h_{ijk} \in \mathbb{R}^{r_{C_j}}$ and $\ell_{ijk} \in \mathbb{R}^{r_{D_{ij}}}$, respectively. The vector $\epsilon_{ijk} \in \mathbb{R}^p$ consists of uncorrelated, mean zero error terms, with variance σ^2_{ij}. We assume that all random scores have zero means and $\mathrm{cov}(f_{ijk}) = \mathbf{\Sigma}_{f_{ij}}, \mathrm{cov}(g_{ijk}) = \mathbf{\Sigma}_{g_{ij}}, \mathrm{cov}(h_{ijk}) = \mathbf{\Sigma}_{h_{ij}}, \mathrm{cov}(\ell_{ijk}) = \mathbf{\Sigma}_{\ell_{ij}}$, with positive diagonal elements in all covariances. In particular, we assume that $\mathbf{\Sigma}_{f_{11}}$, $\mathbf{\Sigma}_{g_{i1}} (i = 1, 2)$, $\mathbf{\Sigma}_{h_{1j}} (j = 1, 2)$, and $\mathbf{\Sigma}_{\ell_{ij}} (i = 1, 2, j = 1, 2)$ are diagonal matrices.The other covariance matrices can be non-diagonal, implying that components do not have to align across the data sets. Finally, the random terms $\{f_{ijk}, g_{ijk}, h_{ijk}, \ell_{ijk}, \epsilon_{ijk}\}$ are assumed uncorrelated across i, j, and k.

For identifiability, we impose the condition that the orthogonal matrices $\mathbf{A}, \mathbf{B}_1, \mathbf{B}_2, \mathbf{C}_1, \mathbf{C}_2, \mathbf{D}_{ij} (1 \leq i, j \leq 2)$ are also mutually orthogonal. This condition ensures that the components are not movable, e.g., the components in $\mathbf{A}$ can not be transferred to the other matrices without changing the model.

Model (1) can be interpreted as a multivariate ANOVA model. To elaborate, we use the notation $\mathrm{col}(\cdot)$ to denote the column space of a matrix. Then, $\mathrm{col}(\mathbf{A})$ is the space of common components, the space of variation that is shared by all data sets. Next, $\mathrm{col}(\mathbf{B}_1)$ is the space of partially-shared components specific to *in vivo* system, i.e., the two *in vivo* data sets. Thus, it is orthogonal to both the common components, i.e., $\mathrm{col}(\mathbf{A})$, and the space of variation specific to the *in vitro* system, i.e., $\mathrm{col}(\mathbf{B}_2)$. Third, $\mathrm{col}(\mathbf{C}_1)$ is the space of partially-shared components specific to the single cell sequencing technology and $\mathrm{col}(\mathbf{C}_2)$ is for the bulk sequencing method. Finally, $\mathrm{col}(\mathbf{D}_{ij})$ are the individual components specific to each data set.

For data set (i, j), model (1) is a latent factor model with the signal space being the column space of $\mathbf{U}_{ij} = [\mathbf{A}, \mathbf{B}_i, \mathbf{C}_j, \mathbf{D}_{ij}]$, which can be denoted by $\mathrm{col}(\mathbf{U}_{ij})$.

2.2. Model Estimation

We propose a two-stage estimation method. The first stage is to estimate the signal space from each data set separately using principal component analysis (PCA). The second stage combines the signal spaces to extract common subspaces, partially-shared subspaces and individual subspaces, also using PCA.

For the first stage, the key is to estimate the number of components in each data set, i.e., the rank of $\mathrm{col}(\mathbf{U}_{ij})$. We use the BEMA method (Ke et al., 2020) given in Algorithm A.1 in appendix, which fits a distribution to the eigenvalues (putatively) arising from noise using the observed bulk eigenvalues and then determines the number of eigenvalues arising from signals via extrapolation of the fitted distribution. This method yields consistent estimation and worked well in our simulations. Denote by $\hat{\mathbf{U}}_{ij}$ the matrices of the estimated eigenvectors associated with the top eigenvalues from the sample covariance of data set (i, j).

For the second stage, we extract common subspaces, partially-shared subspaces, and individual subspaces, sequentially. Assume that one extracts the subspaces in the following order:

$$\mathbf{A} \to \mathbf{B}_1 \to \mathbf{B}_2 \to \mathbf{C}_1 \to \mathbf{C}_2 \to \mathbf{D}_{11} \to \mathbf{D}_{12} \to \mathbf{D}_{21} \to \mathbf{D}_{22}.$$

Here, to simplify notation, we use the matrices to denote the subspaces.

The estimation of common subspace is based on the following observation. Let $\mathbf{H} = \mathbf{U}_{11}\mathbf{U}_{11}^T + \mathbf{U}_{12}\mathbf{U}_{12}^T + \mathbf{U}_{21}\mathbf{U}_{21}^T + \mathbf{U}_{22}\mathbf{U}_{22}^T$. Then it is easy to show that for any eigenvector of $\mathbf{H}$, $\mathbf{a}^T\mathbf{H}\mathbf{a}$ equals 4 if $\mathbf{a} \in \mathrm{col}(\mathbf{A})$ and is at most 2 otherwise. Thus, there exists a natural gap of eigenvalues associated with common subspace and those of other subspaces for $\mathbf{H}$. From a theoretical perspective, any cut-off value between 2 and 4 for the eigenvalues may lead to consistent estimation. In practice, a scree plot of the eigenvalues of $\hat{\mathbf{H}} = \sum_{i,j} \hat{\mathbf{U}}_{ij}^T \hat{\mathbf{U}}_{ij}$ can be used to determine the dimension of common subspace, i.e. considering the relative amount of variation explained among the top eigenvalues. Once the dimension of common subspace is determined, the eigenvectors of $\hat{\mathbf{H}}$ associated with these top eigenvalues give us an estimate of the space of common components. Let $\hat{\mathbf{A}}$ be the matrix with columns corresponding to the retained eigenvectors of $\hat{\mathbf{H}}$. Note that $\hat{\mathbf{A}}$ may not be a good estimate of $\mathbf{A}$ elementwise, rather, the notation is used to represent that $\mathrm{col}(\hat{\mathbf{A}})$ is an estimate of $\mathrm{col}(\mathbf{A})$ and $\hat{\mathbf{A}}$ is one of many matrix representations of the former.

Next we estimate the partially-shared subspaces. We first project $\hat{\mathbf{U}}_{ij}$ onto orthogonal complement of $\hat{\mathbf{A}}$, i.e., $(\mathrm{I} - \hat{\mathbf{A}}\hat{\mathbf{A}}^T)\hat{\mathbf{U}}_{ij}$. For simplicity, we still denote it by $\hat{\mathbf{U}}_{ij}$, but the common subspace has been removed. We now consider estimation of $\mathrm{col}(\mathbf{B}_1)$, the space of partially-shared subspaces corresponding to the *in vivo* system. With similar arguments as before, now the matrix $\hat{\mathbf{U}}_{11}\hat{\mathbf{U}}_{11}^T + \hat{\mathbf{U}}_{12}\hat{\mathbf{U}}_{12}^T$ has one set of values close to 2 and another set of values close to or smaller than 1. Indeed, the theoretic counterpart of the above matrix has two sets of non-zero eigenvalues: one set of eigenvalues of 2 with associated eigenvectors forming the space $\mathrm{col}(\mathbf{B}_1)$ and another set of eigenvalues of 1 with associated eigenvectors from the individual space of either $\mathrm{col}(\mathbf{D}_{11})$ or $\mathrm{col}(\mathbf{D}_{12})$. Therefore, another scree plot of eigenvalues of $\hat{\mathbf{U}}_{11}\hat{\mathbf{U}}_{11}^T + \hat{\mathbf{U}}_{12}\hat{\mathbf{U}}_{12}^T$ will determine the dimension of subspaces in $\mathrm{col}(\mathbf{B}_1)$ and also give an estimate of the space.

Similar analysis can be used to estimate the other partially-shared subspaces. For example, a spectral analysis of $\hat{\mathbf{U}}_{21}\hat{\mathbf{U}}_{21}^T + \hat{\mathbf{U}}_{22}\hat{\mathbf{U}}_{22}^T$ can be used to estimate $\mathrm{col}(\mathbf{C}_1)$. Further steps follow similarly and thus are omitted. To ensure the estimated spaces are orthogonal, orthogonal projection is always used after each estimation of the spaces of partially-shared subspaces. After the projection of $\hat{\mathbf{U}}_{ij}$ onto the complement of estimates of $\mathrm{col}(\mathbf{B}_i)$ and $\mathrm{col}(\mathbf{C}_j)$, the remaining space is an estimate of the individual space $\mathrm{col}(\mathbf{D}_{ij})$.

A remaining problem lies in the determination of the order which components are estimated. First, it is clear that the components shared by more data sets should be prioritized, and estimated earlier in the process. The problem then lies in the order to follow to extract the partially-shared components. While this is not a problem in our theoretical analysis, in real data analysis, different orders produce different estimates, because of the orthogonal condition imposed on the model and the orthogonal projection used in model estimation. A reasonable method considers the quality of the data sets (e.g., the sample size) and the spaces of shared components with higher perceived quality prioritized in the estimation order.

Algorithm 1. Two-stage linked component analysis for two-way factorial design
Result: Estimated signal spaces $\hat{\mathbf{A}}, \hat{\mathbf{B}}_i, \hat{\mathbf{C}}_j, \hat{\mathbf{D}}_{ij}$.

