BIDIMENSIONAL LINKED MATRIX FACTORIZATION FOR PAN-OMICS PAN-CANCER ANALYSIS

Eric F. Lock[1,a], **Jun Young Park**[2,b], **Katherine A. Hoadley**[3,c]

[1]Division of Biostatistics, School of Public Health, University of Minnesota

[2]Department of Statistical Sciences, Faculty of Arts & Science, University of Toronto

[3]Department of Genetics, Computational Medicine Program, University of North Carolina

Abstract

Several modern applications require the integration of multiple large data matrices that have shared rows and/or columns. For example, cancer studies that integrate multiple omics platforms across multiple types of cancer, *pan-omics pan-cancer analysis*, have extended our knowledge of molecular heterogeneity beyond what was observed in single tumor and single platform studies. However, these studies have been limited by available statistical methodology. We propose a flexible approach to the simultaneous factorization and decomposition of variation across such *bidimensionally linked* matrices, BIDIFAC+. BIDIFAC+ decomposes variation into a series of low-rank components that may be shared across any number of row sets (e.g., omics platforms) or column sets (e.g., cancer types). This builds on a growing literature for the factorization and decomposition of linked matrices which has primarily focused on multiple matrices that are linked in one dimension (rows or columns) only. Our objective function extends nuclear norm penalization, is motivated by random matrix theory, gives a unique decomposition under relatively mild conditions, and can be shown to give the mode of a Bayesian posterior distribution. We apply BIDIFAC+ to pan-omics pan-cancer data from TCGA, identifying shared and specific modes of variability across *four* different omics platforms and 29 different cancer types.

Keywords

Cancer genomics; data integration; low-rank matrix factorization; missing data imputation; nuclear norm penalization

1. Introduction.

Data collection and curation for the Cancer Genome Atlas (TCGA) program, completed in 2018, provide a unique and valuable public resource for comprehensive studies of

[a] elock@umn.edu . [b] junjy.park@utoronto.ca . [c] hoadley@med.unc.edu .

molecular profiles across several types of cancer (Hutter and Zenklusen (2018)). The database includes information from several molecular platforms for over 10,000 tumor samples from individuals representing 33 types of cancer. The molecular platforms capture signal at different omics levels (e.g., the genome, epigenome, transcriptome, and proteome) which are biologically related and can each influence the behavior of the tumor. Thus, when studying molecular signals in cancer, it is often necessary to consider data from multiple omics sources at once. This and other applications have motivated a very active research area in statistical methods for multiomics integration.

A common task in multiomics applications is to jointly characterize the molecular heterogeneity of the samples. Several multiomics methods have been developed for this purpose which can be broadly categorized by: (1) clustering methods that identify molecularly distinct subtypes of the samples (Huo and Tseng (2017), Lock and Dunson (2013), Gabasova, Reid and Wernisch (2017)), (2) factorization methods that identify continuous lower-dimensional patterns of molecular variability (Lock et al. (2013), Argelaguet et al. (2018), Gaynanova and Li (2019)), or methods that combine aspects of (1) and (2) (Shen, Wang and Mo (2013), Mo et al. (2018), Hellton and Thoresen (2016)). These extend classical approaches, such as: (1) k-means clustering and (2) principal components analysis, to the multiomics context, allowing the exploration of heterogeneity that is shared across the different omics sources while accounting for their differences.

Several multiomics analyses have been performed on the TCGA data, including flagship publications for each type of cancer (e.g., see Network et al. (2012, 2014), Verhaak et al. (2010)). These have revealed striking molecular heterogeneity within each classical type of cancer which is often clinically relevant. However, restricting an analysis to a particular type of cancer sacrifices power to detect important genomic changes that are present across more than one cancer type. In 2013, TCGA began the Pan-Cancer Analysis Project, motivated by the observation that "cancers of disparate organs reveal many shared features, and, conversely, cancers from the same organ are often quite distinct" (Weinstein et al. (2013)). Subsequently, several pan-cancer studies have identified important shared molecular alterations for somatic mutations (Kandoth et al. (2013)), copy number (Zack et al. (2013)), mRNA (Hoadley et al. (2014)), and protein abundance (Akbani et al. (2014)). However, a multiomics analysis found that pan-cancer molecular heterogeneity is largely dominated by cell-of-origin and other factors that define the classical cancer types (Hoadley et al. (2018)).

In this study we do not focus on baseline molecular differences between the cancer types. Rather, we focus on whether patterns of variability within each cancer type are shared across cancer types, that is, whether multiomic molecular profiles that drive heterogeneity in one type of cancer also drive heterogeneity in other cancers. Systematic investigations of heterogeneity in a pan-omics and pan-cancer context are presently limited by a lack of principled and computationally feasible statistical approaches for the comprehensive analysis of such data. In particular, the data take the form of *bidimensionally linked matrices*, that is, multiple large matrices that may share row sets (here, defined by the omics platforms) or column sets (here, defined by the cancer types); this is illustrated in Figure 1, and the formal framework is described in Section 2. Such bidimensional integration problems are increasingly encountered in practice, particularly for biomedical applications

that involve multiple omics platforms and multiple sample cohorts that may correspond to different studies, demographic strata, species, diseases, or disease subtypes.

In this article we propose a flexible approach to the simultaneous factorization and decomposition of variation across bidimensionally linked matrices, BIDIFAC+. Our approach builds on a growing literature for the factorization and decomposition of linked matrices, which we review in Section 3. However, previous methods have focused on multiple matrices that are linked in just one dimension (rows or columns) or assume that shared signals must be present across all row sets or column sets. This is limiting for pan-cancer analysis and other applications, where we expect patterns of variation that are shared across some, but not necessarily all, cancer types and omics platforms. With this motivation our proposed approach decomposes variation into a series of low-rank components that may be shared across any number of row sets (e.g., omics platforms) or column sets (e.g., cancer types). We develop a new approach to model selection and new estimating algorithms to accommodate this more flexible decomposition. We establish theoretical results, most notably, concerning the uniqueness of the decomposition without orthogonality constraints which are entirely new for linked matrix decompositions. Moreover, we show how BIDIFAC+ can improve the estimation of underlying structures over existing methods, even in the more familiar context for which matrices are linked in just one dimension.

2. Formal framework and notation.

Here, we introduce our framework and notation for pan-omics pan-cancer data. Let $\mathbf{X}_{ij} : M_i \times N_j$ denote the data matrix for omics data source i and sample set (i.e., cancer type) j for $j = 1, \ldots, J$ and $i = 1, \ldots, I$. Columns of $\mathbf{X}_{ij}$ represent samples, and rows represent variables (e.g., genes, miRNAs, proteins). The sample sets of size $\mathbf{N} = [N_1, \ldots, N_J]$ are consistent across each omics source, and the features measured for each omics source $\mathbf{M} = [M_1, \ldots, M_I]$ are consistent across sample sets. As illustrated in Figure 1, the collection of available data can be represented as a single data matrix $\mathbf{X}.. : M \times N$, where $M = M_1 + \ldots + M_I$ and $N = N_1 + \ldots + N_J$ by horizontally and vertically concatenating its constituent blocks,

$$\mathbf{X}.. = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} & \cdots & \mathbf{X}_{1J} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{X}_{I1} & \mathbf{X}_{I2} & \cdots & \mathbf{X}_{IJ} \end{bmatrix} \text{ where } \mathbf{X}_{ij} \text{ are } M_i \times N_j. \tag{1}$$

Similarly, $\mathbf{X}_{i.}$ defines the concatenation of omics source i across cancer types, and $\mathbf{X}_{.j}$ defines the concatenation of cancer type j across omics sources,

$$\mathbf{X}_{i.} = [\mathbf{X}_{i1} \cdots \mathbf{X}_{iJ}], \quad \mathbf{X}_{.j} = \left[\mathbf{X}'_{1j} \cdots \mathbf{X}'_{Ij} \right]'.$$

The notation $\mathbf{X}_{ij}[\cdot, n]$ defines the nth column of matrix ij, $\mathbf{X}_{ij}[m, \cdot]$ defines the mth row, and $\mathbf{X}_{ij}[m, n]$ defines the entry in row m and column n. In our context, the entries are all quantitative, continuous measurements; missing data are addressed in Section 9.

We will investigate shared or unique patterns of systematic variability (i.e., heterogeneity) among the constituent data blocks. We are not interested in baseline differences between the different omics platforms or sample sets, and after routine preprocessing the data will be centered so that the mean of the entries within each data block, $\mathbf{X}_{ij}$, is 0. Moreover, to resolve the disparate scale of the data blocks, each block will be scaled to have comparable variability, as described in Section 6.1.

In what follows, $\|\mathbf{A}\|_F$ denotes the Frobenius norm for any given matrix so that $\|\mathbf{A}\|_F^2$ is the sum of squared entries in $\mathbf{A}$. The operator $\|\mathbf{A}\|_*$ denotes the nuclear norm of $\mathbf{A}$ which is given by the sum of the singular values in $\mathbf{A}$; that is, if $\mathbf{A} : M \times N$ has ordered singular values $\mathbf{D}[1, 1], \mathbf{D}[2, 2], \dots$, then $\|\mathbf{A}\|_* = \sum_{r=1}^{\min(M, N)} \mathbf{D}[r, r]$.

