Integrative Factorization of Bidimensionally Linked Matrices

Jun Young Park, **Eric F. Lock**

Division of Biostatistics, University of Minnesota School of Public Health, Minneapolis, Minnesota, U.S.A.

SUMMARY:

Advances in molecular "omics" technologies have motivated new methodology for the integration of multiple sources of high-content biomedical data. However, most statistical methods for integrating multiple data matrices only consider data shared *vertically* (one cohort on multiple platforms) or *horizontally* (different cohorts on a single platform). This is limiting for data that take the form of *bidimensionally linked* matrices (e.g., multiple cohorts measured on multiple platforms), which are increasingly common in large-scale biomedical studies. In this paper, we propose BIDIFAC (Bidimensional Integrative Factorization) for integrative dimension reduction and signal approximation of bidimensionally linked data matrices. Our method factorizes the data into (i) globally shared, (ii) row-shared, (iii) column-shared, and (iv) single-matrix structural components, facilitating the investigation of shared and unique patterns of variability. For estimation we use a penalized objective function that extends the nuclear norm penalization for a single matrix. As an alternative to the complicated rank selection problem, we use results from random matrix theory to choose tuning parameters. We apply our method to integrate two genomics platforms (mRNA and miRNA expression) across two sample cohorts (tumor samples and normal tissue samples) using the breast cancer data from TCGA. We provide R code for fitting BIDIFAC, imputing missing values, and generating simulated data.

Keywords

BIDIFAC; Bidimensional data; Cancer genomics; Data integration; Dimension reduction; Principal component analysis

1. Introduction

1.1 Overview

Several recent methodological developments have been motivated by the integration of multiple sources of genetic, genomic, epigenomic, transcriptomic, proteomic, and other omics data. Successful integration of these disparate but related sources is essential for a complete understanding of the molecular underpinnings of human diseases, by providing

essential tools for novel hypotheses (Hawkins et al., 2010) and improving statistical power. For example, TWAS and MetaXcan/PrediXcan have shown improved power for gene-based association testing, by integrating expression quantitative trait loci (eQTL) and genome-wide association (GWAS) data (Gamazon et al., 2015; Gusev et al., 2016). Similarly, the iBAG approach (Wang et al., 2012) has shown improved power for gene-based association by integrating messenger RNA (mRNA) levels with epigenomic data (e.g., DNA methylation). Moreover, recent studies showed that protein-level variations explain additional individual phenotypic differences not explained by the mRNA levels (Wu et al., 2013).

In addition to integrating multiple sources of high-dimensional data, integrating high-dimensional data across multiple patient cohorts can also improve interpretation and statistical power. For example, the integration of genome-wide data from multiple types of cancers can improve classification of oncogenes or tumor suppressors (Kumar et al., 2015) and may improve clinical prognoses (Liu et al., 2018).

Most statistical methods for the integration of high-dimensional matrices apply to data that are linked *vertically* (e.g., one cohort measured with more than one platform, such as mRNA and miRNA) or *horizontally* (e.g., mRNA expression measured for multiple cohorts) (Tseng et al., 2012). However, linked structures in molecular biomedical data are often more complex. In particular, the integration of *bidimensionally* linked data (e.g., more than one heterogeneous groups of subjects measured by more than one platform) is largely unaddressed. In this paper, we propose a new statistical method for the low-rank structural factorization of large bidimensionally linked datasets. This can be used to accomplish three important tasks: missing value imputation; dimension reduction; and the interpretation of lower-dimensional patterns that are shared across matrices or unique to particular matrices.

1.2 Motivating Example

The Cancer Genome Atlas (TCGA) is the most comprehensive and well-curated study of the cancer genome, with data for 6 different *omics* data sources from 11,000 patients representing 33 different cancer tumor types as well as rich clinical phenotypes. We consider integrating a cohort of breast cancer (BRCA) tumor samples, and a cohort of normal adjacent tissue (NAT) samples, from TCGA. NAT samples are often used for differential analyses, e.g., to identify genes with mean differential expression between cancer and normal tissue. However, such analyses do not address the molecular heterogeneity or trans-omic interactions that characterize cancer cells. Noticeably, Aran et al. (2017) conducted a comprehensive study on NAT across different cancers using TCGA and the Genotype-Tissue Expression (GTEx) program data and showed that the expression levels of NAT from breast, colon, liver, lung, and uterine tumors yield different clustering from their respective tumor tissues. In addition, Huang et al. (2016), using TCGA data, suggested that NATs not only serve as controls to tumor tissues but also provide useful information on patients' survival that tumor samples do not. More detailed investigations of the molecular heterogeneity between tumor and NAT tissue are limited by available statistical methods, especially for multi-omics data. Our premise is that comprehensive analysis of multiple *omics* data sources across both tumor tissues and NAT would distinguish the joint signals that are shared across

different *omics* profiles (e.g., mRNA and miRNA) and those that are only attributed to the tumors.

1.3   Existing Methods on Joint Matrix Factorization

Principal components analysis (PCA) and related techniques such as the singular value decomposition (SVD) are popular for the dimension reduction of a single data matrix $\mathbf{X} : m \times n$, resulting in the low-rank approximation $\mathbf{X} \approx \mathbf{U}\mathbf{V}^T$. Here, $\mathbf{U}$ are row loadings and $\mathbf{V}$ are column scores that together explain variation in $\mathbf{X}$. There is also a growing literature on the simultaneous dimension reduction of multiple data matrices $\mathbf{X}_i$ with size $m_i \times n$, which estimate low-rank signals that are jointly shared across data matrices. To capture joint variation, concatenated PCA assumes $\mathbf{X}_i = \mathbf{U}_i\mathbf{V}^T$ for each matrix $\mathbf{X}_i$, i.e., the scores are shared across matrices. The iCluster (Shen et al., 2009) and irPCA (Liu et al., 2016) approaches make this assumption for the integration of multi-source biomedical data. Alternatively, more flexible approaches allow for structured variation that may be shared across matrices or specific to individual matrices. The Joint and Individual Variations Explained (JIVE) method (Lock et al., 2013) decomposes joint and individual low-rank signals across matrices via the decomposition $\mathbf{X}_i = \mathbf{U}_i\mathbf{V}^T + \mathbf{W}_i\mathbf{V}_i^T + \mathbf{E}_i$. In the context of vertical integration, the joint and individual scores $\mathbf{V}$ and $\mathbf{V}_i$ have been applied to risk prediction (Kaplan and Lock, 2017) and clustering (Hellton and Thoresen, 2016) for high-dimensional data. Several related techniques such as AJIVE (Feng et al., 2018) and SLIDE (Gaynanova and Li, 2019) have been proposed (Zhou et al., 2016), as well as extensions that allow the adjustment of covariates (Li and Jung, 2017) or accommodate heterogeneity in the distributional assumptions for different sources (Li and Gaynanova, 2018; Zhu et al., 2018).

The aforementioned methods focus exclusively on data that share a single dimension (i.e., either horizontally or vertically), and extension to matrices that are linked both vertically and horizontally is not straightforward. O'Connell and Lock (2019) decompose shared and individual low-rank structure for three interlinked matrices $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ where $\mathbf{X}$ and $\mathbf{Y}$ are shared vertically and $\mathbf{X}$ and $\mathbf{Z}$ are shared horizontally. However, their approach is not directly applicable to more general forms of bidimensionally linked data, and it suffers from potential convergence to a local minimum of the objective during estimation.

1.4   Our Contribution

We propose the first unified framework to decompose bidimensionally linked matrices into globally shared, horizontally shared (i.e., row-shared), vertically shared (i.e., column-shared), or individual structural components. Our specific aims are to (i) separate shared and individual structures, (ii) separate the shared components into one of globally-shared, column-shared, or row-shared structures, and (iii) maintain the low-rank structures for the signals. Our approach extends soft singular value thresholding (SSVT), i.e., nuclear norm penalization, for a single matrix. It requires optimizing a single convex objective function, which is relatively computationally efficient. It also facilitates a simple and intuitive approach based on random matrix theory for model specification, rather than complex and computationally expensive procedures to select tuning parameters or model ranks. Although our primary focus is bidimensional integration, our approach includes a novel method for

vertical-only or horizontal-only integration as a special case. We show in simulation studies that our method outperforms existing methods, including JIVE.

The rest of this article is organized as follows. Section 2 describes the proposed method, denoted by BIDIFAC (bidimensional integrative factorization) for bidimensionally linked matrices, and addresses estimation, tuning parameter selection and imputation algorithms. Section 3 constructs simulated data under various scenarios and compares our method to existing methods in terms of structural reconstruction error and imputation performance. In Section 4, we apply BIDIFAC to the breast cancer and NAT data obtained from TCGA and illustrate the utility of the model. We conclude with some points of discussions in Section 5.

2. Methods

2.1 Notation and Definitions

Consider a set of pq matrices $\{\mathbf{A}_{ij} : m_i \times n_j \mid i = 1,...,p, j = 1,...,q\}$, which may be concatenated to form the matrix

$$\mathbf{A}_{00} = \begin{bmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1q} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{p1} & \cdots & \mathbf{A}_{pq} \end{bmatrix}, \tag{1}$$

where $\mathbf{A}_{00} : m_0 \times n_0$ with with $m_0 = \sum_{i=1}^{p} m_i$ and $n_0 = \sum_{j=1}^{p} n_j$. Analogously, we define the column-concatenated matrices $\mathbf{A}_{i0} = [\mathbf{A}_{i1},...,\mathbf{A}_{iq}]$ for $i = 1,...,p$ and the row-concatenated matrices $\mathbf{A}_{0j} = \left[\mathbf{A}_{1j}^T, ..., \mathbf{A}_{pj}^T\right]^T$ for $j = 1,...,q$. We first define terms to characterize the relationships among these data matrices.