1. Extract signal space matrix $\mathbf{U}_{ij}$ for each data set,

$$\hat{\mathbf{U}}_{ij} = \text{top } r_{ij} \text{ eigenvector from } \mathbf{Y}_{ij}\mathbf{Y}_{ij}^T,$$

   determine the r_{ij} by the BEMA method (Ke et al., 2020).
2. Follow the order

$$\mathbf{A} \to \mathbf{B}_1 \to \mathbf{B}_2 \to \mathbf{C}_1 \to \mathbf{C}_2 \to \mathbf{D}_{11} \to \mathbf{D}_{12} \to \mathbf{D}_{21} \to \mathbf{D}_{22}.$$

   to identify the specific signal subspaces as follows.
   a. Select the rank of components (see Algorithm 2 for the general design) and extract principal components from $\hat{\mathbf{U}}_{ij}$ of data sets that share specific signal subspaces;
   b. Project all $\hat{\mathbf{U}}_{ij}$ onto the space orthogonal to the extracted signal subspaces;
   c. Repeat the above steps until all signal space has been extracted.
3. Return estimated signal spaces $\hat{\mathbf{A}}, \hat{\mathbf{B}}_i, \hat{\mathbf{C}}_j, \hat{\mathbf{D}}_{ij}$.

2.3. Model for General Design

We extend the above model for a two-way factorial design to a model with a general design. To accommodate the more general setting, the notation here is slightly different from the previous sections. Assume that there are I data sets each containing n_i samples.

Denote by $y_{ij} \in \mathbb{R}^p$ the jth sample for the ith data set. We consider a latent factor model with general design as follows:

$$y_{ij} = \Theta_{i1}\mathbf{V}_1 f_{1ij} + \Theta_{i2}\mathbf{V}_2 f_{2ij} + \ldots \Theta_{iK}\mathbf{V}_K f_{Kij} + \epsilon_{ij}. \tag{2}$$

The orthonormal matrices $\mathbf{V}_k \in \mathbb{R}^{p\times r_k}$ correspond to the kth space of components with associated scores, $f_{kij} \in \mathbb{R}^{r_k}$. So the kth space is of dimension r_k, which is unknown and needs to be determined. The indicator variables, Θ_{ik}, denote if the ith data set contains the kth space of components, and $\epsilon_{ij} \in \mathbb{R}^p$ are random errors. Here the number of spaces, K, and the indicator variables are determined by design. For example, for the two-way factorial design considered above, $K = 9$. Two clear restrictions on the indicator variables are: (1) for each k, there exists at least one i such that $\Theta_{ik} = 1$; (2) there does not exist $k \neq k'$ with identical indicator values across all data sets. Let $\eta_k = \sum_i \Theta_{ik}$, which is the number of data sets sharing the kth space of components. Without loss of generality, we assume that $\eta_1 \geq \cdots \geq \eta_K$.

We assume that all random scores have zero means and $\mathrm{cov}(f_{kij}) = \Sigma_{ki} \in \mathbb{R}^{p\times p}$ with positive diagonal elements in the covariance. In particular, for each k, there exists a predetermined $\Theta_{ik} = 1$ such that Σ_{ki} is a diagonal matrix, i.e., the corresponding random vector f_{kij} has uncorrelated elements. We also assume that all random terms in the model are uncorrelated from each other and across i and j. For model identifiability, we impose that $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_K$ are mutually orthogonal. Let $K_i = \sum_{k=1}^K \Theta_{ik} r_k$, the total dimension of signal subspaces in data set i.

We now extend the model estimation proposed in the previous subsection. The first step remains the same and denote by $\hat{\mathbf{U}}_i$ the matrices of estimated eigenvectors. Consider the second step where the critical issue is the order of components to be extracted. The key idea is that spaces shared by more data sets should be extracted before those shared by fewer data sets. For each k, denote by $\mathcal{I}_k$ the index set for which if $k' \in \mathcal{I}_k$, then $\Theta_{ik'} \leq \Theta_{ik}$ and for at least one i the inequality is strict. Then let α_k be the smallest element in $\mathcal{I}_k$. If $\mathcal{I}_k$ is empty then let α_k be 0. Obviously, $\alpha_k < \eta_k$ for all k. Suppose we have estimated the first k spaces of shared components and have projected $\hat{\mathbf{U}}_i$ onto the orthogonal space of the previously extracted spaces. For simplicity, still denote those spaces by $\hat{\mathbf{U}}_i$. To estimate the $(k+1)$th space, compute $\sum_{i:\Theta_{i(k+1)}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T$ and its eigendecomposition. Let $\hat{\mathbf{V}}_{k+1}$ consists of the resulting eigenvectors with associated eigenvalues larger than $(\eta_{k+1} + \alpha_{k+1})/2$. Then $\mathrm{col}(\hat{\mathbf{V}}_{k+1})$ is our estimate of $\mathrm{col}(\mathbf{V}_{k+1})$

The above estimation procedure is unique if η_k are all distinct. When there exists ties, then the order of estimation can be determined similarly as in the two-way design. One note is that our theoretical derivation remains valid no matter what order of estimation taken within the ties.

After obtaining the components, to investigate or visualize the data at hand, we may project the data sets onto the estimated subspaces, which can be obtained by $\hat{\mathbf{V}}_k^T\mathbf{Y}_i$.

Algorithm 2. Component rank r_k selection for general design
Result: An estimator $\hat{r}_k$ for the number of components for component k.

1. Compute $\sum_{i:\Theta_{ik}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T$, and its eigenvalues, $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_p$.
2. Compute $\eta_1 = \sum_{i:\Theta_{ik}=1} \Theta_{ik}$ and $\eta_2 = \max_{k'} \sum_{i:\Theta_{ik}=1, k'\neq k} \Theta_{ik'}$.
3. Return $\hat{r}_k = \#\{k : \hat{\sigma}_k >= (\eta_1 + \eta_2)/2\}$.

Algorithm 3. Two-stage linked component analysis for a general design
Result: Estimated signal spaces $\hat{\mathbf{V}}_1, \hat{\mathbf{V}}_2, \ldots, \hat{\mathbf{V}}_K$.

1. Extract component space matrix $\hat{\mathbf{U}}_i$ for each data set i,
$$\hat{\mathbf{U}}_i = \text{top } \hat{K}_i \text{ eigenvector from } \mathbf{Y}_i\mathbf{Y}_i^T,$$
where $\hat{K}_i$ is determined by the BEMA method.
2. Follow the order
$$\hat{\mathbf{V}}_1 \to \hat{\mathbf{V}}_2 \cdots \to \hat{\mathbf{V}}_K$$
to identify the specific components as follows:
   a. Select the number based on Algorithm 2 and extract principal components from union of data sets that share those components, $\hat{\mathbf{V}}_k$,
$$\text{select } r_k \text{ based on the eigenvalues of } \sum_{i:\Theta_{ik}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T,$$
and
$$\hat{\mathbf{V}}_k = \text{top } r_k \text{ component from } \sum_{i:\Theta_{ik}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T.$$
   b. Project the signal space, $\hat{\mathbf{U}}_i$, onto the space orthogonal to the extracted spaces $\hat{\mathbf{V}}_k$,
$$(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^T)\hat{\mathbf{U}}_i,$$
   c. Repeat the above steps until no signal space left.
3. Return estimated signal spaces $\hat{\mathbf{V}}_1, \hat{\mathbf{V}}_2, \ldots, \hat{\mathbf{V}}_K$.

3. Theoretical Properties

In this section, we establish the consistency of the proposed estimation methods for the general design proposed in (2). Specifically, we prove the consistency of $\text{col}(\hat{\mathbf{V}}_k)$ for estimating the signal space $\text{col}(\mathbf{V}_k)$, for each k, which implicitly implies the consistency of rank estimation for each signal subspace as well.

We shall use the following notation. For a square matrix $\mathbf{A}$, denote by $\text{tr}(\mathbf{A})$ the sum of diagonal elements in $\mathbf{A}$ and $\|\mathbf{A}\|_2$ the operator norm of $\mathbf{A}$. Let $r_e(\mathbf{A}) = \text{tr}(\mathbf{A})/\|\mathbf{A}\|_2$, the effective rank of $\mathbf{A}$.

Consider the following assumptions.

Assumption 1. $p/n \to \gamma, \gamma > 0$ is a constant.

Assumption 2. The minimal nonzero eigenvalue of $\mathbf{\Sigma}_{ki}$, denoted by σ_{ki}, satisfies $\sigma_{ki} \geq \sigma^2(\sqrt{\gamma} + \tau)$, for a positive constant τ and γ in Assumption 1, for all k and i.

Assumption 3. Assume that the random score f_{kij} for each signal subspace $\mathbf{V}_k$ follows a Gaussian distribution, $f_{kij} \sim N(0, \mathbf{\Sigma}_{ki})$, and the noise vector ϵ_{ij} also follows the Gaussian distribution, $\epsilon_{ij} \sim N(0, \sigma_i^2\mathbf{I})$. as well as the random scores and noises are independent from each other and across subjects.

Assumption 4. Let $\mathbf{\Sigma}_i = \mathbb{E}[y_{ij}^T y_{ij}]$ for $i = 1, \ldots, n$. Assume that these covariance matrices have bounded effective ranks, i.e., $r_e(\mathbf{\Sigma}_i) \leq C$, for all i and some fixed positive constant C.