3.   Existing integrative factorization methods.

There is now an extensive literature on the integrative factorization and decomposition of multiple linked datasets that share a common dimension. Much of this methodology is motivated by multiomics integration, that is, *vertical integration* of multiple matrices $\{\mathbf{X}_{11}, \mathbf{X}_{21}, \dots, \mathbf{X}_{I1}\}$ with shared columns in the setting of Section 2. For example, the Joint and Individual Variation Explained (JIVE) method (Lock et al. (2013), O'Connell and Lock (2016)) decomposes variation into *joint* components that are shared among multiple omics platforms and *individual* components that only explain substantial variability in one platform. This distinction not only simplifies interpretation but also improves accuracy in recovering underlying signals. Accuracy improves because structured individual variation can interfere with finding important joint signal, just as joint structure can obscure important signal that is individual to a data source. The factorized JIVE decomposition is

$$\mathbf{X}_{i1} = \mathbf{U}_i \mathbf{V}^T + \mathbf{U}_i^* \mathbf{V}_i^T + \mathbf{E}_i \quad \text{for } i = 1, \dots, I. \tag{2}$$

Joint structure is represented by the common score matrix $\mathbf{V} : N_1 \times R$ which summarize patterns in the samples that explain variability across multiple omics platforms. The loading matrices $\mathbf{U}_i : M_i \times R$ indicate how these joint scores are expressed in the rows (variables) of platform i. The score matrices $\mathbf{V}_i : N_1 \times R_i$ summarize sample patterns specific to platform i with loadings $\mathbf{U}_i^*$. Model (2) can be equivalently represented as a sum of low-rank matrices

$$\mathbf{X}_{\cdot 1} = \mathbf{S}_{\cdot}^{(0)} + \sum_{i=1}^{I} \mathbf{S}_{\cdot}^{(i)} + \mathbf{E}_{\cdot}, \tag{3}$$

where $\mathbf{S}_{\cdot}^{(0)} = \mathbf{U}.\mathbf{V}^T$ is of rank R and $\mathbf{S}_{\cdot}^{(i)} = [\mathbf{S}_1^{(i)'} \cdots \mathbf{S}_I^{(i)'}]'$ is the matrix of rank R_i, given by the individual structure for platform i and zeros elsewhere,

$$\mathbf{S}_{i'}^{(i)} = \begin{cases} \mathbf{0}_{M_{i'} \times N} & \text{if } i' \neq i, \\ \mathbf{U}_i^* \mathbf{V}_i^T & \text{if } i' = i. \end{cases}$$

Several other methods result in a factorized decomposition similar to that in (2) and (3), including approaches that allow for different distributional assumptions on the constituent matrices (Li and Gaynanova (2018), Zhu, Li and Lock (2020)), nonnegative factorization (Yang and Michailidis (2016)), and the incorporation of covariates (Li and Jung (2017)). The Structural Learning and Integrative Decomposition (SLIDE) method (Gaynanova and Li (2019)) allows for a more flexible decomposition in which some components are only partially shared across a subset of the constituent data matrices. SLIDE extends model (3) to the more general decomposition

$$\mathbf{X}_{\boldsymbol{\cdot}1} = \sum_{k=1}^{K} \mathbf{S}_{\boldsymbol{\cdot}}^{(k)} + \mathbf{E}_{\boldsymbol{\cdot}}, \tag{4}$$

where $\mathbf{S}_{\boldsymbol{\cdot}}^{(k)} = [\mathbf{S}_1^{(k)'} \cdots \mathbf{S}_I^{(k)'}]'$ is a low-rank matrix with nonzero values for some subset of the sources that is identified by a binary matrix $\mathbf{R} : I \times K$ and

$$\mathbf{S}_i^{(i)} = \begin{cases} \mathbf{0}_{M_i \times N} & \text{if } \mathbf{R}[i,k] = 0, \\ \mathbf{U}_i^{(k)}\mathbf{V}^{(k)T} & \text{if } \mathbf{R}[i,k] = 1. \end{cases}$$

Here, $\mathbf{V}^{(k)}$ gives scores that explain variability for only those patterns for the omics sources identified by $\mathbf{R}[\boldsymbol{\cdot}, k]$.

The BIDIFAC approach (Park and Lock (2020)) is designed for the decomposition of bidimensionally linked matrices as in (1). Its low-rank factorization can be viewed as an extension of that for JIVE, decomposing variation into structure that is shared globally (G), across rows (Row), across columns (Col), or individual to the constituent matrices (Ind). Following (3) for JIVE and (4) for SLIDE, its full decomposition can be expressed as

$$\mathbf{X}_{\boldsymbol{\cdot\cdot}} = \mathbf{S}_{\boldsymbol{\cdot\cdot}}^{(G)} + \sum_{i=1}^{I} \mathbf{S}_{\boldsymbol{\cdot\cdot}}^{(i,\,\text{Row})} + \sum_{j=1}^{J} \mathbf{S}_{\boldsymbol{\cdot\cdot}}^{(j,\,\text{Col})} + \sum_{i=1}^{I}\sum_{j=1}^{J} \mathbf{S}_{\boldsymbol{\cdot\cdot}}^{(i,\,j,\,\text{Ind})} + \mathbf{E}_{\boldsymbol{\cdot\cdot}}, \tag{5}$$

where $\mathbf{S}_{ij}^{(G)} = \mathbf{U}_i^{(G)}\mathbf{V}_j^{(G)T}$,

$$\mathbf{S}_{i'j'}^{(i,\,\text{Row})} = \begin{cases} \mathbf{0}_{M_i \times N_j} & \text{if } i' \neq i, \\ \mathbf{U}_i^{(i,\,\text{Row})}\mathbf{V}_j^{(i,\,\text{Row})T} & \text{if } i' = i, \end{cases} \qquad \mathbf{S}_{i'j'}^{(j,\,\text{Col})} = \begin{cases} \mathbf{0}_{M_i \times N_j} & \text{if } j' \neq j, \\ \mathbf{U}_i^{(j,\,\text{Col})}\mathbf{V}_j^{(j,\,\text{Col})T} & \text{if } j' = j, \end{cases}$$

and

$$\mathbf{S}_{i'j'}^{(i,\,j\,\text{Ind})} = \begin{cases} \mathbf{0}_{M_i \times N_j} & \text{if } i' \neq i \text{ or } j' \neq j, \\ \mathbf{U}_i^{(i,\,j,\,\text{Ind})}\mathbf{V}_j^{(i,\,j,\,\text{Ind})T} & \text{if } i' = i \text{ and } j' = j. \end{cases}$$

4. Proposed model.

We consider a flexible factorization of bidimensionally linked data that combines aspects of the BIDIFAC and SLIDE models. Our full decomposition can be expressed as

$$\mathbf{X}.. = \sum_{k=1}^{K} \mathbf{S}_{..}^{(k)} + \mathbf{E}...,\qquad(6)$$

where

$$\mathbf{S}_{..}^{(k)} = \begin{bmatrix} \mathbf{S}_{11}^{(k)} & \mathbf{S}_{12}^{(k)} & \cdots & \mathbf{S}_{1J}^{(k)} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{S}_{I1}^{(k)} & \mathbf{S}_{I2}^{(k)} & \cdots & \mathbf{S}_{IJ}^{(k)} \end{bmatrix}$$

and the presence of each $\mathbf{S}_{ij}^{(k)}$ is determined by a binary matrix of row indicators $\mathbf{R} : I \times K$ and column indicators $\mathbf{C} : J \times K$,

$$\mathbf{S}_{ij}^{(k)} = \begin{cases} \mathbf{0}_{M_i \times N_j} & \text{if } \mathbf{R}[i, k] = 0 \text{ or } \mathbf{C}[j, k] = 0, \\ \mathbf{U}_i^{(k)} \mathbf{V}_j^{(k)T} & \text{if } \mathbf{R}[i, k] = 1 \text{ and } \mathbf{C}[j, k] = 1 . \end{cases}$$

Each $\mathbf{S}_{..}^{(k)}$ gives a low-rank *module* that explains variability within the omics platforms identified by $\mathbf{R}[\cdot, k]$ and the cancer types identified by $\mathbf{C}[\cdot, k]$. By requiring $\mathbf{R}[i, k] = 1$ *and* $\mathbf{C}[j, k] = 1$, the module is nonzero on a contiguous submatrix. There are, in total, $(2^I - 1)(2^J - 1)$ such submatrices, so by default we set $K = (2^I - 1)(2^J - 1)$ and let $\mathbf{R}$ and $\mathbf{C}$ enumerate all possible modules (see Appendix A (Lock, Park and Hoadley (2022))). The SLIDE decomposition (4) is a special case when $J = 1$ or $I = 1$ (i.e., unidimensional integration); the BIDIFAC model (5) is a special case where each column of $\mathbf{R}$ and $\mathbf{C}$ contains either entirely 1's (i.e., all rows or columns included) or just one 1 (i.e., just one row set or column set included). In practice, if the row and column set for a structural module is not included, it may be subsumed into a larger module or broken into separate smaller modules. The matrix $\mathbf{E}..$ is an error matrix whose entries are assumed to be sub-Gaussian with mean 0 and variance 1 after scaling (see Section 6.1).