DEFINITION 1: In the arrangement (1), a set of data matrices with the structure of $\{\mathbf{X}_{ij} | i = 1,...,p, j = 1,...,q\}$ follows a $p \times q$ *bidimensionally linked structure*. The elements of $\{\mathbf{X}_{ij} | j = 1,...,q\}$ are *row-shared* and the elements of $\{\mathbf{X}_{ij} | i = 1,...,p\}$ are *column-shared*. The elements of $\{\mathbf{X}_{ij} | i = 1,...,p, j = 1...,q\}$ are *globally-shared* if every $\mathbf{X}_{ij}$ in the set is column-shared with $\mathbf{X}_{i'j}$ and row-shared with $\mathbf{X}_{ij'}$ for $i' = 1,...,p$ and $j' = 1,...,q$.

2.2 Model Specification

We assume that each matrix is decomposed by $\mathbf{X}_{ij} = \mathbf{S}_{ij} + \mathbf{E}_{ij}$, where $\mathbf{S}_{ij}$ is a low-rank signal matrix and $\mathbf{E}_{ij}$ is a full-rank white noise. We further assume that $\mathbf{S}_{ij}$ can be decomposed as

$$\mathbf{S}_{ij} = \mathbf{G}_{ij} + \mathbf{R}_{ij} + \mathbf{C}_{ij} + \mathbf{I}_{ij}, \tag{2}$$

where $\mathbf{G}_{ij}$ is globally shared structure, $\mathbf{R}_{ij}$ is row-shared structure, $\mathbf{C}_{ij}$ is column-shared structure, and $\mathbf{I}_{ij}$ is individual structure for matrix $\mathbf{X}_{ij}$. The shared nature of the terms are apparent from their factorized forms. Defining the parameter set as $\Theta = \{\mathbf{G}_{00}, \mathbf{R}_{i0}, \mathbf{C}_{0j}, \mathbf{I}_{ij} | i = 1, \cdots, p, j = 1, \cdots, q\}$, we write each term as a product of row *loadings* $\mathbf{U}$ and column *scores* $\mathbf{V}$:

$$\mathbf{G}_{00} = \mathbf{U}_{00}^{(G)}\mathbf{V}_{00}^{(G)T}, \quad \mathbf{R}_{i0} = \mathbf{U}_{i0}^{(R)}\mathbf{V}_{i0}^{(R)T}, \quad \mathbf{C}_{0j} = \mathbf{U}_{0j}^{(C)}\mathbf{V}_{0j}^{(C)T}, \quad \mathbf{I}_{ij} = \mathbf{U}_{ij}^{(I)}\mathbf{V}_{ij}^{(I)T}. \quad (3)$$

The loadings and scores are shared across matrices for the global components $\mathbf{G}$, i.e., row-shared matrices have common global loadings $\mathbf{U}_{i0}^{(G)}$, and column-shared matrices have common global scores $\mathbf{V}_{0j}^{(G)}$. The row-shared structures $\mathbf{R}$ have common loadings, and the column-shared structures $\mathbf{C}$ have common scores. The dimensions of $\mathbf{U}$ and $\mathbf{V}$ depend on the global shared rank r_{00}, column-shared ranks r_{0j}, row-shared ranks r_{i0}, and individual ranks r_{ij}: $\mathbf{U}_{ij}^{(\cdot, \cdot)}$: $m_i \times r_{ij}$ and $\mathbf{V}_{ij}^{(\cdot, \cdot)}$: $n_j \times r_{ij}$ for $i = 0, \cdots, p$, and $j = 0, \cdots q$. A diagram with the proposed notation for 2×2 linked structure is shown in Figure 1.

To understand the factorized forms in (3), it is instructive to consider their interpretation under the motivating example of Section 1.2. Say $\mathbf{X}_{11}$ gives gene (mRNA) expression for tumor samples, $\mathbf{X}_{12}$ gives gene expression for NAT samples, $\mathbf{X}_{21}$ gives miRNA expression for tumor samples, and $\mathbf{X}_{22}$ gives miRNA expression for NAT samples. Then, the loadings of the global structures $\mathbf{G}_{00}$ give trans-omic signatures $\mathbf{U}_{00}^{(G)}$ that explain substantial variability across both tumor and NAT samples with associated scores $\mathbf{V}_{00}^{(G)}$. The loadings of the column-shared structures $\mathbf{C}_{01}$ include trans-omic signatures $\mathbf{U}_{01}^{(C)}$ that explain substantial variability in the tumor samples with associated scores $\mathbf{V}_{01}^{(C)}$ but not the NAT samples. The loadings of the row-shared structures $\mathbf{R}_{10}$ include gene signatures $\mathbf{U}_{10}^{(R)}$ that explain substantial variability across both tumor and NAT samples, but are unrelated to miRNA. The loadings of the individual structures $\mathbf{I}$ include gene signatures unrelated to miRNA that explain variability in only the tumor samples, $\mathbf{U}_{11}^{(I)}$.

Marginally, $\mathbf{G}_{i0} + \mathbf{R}_{i0}$ denotes the components shared by $\mathbf{X}_{i0}$ and $\mathbf{G}_{0j} + \mathbf{C}_{0j}$ denotes the components shared by $\mathbf{X}_{0j}$. In practice, we assume the errors $\mathbf{E}_{ij}$ are Gaussian white noise with mean zero and variance σ_{ij}^2. For vertical integration, $q = 1$ and the $\mathbf{G}_{ij}$ and $\mathbf{R}_{ij}$ terms are redundant with $\mathbf{C}_{ij}$ and $\mathbf{I}_{ij}$, respectively. Thus, we suppress these terms and the decomposition of the model reduces to the "joint and individual" structures as in onedimensional factorization methods, including JIVE. Note, however, that we do not require pairwise orthogonality constraints on $\mathbf{G}_{ij}$, $\mathbf{R}_{ij}$, $\mathbf{C}_{ij}$, and $\mathbf{I}_{ij}$, which will be discussed in the remaining sections.

2.3 Estimation

Without any penalization, estimation of $\boldsymbol{\Theta}$ would have an identifiability issue. We first consider minimizing the sum of squared errors over all matrices with different levels of matrix L_2 penalties on the $\left\{ \mathbf{U}_{ij}^{(\cdot)}, \mathbf{V}_{ij}^{(\cdot)} \right\}$. Our objective function is

$$
\begin{aligned}
f_1\Big(\Big\{&\mathbf{U}_{ij}^{(\cdot\,,)}, \mathbf{V}_{ij}^{(\cdot\,,)}\,\Big|\,i=0,\dots,p,\,j=0,\dots,q,(i,j)\neq(0,0)\Big\}\Big)\\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}\Big\|\mathbf{X}_{ij}-\mathbf{U}_{i0}^{(G)}\mathbf{V}_{0j}^{(G)T}-\mathbf{U}_{i0}^{(R)}\mathbf{V}_{ij}^{(R)T}-\mathbf{U}_{ij}^{(C)}\mathbf{V}_{0j}^{(C)T}-\mathbf{U}_{ij}^{(I)}\mathbf{V}_{ij}^{(I)T}\Big\|_{F}^{2}\\
&\quad+\lambda_{00}\Big(\big\|\mathbf{U}_{00}^{(G)}\big\|_{F}^{2}+\big\|\mathbf{V}_{00}^{(G)}\big\|_{F}^{2}\Big)+\sum_{i=1}^{p}\lambda_{i0}\Big(\big\|\mathbf{U}_{i0}^{(R)}\big\|_{F}^{2}+\big\|\mathbf{V}_{i0}^{(R)}\big\|_{F}^{2}\Big)\\
&\quad+\sum_{j=1}^{q}\lambda_{0j}\Big(\big\|\mathbf{U}_{0j}^{(C)}\big\|_{F}^{2}+\big\|\mathbf{V}_{0j}^{(C)}\big\|_{F}^{2}\Big)+\sum_{i=1}^{p}\sum_{j=1}^{q}\lambda_{ij}\Big(\big\|\mathbf{U}_{ij}^{(I)}\big\|_{F}^{2}+\big\|\mathbf{V}_{ij}^{(I)}\big\|_{F}^{2}\Big),
\end{aligned}
\tag{4}
$$

where $\|\cdot\|_F$ denotes the Frobenious norm and each λ_{ij} is a non-negative penalty factor. Following Section 2.1,

$\mathbf{U}_{0j}^{(C)}=\Big[\mathbf{U}_{1j}^{(C)T},\cdots,\mathbf{U}_{pj}^{(C)T}\Big]$, $\mathbf{V}_{i0}^{(R)}=\Big[\mathbf{V}_{i1}^{(R)},\cdots,\mathbf{V}_{iq}^{(R)}\Big]$, $\mathbf{U}_{00}^{(G)}=\Big[\mathbf{U}_{10}^{(G)T},\dots,\mathbf{U}_{p0}^{(G)T}\Big]^{T}$, and $\mathbf{V}_{00}^{(G)}$ We $=\Big[\mathbf{V}_{01}^{(G)},\cdots\mathbf{V}_{0q}^{(G)}\Big]$.

upper bound the ranks by setting $r_{ij}=\min(m_i,\,n_j)$ for $i=0,\dots,p$ and $j=0,\dots,q$; the actual ranks of the solution may be lower, as discussed below. The objective function (4) is a convex function of each of $\mathbf{U}_{ij}^{(\cdot\,,)}$ and $\mathbf{V}_{ij}^{(\cdot\,,)}$, given all the others fixed. One may use alternating least squares (ALS) with a matrix L_2 penalty to iteratively update each of $\Big\{\mathbf{U}_{00}^{(G)},\mathbf{V}_{00}^{(G)}\Big\},\Big\{\mathbf{U}_{i0}^{(R)},\mathbf{V}_{i0}^{(R)}\Big\},\Big\{\mathbf{U}_{0j}^{(C)},\mathbf{V}_{0j}^{(C)}\Big\},\Big\{\mathbf{U}_{ij}^{(I)},\mathbf{V}_{ij}^{(I)}\Big\}$ until convergence.