Assumptions 1 and 2 are needed for the consistency of the BEMA method for estimating the rank of the signal space for each data matrix in the first step of model estimation (Ke et al., 2020). Assumptions 3 and 4 are needed to establish the consistency of the sample covariance matrix estimators (Bunea & Xiao, 2015). When Assumption 4 is invalid, such as in the instance of a high dimensional spiked covariance matrix, additional sparsity assumptions are needed and then the sample covariances have to be replaced by sparsity-inducing estimators in our estimation method; see the discussion section for more details.

For two signal subspaces $\mathbf{V}$ and $\hat{\mathbf{V}}$, define convergence in space as

$$\mathbf{V} \xrightarrow[\text{in space}]{} \hat{\mathbf{V}} \iff \sin\Theta(\hat{\mathbf{V}}, \mathbf{V}) \to 0.$$

Let $\mathbf{V}, \hat{\mathbf{V}} \in \mathbb{R}^{p\times d}$ both have orthonormal columns, then the vector of d principal angles between their column spaces is given by $(cos^{-1}\sigma_1, \ldots, cos^{-1}\sigma_d)^T$, where $\sigma_1 \geq \cdots \geq \sigma_d$ are the singular values of $\hat{\mathbf{V}}^T\mathbf{V}$. Let $\Theta(\hat{\mathbf{V}}, \mathbf{V})$ denote $d \times d$ diagonal matrix whose jth diagonal entry is the jth principal angle, and let $\sin\Theta(\hat{\mathbf{V}}, \mathbf{V})$ be defined entry-wise.

Before giving the proof, we first state the Davis-Kahan perturbation theory, proposed in Yu et al. (2015).

Lemma 1 (Davis-Kahan). (Yu et al., 2015) Let $\mathbf{\Sigma}, \hat{\mathbf{\Sigma}} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_p$ respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) > 0$, where $\lambda_0 := \infty$ and $\lambda_{p+1} := -\infty$. Let $d := s - r + 1$, and let $\mathbf{V} = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{\mathbf{V}} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\mathbf{\Sigma}v_j = \lambda_j v_j$ and $\hat{\mathbf{\Sigma}}\hat{v}_j = \hat{\lambda}_j\hat{v}_j$ for $j = r, r+1, \ldots, s$. Then

$$\|\sin\Theta(\hat{\mathbf{V}}, \mathbf{V})\|_F \leq \frac{2\min(d^{1/2}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{op}, \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F)}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}.$$

Moreover, there exists an orthogonal matrix $\hat{\mathbf{O}} \in \mathbb{R}^{d\times d}$ such that

$$\|\hat{\mathbf{V}}\hat{\mathbf{O}} - \mathbf{V}\|_F \leq \frac{2^{3/2}\min(d^{1/2}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{op}, \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F)}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}.$$

Lemma 1 bounds the difference between the population eigenvector and the estimated eigenvector based on the population and estimated covariance matrix, and will be used in Lemma 3. Next, we list several key lemmas that are critical in proving the consistency of our proposed estimator; a few additional lemmas are listed in the appendix. The main result is given in Theorem 1.

Lemma 2. Under Model (2), Assumptions 1, 2, 4 and $\hat{K}_i$ is from BEMA(Ke et al., 2020) used in Algorithm A.1 with $\alpha \in (0, 1/2)$ a fixed constant and $\beta \to 0$ at a properly rate as $n \to \infty$, then

$$\hat{K}_i \xrightarrow{P} \sum_{k=1}^{K} \Theta_{ik} r_k.$$

Proof. This is a direct application of Lemma 2 in Ke et al. (2020). □

Lemma 3. Let $\hat{\mathbf{\Sigma}}_i$ be the sample covariance matrix for the ith data set, $i = 1, \ldots, n$. Under Model (2) and the assumptions in Lemma 2, we have

$$\|\hat{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i\|_F \xrightarrow{P} 0,$$

and the top $\hat{K}_i$ eigenvectors $\hat{\mathbf{U}}_i$ of $\hat{\mathbf{\Sigma}}_i$ converge in probability to the top $\sum_{k=1}^{K} \Theta_{ik} r_k$ eigenvectors $\mathbf{U}_i$ of population covariance matrix $\mathbf{\Sigma}_i$, i.e.,

$$\hat{\mathbf{U}}_i \xrightarrow[\text{in space}]{P} \mathbf{U}_i.$$

Proof. By Theorem 2.1 in Bunea & Xiao (2015), there exists a constant c_1 such that

$$\|\hat{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i\|_F \leq 2c_1 \cdot \|\mathbf{\Sigma}_i\|_2 \cdot r_e(\mathbf{\Sigma}_i) \cdot \sqrt{\frac{\ln n}{n}},$$

with probability $1 - 5n^{-1}$. Thus, when $r_e(\mathbf{\Sigma}_i) \leq C$, $\|\hat{\mathbf{\Sigma}}_i - \mathbf{\Sigma}_i\|_F \xrightarrow{P} 0$. Then, by Lemma 1,

$$\hat{\mathbf{U}}_i \xrightarrow[\text{in space}]{P} \mathbf{U}_i,$$

which completes the proof. □

Theorem 1. Under Model (2), if Assumptions 1 - 4 hold, then

$$\hat{\mathbf{V}}_k \xrightarrow[\text{in space}]{P} \mathbf{V}_k, \text{ for all } k.$$

Proof. By recursively applying Lemmas 2, 3, and A.1, A.2, A.3 in the appendix, consistency is proved. □

4. Simulation Studies

4.1. Settings

Consider the two-way model (1) and let $r_A = r_{B_1} = r_{B_2} = r_{C_1} = r_{C_2} = r_{D_1} = r_{D_2} = r_{D_3} = r_{D_4} = 2$, so that the spaces of common components, partially-shared components, as well as individual components are all of dimension 2. We sample the orthonormal matrices $(\mathbf{A}, \mathbf{B}_i, \mathbf{C}_j, \mathbf{D}_{ij})$ from a Stiefel manifold and also generate the random scores and noises from normal distributions. For the diagonal covariance matrices $(\Sigma_{f_{11}}, \Sigma_{g_{11}}, \Sigma_{h_{11}}, \Sigma_{\ell_{11}}, \Sigma_{\ell_{12}}, \Sigma_{\ell_{21}}, \Sigma_{\ell_{22}})$, we sample the diagonal elements from a continuous uniform distribution on $[1, 2]$, then multiply them by the dimension of the covariance matrix p. A scale parameter is also multiplied to the eigenvalues to define the signal to noise ratio. For the other covariance matrices, we randomly rotate $\Sigma_{f_{11}}$ to $\Sigma_{f_{ij}}$ with a rotation angle $\pi/3$. Similarly, we randomly rotate $\Sigma_{g_{11}}$ and $\Sigma_{h_{11}}$ to $\Sigma_{g_{ij}}$ and $\Sigma_{h_{ij}}$ respectively with a rotation angle $\pi/2$. We generate noises for each data set with $\epsilon_{ijk} \sim N(0, \sigma^2)$, where $\sigma^2 = 1$. The case with heterogeneous noises between data sets or within the same data set is considered in a sensitivity analysis later.

We set $n_{ij} = n$ for all data sets and consider two cases as follows:

1. case 1: a high dimensional regime, $p = 500, n = 50$;

2. case 2: a low dimensional regime, $p = 50, n = 500$.

Finally, we define the signal to noise ratio (SNR) of the covariance matrix Σ_i as

$$\mathrm{SNR}_i = \frac{\mathrm{tr}(\mathbf{\Sigma}_i)}{p\sigma^2} - 1.$$

We consider three conditions on the SNR (the same for all data sets), $\mathrm{SNR} \in \{0.2, 1, 5\}$. Therefore, there are a total number of 6 model conditions. Under each model condition, 1000 simulations are conducted.

4.2. Evaluation Metrics

To evaluate the performance of our method, consider two metrics: percentage of variance explained (PVE) and space alignment (SA). The PVE metric measures the fraction of the expected variation explained by the estimated subspaces. Suppose for data set (i, j), the estimated subspaces are $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}_i$, $\hat{\mathbf{C}}_j$ and $\hat{\mathbf{D}}_{ij}$, and the population subspaces are $\mathbf{A}, \mathbf{B}_i, \mathbf{C}_j, \mathbf{D}_{ij}$, then the PVE metric is

$$\mathrm{PVE}(i,j) = \frac{\mathrm{tr}(\hat{\mathbf{A}}^T\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T\hat{\mathbf{A}} + \hat{\mathbf{B}}_i^T\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T\hat{\mathbf{B}}_i + \hat{\mathbf{C}}_j^T\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T\hat{\mathbf{C}}_j + \hat{\mathbf{D}}_{ij}^T\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T\hat{\mathbf{D}}_{ij})}{\mathrm{tr}(\mathbf{A}^T\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T\mathbf{A} + \mathbf{B}_i^T\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T\mathbf{B}_i + \mathbf{C}_j^T\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T\mathbf{C}_j + \mathbf{D}_{ij}^T\mathbf{Y}_{ij}\mathbf{Y}_{ij}^T\mathbf{D}_{ij})}.$$

The PVE metric always lies in $[0, 1]$, with 0 indicating no discovered signals, while 1 implying that the estimated subspaces precisely capture the population subspaces. Convincingly, the PVE metric can be easily extended for a model with general design.