Let the rank of each module be rank $(\mathbf{S}_{..}^{(k)}) = R_k$ so that the dimensions of the nonzero components of the factorization are $\mathbf{U}_i^{(k)} : M_i \times R_k$ and $\mathbf{V}_j^{(k)} : N_j \times R_k$. The rth component of the kth module gives a (potentially multiomic) molecular profile $\{\mathbf{U}_i^{(k)}[\cdot, r] : \mathbf{R}[i, k] = 1\}$ that explains variability within those cancer types, defined by $\mathbf{C}[\cdot, k]$ with corresponding sample scores $\{\mathbf{V}_j^{(k)}[r, \cdot] : \mathbf{C}[j, k] = 1\}$.

5.   Objective function.

To estimate model (6), we minimize a least squares criterion with a structured nuclear norm penalty,

$$\underset{\{\mathbf{S}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}}{\operatorname{argmin}} \frac{1}{2}\left\|\mathbf{X}_{\cdot\cdot} - \sum_{k=1}^{K}\mathbf{S}_{\cdot\cdot}^{(k)}\right\|_F^2 + \sum_{k=1}^{K}\lambda_k\|\mathbf{S}_{\cdot\cdot}^{(k)}\|_* \tag{7}$$

subject to $\mathbf{S}_{ij}^{(k)} = \mathbf{0}_{M_i \times N_j}$ if $\mathbf{R}[i, k] = 0$ or $\mathbf{C}[j, k] = 0$. The choice of the penalty parameters $\{\lambda_k\}_{k=1}^{K}$ is critical and must satisfy the conditions of Proposition 1 to allow for nonzero estimation of each module.

PROPOSITION 1.

Under objective (7), *the following are necessary to allow for each* $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}$ *to be nonzero:*

1.   *If for $k' \neq k$ the rows and columns of module k' are contained within those for module k, $\mathbf{R}[i, k] - \mathbf{R}[i, k'] \geq 0 \forall i$ and $\mathbf{C}[j, k] - \mathbf{C}[j, k'] \geq 0 \forall j$, then $\lambda_k > \lambda_{k'}$.*

2.   *If $\mathcal{I}_k \subset \{1, ..., k-1, k+1, ..., K\}$ is any subset of modules that together cover the rows and columns of module k, $\sum_{j \in \mathcal{I}_k}\mathbf{R}[\mathbin{|}\mathbin{|}j] \sqsupseteq r \mathbin{|} \mathbf{R}[\mathbin{|}\mathbin{|}k]$ and $\sum_{j \in \mathcal{I}_k}\mathbf{C}[\mathbin{|}\mathbin{|}j] \sqsupseteq c \mathbin{|} \mathbf{C}[\mathbin{|}\mathbin{|}k]$ for positive integers r and c, then $\lambda_k < \sum_{j \in \mathcal{I}_k}\lambda_j$.*

We determine the λ_k's by random matrix theory, motivated by two well-known results for a single matrix that we repeat here in Propositions 2 and 3.

PROPOSITION 2 (Mazumder, Hastie and Tibshirani (2010)).

Let $\mathbf{UDV}^T$ be the SVD of a matrix $\mathbf{X}$. The approximation $\mathbf{A}$ that minimizes

$$\frac{1}{2}\|\mathbf{X} - \mathbf{A}\|_F^2 + \lambda\|\mathbf{A}\|_* \tag{8}$$

is $\mathbf{A} = \mathbf{U}\tilde{\mathbf{D}}\mathbf{V}^T$, where $\tilde{\mathbf{D}}$ is diagonal with entries $\tilde{\mathbf{D}}[r, r] = \max(\mathbf{D}[r, r] - \lambda, 0)$.

PROPOSITION 3 (Rudelson and Vershynin (2010)).

Let $D[1, 1]$ be the largest singular value of a matrix $\mathbf{E} : M \times N$ of independent entries with mean 0, variance σ^2, and finite fourth moment. Then, $\mathbf{D}[1, 1] \approx \sigma(\sqrt{M} + \sqrt{N})$ as $M, N \to \infty$, and if the entries of $\mathbf{E}$ are Gaussian $\mathbb{E}(\mathbf{D}[1, 1]) \leq \sigma(\sqrt{M} + \sqrt{N})$ for any M, N.

Fixing $\lambda = \sigma(\sqrt{M} + \sqrt{N})$ is a reasonable choice for the matrix approximation task in (8), because it keeps only those components r whose signal is greater than that expected for independent error by Proposition 3: $\mathbf{D}[r, r] > \sigma(\sqrt{M} + \sqrt{N})$ (Shabalin and Nobel (2013)). For our context, $\sigma = 1$ after normalizing, as discussed in Section 6.1, and thus we set $\lambda_k = \sqrt{\mathbf{R}[\cdot, k] \cdot \mathbf{M}} + \sqrt{\mathbf{C}[\cdot, k] \cdot \mathbf{N}}$, where $\mathbf{R}[\cdot, k] \cdot \mathbf{M} \times \mathbf{C}[\cdot, k] \cdot \mathbf{N}$ gives the dimensions of the nonzero submatrix for $\mathbf{S}_{\cdot\cdot}^{(k)}$,

$$\mathbf{R}[\,\cdot\,,k]\cdot\mathbf{M} = \sum_{i=1}^{I} M_i \mathbf{R}[i,k] \quad \text{and} \quad \mathbf{C}[\,\cdot\,,k]\cdot\mathbf{N} = \sum_{j=1}^{J} N_j \mathbf{C}[j,k].$$

Our choice of λ_k is motivated to distinguish signal from noise in module $\mathbf{S}_{\cdot\cdot}^{(k)}$, conditional on the other modules. Moreover, it is guaranteed to satisfy the necessary conditions in Proposition 1, which we establish in Proposition 4.

PROPOSITION 4.

Setting $\lambda_k = \sqrt{\mathbf{R}[\,\cdot\,,k]\cdot\mathbf{M}} + \sqrt{\mathbf{C}[\,\cdot\,,k]\cdot\mathbf{N}}$ *in* (7) *satisfies the necessary conditions of* Proposition 1.

A similarly motivated choice of penalty weights is used in the BIDIFAC method which solves an equivalent objective under the restricted scenario where the columns of $\mathbf{R}$ and $\mathbf{C}$ are fixed and contain either entirely 1's (i.e., all rows or columns included) or just one 1 (i.e., just one row set or column set included). Thus, we call our more flexible approach BIDIFAC+.

It is often infeasible to explicitly consider each of the $K = (2^I - 1)(2^J - 1)$ possible modules in (7), and the solution is often sparse, with $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)} = \mathbf{0}$ for several k. Thus, in practice, we also optimize over the row and column sets $\mathbf{R}$ and $\mathbf{C}$ for some smaller number of modules $\tilde{K} \ll K$,

$$\operatorname*{argmin}_{\mathbf{R},\mathbf{C},\{\mathbf{S}_{\cdot\cdot}^{(k)}\}_{k=1}^{\tilde{K}}} \frac{1}{2}\left\|\mathbf{X}_{\cdot\cdot} - \sum_{k=1}^{\tilde{K}} \mathbf{S}_{\cdot\cdot}^{(k)}\right\|_F^2 + \sum_{k=1}^{\tilde{K}} (\sqrt{\mathbf{M}\cdot\mathbf{R}[\,\cdot\,,k]} + \sqrt{\mathbf{N}\cdot\mathbf{C}[\,\cdot\,,k]})\|\mathbf{S}_{\cdot\cdot}^{(k)}\|_* . \quad (9)$$

Note that if $\tilde{K}$ is greater than the number of nonzero modules, then the solution to (9) is equivalent to the solution to (7) in which $\mathbf{R}$ and $\mathbf{C}$ are fixed and enumerate all possible modules. If $\tilde{K}$ is not greater than the number of nonzero modules, then the solution to (9) can still be informative, as the set of $\tilde{K}$ modules that together give the best structural approximation via (7). In this case it helps to order the estimated modules by variance explained; if the $\tilde{K}$ th module still explains substantial variance, consider increasing $\tilde{K}$. Moreover, we suggest assessing the sensitivity of results to different values of $\tilde{K}$.

6. Estimation.

6.1. Scaling.

We center each dataset $\mathbf{X}_{ij}$ to have mean 0 and scale each dataset to have residual variance var($\mathbf{E}_{ij}$) of approximately 1. Such scaling requires an estimate of the residual variance for each dataset. By default, we use the median absolute deviation estimator $\hat{\sigma}_{\mathrm{MAD}}^2$ of Gavish and Donoho (2017), which is motivated by random-matrix theory under the assumption that $\mathbf{X}_{ij}$ is composed of low-rank structure and mean 0 independent noise of variance σ^2. We

estimate $\hat{\sigma}^2_{\text{MAD}}$ for the unscaled data $\mathbf{X}^{\text{unscaled}}_{ij}$ and set $\mathbf{X}_{ij} = \mathbf{X}^{\text{unscaled}}_{ij}/\hat{\sigma}_{\text{MAD}}$. An alternative approach is to scale each dataset to have overall variance 1, $\text{var}(\mathbf{X}_{ij}) = 1$, which is more conservative because $\text{var}(\mathbf{E}..) \leq \text{var}(\mathbf{X}_{ij})$; thus, this approach results in relatively larger λ_k in the objective function and leads to sparser overall ranks. Yet another approach is to normalize each data block to have unit Frobenius norm, as in Lock et al. (2013). However, our default choice of penalty parameters λ_k in Section 5 is theoretically motivated by the assumption that residual variances are the same across the different data blocks.