Alternatively, we reformulate (4) and motivate our model using nuclear norm penalties. A matrix $\mathbf{A}:m\times n$ with ordered singular values $\delta_1,\delta_2,\dots$ has nuclear norm $\|\mathbf{A}\|_*=\sum_{i=1}^{\min\{m,n\}}\delta_i$. We first present a well-known result on the equivalence of nuclear norm penalization and matrix factorization for a single matrix in Proposition 1.

PROPOSITION 1: (Mazumder et al., 2010) For a matrix $\mathbf{X}:m\times n$,

$$
\min_{\mathbf{U}:n\times r,\,\mathbf{V}:m\times r}\|\mathbf{X}-\mathbf{U}\mathbf{V}^{T}\|_{F}^{2}+\lambda\Big(\|\mathbf{U}\|_{F}^{2}+\|\mathbf{V}\|_{F}^{2}\Big)=\min_{\mathbf{Z}:r(\mathbf{Z})\leqslant r}\|\mathbf{X}-\mathbf{Z}\|_{F}^{2}+2\lambda\|\mathbf{Z}\|_{*}
\tag{5}
$$

where $r=\min(m,n)$. Moreover, $\widehat{\mathbf{Z}}=\widehat{\mathbf{U}}\widehat{\mathbf{V}}^{T}$, where $\widehat{\mathbf{Z}}$ solves the right-hand side of (5) and $\{\widehat{\mathbf{U}},\widehat{\mathbf{V}}\}$ solves the left-hand side of (5).

Proposition 1 depends on Lemma 1, which is also shown in Mazumder et al. (2010).

LEMMA 1: *For a matrix* $\mathbf{Z}:m\times n$, $\displaystyle\min_{\substack{\mathbf{U}\mathbf{V}^{T}=\mathbf{Z}\\\mathbf{U}:m\times\min(m,n)\\\mathbf{V}:m\times\min(m,n)}}\Big(\|\mathbf{U}\|_{F}^{2}+\|\mathbf{V}\|_{F}^{2}\Big)=2\|\mathbf{Z}\|_*.$

We extend the nuclear norm objective in (5) to our context as follows:

$$f_2(\Theta) = \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{q} \|\mathbf{X}_{ij} - \mathbf{G}_{ij} - \mathbf{R}_{ij} - \mathbf{C}_{ij} - \mathbf{I}_{ij}\|_F^2$$

$$+ \lambda_{00}\|\mathbf{G}_{00}\|_* + \sum_{i=1}^{p} \lambda_{i0}\|\mathbf{R}_{i0}\|_* + \sum_{j=1}^{q} \lambda_{0j}\|\mathbf{C}_{0j}\|_* + \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_{ij}\|\mathbf{I}_{ij}\|_*.$$

(6)

Following Section 2.1, $\mathbf{R}_{i0} = [\mathbf{R}_{i1}, \cdots, \mathbf{R}_{iq}]$, $\mathbf{C}_{0j} = [\mathbf{C}_{1j}^T, \cdots \mathbf{C}_{pj}^T]^T$, and $\mathbf{G}_{00} = [\mathbf{G}_{01}, \cdots, \mathbf{G}_{0q}]$. Theorem 1 establishes the equivalence of (4) and (6), with proof in Web Appendix B.

THEOREM 1: *Let*

$$\widehat{\Theta}_1 = \{\widehat{\mathbf{U}}_{i0}^{(G)}\widehat{\mathbf{V}}_{0j}^{(G)T}, \widehat{\mathbf{U}}_{i0}^{(R)}\widehat{\mathbf{V}}_{ij}^{(R)T}, \widehat{\mathbf{U}}_{ij}^{(C)}\widehat{\mathbf{V}}_{0j}^{(C)T}, \widehat{\mathbf{U}}_{ij}^{(I)}\widehat{\mathbf{V}}_{ij}^{(I)T} \Big| i = 1\ldots p, j = 1, \ldots q\}$$

minimize (4). *Then,*

$$\widehat{\Theta}_2 = \left\{\widehat{\mathbf{G}}_{ij}, \widehat{\mathbf{R}}_{ij}, \widehat{\mathbf{C}}_{ij}, \widehat{\mathbf{I}}_{ij} \Big| i = 1\ldots p, j = 1, \ldots q\right\}$$

minimizes (6), *where*
$\widehat{\mathbf{G}}_{ij} = \widehat{\mathbf{U}}_{i0}^{(G)}\widehat{\mathbf{V}}_{0j}^{(G)T}, \widehat{\mathbf{R}}_{ij} = \widehat{\mathbf{U}}_{i0}^{(R)}\widehat{\mathbf{V}}_{ij}^{(R)T}, \widehat{\mathbf{C}}_{ij} = \widehat{\mathbf{U}}_{ij}^{(C)}\widehat{\mathbf{V}}_{0j}^{(C)T}, \text{ and } \widehat{\mathbf{I}}_{ij} = \widehat{\mathbf{U}}_{ij}^{(I)}\widehat{\mathbf{V}}_{ij}^{(I)T}.$

The objective (6) has several advantages. First, the function is convex, which we state in Theorem 2 and prove in Web Appendix B.

THEOREM 2: *The objective $f_2(\cdot)$ in (6) is convex over its domain.*

Fortunately, minimizing one term ($\mathbf{G}_{00}$, $\mathbf{R}_{i0}$, $\mathbf{C}_{0j}$, or $\mathbf{I}_{ij}$) with the others fixed is straightforward via soft singular value thresholding (SSVT). We state the well-known equivalence between nuclear norm penalization and SSVT in Proposition 2; for a proof see Mazumder et al. (2010).

PROPOSITION 2: If the SVD of $\mathbf{X}$ is $\mathbf{U}_X \mathbf{D}_X \mathbf{V}_X^T$ and $\mathbf{D}_X$ has diagonal entries $\delta_1 \geqslant \cdots \geqslant \delta_r \geqslant 0$, the solution for $\mathbf{Z}$ in (5) is equal to $\widehat{\mathbf{Z}} = \mathbf{U}_X \mathbf{D}_X(\lambda)\mathbf{V}_X^T$, where $\mathbf{D}_X(\lambda)$ is a diagonal matrix with $\delta_1, \ldots, \delta_r$ replaced by $\max(\delta_1 - \lambda, 0), \ldots, \max(\delta_r - \lambda, 0)$, respectively.

We use an *iterative soft singular value thresholding* (ISSVT) algorithm to solve (6), applying Proposition 2 to the appropriate residual matrices for $\mathbf{G}_{00}$, $\mathbf{R}_{i0}$, $\mathbf{C}_{0j}$, or $\mathbf{I}_{ij}$. For example, $\widehat{\mathbf{I}}_{ij}$ and $\widehat{\mathbf{R}}_{i0}$ are obtained by soft-thresholding the singular values of $\mathbf{X}_{ij} - \widehat{\mathbf{G}}_{ij} - \widehat{\mathbf{R}}_{ij} - \widehat{\mathbf{C}}_{ij}$ and $\mathbf{X}_{i0} - \widehat{\mathbf{G}}_{i0} - \widehat{\mathbf{C}}_{i0} - \widehat{\mathbf{I}}_{i0}$ towards 0, respectively. This iterative algorithm is guaranteed to converge to a coordinatewise-minimum, and convexity implies that it will be a global minimum if it is a local minimum. In practice, we find that iterative algorithms to solve either (4) or (6) converge to the same solution and are robust to their initial values. The detailed algorithm of ISSVT is provided in Web Appendix A. The

resulting global, column-shared, row-shared and individual terms of the decomposition will have reduced rank depending on the penalty factors λ_{ij}. Moreover, this relation between the penalty factors and the singular values motivates a straightforward choice of tuning parameters using random matrix theory described in Section 2.6.

ISSVT can also be represented as a blockwise coordinate descent algorithm for the ALS objective (4) (see Web Appendix A), and it converges faster than ALS, so we use it as our default algorithm. However, the ALS approach may be extended to certain related contexts. For example, it can incorporate sparsity in the loadings via an additional penalty (e.g., L_1) on $\mathbf{U}_{ij}^{(\,;\,)}$. Also, with only three linked matrices $\mathbf{X}_{11}$, $\mathbf{X}_{12}$ and $\mathbf{X}_{21}$, as in O'Connell and Lock (2019), (6) cannot properly construct the globally shared components. The formulation of (4) with an ALS algorithm can handle such cases, where the linked matrices do not form a complete 2-way grid.