The space alignment (SA) metric is used to measure the alignment of two spaces, for instance, $\hat{\mathbf{A}}$ is the estimated common subspace and $\mathbf{A}$ is the population common subspace, then the SA between them is

$$\mathrm{SA}(\mathbf{A}, \hat{\mathbf{A}}) = \frac{\|\hat{\mathbf{A}}^T\mathbf{A}\|_*}{\max\{\mathrm{rank}(\mathbf{A}), \mathrm{rank}(\hat{\mathbf{A}})\}}.$$

Here $\|\cdot\|_*$ denotes the nuclear norm of a matrix, and also always lies in $[0, 1]$, with 1 indicating that the two spaces align perfectly, while a 0 indicates that the two spaces are orthogonal.

4.3. Simulation Results

Table 1 gives the empirical performance of BEMA under our model setting. Except in the low signal to noise scenario in the high dimensional regime, BEMA estimates the number of spikes, indicating that it serves as an excellent vehicle for rank selection.

Next, we illustrate the performance of the proposed rank estimation algorithm (Algorithm 2) for common/partially-shared/individual subspaces for one simulated data set, under the high dimensional regime and with SNR equal to 0.5; see supplementary material. We used 3 as the cutoff for the common subspace, 1.5 for the partial and 0.5 for the individual.

Finally, we evaluate the accuracy of the estimated subspaces by our method. Fig. 2 and Fig. 3 gives the results with the evaluation metrics SA and PVE, respectively. In summary, all spaces of components can be accurately estimated. As expected, the common subspaces are estimated most accurately, followed by partial subspaces, and the individual subspaces. This is expected as common subspaces are shared by all data sets, while individual subspaces belong exclusively to each data set, thus there is simply more data to estimate the shared subspaces.

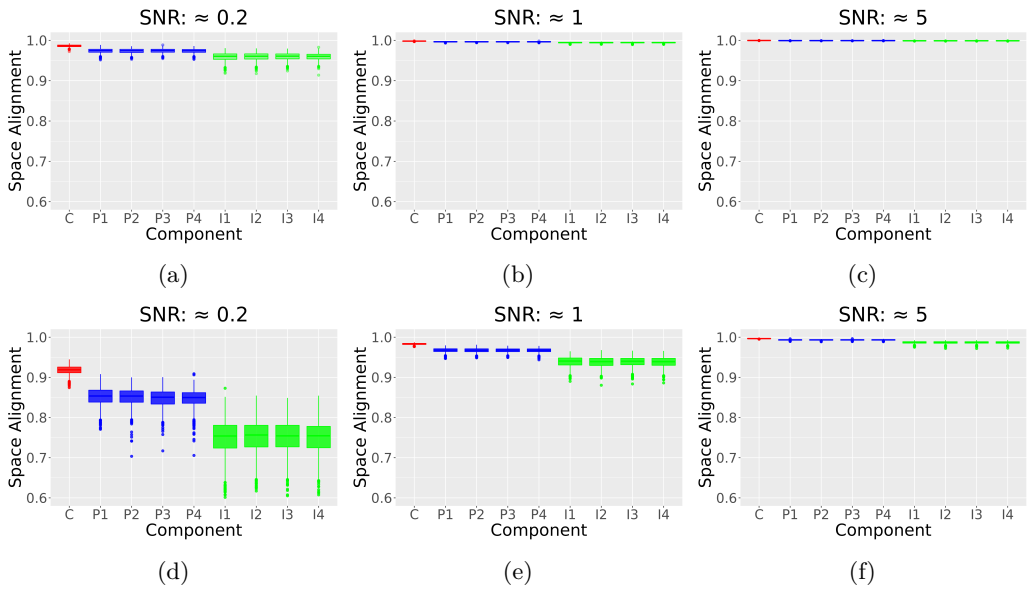

Fig. 2: Boxplots for space alignment (SA) in both the low and high dimensional regimes. In the x-axis, "C" refers to "Common", "P" refers to "Partial" and "I" refers to "Individual". Subfigures (a), (b) and (c) belong to the low dimensional regime with $p = 50, n_1 = n_2 = n_3 = n_4 = 500$, while subfigures (d), (e), (f) refer to the high dimensional, with $p = 500, n_1 = n_2 = n_3 = n_4 = 50$. SNR in both regimes varies from 0.5 to 5.

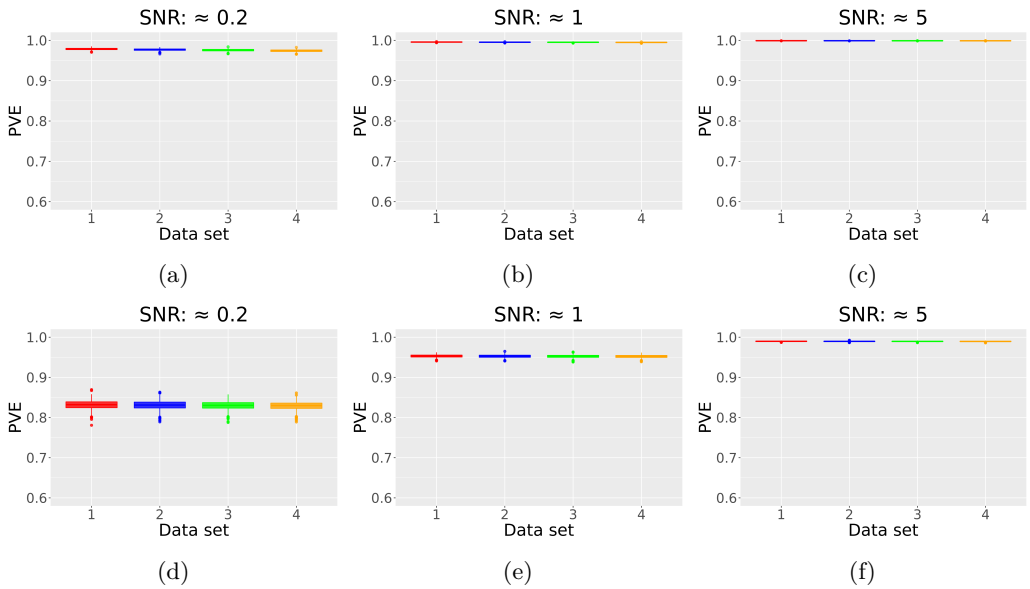

Fig. 3: Boxplots for the percentage of variance explained (PVE) result in both the low dimensional and high dimensional regimes. Subfigures (a), (b), (c) belongs to the low dimensional regime while subfigures (d), (e), (f) belong to the high dimensional regime.

Low dimensional regime: $p = 50, n = 500$			
SNR	≈ 0.2	≈ 1	≈ 5
data set-1	1.000	1.000	1.000
data set-2	0.998	1.000	1.000
data set-3	0.999	1.000	1.000
data set-4	0.999	1.000	1.000
High dimensional regime: $p = 500, n = 50$			
SNR	≈ 0.2	≈ 1	≈ 5
data set-1	0.419	1.000	1.000
data set-2	0.402	1.000	1.000
data set-3	0.401	1.000	1.000
data set-4	0.418	1.000	1.000

Table 1: Proportion of accurately estimated ranks for each data set under various simulation settings.

4.4. Sensitivity Analysis

We conduct a sensitivity analysis when the model assumptions are violated or when the number of components in the first step of the proposed algorithm is overestimated. Specifically, three main categories of sensitivity analysis are conducted:

1. Heterogeneous noises between data sets: $\epsilon_{ijk} \sim N(0, \sigma_{ij}^2)$, where $\sigma_{11}^2 = 1$, $\sigma_{12}^2 = 2$, $\sigma_{21}^2 = 3$, $\sigma_{22}^2 = 4$.
2. Heterogeneous noises within data sets: $\epsilon_{ijk} \sim N(0, \sigma_{ijk}^2)$, where $\sigma_{ijk}^2 \sim \text{uniform}(0.5, 1.5)$.
3. Overselecting the number of components for each data set.

For the third category, we manually over-select one more component for each data set, and then follow the second step using the proposed method. The results are provided in the appendix. For the first two categories of sensitivity analysis, supplementray material show that the accuracy of the estimated subspace is still reasonable, as shown by the high values of the SA and PVE metrics.

When the number of components for each data set is over-selected by one, supplementary material shows that the overselected component is usually estimated in the individual component, while the common and partial components are not affected. In applications, focus is usually on the common and partially shared components, rather than the individual components, so that we consider this to be a advantage of the proposed method. That is, shared components seem less likely to be affected by over-selected components components in the first step.

5. Experimental Data Analysis

5.1. Analysis of Four Data Sets

We apply the proposed 2s-LCA to the four brain tissue data sets focused on brain development. As described in Section 1, the data sets were collected using two technologies, bulk and single cell RNA sequencing, and under two cellular environments, *in vivo* and *in vitro*. Naturally, this exhibits a two-way design as in model (1), which includes one

subspace of common components, four subspaces of partially shared components, and four subspaces of individual components. As two dimensional visualizations of the data are of interest, a natural choice of the number of components for each subspace is 2.

Due to the heterogeneity of data sets, before joint analysis, it might be necessary to normalize the data to make the data points comparable (Leek et al., 2010). Because of the distributed nature of our algorithm, normalization across data sets is not needed, but normalization within each data set is still necessary. For each data set, we therefore center the expression value for each gene, center and scale expression level of every sample by the standard deviation across genes, to mitigate technical effects.