6.2. Optimization algorithm: Fixed modules.

We estimate across all modules $k = 1, \ldots, K$, simultaneously, by iteratively optimizing the objectives in Section 5. First, assume the row and column inclusions for each module, defined by $\mathbf{R}$ and $\mathbf{C}$, are fixed, as in objective (7). To estimate $\mathbf{S}^{(k)}_{..}$, given the other modules $\{\mathbf{S}^{(k')}_{..}\}_{k' \neq k}$, we can apply the soft-singular value estimator in Proposition 2 to the residuals on the submatrix, defined by $\mathbf{R}[\cdot, k]$ and $\mathbf{C}[\cdot, k]$. The iterative estimation algorithm proceeds as follows:

1.   *Initialize* $\hat{\mathbf{S}}^{(k)}_{..} = \mathbf{0}_{M \times N}$ *for* $k = 1, \ldots, K$.

2.   For $k = 1, \ldots, K$:

     a.   Compute the residual matrix $\mathbf{X}^{(k)}_{..} = \mathbf{X}.. - \sum_{k' \neq k} \hat{\mathbf{S}}^{(k')}_{..}$.

     b.   Set $\mathbf{X}^{(k)}_{ij} = \mathbf{0}_{M_i \times N_j}$ where $\mathbf{R}[i, k] = 0$ or $\mathbf{C}[j, k] = 0$.

     c.   Compute the SVD of $\mathbf{X}^{(k)}_{..}, \mathbf{X}^{(k)}_{..} = \mathbf{U}^{(k)}_{.}\mathbf{D}^{(k)}\mathbf{V}^{(k)}_{.}$.

     d.   Update $\hat{\mathbf{S}}^{(k)}_{..} = \mathbf{U}^{(k)}_{.}\hat{\mathbf{D}}^{(k)}\mathbf{V}^{(k)}_{.}$ where $\hat{\mathbf{D}}[r,r] = \max(\mathbf{D}[r,r] - \lambda_k, 0)$ for $r = 1, 2, \ldots$.

3.   Repeat step 2. until convergence of the objective function (9).

Step 2(d) minimizes the objective (7) for $\hat{\mathbf{S}}^{(k)}_{..}$, given $\{\mathbf{S}^{(k')}_{..}\}_{k' \neq k}$, by Proposition 2. In this way we iteratively optimize (7) over the K modules $\{\mathbf{S}^{(k)}_{..}\}^K_{k=1}$ until convergence.

6.3. Optimization algorithm: Dynamic modules.

If the row and column inclusions $\mathbf{R}$ and $\mathbf{C}$ are not predetermined, we incorporate additional substeps to estimate the nonzero submatrix, defined by $\mathbf{R}[\cdot, k]$ and $\mathbf{C}[\cdot, k]$, for each module to optimize (9). We use a dual forward-selection procedure to iteratively determine the optimal row-set $\mathbf{R}[\cdot, k]$ with columns $\mathbf{C}[\cdot, k]$ fixed, and the column-set $\mathbf{C}[\cdot, k]$ with rows $\mathbf{R}[\cdot, k]$ fixed, until convergence prior to estimating each $\mathbf{S}^{(k)}_{..}$. The iterative estimation algorithm proceeds as follows:

1.   Initialize $\hat{\mathbf{S}}^{(k)}_{..} = \mathbf{0}_{M \times N}$ for $k = 1, \ldots, K$.

2.   Initialize $\hat{\mathbf{C}}[j, k] = 1$ for $j = 1, \ldots, J$.

3.   For $k = 1, \ldots, K$:

    a.   Compute the residual matrix $\mathbf{X}_{\cdot\cdot}^{(k)} = \mathbf{X}_{\cdot\cdot} - \sum_{k' \neq k} \hat{\mathbf{S}}_{\cdot\cdot}^{(k')}$

    b.   Update $\hat{\mathbf{R}}[\cdot, k]$ and $\hat{\mathbf{C}}[\cdot, k]$, as follows:

        i.   With $\hat{\mathbf{C}}[\cdot, k]$ fixed, update $\hat{\mathbf{R}}[\cdot, k]$ by forward selection, beginning with $\hat{\mathbf{R}}[\cdot, k] = \mathbf{0}$ and iteratively adding rows i ($\hat{\mathbf{R}}[i, k] = 1$) to minimize the objective (9).

        ii.   With $\hat{\mathbf{R}}[\cdot, k]$ fixed, update $\hat{\mathbf{C}}[\cdot, k]$ by forward selection, beginning with $\hat{\mathbf{C}}[\cdot, k] = 0$ and iteratively adding columns j ($\hat{\mathbf{C}}[i, k] = 1$) to minimize the objective (9).

        iii.   Repeat steps i. and ii. until convergence of the chosen row and column sets $\hat{\mathbf{C}}[\cdot, k]$ and $\hat{\mathbf{R}}[\cdot, k]$.

    c.   Set $\mathbf{X}_{ij}^{(k)} = \mathbf{0}_{M_i \times N_j}$, where $\hat{\mathbf{R}}[i, k] = 0$ or $\hat{\mathbf{C}}[j, k] = 0$.

    d.   Compute the SVD of $\mathbf{X}_{\cdot\cdot}^{(k)}$, $\mathbf{X}_{\cdot\cdot}^{(k)} = \mathbf{U}_{\cdot}^{(k)}\mathbf{D}^{(k)}\mathbf{V}_{\cdot}^{(k)}$.

    e.   Update $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)} = \mathbf{U}_{\cdot}^{(k)}\hat{\mathbf{D}}^{(k)}\mathbf{V}_{\cdot}^{(k)}$, where $\hat{\mathbf{D}}[r, r] = \max(\mathbf{D}[r, r] - \lambda_k, 0)$ for $r = 1, 2, \ldots$.

4.   Repeat step 3. until convergence of the objective function.

For the steps in 3b), objective (9) can be efficiently computed using the singular values for the residual submatrix, given by $\hat{\mathbf{R}}[\cdot, k]$ and $\hat{\mathbf{C}}[\cdot, k]$, without reestimating $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}$ completely at each substep.

6.4. Convergence and tempered regularization.

The algorithms in Sections 6.2 and 6.3 both monotonically decrease the objective function at each substep, and thus both converge to a coordinatewise optimum. For (7) the objective function and solution space are both convex, and the algorithm with fixed modules tends to converge to a global optimum, as observed for the BIDIFAC method Park and Lock (2020). However, the stepwise updating of $\mathbf{R}$ and $\mathbf{C}$ in Section 6.3 can get stuck at coordinatewise optima, analogous to stepwise variable selection in a predictive model. In practice, we find that the convergence of the algorithm improves substantially if the initial iterations use a high nuclear norm penalty that gradually decreases to the desired level of penalization. Thus, in our implementation for the first iteration the penalties are set to $\tilde{\lambda}_k = \alpha\lambda_k$ for $k = 1, \ldots, K$ and some $a > 1$. The penalties then gradually decrease over each subsequent iteration of the algorithm before reaching the desired level of regularization ($a = 1$).

7.   Uniqueness.

Here, we consider the uniqueness of the decomposition in (4) under the objective (7). To account for permutation invariance of the K modules, throughout this section we assume that $\mathbf{R}$ and $\mathbf{C}$ are fixed and that they enumerate all of the $K = (2^I - 1)(2^J - 1)$ possible modules. Note that the solution may still be sparse, with $\mathbf{S}_{\cdot\cdot}^{(k)} = \mathbf{0}$ for some or most of the modules. Without loss of generality, we fix $\mathbf{R}$ and $\mathbf{C}$, as in Supplementary Material Appendix A (Lock, Park and Hoadley (2022)). Then, let $\mathbb{S}_{\hat{\mathbf{X}}}$ be the set of possible decompositions that yield a given approximation $\hat{\mathbf{X}}_{\cdot\cdot}$,

$$\mathbb{S}_{\hat{\mathbf{X}}} = \left\{ \{\mathbf{S}_{\cdot\cdot}^{(k)}\}_{k=1}^{K} \mid \hat{\mathbf{X}}_{\cdot\cdot} = \sum_{k=1}^{K} \mathbf{S}_{\cdot\cdot}^{(k)} \right\}.$$

If either $I > 1$ or $J > 1$, then the cardinality of $\mathbb{S}_{\hat{\mathbf{X}}}$ is infinite; that is, there are an infinite number of ways to decompose $\hat{\mathbf{X}}_{\cdot\cdot}$. Thus, model (4) is clearly not identifiable, in general, even in the no-noise case $\mathbf{E}_{\cdot\cdot} = \mathbf{0}$. However, optimizing the structured nuclear norm penalty in (7) may uniquely identify the decomposition; let $f_{\text{pen}}(\cdot)$ give this penalty,

$$f_{\text{pen}}(\{\mathbf{S}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}) = \sum_{k=1}^{K} (\sqrt{\mathbf{R}[\,\cdot\,, k] \cdot \mathbf{M}} + \sqrt{\mathbf{C}[\,\cdot\,, k] \cdot \mathbf{N}}) \|\mathbf{S}_{\cdot\cdot}^{(k)}\|_* .$$

Proposition 5 gives an equivalence of the left and right singular vectors for any two decompositions that minimize $f_{\text{pen}}(\cdot)$.

PROPOSITION 5.