2.4 Data Pre-Processing

In practice, the data matrices $\mathbf{X}_{ij}$ may have very different levels of variability or be measured on different scales. Thus, a straightforward application of the objective (4) or (6) is not appropriate without further processing. By default we center the matrices to have mean 0, so that each matrix has the same baseline. To resolve issues of scale, we propose dividing each data matrix $\mathbf{X}_{ij}$, $i, j > 0$ by an estimate of the square root of its noise variance $\hat{\sigma}_{ij}$, denoted by $\mathbf{X}_{ij,scale}$. We discuss estimating the Gaussian noise variance in Section 2.6. After scaling, each matrix has homogeneous unit noise variance, which motivates the proposed penalties. After all components, denoted by $\mathbf{G}_{ij,scale}$, $\mathbf{R}_{ij,scale}$, $\mathbf{C}_{ij,scale}$, and $\mathbf{I}_{ij,scale}$, are estimated, we transform the results back to the original scale by multiplying each matrix by $\hat{\sigma}_{ij}$.

We comment on estimating the noise variance σ_{ij}^2. Without any signal or with a very weak signal, the standard deviation of $\text{vec}(\mathbf{X}_{ij})$, denoted as $\hat{\sigma}_{ij}^{SD}$, provides a nearly unbiased estimate of σ_{ij}. However, this estimate is biased and overly conservative with a high signal-to-noise ratio. An alternative is to use random matrix theory, and estimate σ_{ij} by minimizing the Kolmogorov-Smirnov distance between the theoretical and empirical distribution functions of singular values, as in Shabalin and Nobel (2013). Their estimate $\hat{\sigma}_{ij}^{KS}$ is based on grid-search on a candidate set of σ. Recently, Gavish and Donoho (2017) proposed another estimator $\hat{\sigma}_{ij}^{MAD}$ based on random matrix theory, which is defined as the median of the singular values of $\mathbf{X}_{ij}$ divided by the square root of the median of the Marcenko-Pastur distribution. Our simulations, not shown here, revealed that both $\hat{\sigma}_{ij}^{KS}$ and $\hat{\sigma}_{ij}^{MAD}$ well approximate the standard deviation of a true noise matrix when the data matrix consists of low rank signal, and we use $\hat{\sigma}_{ij}^{MAD}$ as a default throughout this paper for its simplicity. From here, we assume that BIDIFAC is applied to the data matrices with $\sigma_{ij}^2 = 1$.

2.5 Summarizing Results

Given that the mean of each matrix is 0, we propose proportion of variance explained $\left(R_{\mathbf{X}_{ij}}^2(\,\cdot\,) \right)$ as a summary statistic. For example,

$$R_{\mathbf{X}_{ij}}^2(\mathbf{G}_{ij}) = 1 - \frac{\|\mathbf{X}_{ij} - \mathbf{G}_{ij}\|_F^2}{\|\mathbf{X}_{ij}\|_F^2} \tag{7}$$

provides a measure of the proportion of variability explained by the globally shared component. However, because orthogonality is not explicitly enforced in BIDIFAC, this equality does not hold in general and $R^2(\cdot)$ is not necessarily additive across terms (e.g., $R_{\mathbf{X}_{ij}}^2(\mathbf{G}_{ij} + \mathbf{C}_{ij}) \neq R_{\mathbf{X}_{ij}}^2(\mathbf{G}_{ij}) + R_{\mathbf{X}_{ij}}^2(\mathbf{C}_{ij})$).

2.6 Selecting Tuning Parameters

The performance of the proposed method depends heavily on the choice of tuning parameters. In the literature, there are several approaches to select ranks in the context of vertical integration, including permutation testing (Lock et al., 2013), BIC (O'Connell and Lock, 2016), and cross-validation (Li and Jung, 2017). In our context, the issue of rank selection is analogous to selecting the tuning parameters λ_{ij}. Although cross-validation is a natural way of selecting tuning parameters in penalized regression, our objective involves too many parameters ($1+p+q+pq$) to be computationally feasible. Moreover, despite the rich literature on cross-validating SVD for a single matrix, it is not clear how to define the training and test sets (e.g., randomly select cells, rows, columns, a whole matrix, etc). A general description of the difficulties in cross-validating matrices is provided by Owen and Perry (2009).

Admitting that cross validation in BIDIFAC is not straightforward and inefficient, we provide an alternative approach to select the tuning parameters based on random matrix theory. We first construct necessary conditions for each element of $\widehat{\Theta}$ to be nonzero.

PROPOSITION 3: The following conditions are necessary to allow for non-zero $\widehat{\mathbf{G}}_{00}, \widehat{\mathbf{R}}_{i0}, \widehat{\mathbf{C}}_{0j}$, and $\widehat{\mathbf{I}}_{ij}$:

1. $\max_j \lambda_{ij} < \lambda_{i0} < \sum_j \lambda_{ij}$ for $i = 1, \ldots, p$ and $\max_i \lambda_{ij} < \lambda_{0j} < \sum_i \lambda_{ij}$ for $j = 1, \ldots, q$

2. $\max_i \lambda_{i0} < \lambda_{00} < \sum_i \lambda_{i0}$ and $\max_j \lambda_{0j} < \lambda_{00} < \sum_j \lambda_{0j}$.

We provide a proof of Proposition 3 in Web Appendix B. Without loss of generality, suppose that each cell of $\mathbf{X}_{ij}$ has Gaussian noise with unit variance ($\sigma_{ij}^2 = 1$), which is independent within each matrix and across matrices. Based on random matrix theory, we propose using the following penalty factors:

$$\lambda_{ij} = \sqrt{m_i} + \sqrt{n_j}, \quad \text{where} \quad i = 0, \ldots, p, \quad \text{and} \quad j = 0, \ldots, q. \tag{8}$$

It is straightforward to show that our choice of tuning parameters meets the necessary requirement. Also, under the aforementioned assumptions, $\sqrt{m_i} + \sqrt{n_j}$ provides a tight upper bound for the largest singular value of $\mathbf{E}_{ij}$ (Rudelson and Vershynin, 2010). Thus, without any shared structure, the motivation for (8) is apparent by considering the penalty as a soft-thresholding operator on the singular values in Proposition 2. The penalty (8) is also used in

the single matrix reconstruction method in Shabalin and Nobel (2013). The penalties for the shared components are decided analogously because the stack of column-/row-shared matrices are also Gaussian random matrices with unit noise variance. For example, the penalty for $\mathbf{R}_{i0}$ is determined by an estimate of the largest singular value of its concatenated noise matrix $\mathbf{E}_{i0}$ with unit variance: $\sqrt{m_i} + \sqrt{n_0}$.

2.7 Imputation

A convenient feature of BIDIFAC is its potential for missing value imputation. For single block data, PCA or related low-rank factorizations can be used to impute missing values by iteratively updating missing entries with their low-rank approximation, and this approach has proven to be very accurate in many applications (Kurucz et al., 2007). An analogous algorithm has been used for imputation with joint matrix factorization (O'Connell and Lock, 2019), and this approach readily extends to BIDIFAC. Importantly, this allows for the imputation of data that are missing an entire row or column within a block, via an expectation-maximization (EM) approach. Our imputation algorithm is presented below.

1. Let $\mathcal{I}_{ij} = \{(r, s) | \mathbf{X}_{ij}[r, s] \text{ is missing} \}$. For each matrix, initialize the missing values by the column- and/or row-wise mean and denote the initial matrix by $\widehat{\mathbf{X}}_{ij}^{(old)}$.

2.
   a. Maximization: Apply BIDIFAC to $\left\{ \widehat{\mathbf{X}}_{ij}^{(old)} \right\}_{i, j = 1}^{p, q}$.

   b. Expectation: For $(r, s) \in \mathcal{I}_{ij}$, replace $\widehat{\mathbf{X}}_{ij}^{(old)}[r, s]$ with $\widehat{\mathbf{S}}_{ij}[r, s]$ and denote the imputed matrix by $\widehat{\mathbf{X}}_{ij}^{(new)}$.

3. If $\sum_{i, j = 1}^{p, q} \sum_{(r, s) \in \mathcal{I}_{ij}} \left| \widehat{\mathbf{X}}_{ij}^{(old)}[r, s] - \widehat{\mathbf{X}}_{ij}^{(new)}[r, s] \right|^2 < \epsilon$, the algorithm converges. If not, reapply (2) after replacing $\widehat{\mathbf{X}}_{ij}^{(old)}$ by $\widehat{\mathbf{X}}_{ij}^{(new)}$.

To improve computational efficiency, each maximization step uses the $\widehat{\Theta}$ from the previous maximization step as starting values. The imputation scheme can be used to impute entries or entire rows or columns of the constituent data matrices; however, in all cases the missing entities (entries, rows or columns) must be missing at random.

Our algorithm can be considered a regularized EM algorithm, using a model-based motivation for the objective function (4). Because each component is estimated as a product of two matrices, it is naturally translated as a probabilistic matrix factorization. The unpenalized objective with $\lambda_{ij} = 0$ for all i, j maximizes a Gaussian likelihood model if the noise variances are the same across the matrices (i.e., $\sigma^2 = \sigma_{11}^2 = \cdots = \sigma_{pq}^2$). With penalization, the approach is analogous to maximizing the posterior distribution in a Bayesian context with Gaussian prior on the terms of $\mathbf{U}_{ij}^{(\cdot)}$ and $\mathbf{V}_{ij}^{(\cdot)}$. Specifically, if the entries of $\mathbf{U}_{ij}^{(\cdot)}$ and $\mathbf{V}_{ij}^{(\cdot)}$ are independent Gaussian with mean 0 and variance σ^2/λ_{ij}, then minimizing (4) is analogous to finding the posterior mode (Mnih and Salakhutdinov, 2008). This provides a theoretical foundation for the iterative imputation algorithm based on

expectation-maximization (EM) algorithm. It is closely related to softImpute, proposed by Mazumder et al. (2010), a computationally efficient imputation algorithm for a single matrix based on Proposition 1. Prediction intervals for the imputed values may be obtained via a resampling approach, as described in Web Appendix E of O'Connell and Lock (2019).