Fig. 4 (a) shows the projection of every data set onto its top two eigenvectors–a separate analysis. We then apply 2s-LCA to jointly analyze the four data sets and obtain common, partially-shared, and individual subspaces, each of which has rank 2. Then, we project every data set onto these subspaces to investigate biological processes associated with the scores; see Fig. 4 (b) for plots of the scores for common and partially-shared subspaces. Note that the axes in each plot have different ranges and to facilitate interpretation, we have rotated the scores associated with the common subspace (top panel of Fig. 4 (a)) simultaneously for all four data sets.

To evaluate and interpret results, we shall use the DCX gene, which is turned on as neural progenitors transition to being neurons. According to of Health et al. (2013), "The DCX gene provides instructions for producing a protein called doublecortin. This protein is involved in the movement of nerve cells (neurons) to their proper locations in the developing brain, a process called neuronal migration." The $\log_2$ transformed expression level of the DCX gene is used to color the cells, with the dark color indicating low expression level corresponding to neural progenitors and the yellow color indicating high expression level corresponding to neurons.

We first focus on the comparison of the separate components with the common components; see Fig. 4 (a) and the top row of panels in Fig. 4. For the separate analysis, the plots show that cells with different expression levels of the DCX gene are clustered and separated; however, the extent to which the scores and gene loadings align across the data sets is unclear. In contrast to the separate analysis, the proposed 2s-LCA produces jointly derived components, allowing direct assessment of how effects across the different data sets align. The plots for common components show that cells with similar expression levels in the DCX gene tend to cluster together, which means that the local structure of the cells is preserved. In addition, by projecting the data sets onto the same subspace shared among them, a similar pattern can be observed among data sets, indicating shared global structure across the data sets. For example, the x-axis along which DCX expression increases during fetal development as neural progenitors become neurons (blue arrows) can be seen precisely aligned across the common components in each of the four data sets. Moreover, a visual comparison of the separate components with those of the proposed joint method suggests that some of the order defined in the separate components is preserved in the joint analysis (in the sense of the overall shape of the distribution of cells and clustering of cells with similar expression levels), consistent with the knowledge that these four diverse experiments capture common molecular elements of neurogenesis. In addition to the visual comparison, we also compute the ratio of the variances explained by the common components versus those explained by the separate analysis for each data set. The ratios are: (1) van de Leemput: 0.195, (2) Yao: 0.524 (3) BrainSpan: 0.382, (4) Nowakowski: 0.538. The values confirm our observation made from the visual comparison above.

We have also evaluated the common components using the time of the samples: days of neural differentiation for the *in vitro* data sets and age (years) for the *in vivo* data sets; see Fig. 5 which is a recoloring of Fig. 4. Interestingly the y-axis (red arrows) appears to be a common temporal dimension orthogonal to the DCX expression component, which is along the x-axis in Fig. 4. As developmental time progresses, it appears that the cellular identities of the neural precursors and their post-mitotic neuronal progeny become less distinct, i.e. the clusters of low and high DCX expressing cells begin to merge. The cellular basis underlying this dimension is not known and might be of significant biological interest to explore further.

We next evaluate the partially-shared components and focus on the components shared only by the two *in vitro* data sets; see the left two plots on the second row of Fig. 4 (*b*). In both plots, those few cells circled in blue are pluripotent cells, which are present only in the *in vitro* data sets as they would have disappeared prior to the developmental time points measured in the *in vivo* system. This result demonstrates that the proposed joint decomposition method captures known biological effects unique to the *in vitro* experimental paradigm and not represented in the *in vivo* data sets.

5.2. A Validation Study

To validate the common components defined by the joint 2s-LCA decomposition (top row of panels in Fig. 4 and Fig. 5, we use 8 additional single cell RNA-seq data sets from 5 studies (Liu et al., 2016; Darmanis et al., 2015; Onorati et al., 2016; Camp et al., 2015; Yao et al., 2017). For those data sets, raw RNA-Seq reads were downloaded from the Sequence Read Archive at the National Center for Biotechnology Information (NCBI). These were analyzed using the Rail-RNA spliced aligner (Nellore et al., 2017) and outputs were fed to the recount2/Snaptron software system (Collado-Torres et al., 2017; Wilks et al., 2018) to generate the gene-by-sample count tables used here. We project the 8 additional data sets on the found common and *in vitro* components through the "projectR" package (Sharma et al., 2020). Projection of these data onto the common components clearly demonstrates recapitulation of the biological effect captured in the first component that is aligned with DCX expression and neurogenesis, as well as the second that is aligned with time (Fig. 6(a) and (b), blue and red arrows, respectively).

The *in vitro* components generated by 2s-LCA (first 2 panels in the second row of Fig. 4(b) and Fig. 5(b)) were validated by projecting data from another bulk RNA-seq study of neural differentiation (Burke et al., 2020) onto these components. This replicated the segregation of pluripotent stem cells with high values for the first component, away from other cells in this study that proceeded through neural progenitor and neuronal states, validating the identification of this cell type specific to the *in vitro* studies (Fig. 7, black circles).

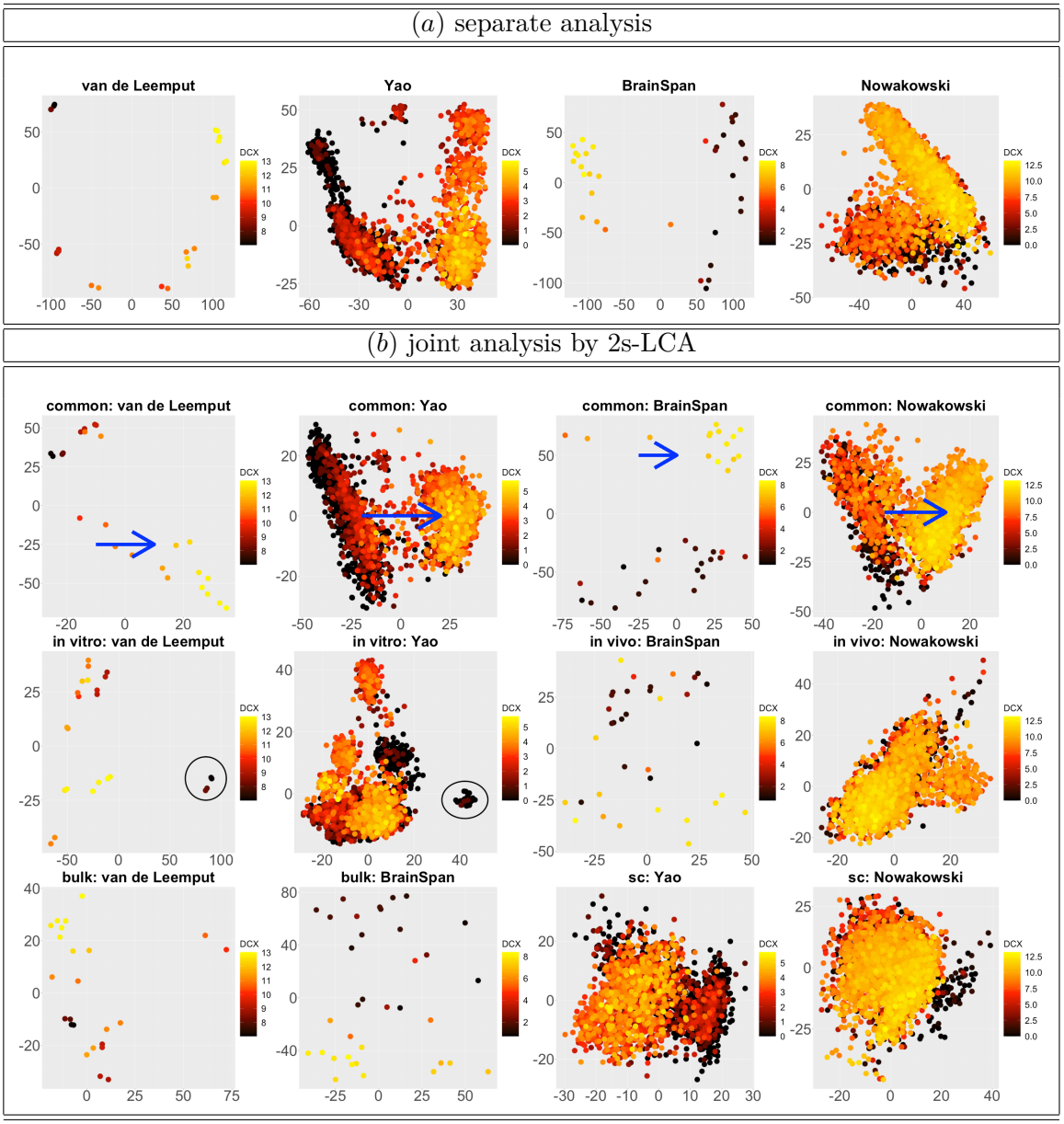

Fig. 4: (a) Scatterplots of scores corresponding to the top two principal components of each data set by separate PCA and (b) Scatterplots of scores of each data set corresponding to common components (top panel), partially-shared components associated with environments (middle panel: *in vitro* on the left two plots and *in vivo* on the right two plots), and partially-shared components associated with technologies (bottom panel: bulk on the left two plots and single cell on the right two plots) by the proposed 2s-LCA. The 4 columns in both parts correspond to the data sets: (1) van de Leemput: *in vitro* + bulk; (2) Yao: *in vitro* + single cell; (3) BrainSpan: *in vivo* + bulk; and (4) Nowakowski: *in vivo* + single cell. Each point corresponds to either one tissue sample or one cell and is colored by the $\log_2$ transformed expression level of the DCX gene. Blue arrows indicate alignment of the first common component (x-axis in top row of panels in (b) with DCX expression (neurogenesis) in all four data sets. Black circles indicate pluripotent stem cells, which are present only in the *in vitro* data sets.