Take two decompositions $\{\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K} \in \mathbb{S}_{\hat{\mathbf{X}}}$, *and* $\{\tilde{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K} \in \mathbb{S}_{\hat{\mathbf{X}}}$, *and assume that both minimize the structured nuclear norm penalty,*

$$f_{\text{pen}}(\{\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}) = f_{\text{pen}}(\{\tilde{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}) = \min_{\mathbb{S}_{\hat{\mathbf{X}}}} f_{\text{pen}}(\{\mathbf{S}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}).$$

Then, $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)} = \mathbf{U}_{\cdot}^{(k)} \hat{\mathbf{D}}_{\cdot}^{(k)} \mathbf{V}_{\cdot}^{(k)T}$ *and* $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)} = \mathbf{U}_{\cdot}^{(k)} \tilde{\mathbf{D}}_{\cdot}^{(k)} \mathbf{V}_{\cdot}^{(k)T}$, *where* $\mathbf{U}_{\cdot}^{(k)}: M \times R_k$ *and* $\mathbf{V}_{\cdot}^{(k)}: N \times R_k$ *have orthonormal columns and* $\hat{\mathbf{D}}_{\cdot}^{(k)}$ *and* $\tilde{\mathbf{D}}_{\cdot}^{(k)}$ *are diagonal.*

The proof of Proposition 5 uses two novel lemmas (see Supplementary Material Appendix B (Lock, Park and Hoadley (2022))): one establishing that $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}$ and $\tilde{\mathbf{S}}_{\cdot\cdot}^{(k)}$ must be additive in the nuclear norm, $\|\hat{\mathbf{S}}_{\cdot\cdot}^{(k)} + \tilde{\mathbf{S}}_{\cdot\cdot}^{(k)}\|_* = \|\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\|_* + \|\tilde{\mathbf{S}}_{\cdot\cdot}^{(k)}\|_*$ and a general result establishing that any two matrices that are additive in the nuclear norm must have the equivalence in Proposition 5.

Proposition 5 implies that left or right singular vectors of $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}$ ($\hat{\mathbf{D}}^{(k)}[r,r] > 0$) are either shared with $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}$ (if $\tilde{\mathbf{D}}^{(k)}[r,r] > 0$) or orthogonal to $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}$ (if $\tilde{\mathbf{D}}^{(k)}[r,r] = 0$). For uniqueness, one must establish that $\hat{\mathbf{D}}^{(k)}[r,r] = \tilde{\mathbf{D}}^{(k)}[r,r]$ for all k and r. Theorem 1 gives sufficient conditions for uniqueness of the decomposition.

THEOREM 1.

Consider $\{\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K} \in \mathbb{S}_{\hat{\mathbf{X}}}$, *and let* $\mathbf{U}_{\cdot}^{(k)}\hat{\mathbf{D}}^{(k)}\mathbf{V}_{\cdot}^{(k)T}$ *give the SVD of* $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}$ *for* $k = 1, \ldots, K$. *The following three properties uniquely identify* $\{\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}$:

1. $\{\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}$ *minimizes* $f_{\text{pen}}(\cdot)$ *over* $\mathbb{S}_{\hat{\mathbf{X}}}$,

2. $\{\hat{\mathbf{U}}_i^{(k)}[\cdot, r]: \mathbf{R}[i,k] = 1 \text{ and } \hat{\mathbf{D}}^{(k)}[r,r] > 0\}$ *are linearly independent for* $i = 1, \ldots I$,

3. $\{\hat{\mathbf{V}}_j^{(k)}[\cdot, r]: \mathbf{C}[j,k] = 1 \text{ and } \hat{\mathbf{D}}^{(k)}[r,r] > 0\}$ *are linearly independent for* $j = 1, \ldots, J$.

The linear independence conditions (2. and 3.) are, in general, not sufficient, and several related integrative factorization methods, such as JIVE (Lock et al. (2013)) and SLIDE (Gaynanova and Li (2019)), achieve identifiability via stronger orthogonality conditions across the terms of the decomposition. Theorem 1 implies that orthogonality is not necessary under the penalty $f_{\text{pen}}(\cdot)$. Conditions 2 and 3 are straightforward to verify for any $\{\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}$, and they will generally hold whenever the ranks in the estimated factorization are small relative to the dimensions $\{M_i\}_{i=1}^{I}$ and $\{N_j\}_{j=1}^{J}$. Moreover, the conditions of Theorem 1 are only sufficient for uniqueness; there may be cases for which the minimizer of $f_{\text{pen}}(\cdot)$ is unique and the terms of its decomposition are not linearly independent. Theorem 1 implies uniqueness of the BIDIFAC decomposition (Park and Lock (2020)) under linear independence as a special case which is a novel result.

8. Bayesian interpretation.

Express the BIDIFAC+ model (6) in factorized form

$$\mathbf{X}_{\cdot\cdot} = \sum_{k=1}^{K} \mathbf{U}_{\cdot}^{(k)}\mathbf{V}_{\cdot}^{(k)T} + \mathbf{E}_{\cdot\cdot}, \tag{10}$$

where

$$\mathbf{U}_{\cdot}^{(k)'} = [\mathbf{U}_1^{(k)'} \cdots \mathbf{U}_I^{(k)'}]' \quad \text{with } \mathbf{U}_i^{(k)} = M_1 \times R_k \text{ and } \mathbf{U}_i^{(k)} = \mathbf{0}_{M_1 \times R_k} \text{ if } \mathbf{R}[i,k] = 1 \tag{11}$$

for all i and k, and

$$\mathbf{V}_{\cdot}^{(k)} = [\mathbf{V}_1^{(k)} \cdots \mathbf{V}_J^{(k)}] \quad \text{with } \mathbf{V}_i^{(k)} = N_j \times R_k \text{ and } \mathbf{V}_i^{(k)} = \mathbf{0}_{N_j \times R_k} \text{ if } \mathbf{C}[j,k] = 1 \tag{12}$$

for all j and k. The structured nuclear norm objective (7) can also be represented by L_2 penalties on the factorization components $\mathbf{U}_.^{(k)}$ and $\mathbf{V}_.^{(k)}$. We formally state this equivalence in Proposition 6 which extends analogous results for a single matrix (Mazumder, Hastie and Tibshirani (2010)) and for the BIDIFAC framework (Park and Lock (2020)).

PROPOSITION 6.

Fix $\mathbf{R}$ *and* $\mathbf{C}$. *Let* $\{\hat{\mathbf{U}}_.^{(k)}\}_{k=1}^{K}$ *and* $\{\hat{\mathbf{V}}_.^{(k)}\}_{k=1}^{K}$ *minimize*

$$\left\| \mathbf{X}_{..} - \sum_{k=1}^{K} \mathbf{U}_.^{(k)}, \mathbf{V}_.^{(k)} \right\|_F^2 + \sum_{k=1}^{K} \lambda_k (\|\mathbf{U}_.^{(k)}\|_F^2 + \|\mathbf{V}_.^{(k)}\|_F^2) \tag{13}$$

with the restrictions (11) *and* (12). *Then,* $\{\hat{\mathbf{S}}_{..}^{(k)}\}_{k=1}^{K}$ *solves* (7), *where* $\hat{\mathbf{S}}_{..}^{(k)} = \hat{\mathbf{U}}_.^{(k)} \hat{\mathbf{V}}_.^{(k)T}$ *for* $k = 1, \ldots, K$.

From (13), it is apparent that our objective gives the mode of a Bayesian posterior with normal priors on the errors and the factorization components, as stated in Proposition 7.

PROPOSITION 7.

Let the entries of $\mathbf{E}_{..}$ *be independent* Normal(0, 1), *the entries of* $\mathbf{U}_i^{(k)}$ *be independent* Normal(0, τ^2) *if* $\mathbf{R}[i, k] = 1$, *and the entries of* $\mathbf{V}_j^{(k)}$ *be independent* Normal $(0, \tau_k^2)$ *if* $\mathbf{C}[j, k] = 1$, *where* $\tau_k^2 = 1/\lambda_k$. *Then,* (13) *is proportional to the log of the joint likelihood*

$$p(\mathbf{X}_{..}, \{\mathbf{U}_.^{(k)}\}_{k=1}^{K}, \{\mathbf{V}_.^{(k)}\}_{k=1}^{K} \mid \mathbf{R}, \mathbf{C}).$$

9. Missing data imputation.

The probabilistic formulation of the objective, described in Section 8, motivates a modified expectation-maximization (EM)-algorithm approach to impute missing data. Let $\mathcal{M}$ index observations in the full dataset $\mathbf{X}_{..}$ that are unobserved: $\mathcal{M} = \{(m, n) : \mathbf{X}_{..}[m, n]$ is missing$\}$. Our iterative algorithm for missing data imputation proceeds as follows:

1. Initialize $\hat{\mathbf{X}}_{..}$ by

$$\hat{\mathbf{X}}_{..}[m, n] = \begin{cases} \mathbf{X}_{..}[m, n] & \text{if } (m, n) \notin \mathcal{M}, \\ 0 & \text{if } (m, n) \in \mathcal{M}. \end{cases}$$

2. M-step: Estimate $\{\hat{\mathbf{S}}_{..}^{(k)}\}_{k=1}^{K}$ by optimizing (7) for $\hat{\mathbf{X}}_{..}$.

3. E-step: Update $\hat{\mathbf{X}}_{..}$ by $\hat{\mathbf{X}}_{..}[m, n] = \sum_{k=1}^{K} \hat{\mathbf{S}}_{..}^{(k)}[m, n]$.