3. Simulation Studies

3.1 Simulation Setup

In this section, we compare our model to existing factorization methods using simulated data. Because competing approaches apply only to uni-dimensionally linked matrices (vertical or horizontal), we constructed two simulation designs with $p = q = 2$ for proper comparison. Design 1 does not include any row-shared or globally shared structure and may be considered as two separate sets of vertically-linked matrices. Design 2 includes all linked structures that are represented in our model: global, column-shared, row-shared and individual.

For each simulation design, we used existing methods to compare performance. First, we fit 2 separate JIVE models to $\{\mathbf{X}_{11}, \mathbf{X}_{21}\}$ and $\{\mathbf{X}_{12}, \mathbf{X}_{22}\}$, using both (i) true marginal ranks for joint and individual components and (ii) rank selection based on permutation testing, denoted by JIVE(T) and JIVE(P) respectively. The "true marginal rank" in this context means $r(\mathbf{G}_{ij} + \mathbf{C}_{ij})$ for the joint component and $r(\mathbf{R}_{ij} + \mathbf{I}_{ij})$ for the individual components given that j is fixed. We similarly apply the AJIVE and SLIDE methods, where we used the rank of $\mathbf{S}_{ij}$ as the initial rank for AJIVE. We also consider an approach that is analogous to BIDIFAC but reduced to vertical integration only, i.e., with $\widehat{\mathbf{G}}_{ij}$ and $\widehat{\mathbf{R}}_{ij}$ set to $\mathbf{0}_{m_i \times n_j}$, denoted as UNIFAC in this paper. We fit UNIFAC to $\{\mathbf{X}_{11}, \mathbf{X}_{21}\}$ and $\{\mathbf{X}_{12}, \mathbf{X}_{22}\}$ separately. For notational simplicity, we denote the joint and individual components estimated by JIVE(P), JIVE(T), AJIVE, SLIDE and UNIFAC by $\widehat{\mathbf{C}}_{ij}$ and $\hat{\mathbf{I}}_{ij}$. We also applied soft singular value thresholding for each single matrix with the corresponding imputation algorithm, softImpute, with tuning parameters decided as in Section 2.6 $\left(\lambda_{ij} = \hat{\sigma}_{ij} \cdot (\sqrt{m_i} + \sqrt{n_j})\right)$. We denote this approach by SVD(soft). We also consider the performance of the hard-thresholding low-rank approximation of each matrix, denoted by SVD(T), using the true marginal rank for a single matrix ($r(\mathbf{S}_{ij})$). Similar to the above, SVD(soft) and SVD(T) estimate $\mathbf{I}_{ij}$ (or, equivalently, $\mathbf{S}_{ij}$) components only and assume $\widehat{\mathbf{G}}_{ij} = \widehat{\mathbf{R}}_{ij} = \widehat{\mathbf{C}}_{ij} = \mathbf{0}_{m_i \times n_j}$. BIDIFAC, UNIFAC, and SVD(soft) are soft-thresholding methods, while JIVE(T), JIVE(P), and SVD(T) are based on hard-thresholding.

In our simulation studies, the number of rows and columns for each matrix was set to 100: $m_i = n_j = 100$. The rank of the total signal in each matrix, $\mathbf{S}_{ij} = \mathbf{G}_{ij} + \mathbf{R}_{ij} + \mathbf{C}_{ij} + \mathbf{I}_{ij}$, was 10. This total rank was distributed across each of the 4 terms (or the two terms $\mathbf{C}_{ij}$ and $\mathbf{I}_{ij}$ for Design 1) via a multinomial distribution with equal probabilities. For clarity of the simulation studies and to allow for comparison with other methods, we enforced orthogonality among the shared structures, both within each matrix (i.e., $\mathbf{G}_{11}$ and $\mathbf{C}_{11}$ are orthogonal) and across matrices (i.e., $\mathbf{G}_{11}$ and $\mathbf{G}_{12}$ are orthogonal). Note, however, that our model does not enforce orthogonality when estimating parameters.

Each signal matrix, $\mathbf{S}_{ij}$, was generated by applying SVD to a Gaussian random matrix with mean 0 and unit variance, denoted as $\mathbf{Y} = \mathbf{U}_Y \mathbf{D}_Y \mathbf{V}_Y^T$. Then $\mathbf{U}_Y$ and $\mathbf{V}_Y$ were rearranged accordingly to guarantee orthogonality and a 2×2 linked structure. The singular values were randomly permuted within each shared component to allow for heterogeneity in the size of the joint signal across matrices, e.g., $\mathbf{C}_{11}$ and $\mathbf{C}_{21}$ have the same loadings and scores but with different order. Each signal matrix $\mathbf{S}_{ij}$ was standardized to have $\|\mathbf{S}_{ij}\|_F = 1$. Finally, independent Gaussian noise was added to each signal matrix, where the noise variance was decided by the signal-to-noise ratio (SNR) defined by $1/\left(\sigma_{ij} \cdot \sqrt{m_i \cdot n_j}\right)$. We first considered three SNRs in our simulation studies: 0.5, 1, and 2. For the true structure $\mathbf{S}_{ij}$ the expected value of $R_{\mathbf{X}_{ij}}^2(\mathbf{S}_{ij})$ is $\mathrm{SNR}^2/\left(1 + \mathrm{SNR}^2\right)$, when SNRs are the same for all matrices. We also considered a scenario where the SNR is randomly selected separately for the different matrices, uniformly between 0.5 and 2, to accommodate heterogeneous noise variances.

We compared the performance of our method and the competing methods from two perspectives: prediction error and imputation performance. We computed prediction error for each term in the decomposition ($\mathbf{G}$, $\mathbf{R}$, $\mathbf{C}$ or $\mathbf{I}$) as the relative reconstruction error:

$$\mathrm{PredErr}(\widehat{\mathbf{G}}) = \frac{\sum_{i=1, j=1}^{p, q} \left\|\mathbf{G}_{ij} - \widehat{\mathbf{G}}_{ij}\right\|_F^2}{\sum_{i=1, j=1}^{p, q} \left\|\mathbf{G}_{ij}\right\|_F^2}. \tag{9}$$

For a fair comparison between our method and the existing methods for uni-dimensionally shared matrices, we also report $\mathrm{PredErr}(\widehat{\mathbf{G}} + \widehat{\mathbf{C}})$, $\mathrm{PredErr}(\widehat{\mathbf{R}} + \widehat{\mathbf{I}})$. Finally, we also report $\mathrm{PredErr}(\widehat{\mathbf{S}})$ to evaluate the overall signal reconstruction performances.

For imputation, we considered three scenarios where in each matrix (i) 200 randomly selected cells are missing, (ii) 2 columns are missing, and (iii) 2 rows are missing. To foster borrowing information from the shared structures in (ii) and (iii), there was no overlapping row/column missing simultaneously in shared matrices. To reduce computation cost, imputation using JIVE(P) used fixed ranks determined by the complete data, resulting in slightly inflated imputation performances. We evaluated the imputation performance using the scaled reconstruction error for missing cells, defined by

$$\mathrm{ImputeErr} = \frac{\sum_{i=1, j=1}^{p, q} \sum_{(r, s) \in \mathscr{I}_{ij}} \left|\mathbf{S}_{ij}[r, s] - \widehat{\mathbf{S}}_{ij}[r, s]\right|^2}{\sum_{i=1, j=1}^{p, q} \sum_{(r, s) \in \mathscr{I}_{ij}} \left|\mathbf{S}_{ij}[r, s]\right|^2}. \tag{10}$$

3.2   Results

We repeated each simulation 200 times and averaged the performance. The results are shown in Tables 1 and 2. We summarize the results below from a few perspectives.

On performance of BIDIFAC: BIDIFAC was competitive against all compared models in separating signals into multiple shared structures. In Design 1 where $\|\mathbf{G}_{ij}\|_F^2 = \|\mathbf{R}_{ij}\|_F^2 = 0$,

the loss from the model misspecification was negligible when comparing the prediction errors for column-shared matrices (**C**) and the sum of global and column-shared matrices (**G** + **C**), as well as individual matrices (**I**) and the sum of row and individual matrices (**R** + **I**). The negligible effect of misspecification can also be seen when compared to UNIFAC, where the global or row-shared structures are ignored. In Design 2, where BIDIFAC is the only method estimating global (**G**) and row-shared (**R**) components, the prediction errors were comparable to the other components regardless of SNR. Even when error variances differ by each matrix, BIDIFAC was not affected severely.

On soft and hard thresholding: When SNR is low, matrix completion via soft thresholding generally outperformed hard-thesholding approaches (JIVE, SLIDE and SVD(T)). It is easily seen by comparing SVD(soft) to SVD(T) or comparing JIVE(T)/JIVE(P)/SLIDE to UNIFAC. Especially, when SNR= 0.5 in Design 1, JIVE(T) overfitted severely even when true ranks are given. JIVE(P) performed even worse, due to erroneous rank selection. SLIDE suffered less from overfitting, but this was because it selected 0 rank in most simulated data. As SNR increases, SLIDE, JIVE(P) and JIVE(T) outperformed UNIFAC and BIDIFAC in estimating **G**+**C** and **R**+**I** in both designs. This result is intuitive: soft-thresholding prevents over-fitting when the SNR is low, but over-penalizes the estimated signal when SNR is high. We found that AJIVE did not perform well unless SNR is 2, where it provides provides similar results to JIVE with the true ranks, JIVE(T).