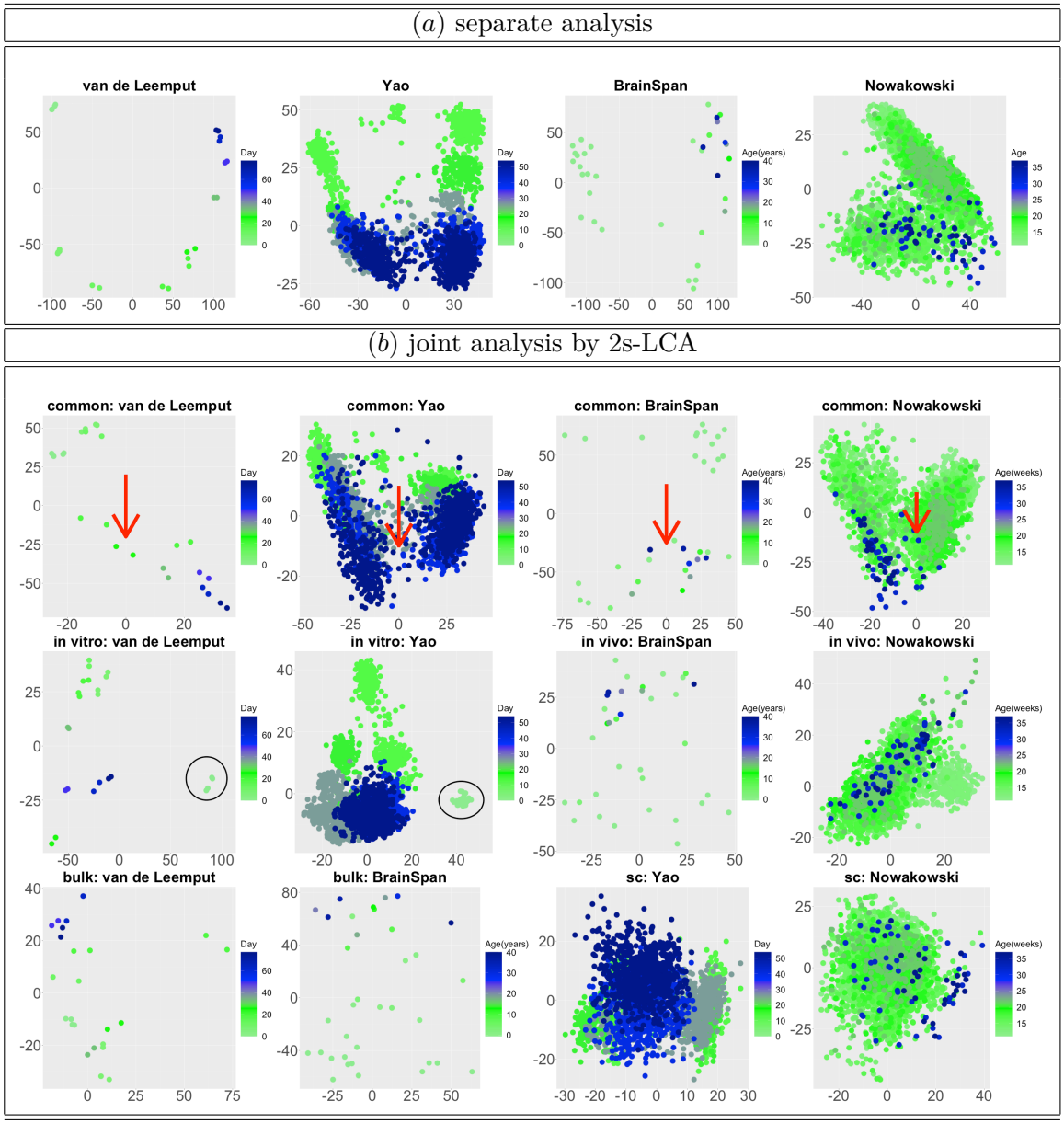

Fig. 5: This figure is a re-coloring of the data shown in Fig. 4, in order to show effects across time. Each point corresponds to either one tissue sample or one cell and is colored by days of neural differentiation for the *in vitro* data sets and age in years or gestational weeks for the *in vivo* data sets. Red arrows indicate alignment of the second common component (y-axis in top row of panels in (b) with developmental time in all four data sets. Black circles indicate pluripotent stem cells, which are present only in the *in vitro* data sets.

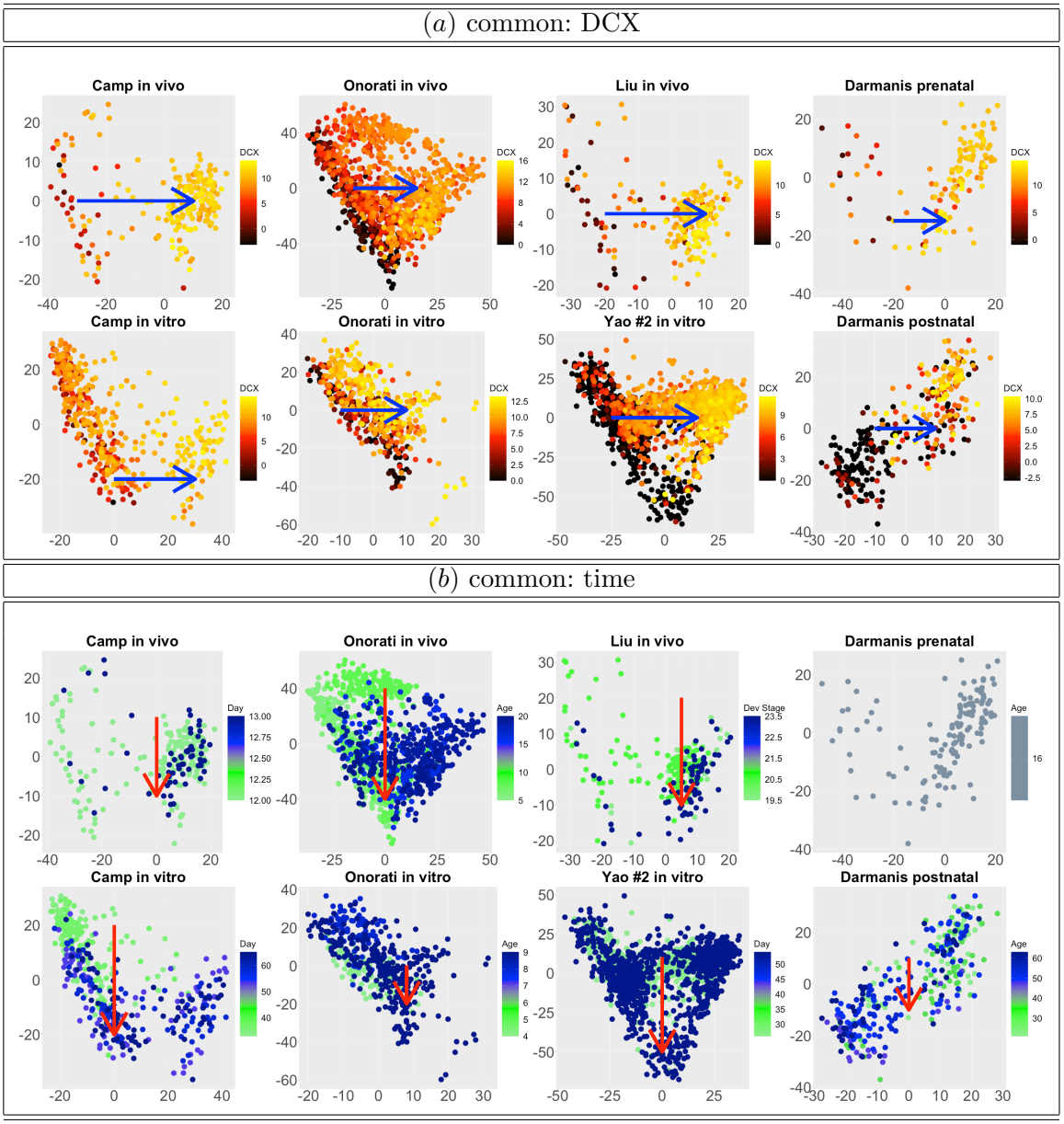

Fig. 6: Biological validation by projecting additional data sets onto common components obtained from 2s-LCA. (*a*) Projection of 8 additional scRNA-seq data sets onto the common components with cells colored by the log2 transformed expression level of the DCX gene. Blue arrows indicate alignment of the first common component with DCX expression (neurogenesis). (*b*) Projection of the same data sets onto the common components with cells colored by time: days of neural differentiation for the *in vitro* data sets and age in years or gestational weeks for the *in vivo* data sets. Red arrows indicate alignment of the second common component with developmental time. The Darmanis et al. (2015) prenatal study did not specify the exact age of the 4 fetal tissue donors used in their prenatal study, indicating only 16-18 gestational weeks for all samples.