4. Repeat steps 2 and 3 until convergence.

Analogous approaches to imputation for other low-rank factorization techniques have been proposed (Kurucz, Benczúr and Csalogány (2007), O'Connell and Lock (2019), Park and Lock (2020), Mazumder, Hastie and Tibshirani (2010)). Due to centering, initializing missing data to 0 in step 1 is equivalent to starting with mean imputation which is used by other SVD-based imputation approaches (Mazumder, Hastie and Tibshirani (2010), Kurucz, Benczúr and Csalogány (2007)); however, in practice, random initializations can also be used. The M-step maximizes the joint density for the model in Proposition 7, where $\hat{\mathbf{S}}_{\cdot\cdot}^{(k)} = \hat{\mathbf{U}}_{\cdot}^{(k)}\hat{\mathbf{V}}_{\cdot}^{(k)T}$ for $k = 1, \dots, K$. The E-step updates the log joint density with its conditional expectation over $\{\mathbf{X}_{\cdot\cdot}[m, n]: (m, n) \in \mathcal{M}\}$ by an argument analogous to that in Zhang et al. (2005). Thus, the approach is an EM-algorithm for maximum a posteriori imputation under the model in Section 8, and it is also a direct block coordinate descent of objective (7) over $\{\mathbf{S}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}$ and $\{\mathbf{X}_{\cdot\cdot}[m, n]: (m, n) \in \mathcal{M}\}$. Crucially, for our context the method can be used to impute data that may be missing from an entire column or an entire row of each $\mathbf{X}_{ij}$ or, in certain cases, can even be used to impute an entire matrix $\mathbf{X}_{ij}$ based on joint structure.

10. Application to TCGA data.

10.1. Data acquisition and preprocessing.

Our data were curated for the TCGA Pan-Cancer Project and were used for the pan-cancer clustering analysis described in Hoadley et al. (2018). We used data from four ($I = 4$) omics sources: (1) batch-corrected RNA-Seq data capturing (mRNA) expression for 20, 531 genes, (2) batch-corrected miRNA-Seq data capturing expression for 743 miRNAs, (3) between-platform normalized data from the Illumina 27K and 450K platforms capturing DNA methylation levels for 22,601 CpG sites, and (4) batch-corrected reverse-phase protein array data capturing abundance for 198 proteins. These data are available for download at https://gdc.cancer.gov/about-data/publications/PanCan-CellOfOrigin [accessed 11/19/2019]. We consider data for $N = 6973$ tumor samples from different individuals with all four omics sources available; these tumor samples represent $J = 29$ different cancer types, listed in Table 1.

We log-transformed the counts for the RNA-Seq and miRNA-Seq sources. To remove baseline differences between cancer types, we center each data source to have mean 0 across all rows for each cancer type,

$$\mathrm{mean}(\mathbf{X}_{ij}[m, \cdot]) = 0 \quad \text{for all } i, j, m.$$

We filter to the 1000 genes and the 1000 methylation CpG probes that have the highest standard deviation after centering, leaving $M_1 = 1000$ genes, $M_2 = 743$ miRNAs, $M_3 = 1000$ CpGs, and $M_4 = 198$ proteins. Lastly, to account for differences in scale we standardize so that each variable has standard deviation 1,

$$\mathrm{SD}(\mathbf{X}_{i\cdot}[m, \cdot]) = 1 \quad \text{for all } i, m.$$

This is a conservative alternative to scaling by an estimate of the noise variance, as mentioned in Section 6.1.

10.2. Factorization results.

We apply the BIDIFAC+ method to the complete-case data with $I = 4$ omics sources and $J = 29$ cancer types. We simultaneously estimate a maximum of $K = 50$ low-rank modules; all modules are nonzero, but the variation explained by the smaller modules are negligible.

Figure 2 gives the total variance explained by each module, $\|\hat{\mathbf{S}}_{..}^{(k)}\|_F^2$, for $k = 1, \ldots, 50$ in decreasing order. The top 15 modules, ordered by total variance, explained are given in Table 2, and all 50 modules are given in the Supplementary Material Spreadsheet S1 (Lock, Park and Hoadley (2022)). The first module explains global variation, with all cancer types and all omics sources included. Other modules that explain substantial variability across all or almost all cancer types are specific to each omics source: miRNA (Module 2), methylation (Module 3), gene expression (Module 5), and Protein (Module 8).

The module that explains the fourth most variation (Module 4) identifies structure in the genes and DNA methylation that explains variation in 22 of the 29 cancer types; we focus on this module as an illustrative example. The cancer types *not* included in Module 4 are BRCA (breast), CESC (cervical), OV (ovarian), PRAD (prostate), TGCT (testicular), UCEC (uterine endometrial), and UCS (uterine). Interestingly, all tumor types that were excluded were cancers specific to either males or females (or heavily skewed in BRCA); while cancer types included have both sexes. Figure 3 shows that Module 4 is indeed dominated by a single component that corresponds to molecular differences between the sexes. The gene loadings for this component are negligible, except for those on the Y chromosome and two genes on the X chromosome that are responsible for X-inactivation in females, *XIST* and *TSIX*; the methylation loadings are negligible, except for those in the X chromosome. These results are an intuitive illustration of the method, revealing a multiomic molecular signal that explains heterogeneity in some cancer types, but not all cancer types (only those that have both males and females).

The module that explains the sixth most variation (Module 6) identifies structure across all four omics sources that explains variation in the breast cancer (BRCA) samples only. Figure 4 shows that the first two components in this module are driven primarily by distinctions between the PAM50 molecular subtypes for BRCA (Network et al. (2012)). Thus, our analysis suggests that molecular signals that distinguish these subtypes are present across all four omics sources, but that these signals do not explain substantial variation within any other type of cancer considered.

Several other modules explain variability in just one type of type of cancer, including LGG (Module 7: mRNA, miRNA and Protein), THCA (Module 9), UCEC (Module 16), and PRAD (Modules 18 and 19). Module 12, which is specific to LGG methylation, reveals distinct clustering by mutation status of the *IDH* genes (see Figure 5). IDH mutations have been shown to lead to a distinct CpG-island hypermethylated phenotype (Noushmehr et al. (2010)). Other modules explain variability in multiple cancer types that share similarities regarding their origin or histology. For example, Module 14 explains variability within

the three kidney cancers (KICH, KIRC, and KIRP), and digestive and gastrointestinal cancers (CORE, ESCA, PAAD, STAD) are represented in Modules 25 (methylation) and 28 (mRNA).

To assess sensitivity of the results to the maximum number of modules $\tilde{K} = 50$, we also performed the decomposition with $\tilde{K} = 25, 30, 35, 40,$ or 45. The resulting decompositions share several similarities. Nine of the 15 modules in Table 2 have a precise match in each of the six decompositions (Modules 1, 2, 4, 5, 6, 7, 9, 12, 14). Entrywise correlations for the terms $\mathbf{S}_{..}^{(k)}$ between matched modules ranged from $r = 0.88$ to $r = 0.99$. Other modules had slight variations to the cancer types and omics sources included. The modules for different $\tilde{K}$ are provided as separate tabs in Supplementary Material Spreadsheet S1 (Lock, Park and Hoadley (2022)). The overall structural approximations $\sum_{k=1}^{\tilde{K}} \mathbf{S}_{..}^{(k)}$ were very similar across decompositions, with entrywise correlations all greater than $r = 0.99$.

10.3. Missing data imputation.

To assess the accuracy of missing data imputation using BIDIFAC+, we hold out observed entries, rows, and columns of each dataset in the pan-omics pan-cancer and impute them using the approach in Section 9. We randomly set 100 columns (samples) to missing for each of the *four* omic platforms, and we randomly set 100 rows (features) to missing for each of the 29 cancer types. We then randomly set 5000 of the values remaining in the joint matrix $\mathbf{X}_{..}$ to missing. We impute missing values using BIDIFAC+, as described in Section 9, and, for comparison, we use an analogous approach to imputation using four other low-rank factorizations: (1) soft-threshold (nuclear norm) SVD of the joint matrix $\mathbf{X}_{..}$, (2) soft-threshold SVD of each matrix $\mathbf{X}_{ij}$ separately, (3) hard-threshold SVD (SVD approximation using the first R singular values) of $\mathbf{X}_{..}$, (4) hard-threshold SVD of each $\mathbf{X}_{ij}$ separately. For the soft-thresholding SVD methods, the penalty factor is estimated by random matrix theory, as in Section 5. For the hard-thresholding methods the ranks are determined by cross-validation by minimizing imputation error on an additional held-out cohort of the same size.

We consider the imputation error under the different methods, broken down by: (1) observed values, (2) values that are missing but have the rest of their row and column present (entrywise missing), (3) values that are missing their entire row, (4) values that are missing their entire column, and (5) values that are missing both their row and their column. For a given set of values $\mathcal{M}$, we compute the relative squared error as

$$\text{RSE} = \frac{\sum_{(m, n) \in \mathcal{M}} (\mathbf{X}_{..}[m, n] - \hat{\mathbf{X}}_{..}[m, n])^2}{\sum_{(m, n) \in \mathcal{M}} \mathbf{X}_{..}[m, n]^2},$$

where $\hat{\mathbf{X}}_{..}$ is the structural approximation resulting from the given method. Table 3 gives the RSE for each method and for each missing condition. Imputation by BIDIFAC+ outperforms the other methods for each type of missingness, illustrating the advantages of decomposing

joint and individual structures. The hard-thresholding approaches have much less error for the observed data than for the missing data, due to overfitting of the signal.