On overall signal recovery: In Design 1, JIVE(T) can be considered as the gold standard for overall signal recovery as it reflects the true joint and individual rank structures, which are unknown in practice. Except when SNR is 0.5 (which is explained by the difference between soft and hard thresholding), JIVE(T), JIVE(P) and SVD(soft) performed better than BIDIFAC in recovering the overall signals. The overall signal recovery of BIDIFAC was the same as UNIFAC and better than SVD(soft), revealing that BIDIFAC and UNIFAC obtained additional power from the column-shared matrices and the effect of model misspecification of BIDIFAC is negligible. In Design 2, BIDIFAC performed better than UNIFAC and SVD(soft) as it closely matched the data generating process.

On imputation performance: JIVE(T) was the winner in both designs in imputing missing cells and columns except when SNR is 0.5. However, recall that JIVE(T) uses the true ranks which are unknown in practice. BIDIFAC performed close to or even better than JIVE(P) in both designs, revealing that signal detection does not necessarily guarantee imputation performance as it is also affected by appropriate rank selection and detecting shared structures. In Design 1, it is not surprising that all models suffered when an entire row is missing, as there was no row-shared structure. In Design 2, BIDIFAC was the only method that successfully imputed missing rows.

To summarize, the performance of BIDIFAC was promising in simulation studies even when it was misspecified. It appropriately separated signals into linked structures and did not overfit for low SNR. Among the compared models, BIDIFAC is the only model that well accommodates the cases where a whole data matrix is missing or both a whole column and a whole row is missing. Acknowledging that it is not possible to obtain the true ranks of shared and individual structures, BIDIFAC performed the best across different scenarios.

4. Data Analysis

4.1 TCGA Breast Cancer Data

In this section, we apply our method to breast cancer data from TCGA(Cancer Genome Atlas Network, 2012). We integrate mRNA and miRNA profiles of the tumor samples and NATs, which, to our knowledge, has not been previously investigated in a fully unified framework. The data used here are freely available from TCGA. Specifically, we obtained the level III raw count of mRNA (RNASeq-V2) and miRNA (miRNA-Seq) data using the R package TCGA2STAT (Wan et al., 2015). In our analysis we first removed the tumor data of those samples with matched NAT, so that tumor and NAT data correspond to two independent cohorts. We also filtered mRNAs and miRNAs with more than half zero counts for all individuals. Then we took log(1 + count) and centered each of mRNA and miRNA profiles to the mean of tumors and NATs, so that the mean of each row of each matrix is 0. Lastly, we selected the 500 mRNAs and miRNAs with maximum variability. The resulting data had 500 mRNA and miRNA profiles for 660 tumor samples and 86 NATs.

Breast cancer tumor samples are classified into 5 intrinsic subtypes, based on expression levels of 50 pre-defined genes: Luminal A (LumA), Luminal B (LumB), HER2-enriched (HER2), Basal-enriched (Basal), and Normal-like tumors (Ciriello et al., 2015). In our TCGA data, 419 out of 660 tumor samples had the labeled subtypes.

4.2 Results

We first applied BIDIFAC to the processed data until convergence. We summarize the proportion of explained variance, as well as the estimated ranks of the components, in Table 3. The difference between $R^2_{\mathbf{X}_{ij}}(\mathbf{S}_{ij}) - R^2_{\mathbf{X}_{ij}}(\mathbf{G}_{ij} + \mathbf{R}_{ij} + \mathbf{C}_{ij})$ among tumors and NATs suggests that most of the variability of the miRNA and mRNA profiles were attributed to shared structures. In particular, more than a half of the variability of both the mRNA and miRNA profiles from the NAT is attributed to the row-shared components (Global+Row); this makes sense, as tumor cells are derived from normal tissue and thus we expect many of the same patterns of variability that are present in normal tissue to also be present in tumors. Between 30–45% of the variability is explained by the shared structure between miRNA and mRNA across the four matrices. There is a large difference of sample sizes between tumor and normal tissues, which may have also affected the estimate of proportion of variance explained. Even though the rank of the estimated signals in NAT was close to the rank of the data matrix (86) linear independence among the terms of the decomposition was preserved in our decomposition.

We compared BIDIFAC to existing methods on one-dimensionally linked matrices, including UNIFAC, JIVE(P) and SLIDE. We used two criteria to compare models: subtype classification and survival analysis. Specifically, we hypothesize that accounting for multiple omics profiles and removing the shared variations between tumors and normal tissues would increase the biological interpretation of cancer subtypes and patient's survival. Thus, we focused on column-shared and individual components estimated by BIDIFAC and compared it to other methods using tumor mRNA and miRNA only.

We first compared how the estimated components are well distinguished by breast cancer subtypes. To summarize the subtype distinctions, we used the SWISS (Standardized WithIn Class Sum of Squares) score (Cabanski et al., 2010). Interpreted similar to ANOVA, a lower SWISS score implies more clear subtype distinction. We restricted the attention to the tumor samples with labeled subtypes. As summarized in Table 4, the SWISS score of $\mathbf{C} + \mathbf{I}$ of UNIFAC was superior to JIVE and SLIDE, suggesting that soft-thresholding better uncovered subtype heterogeneity when integrating mRNA and miRNA. Compared to UNIFAC, BIDIFAC obtained a slightly better SWISS score with lower rank of the $\mathbf{C} + \mathbf{I}$. The scatterplot of the principal components from the column-shared structures of BIDIFAC, shown in Figure 2, reveals that the subtypes are well distinguished. The column-shared structure from JIVE(P) had the lowest SWISS score overall, which may be because of its extremely low rank.

Using overall survival data for patients with tumors, we applied the Cox proportional hazards model (PHM) using scores from the estimated components as predictors and used the score test to assess its significance. The results are also shown in Table 4. At $a = 0.05$, all models suggested that scores from $\mathbf{C} + \mathbf{I}$ were associated. Narrowing down the scope to each structural component, we also found that the individual components of UNIFAC and BIDIFAC, even though not indicative of subtype distinction in SWISS scores, could provide additional information on patients' survival. BIDIFAC, accounting for normal tissues, provided reduced rank for individual components compared to UNIFAC.

5 Conclusion

In this paper we propose BIDIFAC, the first unified framework to handle multiple matrices that are shared both vertically and horizontally. In contrast to existing methods on joint matrix factorization, we provide the estimator based on soft-thresholding and nuclear norm penalization, where the complicated problem of selecting tuning parameters is alleviated by using the well-known result from random matrix theory. We conducted extensive simulation studies to show the efficiency and flexibility of BIDIFAC compared to existing methods. We applied our method to TCGA breast cancer data, where mRNA and miRNA profiles are obtained separately from both tumor tissues and normal tissues adjacent to tumors (NATs). From this application we conclude that (i) patterns of variability in normal tissue are largely also present in tumor tissue for both mRNA and miRNA, (ii) patterns that are associated with survival and clinical subtypes across both mRNA and miRNA are largely *not* present in normal tissue, and (iii) patterns that distinguish the clinical subtypes are shared by mRNA and miRNA. Existing vertical integration methods establish (iii), but not (i) or (ii). In addition to the integration of tumor and NAT data, this methodology may be applied to a growing number of applications with bidimensionally linked matrices. A particularly intriguing potential application is the integration of "pan-omics pan-cancer" data; that is, the integration of multi-omic data for samples from multiple types of cancers (as defined by their tissue-of-origin).

We describe several limitations of our method. These limitations primarily relate to use of the Frobenious norm in the objective and applying random matrix theory to select tuning parameters. Importantly, our model assumes normality of each matrix to select appropriate

tuning parameters, which may be violated in practice. For example, application to SNPs data would be limited as each cell of a matrix takes discrete values (e.g., 0,1, 2). Similarly, we do not consider the case where one or more matrices are binary, which is common for biomedical data. Also, our approach is sensitive to outliers, due to the use of the Frobenious norm in the objective function. Even when the Gaussian assumption holds, our model may be extended by adopting variable sparsity or overcoming deficiencies of soft-thresholding, as neither soft-thresholding nor hard-thresholding provides optimal signal recovery (Shabalin and Nobel, 2013). Empirical approaches (e.g., cross-validation) may also be used to select the tuning parameters. Although our convex objective guarantees convergence to a stationary point that is empirically consistent for different starting values, the theoretical identifiability properties of the resulting decomposition deserves further study. Lastly, our approach does not inherently model uncertainty in the underlying structural decomposition and missing value imputations; the result gives the mode of a Bayesian posterior, and extensions to fully Bayesian approaches are worth considering.

Here we have focused on capturing four types of low-rank signals in bidimensional data: global, column-shared, row-shared, and individual. Other joint signals are possible. For example, when $p = 3$, $q = 1$, it is possible that $\mathbf{X}_{11}$ and $\mathbf{X}_{21}$ share a signal that is not present in $\mathbf{X}_{31}$. Similarly, our method would suffer when $\mathbf{X}_{11}$, $\mathbf{X}_{12}$, $\mathbf{X}_{21}$ share a signal that is not present in $\mathbf{X}_{22}$, though this is perhaps less likely for most data applications. It is straightforward to extend our framework to accommodate these or other structures, with appropriate additions to the objective (4).

An interesting extension of our work is the factorization of higher-order arrays (i.e., tensors). For example, integrating multiple *omics* profiles for multiple cohorts that are measured at multiple time points would give additional insights on the linked structure in a longitudinal manner that cross-sectional designs cannot provide.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

REFERENCES

Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, and et al. (2017). Comprehensive Analysis of Normal Adjacent to Tumor Transcriptomes. Nature Communications 8, 1077.