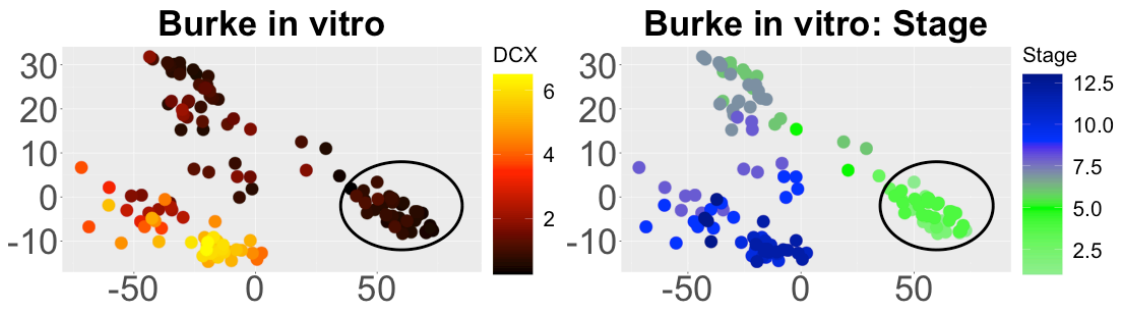

Fig. 7: Projection of an additional *in vitro* bulk RNA-seq data set onto the components shared only by the *in vitro* data sets. Black circles indicate pluripotent stem cells, which are present only in the *in vitro* data sets. Stages: 1-3 (pluripotency), 4-9 (neural progenitors), 10-13 (neurons).

6. Discussion

In this paper, we proposed two-stage linked component analysis (2s-LCA) for the joint analysis of multiple data sets that are independent but have shared underlying structure resulting from common biological processes and/or shared measurement technologies. The proposed method extracts signal spaces that can be characterized as common, partially-shared or individual, which enhances the understanding of the underlying biology between the data sets. The proposed method is general, as it deals with general design in terms of how the data sets are related. We proved the consistency of the proposed method and illustrated its empirical performance via simulation studies. Our experimental data results indicate that the 2s-LCA joint decomposition is a powerful tool to define shared molecular dynamics across biologically related data sets, while avoiding unwanted artifacts.

The proposed method is valid even for high dimensional data, as long as the sample covariance matrices are consistent estimators. For the four experimental data sets, we found that the trace and top eigenvalue of the sample covariance matrix for each data set comparable, hence it is not unreasonable to assume that the population covariance matrices are of reduced effective ranks, for which the sample covariance matrices are consistent.

The proposed method can be easily extended to high dimensional data sets, where sparsity is necessary or desired. In such cases, the sample covariance matrix estimator used in the proposed method can be replaced by any consistent covariance estimator, e.g., banded covariance estimators (Bickel et al., 2008b; Bien & Tibshirani, 2011; Bien et al., 2016) and thresholding-based covariance estimators (Bickel et al., 2008a). Then the consistency of the proposed 2s-LCA can still be established.

The second step of the proposed method relies on accurate estimation of the number of signals extracted from each data set. We showed via a sensitivity analysis that the proposed method can be robust to overestimation of the numbers of signals for estimating shared signals.

The code to conduct the two-stage linked component analysis (2s-LCA) can be found in the PJD package (https://github.com/CHuanSite/PJD) we have developed. The PJD package also contains other tools for integrative data analysis, including the integration of multiple matrices from different species and measurement

technologies. All the experimental data used to illustrate the use of 2s-LCA can be explored at the individual gene level through the NeMO Analytics portal at https://nemoanalytics.org/p?l=ChenEtAl2021&g=DCX.

Supplementary Material

Supplementary material available at Biometrika online includes additional results for simulations, e.g., scree plot for the selection of number of component, sensitivity analysis under heterogeneous noise variance and mis-selection of number of spikes, and additional results for real data analysis, e.g., 2D visualization of rotated separate components.

Appendix

Technical Details

Algorithm A.1. BEMA: standard spiked covariance model space (Ke et al., 2020).
Result: An estimate for the number of spikes, $\hat{K}$.

1. α and β tuning parameters are set to default value, $\alpha = 0.2, \beta = 0.1$.
2. Write $\tilde{p} = \min(n, p)$. For each $\alpha\tilde{p} \leq k \leq (1-\alpha)\tilde{p}$, obtain q_k, the $k/\tilde{p}$-upper-quantile of the MP distribution associated with $\sigma^2 = 1$ and $\gamma_n = p/n$. Compute

$$\hat{\sigma}^2 = \frac{\sum_{\alpha\tilde{p}\leq k\leq(1-\alpha)\tilde{p}} q_k \hat{\lambda}_k}{\sum_{\alpha\tilde{p}\leq k\leq(1-\alpha)\tilde{p}} q_k^2}.$$

3. Obtain $t_{1-\beta}$, the $(1-\beta)$-quantile of Tracy-Widom distribution. Estimate K by

$$\hat{K} = \#\{1 \leq k \leq \tilde{p} : \hat{\lambda}_k > \hat{\sigma}^2[(1+\sqrt{\gamma_n})^2 + t_{1-\beta} \cdot n^{-2/3}\gamma_n^{-1/6}(1+\sqrt{\gamma_n})^{4/3}]\}.$$

Marchenko-Pastur (MP) Distribution: For a given parameter $\gamma > 0$, the Marcenko-Pastur (MP) distribution is defined by the density

$$f_\gamma(x;\sigma^2) = \frac{1}{2\pi\sigma^2}\frac{1}{x(\gamma\wedge 1)}\sqrt{(x-\sigma^2 h_-)(\sigma^2 h_+ - x)} \cdot 1\{\sigma^2 h- < x < \sigma^2 h_+\},$$

where $h_\pm = (1 \pm \sqrt{\gamma})$. We let $F_\gamma(x;\sigma)$ denote its cumulative distribution function.

Tracy-Widom Distribution: The Tracy-Widom distribution is defined as the limit:

$$F_2(s) = \lim_{n\to\infty} \text{Prob}((\lambda_{\max} - \sqrt{2n})(\sqrt{2n})n^{1/6} \leq s)$$

Where $\lambda_{\max}$ denotes the largest eigenvalue of the random Hermitian matrix.

Lemma A.1. Under Model 2, Assumption 1, 2, 4, and based on Lemma 2, 3, we have for any k,

$$\sum_{i:\Theta_{ik}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T \xrightarrow{P} \sum_{i:\Theta_{ik}=1} \mathbf{U}_i\mathbf{U}_i^T.$$

Proof. To prove it, based on the Davis-Kahan Theorem 1, there exists one rotation matrix O, such that

$$\begin{aligned}\|\mathbf{U}_i\mathbf{U}_i^T - \hat{\mathbf{U}}_\mathbf{i}\hat{\mathbf{U}}_\mathbf{i}{}_i^T\|_F &= \|\mathbf{U}_i\mathbf{U}_\mathbf{i}{}^T - \hat{\mathbf{U}}_\mathbf{i}OO^T\hat{\mathbf{U}}_i^T\|_F \\ &= \|\mathbf{U}_i\mathbf{U}_i^T - \hat{\mathbf{U}}_iOO^T\hat{\mathbf{U}}_i^T\|_F \\ &= \|\mathbf{U}_i\mathbf{U}_i^T - \mathbf{U}_iO\hat{\mathbf{U}}_i^T + \mathbf{U}_iO\hat{\mathbf{U}}_i^T + \hat{\mathbf{U}}_iOO^T\hat{\mathbf{U}}_i^T\|_F \\ &\leq \|\mathbf{U}_i\|_F\|\mathbf{U}_\mathbf{i}{}^T - O\hat{\mathbf{U}}_i^T\|_F + \|\mathbf{U}_iO - \hat{\mathbf{U}}_\mathbf{i}\|_F\|O^T\hat{\mathbf{U}}_i\|_F.\end{aligned}$$

According to Yu et al. (2015),

$$\|\mathbf{U}_i\mathbf{U}_i^T - \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T\|_F \xrightarrow{P} 0.$$

when $\hat{\Sigma}_i \xrightarrow{P} \Sigma_i$ in Frobenius Norm. This also indicates that

$$\sum_{i:\Theta_{ik}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T \xrightarrow{P} \sum_{i:\Theta_{ik}=1} \mathbf{U}_i\mathbf{U}_i^T.$$

in the Frobenius norm. □

Lemma A.2. Under Model 2, Assumption 1, 2, 4 and based on Lemma 2, 3, A.1, by choosing the eigenvectors $\hat{\mathbf{V}}_k$ of $\sum_{i,\Theta_{ik}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T$ with eigenvalues larger than $(\eta_1+\eta_2)/2$, where $\eta_1 = \sum_{i,\Theta_{ik}=1} \Theta_{ik}$ and $\eta_2 = \max_{k'} \sum_{i:\Theta_{ik}=1,k\neq k} \Theta_{ik'}$, then

$$\hat{\mathbf{V}}_k \xrightarrow[\text{in space}]{P} \mathbf{V}_k$$

Proof. Note that

$$\sum_{i:\Theta_{ik}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T \xrightarrow{P} \sum_{i:\Theta_{ik}=1} \mathbf{U}_i\mathbf{U}_i^T.$$

Then based on the continuity of the eigenvalue operator, we have that

$$\hat{r}_k = \text{Number of eigenvalues of } \sum_{i:\Theta_{ik}=1} \hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^T \text{ larger than } (\eta_1+\eta_2)/2$$

$\hat{r}_k$ is consistent for r_k, so that

$$\hat{\mathbf{V}}_k \xrightarrow[\text{in space}]{P} \mathbf{V}_k.$$

Lemma A.3. Under Model 2, the Lemma 1, 2, 4, and based on Lemma 2, 3, A.1, A.2, we have that

$$(\mathbf{I} - \hat{\mathbf{V}}_k\hat{\mathbf{V}}_k^t)\hat{\mathbf{U}}_i \xrightarrow[\text{in space}]{P} (\mathbf{I} - \mathbf{V}_k\mathbf{V}_k^t)\mathbf{U}_i.$$

Proof. Note that in Lemma A.2, we have that

$$\hat{\mathbf{V}}_k \xrightarrow[\text{in space}]{P} \mathbf{V}_k,$$

and in Lemma A.1, we have that

$$\hat{\mathbf{U}}_k \xrightarrow[\text{in space}]{P} \mathbf{U}_k,$$

Then, according to the Continuous Mapping Theorem (Durrett, 2019),

$$(\mathbf{I} - \hat{\mathbf{V}}_k \hat{\mathbf{V}}_k^t)\hat{\mathbf{U}}_i \xrightarrow[\text{in space}]{P} (\mathbf{I} - \mathbf{V}_k \mathbf{V}_k^t)\mathbf{U}_i.$$

holds. □

References

Bickel, P. J., Levina, E. et al. (2008a). Covariance regularization by thresholding. The Annals of Statistics 36, 2577–2604.