11. Simulation studies.

11.1. Vertically linked simulations.

We conduct a simulation study to assess the accuracy of the BIDIFAC+ decomposition in the context of vertical integration, where there is a single shared column set ($J = 1$). For all scenarios we simulate data according to model (4), wherein the entries of the residual noise $\mathbf{E}_{..}$ are generated independently from a Normal(0, 1) distribution and the entries of each $\mathbf{U}_i^{(k)}$ and $\mathbf{V}^{(k)}$ are generated independently from a Normal (0, σ^2) distribution.

We first consider a scenario with $I = 3$ matrices, each of dimension 100×100 ($N = 100$ and $M_1 = M_2 = M_3 = 100$), with low-rank modules that are shared jointly, shared across each pair of matrices and individual to each matrix,

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}. \tag{14}$$

We consider a "low-rank" and a "high-rank" condition across three different signal-to-noise levels. For the low-rank condition, each of the seven modules has rank $R = 1$; for the high-rank condition, each module has rank $R = 5$. The variance of the factorized signal component, σ^2 is set to be $\sqrt{1/2}$, 1 or $\sqrt{10}$ so that the signal-to-noise ratio (s2n) of each components is 1/2, 1, or 10, respectively.

For each condition we apply four approaches to uncover the underlying decomposition:

1. BIDIFAC+, with $\mathbf{R}$ given by (14), as in the true generative model,

2. BIDIFAC+, with $\mathbf{R}$ estimated,

3. SLIDE, with $\mathbf{R}$ and the true ranks of each module ($R = 1$ or $R = 5$) provided,

4. SLIDE, with $\mathbf{R}$ and the ranks of each module estimated via the default cross-validation scheme.

We use SLIDE as the basis of comparison with BIDIFAC+, because it is the only other method that is designed to recover each term in the decomposition and it generally outperforms other vertically linked decomposition methods (Gaynanova and Li (2019), Park and Lock (2020)). For each case we compute the mean relative squared error (RSE) in recovering each module of the decomposition:

$$\text{RSE} = \frac{1}{K} \sum_{k=1}^{K} \frac{\| \mathbf{S}_{..}^{(k)} - \hat{\mathbf{S}}_{..}^{(k)} \|_F^2}{\| \mathbf{S}_{..}^{(k)} \|_F^2}. \tag{15}$$

The mean RSE for each condition and under each approach is shown in Table 4, broken down by the global module, pairwise modules, and individual modules. BIDIFAC+ generally outperforms or performs similarly to SLIDE, even when the true ranks are used for the SLIDE implementation (the ranks are never fixed for BIDIFAC+). An exception is when the ranks and s2n ratio are small (rank = 1, s2n = 0.5), where BIDIFAC+ tends to overshrink the signal. BIDIFAC+ performs particularly well, relative to SLIDE, when the rank is large and s2n is high. One likely reason for this improvement is that the SLIDE model necessarily restricts the factorized components $\mathbf{U}_i^{(k)}$ and $\mathbf{V}^{(k)}$ to be mutually orthogonal, whereas BIDIFAC+ has no such constraint. This restriction can be limiting when decomposing generated signals that are independent but not orthogonal (Park and Lock (2020)). Moreover, when estimating the ranks, the SLIDE model can drastically underperform relative to using the true ranks. The results for BIDIFAC+, when fixing the true modules $\mathbf{R}$ vs. estimating $\mathbf{R}$, are nearly identical; because all possible modules are present for this scenario, the two approaches are very similar despite subtle differences in the algorithms.

We consider another scenario with a larger number of matrices ($I = 10$), each of dimension 100×100 ($N = 100$, $M_1 = \cdots = M_{10} = 100$) and sparsely distributed modules. We generate 10 low rank modules out of $2^{10} - 1 = 1023$ possibilities that are present on (1) X_{11} only, (2) X_{11} and X_{21}, (3) X_{11}, X_{21}, and X_{31}, etc. We again consider low-rank ($R = 1$) and high-rank ($R = 5$) scenarios for all modules, and three signal-to-noise levels 0.5, 1, and 10. The resulting mean RSE (15) over all modules, for each approach, is shown in Table 5. Here, BIDIFAC+ with fixed true $\mathbf{R}$ generally performs better than estimating $\mathbf{R}$; however, these gains are modest for most scenarios, suggesting the BIDIFAC+ generally does a good job of identifying which of the 1023 possible modules are nonzero.

11.2. Missing data simulation.

Here, we assess the performance of missing data imputation for a 3×3 grid of matrices ($I = J = 3$) with $\mathbf{X}_{ij}$: 100×100 for each i, j. Data $\mathbf{X}_{..}$ are generated as in model (6), with one fully shared module ($\mathbf{R}[\cdot, 1] = \mathbf{C}[\cdot, 1] = [1, 1, 1]$) and modules specific to each of the 9 matrices ($\mathbf{R}[\cdot, 1] = \mathbf{C}[\cdot, 1] = [1, 0, 0]$, etc.). All modules have rank 5. The entries of residual noise are generated independently from a Normal(0, 1) distribution, and the entries of each $\mathbf{U}_i^{(k)}$ and $\mathbf{V}_i^{(k)}$ are generated independently from a Normal $(0, \sigma_k^2)$ distribution. We consider different levels of the joint signal strength $\sigma_j^2 := \sigma_1^2$ and the matrix-specific (i.e., individual) signal strength $\sigma_i^2 := \sigma_2^2 = \cdots = \sigma_{10}^2$. We further consider scenarios with different kinds of missingness: (1) 1/9 of the entries in $\mathbf{X}_{..}$ missing at random, (2) 1/9 of the columns in each $\mathbf{X}_{ij}$ entirely missing at random, and (3) one of the nine datasets $\mathbf{X}_{ij}$ entirely missing.

For each scenario we impute missing values using the same approaches used in Section 10.3. For hard SVD imputation we use the true rank of the underlying structure. We consider three versions of BIDIFAC+ imputation: (1) initializing by setting missing values to 0, (2) initializing by generating missing values randomly from a N(0,1) distribution, and (3) initializing at 0 and fixing the modules $\mathbf{R}$ and $\mathbf{C}$ to match the data generation (i.e., one fully

shared module and one module specific to each matrix). The resulting RSEs for the missing signal,

$$\text{RSE} = \frac{\sum_{(m,n) \in \mathcal{M}} (\sum_{k=1}^{10} \mathbf{S}_{\cdot\cdot}^{(k)}[m,n] - \hat{\mathbf{X}}\|[m,n])^2}{\sum_{(m,n) \in \mathcal{M}} (\sum_{k=1}^{10} \mathbf{S}_{\cdot\cdot}^{(k)}[m,n])^2},$$

are shown in Table 6, averaged over 100 replications. Imputation by BIDIFAC+ is flexible and performs relatively well across the different scenarios. As expected, the accuracy of columnwise and blockwise imputation depends strongly on the relative strength of the joint signal. Initializing missing values to $\mathbf{0}$ or randomly generated values gave very similar results, suggesting that either approach is reasonable in practice. BIDIFAC+ using the true $\mathbf{R}$ and $\mathbf{C}$ is comparable to estimating $\mathbf{R}$ and $\mathbf{C}$ across most scenarios. An exception is for block-missing data (i.e., missing an entire $\mathbf{X}_{ij}$); in this context, there is some ambiguity on whether shared signals, defined by $\hat{\mathbf{R}}$ and $\hat{\mathbf{C}}$, include the missing block, and so the performance with the true $\mathbf{R}$ and $\mathbf{C}$ is slightly better. Also, as for other blocks, values for the missing block are imputed under the assumption of error variance 1 and mean 0; in practice, they cannot be transformed back to their original measurement scale without prior knowledge of the mean and error variance.

11.3. Application-motivated simulation.

Here, we assess the recovery of the underlying structure and the accuracy of the decomposition into shared components for a bidimensionally linked scenario that reflects our motivating application in Section 10. We generate data by taking the estimated decomposition from Section 10.2 and adding independent noise to it. That is, we simulate

$$\tilde{\mathbf{X}}_{\cdot\cdot} = \sum_{k=1}^{50} \alpha \hat{\mathbf{S}}_{\cdot\cdot}^{(k)} + \tilde{\mathbf{E}}_{\cdot\cdot},$$

where $\{\hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\}_{k=1}^{K}$ is the estimated decomposition from Section 10.2, the entries of $\tilde{\mathbf{E}}_{\cdot\cdot}$ are independent Normal(0, 1), and $a > 0$ is a parameter that controls the total signal-to-noise ratio. We consider three total signal-to-noise ratios, defined by

$$\text{s2n} = \text{var}\left(\sum_{k=1}^{50} \alpha \hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\right) / \text{var}(\tilde{\mathbf{E}}) = \text{var}\left(\sum_{k=1}^{50} \alpha \hat{\mathbf{S}}_{\cdot\cdot}^{(k)}\right),$$

s2n = 0.2, 0.5, and 5. The scenario with s2n = 0.5 corresponds most closely to the real data, for which the ratio of the estimated signal variance over the residual variance is 0.552. For each scenario we estimate the underlying decomposition using BIDIFAC+ with the true $\mathbf{R}$ and $\mathbf{C}$ fixed, and using BIDIFAC+ with estimated modules $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{C}}$ and $\tilde{K} = 50$. In each case we compute the RSE as follows:

$$\text{RSE} = \frac{1}{50} \sum_{k=1}^{50} \frac{\|\tilde{\mathbf{S}}_{..}^{(k)} - \alpha \hat{\mathbf{S}}_{..}^{(k)}\|_F^2}{\|\alpha \mathbf{S}_{..}^{(k)}\|_F^2}. \tag{16}$$