Cabanski CR, Qi Y, Yin X, Bair E, Hayward MC, Fan C, and et al. (2010). Swiss made: Standardized within Class Sum of Squares to Evaluate Methodologies and Dataset Elements. PloS One 5, e9905.

Cancer Genome Atlas Network. (2012). Comprehensive Molecular Portraits of Human Breast Tumours. Nature 490, 61. [PubMed: 23000897]

Ciriello G, Gatza ML, Beck AH, Wilkerson MD, Rhie SK, Pastore A, and et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell 163, 506–519. [PubMed: 26451490]

Feng Q, Jiang M, Hannig J, and Marron JS (2018). Angle-based Joint and Individual Variation Explained. Journal of Multivariate Analysis 166, 241–265.

Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, and et al. (2015). A Gene-based Association Method for Mapping Traits Using Reference Transcriptome Data. Nature Genetics 47, 1091. [PubMed: 26258848]

Gavish M, and Donoho DL (2017). Optimal Shrinkage of Singular Values. IEEE Transactions on Information Theory 63, 2137–2152.

Gaynanova I, and Li G (2019). Structural Learning and Integrative Decomposition of Multi-View Data. Biometrics In press.

Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, and et al. (2016). Integrative Approaches for Large-scale Transcriptome-wide Association Studies. Nature Genetics 48, 245. [PubMed: 26854917]

Hawkins RD, Hon GC, and Ren B (2010). Next-generation Genomics: an Integrative Approach. Nature Reviews Genetics 11, 476.

Hellton KH, and Thoresen M (2016). Integrative Clustering of High-dimensional Data with Joint and Individual Clusters. Biostatistics 17, 537–548. [PubMed: 26917056]

Huang X, Stern DF, and Zhao H (2016). Transcriptional Profiles from Paired Normal Samples Offer Complementary Information on Cancer Patient Survival - Evidence from TCGA Pan-cancer Data. Scientific Reports 6, 20567. [PubMed: 26837275]

Kaplan A, and Lock EF (2017). Prediction with Dimension Reduction of Multiple Molecular Data Sources for Patient Survival. Cancer Informatics 16, 1–11. [PubMed: 28096648]

Kumar RD, Searleman AC, Swamidass SJ, Griffith OL, and Bose R (2015). Statistically Identifying Tumor Suppressors and Oncogenes from Pan-cancer Genome-sequencing Data. Bioinformatics 31, 3561–3568. [PubMed: 26209800]

Kurucz M, Benczr AA, and Csalogny K (2007). Methods for Large Scale SVD with Missing Values. In Proceedings of KDD Cup and Workshop 12, 3138.

Li G, and Jung S (2017). Incorporating Covariates into Integrated Factor Analysis of Multi-view Data. Biometrics 73, 1433–1442. [PubMed: 28407218]

Li G, and Gaynanova I (2018). A General Framework for Association Analysis of Heterogeneous Data. The Annals of Applied Statistics 12, 1700–1726.

Liu B, Shen X, and Pan W (2016). Integrative and Regularized Principal Component Analysis of Multiple Sources of Data. Statistics in Medicine 35, 2235–2250. [PubMed: 26756854]

Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, and et al. (2018). An Integrated TCGA Pan-cancer Clinical Data Resource to Drive High-quality Survival Outcome Analytics. Cell 173, 400–416. [PubMed: 29625055]

Lock EF, Hoadley KA, Marron JS, and Nobel AB (2013). Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types. The Annals of Applied Statistics 7, 523. [PubMed: 23745156]

Mazumder R, Hastie T, and Tibshirani R (2010). Spectral Regularization Algorithms for Learning Large Incomplete Matrices. Journal of Machine Learning Research 11, 2287–2322. [PubMed: 21552465]

Mnih A, and Salakhutdinov RR (2008). Probabilistic Matrix Factorization. In Advances in Neural Information Processing Systems 1257–1264.

O'Connell MJ, and Lock EF (2016). R.JIVE for Exploration of Multi-source Molecular Data. Bioinformatics 32, 2877–2879. [PubMed: 27273669]

O'Connell MJ, and Lock EF (2019). Linked Matrix Factorization. Biometrics In Press.

Owen AB, and Perry PO (2009). Bi-cross-validation of the SVD and the Nonnegative Matrix Factorization. The Annals of Applied Statistics 3, 564–594.

Rudelson M, and Vershynin R (2010). Non-asymptotic Theory of Random Matrices: Extreme Singular Values. In Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. IIIV: Invited Lectures (pp. 1576–1602),

Shabalin AA, and Nobel AB (2013). Reconstruction of a Low-rank Matrix in the Presence of Gaussian Noise. Journal of Multivariate Analysis 118, 67–76.

Shen R, Olshen AB, and Ladanyi M (2009). Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis. Bioinformatics 25, 2906–2912. [PubMed: 19759197]

Tseng GC, Ghosh D, and Feingold E (2012). Comprehensive Literature Review and Statistical Considerations for Microarray Meta-analysis. Nucleic Acids Research 40, 3785–3799. [PubMed: 22262733]

Wan Y, Allen GI, Anderson ML, and Liu Z (2015). TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R. R package version 1.2. https://CRAN.R-project.org/package=TCGA2STAT

Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, and Do KA (2012). iBAG: Integrative Bayesian Analysis of High-dimensional Multiplatform Genomics Data. Bioinformatics 29, 149–159. [PubMed: 23142963]

Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, and et al. (2013). Variation and Genetic Control of Protein Abundance in Humans. Nature 499, 79. [PubMed: 23676674]

Zhou G, Cichocki A, Zhang Y, and Mandic DP (2016). Group Component Analysis for Multiblock Data: Common and Individual Feature Extraction. IEEE Transactions on Neural Networks and Learning Systems 27, 2426–2439. [PubMed: 26529787]

Zhu H, Li G, and Lock EF (2018). Generalized Integrative Principal Component Analysis for Multi-type Data with Block-wise Missing Structure. Biostatistics In Press.

Figure 1.
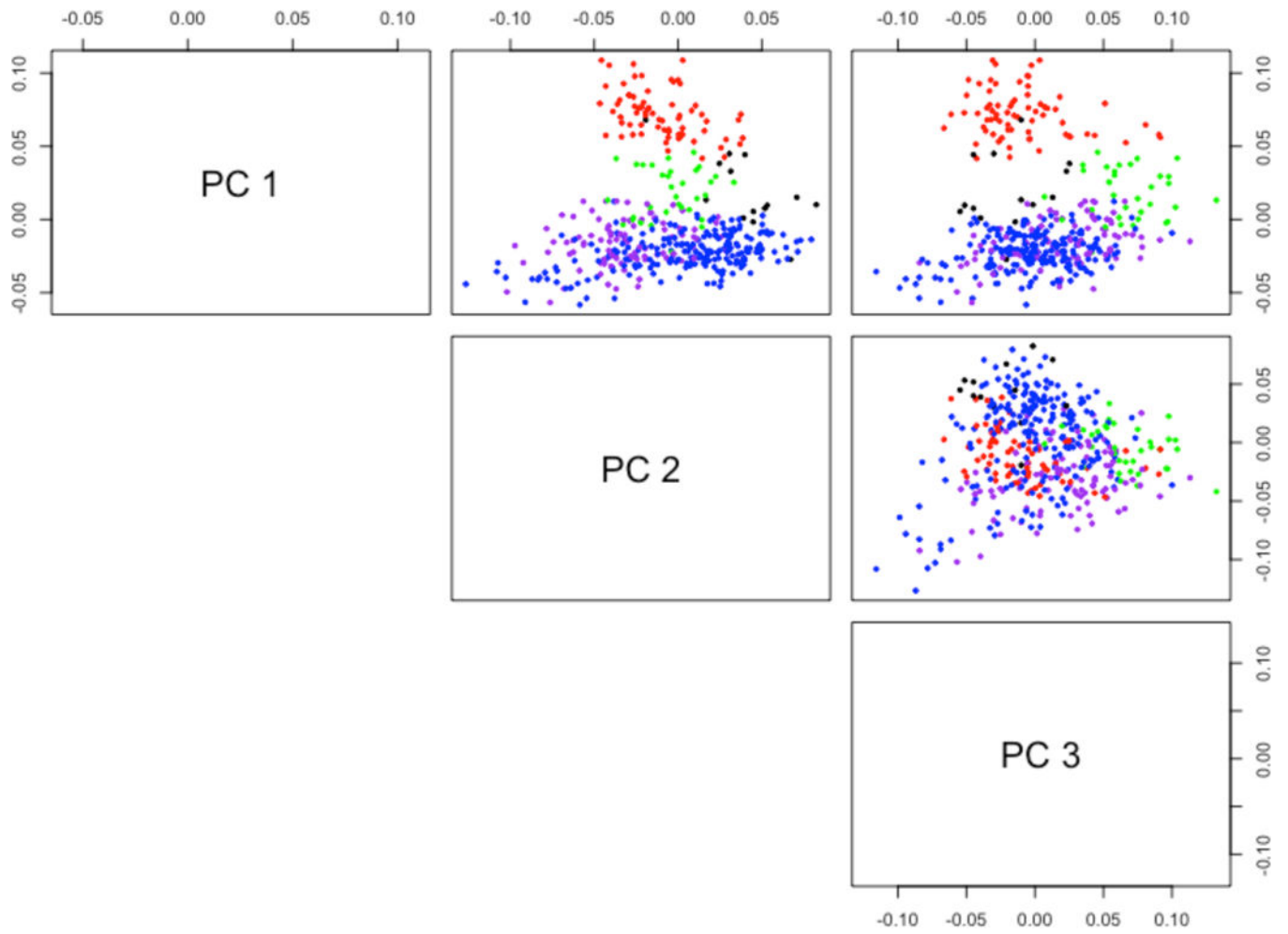Overview of the proposed method, using the 2×2 bidimensionally linked structure.