Bickel, P. J., Levina, E. et al. (2008b). Regularized estimation of large covariance matrices. The Annals of Statistics 36, 199–227.

Bien, J., Bunea, F. & Xiao, L. (2016). Convex banding of the covariance matrix. Journal of the American Statistical Association 111, 834–845.

Bien, J. & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. Biometrika 98, 807–820.

BrainSpan, B. (2011). Atlas of the developing human brain. Secondary BrainSpan: Atlas of the Developing Human Brain .

Bunea, F. & Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. Bernoulli 21, 1200–1230.

Burke, E. E., Chenoweth, J. G., Shin, J. H., Collado-Torres, L., Kim, S.-K., Micali, N., Wang, Y., Colantuoni, C., Straub, R. E., Hoeppner, D. J. et al. (2020). Dissecting transcriptomic signatures of neuronal differentiation and maturation using ipscs. Nature communications 11, 1–14.

Camp, J. G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M. et al. (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proceedings of the National Academy of Sciences 112, 15672–15677.

Clark, B. S., Stein-O'Brien, G. L., Shiau, F., Cannon, G. H., Davis-Marcisak, E., Sherman, T., Santiago, C. P., Hoang, T. V., Rajaii, F., James-Esposito, R. E. et al. (2019). Single-cell rna-seq analysis of retinal development identifies nfi factors as regulating mitotic exit and late-born cell specification. Neuron 102, 1111–1126.

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B. & Leek, J. T. (2017). Reproducible rna-seq analysis using recount2. Nature biotechnology 35, 319–321.

Crainiceanu, C. M., Caffo, B. S., Luo, S., Zipunnikov, V. M. & Punjabi, N. M. (2011). Population value decomposition, a framework for the analysis of image populations. Journal of the American Statistical Association 106, 775–790.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Gephart, M. G. H., Barres, B. A. & Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences 112, 7285–7290.

Durrett, R. (2019). Probability: theory and examples, vol. 49. Cambridge university press.

Feng, Q., Jiang, M., Hannig, J. & Marron, J. (2018). Angle-based joint and individual variation explained. Journal of multivariate analysis 166, 241–265.

Flury, B. K. (1987). Two generalizations of the common principal component model. Biometrika 74, 59–69.

Flury, B. N. (1984). Common principal components in k groups. Journal of the American Statistical Association 79, 892–898.

Gao, L. L., Bien, J. & Witten, D. (2020). Are clusterings of multiple data views independent? Biostatistics 21, 692–708.

Gaynanova, I. & Li, G. (2017). Structural learning and integrative decomposition of multi-view data. arXiv preprint arXiv:1707.06573 .

Johnson, K. E., Howard, G. R., Morgan, D., Brenner, E., Gardner, A. L., Durrett, R. E., Mo, W., Al'Khafaji, A. M., Sontag, E., Jarrett, A. M. et al. (2020). Integrating multimodal data sets into a mathematical framework to describe and predict therapeutic resistance in cancer. bioRxiv .

Ke, Z. T., Ma, Y. & Lin, X. (2020). Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. arXiv preprint arXiv:2006.00436 .

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics 11, 733–739.

Li, G., Gaynanova, I. et al. (2018a). A general framework for association analysis of heterogeneous data. The Annals of Applied Statistics 12, 1700–1726.

Li, G. & Jung, S. (2017). Incorporating covariates into integrated factor analysis of multi-view data. Biometrics 73, 1433–1442.

Li, G., Liu, X. & Chen, K. (2018b). Integrative multi-view reduced-rank regression: Bridging group-sparse and low-rank models. arXiv preprint arXiv:1807.10375 .

Liu, S. J., Nowakowski, T. J., Pollen, A. A., Lui, J. H., Horlbeck, M. A., Attenello, F. J., He, D., Weissman, J. S., Kriegstein, A. R., Diaz, A. A. et al. (2016). Single-cell analysis of long non-coding rnas in the developing human neocortex. Genome biology 17, 1–17.

Lock, E. F., Hoadley, K. A., Marron, J. S. & Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. The annals of applied statistics 7, 523.

Micali, N., Kim, S.-K., Diaz-Bustamante, M., Stein-O'Brien, G., Seo, S., Shin, J.-H., Rash, B. G., Ma, S., Wang, Y., Olivares, N. A. et al. (2020). Variation of human neural stem cells generating organizer states in vitro before committing to cortical excitatory or inhibitory neuronal fates. Cell reports 31, 107599.

Min, E. J. & Long, Q. (2020). Sparse multiple co-inertia analysis with application to integrative analysis of multi-omics data. BMC bioinformatics 21, 1–12.

Nathoo, F. S., Kong, L., Zhu, H. & Initiative, A. D. N. (2019). A review of statistical methods in imaging genetics. Canadian Journal of Statistics 47, 108–131.

Nellore, A., Collado-Torres, L., Jaffe, A. E., Alquicira-Hernández, J., Wilks, C., Pritt, J., Morton, J., Leek, J. T. & Langmead, B. (2017). Rail-rna: scalable analysis of rna-seq splicing and coverage. Bioinformatics 33, 4033–4040.

Nowakowski, T. J., Bhaduri, A., Pollen, A. A., Alvarado, B., Mostajo-Radji, M. A., Di Lullo, E., Haeussler, M., Sandoval-Espinosa, C., Liu, S. J., Velmeshev, D. et al. (2017). Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science 358, 1318–1323.

of Health, N. I. et al. (2013). Genetics home reference. Retrieved August 8, 2013.

Onorati, M., Li, Z., Liu, F., Sousa, A. M., Nakagawa, N., Li, M., Dell'Anno, M. T., Gulden, F. O., Pochareddy, S., Tebbenkamp, A. T. et al. (2016). Zika virus disrupts phospho-tbk1 localization and mitosis in human neuroepithelial stem cells and radial glia. Cell reports 16, 2576–2592.

Orvis, J., Gottfried, B., Kancherla, J., Adkins, R. S., Song, Y., Dror, A. A., Olley, D., Rose, K., Chrysostomou, E., Kelley, M. C. et al. (2020). gear: gene expression analysis resource portal for community-driven, multi-omic data exploration. bioRxiv .

Park, J. Y. & Lock, E. F. (2020). Integrative factorization of bidimensionally linked matrices. Biometrics 76, 61–74.

Richardson, S., Tseng, G. C. & Sun, W. (2016). Statistical methods in integrative genomics. Annual review of statistics and its application 3, 181–209.

Sharma, G., Colantuoni, C., Goff, L. A., Fertig, E. J. & Stein-O'Brien, G. (2020). projectr: an r/bioconductor package for transfer learning via pca, nmf, correlation and clustering. Bioinformatics 36, 3592–3593.

Stein-O'Brien, G. L., Clark, B. S., Sherman, T., Zibetti, C., Hu, Q., Sealfon, R., Liu, S., Qian, J., Colantuoni, C., Blackshaw, S. et al. (2019). Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. Cell systems 8, 395–411.

van de Leemput, J., Boles, N. C., Kiehl, T. R., Corneo, B., Lederman, P., Menon, V., Lee, C., Martinez, R. A., Levi, B. P., Thompson, C. L. et al. (2014). Cortecon: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. Neuron 83, 51–68.

Wang, B., Luo, X., Zhao, Y. & Caffo, B. (2019). Semiparametric partial common principal component analysis for covariance matrices. bioRxiv , 808527.

Wang, M. & Allen, G. I. (2019). Integrative generalized convex clustering optimization and feature selection for mixed multi-view data. arXiv preprint arXiv:1912.05449 .

Wilks, C., Gaddipati, P., Nellore, A. & Langmead, B. (2018). Snaptron: querying splicing patterns across tens of thousands of rna-seq samples. Bioinformatics 34, 114–116.

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

Yao, Z., Mich, J. K., Ku, S., Menon, V., Krostag, A.-R., Martinez, R. A., Furchtgott, L., Mulholland, H., Bort, S., Fuqua, M. A. et al. (2017). A single-cell roadmap of lineage bifurcation in human esc models of embryonic brain development. Cell stem cell 20, 120–134.

Yu, Y., Wang, T. & Samworth, R. J. (2015). A useful variant of the davis–kahan theorem for statisticians. Biometrika 102, 315–323.