Figure 2.
The top three principal components of the column-shared structure from BIDIFAC, colored by subtype (red: basal, green: HER2, blue: luminal A, purple: luminal B, black: normal).

Table 1

Summary of the simulation studies for Design 1, where prediction and imputation errors are computed using Equations (9) and (10) respectively.

		Prediction							Imputation		
SNR	Model	*G*	*R*	*C*	*I*	*G* + *C*	*R* + *I*	*S*	Cell	Column	Row
0.5	BIDIFAC	—	—	0.81	0.87	0.80	0.87	0.81	0.86	0.95	1.00
	UNIFAC	—	—	0.80	0.87	0.80	0.87	0.81	0.86	0.94	1.00
	JIVE(T)	—	—	1.01	1.40	1.01	1.40	1.05	1.49	1.13	1.02
	JIVE(P)	—	—	1.09	2.67	1.09	2.67	1.33	2.05	1.13	1.02
	AJIVE(T)	—	—	4.58	1.00	4.58	1.00	1.48	—	—	—
	SLIDE	—	—	0.97	1.02	0.97	1.02	0.84	—	—	—
	SVD(T)	—	—	1.00	3.61	1.00	3.61	1.20	1.87	1.02	1.02
	SVD(soft)	—	—	1.00	0.89	1.00	0.89	0.85	0.90	1.00	1.00
1	BIDIFAC	—	—	0.35	0.41	0.35	0.41	0.36	0.43	0.80	1.00
	UNIFAC	—	—	0.35	0.41	0.35	0.41	0.36	0.43	0.79	1.00
	JIVE(T)	—	—	0.18	0.23	0.18	0.23	0.19	0.25	0.63	1.01
	JIVE(P)	—	—	0.47	0.54	0.47	0.54	0.23	0.40	0.78	1.01
	AJIVE(T)	—	—	1.98	1.00	1.98	1.00	0.30	—	—	—
	SLIDE	—	—	0.19	0.22	0.19	0.22	0.19	—	—	—
	SVD(T)	—	—	1.00	1.73	1.00	1.73	0.22	0.30	1.01	1.01
	SVD(soft)	—	—	1.00	0.66	1.00	0.66	0.40	0.49	1.00	1.00
2	BIDIFAC	—	—	0.11	0.13	0.11	0.13	0.11	0.16	0.65	1.00
	UNIFAC	—	—	0.11	0.13	0.11	0.13	0.11	0.16	0.65	1.00
	JIVE(T)	—	—	0.04	0.05	0.04	0.05	0.04	0.05	0.54	1.01
	JIVE(P)	—	—	0.35	0.32	0.35	0.32	0.05	0.14	0.81	1.01
	AJIVE(T)	—	—	0.07	0.08	0.07	0.08	0.04	—	—	—
	SLIDE	—	—	0.05	0.06	0.05	0.06	0.06	—	—	—
	SVD(T)	—	—	1.00	1.37	1.00	1.37	0.05	0.06	1.01	1.01
	SVD(soft)	—	—	1.00	0.74	1.00	0.74	0.13	0.18	1.00	1.00
Mixed	BIDIFAC	—	—	0.37	0.41	0.37	0.41	0.31	0.38	0.80	1.00
	UNIFAC	—	—	0.37	0.41	0.37	0.41	0.31	0.36	0.77	1.00
	JIVE(T)	—	—	0.20	0.27	0.20	0.27	0.22	0.29	0.64	1.01
	JIVE(P)	—	—	0.55	0.73	0.55	0.73	0.29	0.47	0.83	1.01
	AJIVE(T)	—	—	2.28	0.92	2.28	0.92	0.64	—	—	—
	SLIDE	—	—	0.51	0.53	0.51	0.53	0.38	—	—	—
	SVD(T)	—	—	1.00	1.78	1.00	1.78	0.74	0.35	1.01	1.01
	SVD(soft)	—	—	1.00	0.70	1.00	0.70	0.59	0.43	1.00	1.00

Table 2

Summary of the simulation studies for Design 2.

SNR	Model	Prediction							Imputation		
		G	R	C	I	G + C	R + I	S	Cell	Column	Row
0.5	BIDIFAC	0.67	0.60	0.81	0.89	0.73	0.83	0.76	0.81	0.91	0.91
	UNIFAC	1.00	1.00	0.85	0.91	0.80	0.87	0.82	0.86	0.95	1.00
	JIVE(T)	1.00	1.00	3.05	3.86	1.00	1.38	1.04	1.47	1.14	1.02
	JIVE(P)	1.00	1.00	2.60	6.58	1.12	2.51	1.36	2.13	1.17	1.02
	AJIVE(T)	1.00	1.00	10.74	1.00	4.05	1.00	1.47	—	—	—
	SLIDE	1.00	1.00	1.00	1.27	0.96	0.97	0.85	—	—	—
	SVD(T)	1.00	1.00	1.00	8.51	1.00	3.32	1.19	1.86	1.02	1.02
	SVD(soft)	1.00	1.00	1.00	0.95	1.00	0.89	0.86	0.90	1.00	1.00
1	BIDIFAC	0.26	0.26	0.34	0.42	0.30	0.37	0.32	0.39	0.77	0.77
	UNIFAC	1.00	1.00	0.65	0.68	0.34	0.41	0.36	0.44	0.79	1.00
	JIVE(T)	1.00	1.00	1.62	1.76	0.18	0.23	0.19	0.25	0.63	1.01
	JIVE(P)	1.00	1.00	1.72	2.18	0.45	0.48	0.23	0.40	0.77	1.01
	AJIVE(T)	1.00	1.00	5.77	1.00	1.76	1.00	0.30	—	—	—
	SLIDE	1.00	1.00	0.99	0.93	0.19	0.22	0.19	—	—	—
	SVD(T)	1.00	1.00	1.00	4.82	1.00	1.56	0.22	0.29	1.01	1.01
	SVD(soft)	1.00	1.00	1.00	1.19	1.00	0.63	0.40	0.49	1.00	1.00
2	BIDIFAC	0.09	0.09	0.11	0.14	0.10	0.12	0.10	0.14	0.63	0.64
	UNIFAC	1.00	1.00	0.76	0.76	0.11	0.13	0.11	0.16	0.65	1.00
	JIVE(T)	1.00	1.00	1.33	1.39	0.04	0.05	0.04	0.06	0.54	1.01
	JIVE(P)	1.00	1.00	1.42	1.70	0.29	0.25	0.05	0.14	0.78	1.01
	AJIVE(T)	1.00	1.00	1.39	1.57	0.05	0.06	0.04	—	—	—
	SLIDE	1.00	1.00	1.02	1.19	0.05	0.06	0.06	—	—	—
	SVD(T)	1.00	1.00	1.00	4.07	1.00	1.22	0.05	0.06	1.01	1.01
	SVD(soft)	1.00	1.00	1.00	2.00	1.00	0.67	0.13	0.18	1.00	1.00
Mixed	BIDIFAC	0.30	0.28	0.37	0.48	0.31	0.35	0.27	0.33	0.77	0.78
	UNIFAC	1.00	1.00	0.69	0.79	0.36	0.40	0.31	0.37	0.77	1.00
	JIVE(T)	1.00	1.00	1.65	1.83	0.20	0.27	0.22	0.29	0.64	1.01
	JIVE(P)	1.00	1.00	1.85	2.53	0.53	0.65	0.30	0.46	0.87	1.01
	AJIVE(T)	1.00	1.00	6.35	1.11	2.07	0.90	0.60	—	—	—
	SLIDE	1.00	1.00	1.01	1.37	0.49	0.49	0.34	—	—	—
	SVD(T)	1.00	1.00	1.00	4.94	1.00	1.61	0.25	0.36	1.01	1.01
	SVD(soft)	1.00	1.00	1.00	1.39	1.00	0.66	0.36	0.43	1.00	1.00

Table 3

Proportion of variance explained (Equation (7)) for each component, estimated by BIDIFAC using the breast cancer data. The parentheses denote the rank of the estimated components.

	Global	Global+Row	Global+Col	Global+Row+Col	Signal
Tumor mRNA	0.14 (34)	0.32 (68)	0.45 (93)	0.58 (127)	0.67 (173)
NAT mRNA	0.23 (34)	0.50 (68)	0.44 (4l)	0.66 (75)	0.78 (83)
Tumor miRNA	0.09 (34)	0.46 (67)	0.30 (93)	0.63 (126)	0.76 (175)
NAT miRNA	0.13 (34)	0.66 (67)	0.24 (4l)	0.75 (74)	0.76 (79)

Table 4

Summary of the SWISS scores and p-values of the score test of the estimated components from the Cox proportional hazards model, including tumors only. 'Rank' refers to the rank of the estimated components.

Model	Components	Rank	SWISS	*p*-value
BIDIFAC	Signal	173	0.54	0.002
	Global	34	0.69	0.046
	Row	34	0.75	0.085
	Col+Indiv	105	0.52	0.003
	Col	59	0.48	0.029
	Indiv	46	0.79	0.003
UNIFAC	Col+Indiv	137	0.54	0.001
	Col	75	0.49	0.012
	Indiv	62	0.73	0.001
JIVE(P)	Col+Indiv	35	0.64	0.020
	Col	3	0.31	0.007
	Indiv	32	0.90	0.114
SLIDE	Col+Indiv	58	0.67	0.002
	Col	16	0.52	0.007
	Indiv	42	0.96	0.061