When computing RSE, we permute the 50 modules so that $\tilde{\mathbf{R}}[\cdot, k] = \mathbf{R}[\cdot, k]$ and $\tilde{\mathbf{C}}[\cdot, k] = \mathbf{C}[\cdot, k]$, wherever possible, and set $\tilde{\mathbf{S}}_{..}^{(k)} = \mathbf{0}$ if $\tilde{\mathbf{R}}[\cdot, k] \neq \mathbf{R}[\cdot, k]$ and $\tilde{\mathbf{C}}[\cdot, k] \neq \mathbf{C}[\cdot, k]$. We also compute the relative overall signal recovery (ROSR) as

$$\text{ROSR} = \frac{\|\sum_{k=1}^{K} \hat{\mathbf{S}}_{..}^{(k)} - \sum_{k=1}^{K} \alpha \hat{\mathbf{S}}_{..}^{(k)}\|_F^2}{\|\sum_{k=1}^{K} \alpha \mathbf{S}_{..}^{(k)}\|_F^2}. \tag{17}$$

The results are shown in Table 7 and demonstrate that the underlying decomposition is recovered reasonably well in most scenarios. However, the RSE for estimated modules is often substantially more than the RSE using the true modules, as the row and column sets defining the modules can be estimated incorrectly. Moreover, the overall signal recovery error (ROSR) is, generally, substantially less than the mean error in recovering each module (RSE), demonstrating how the decomposition can be estimated incorrectly, even if the overall signal is estimated with high accuracy.

12. Discussion.

The successful integration of multiple large sources of data is a pivotal challenge for many modern analysis tasks. While several approaches have been developed, they largely do not apply to the context of bidimensionally linked matrices. BIDIFAC+ is a flexible approach for dimension reduction and decomposition of shared structures among bidimensionally linked matrices which is competitive with alternative methods that integrate over a single dimension (rows or columns). Here, we have focused primarily on the accuracy of the estimated decomposition and exploratory analysis of the results. BIDIFAC+ may also be used for other tasks, such as missing data imputation or as a dimension reduction step preceding statistical modeling (e.g., as in principal components regression). For these other tasks it is desirable to model statistical uncertainty and fully Bayesian extensions that capture the full posterior distribution about the mode in Section 8 are potentially very useful. Moreover, while we have explored the uniqueness of the decomposition under BIDFAC+, it is worthwhile to establish conditions that are both necessary and sufficient for its identifiability.

An implicit assumption of the BIDIFAC+ framework is that shared structures are present over complete submatrices. However, it is conceivable that structured variation may take other partially shared forms. For example, a pattern of variation in DNA methylation may exist across several cancer types but only regulate gene expression in some of those cancer types. Allowing for such shared structures, which do not exist over a complete submatrix, is an interesting direction of future work for which the separable form of the objective (13)

may be useful. Moreover, the approach may be extended to account for different layers of granularity in the row and column sets; for example, it would be interesting to also identify shared and specific patterns of variability among known subtypes within each cancer type. We select penalty terms by random matrix theory which requires assumptions of independent error and homogenous variance; an alternative strategy is to use the imputation accuracy of held-out data as a metric for parameter selection.

Computing time for BIDIFAC+ can range from $<$ *five* minutes for each simulation in Sections 11.1 and 11.2 to $\approx$ 24 hours until convergence for the pan-omics pan-cancer application in Section 10 and the accompanying simulation in Section 11.3. Thus, computational feasibility must be considered carefully or larger scale problems.

Our application to pan-omics pan-cancer data from TCGA revealed molecular patterns that explain variability across all or almost all types of cancer, both across omics platforms and within each omics platform. However, it also revealed patterns several instances in which patterns are specific to one or a small subset of cancers, and these often show sharp distinctions of previously known molecular subtypes (e.g., for BRCA and LGG). Interestingly, BRCA was the only tumor type that showed up with all four platforms in a module. Together, they strongly separated the Basal-like molecular subtype from other subtypes of breast cancer. This mirrors the analysis of individual data types in Network et al. (2012). The LGG data also split by both histological groups and mutation status, based on BIDFAC+, even though both were not included in the analysis. Module 7 included mRNA, miRNA, and protein and was predominantly driven by codeletion of 1p/19q which is predominantly observed in oligodendrogliomas and is associated with better overall survival. This mirrors the previous TCGA work that showed that the LGG could be predominately split by 1p/19q deletion, IHD1 status (Module 12, for methylation) or TP53 mutation status (Network (2015)). An important insight provided by BIDIFAC+ is that these molecular distinctions are specific to BRCA and LGG, respectively, suggesting that similar phenomena do not account for heterogeneity within other types of cancers. Other modules that are broadly shared across cancer types have potential to reveal relevant molecular signal that are undetectable within a single cancer type, especially those with smaller sample sizes. Beyond exploratory visualization, to systematically investigate the clinical relevance of the underlying modules one can cluster their structure to identify novel subtypes analogous to the approach described in Hellton and Thoresen (2016). Furthermore, one can use the results in a predictive model for a clinical outcome, analogous to the approach described in Kaplan and Lock (2017). We are currently pursuing the use of BIDIFAC+ results for these tasks.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements.

Availability.

A GitHub repository for BIDIFAC+ is available at https://github.com/lockEF/bidifac, and R code for all analyses presented herein is provided in a supplemental zipped folder.

REFERENCES

Akbani R, Ng PKS, Werner HMJ, Shahmoradgoli M, Zhang F, Ju Z, Liu W, Yang J-Y, Yoshihara K et al. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat. Commun 5 3887. 10.1038/ncomms4887 [PubMed: 24871328]

Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W and Stegle O (2018). Multi-Omics Factor Analysis—A framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol 14 e8124. 10.15252/msb.20178124 [PubMed: 29925568]

Gabasova E, Reid J and Wernisch L (2017). Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. PLoS Comput. Biol 13 e1005781. 10.1371/journal.pcbi.1005781 [PubMed: 29036190]

Gavish M and Donoho DL (2017). Optimal shrinkage of singular values. IEEE Trans. Inf. Theory 63 2137–2152. MR3626861 10.1109/TIT.2017.2653801

Gaynanova I and Li G (2019). Structural learning and integrative decomposition of multi-view data. Biometrics 75 1121–1132. MR4041816 10.1111/biom.13108 [PubMed: 31254385]

Hellton KH and Thoresen M (2016). Integrative clustering of high-dimensional data with joint and individual clusters. Biostatistics 17 537–548. MR3603952 10.1093/biostatistics/kxw005 [PubMed: 26917056]

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, NIU B, McLellan MD et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158 929–944. 10.1016/j.cell.2014.06.049 [PubMed: 25109877]

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 173 291–304. [PubMed: 29625048]

Huo Z and Tseng G (2017). Integrative sparse K-means with overlapping group lasso in genomic applications for disease subtype discovery. Ann. Appl. Stat 11 1011–1039. MR3693556 10.1214/17-AOAS1033 [PubMed: 28959370]

Hutter C and Zenklusen JC (2018). The Cancer Genome Atlas: Creating lasting value beyond its data. Cell 173 283–285. 10.1016/j.cell.2018.03.042 [PubMed: 29625045]

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature 502 333–339. 10.1038/nature12634 [PubMed: 24132290]

Kaplan A and Lock EF (2017). Prediction with dimension reduction of multiple molecular data sources for patient survival. Cancer Inform. 16 1–11. [PubMed: 28096648]

Kurucz M, Benczúr AA and Csalogány K (2007). Methods for large scale svd with missing values. In Proceedings of KDD Cup and Workshop 12 31–38.

Li G and Gaynanova I (2018). A general framework for association analysis of heterogeneous data. Ann. Appl. Stat 12 1700–1726. MR3852694 10.1214/17-AOAS1127

Li G and Jung S (2017). Incorporating covariates into integrated factor analysis of multi-view data. Biometrics 73 1433–1442. MR3744555 10.1111/biom.12698 [PubMed: 28407218]

Lock EF and Dunson DB (2013). Bayesian consensus clustering. Bioinformatics 29 2610–2616. [PubMed: 23990412]

Lock EF, Park JY and Hoadley KA (2022). Supplement to "Bidimensional linked matrix factorization for pan-omics pan-cancer analysis." 10.1214/21-AOAS1495SUPPA, 10.1214/21-AOAS1495SUPPB, 10.1214/21-AOAS1495SUPPC

Lock EF, Hoadley KA, Marron JS and Nobel AB (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Ann. Appl. Stat 7 523–542. MR3086429 10.1214/12-AOAS597 [PubMed: 23745156]

Mazumder R, Hastie T and Tibshirani R (2010). Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res 11 2287–2322. MR2719857 [PubMed: 21552465]

Mo Q, Shen R, Guo C, Vannucci M, Chan KS and Hilsenbeck SG (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics 19 71–86. MR3799604 10.1093/biostatistics/kxx017 [PubMed: 28541380]

Network TR (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N. Engl. J. Med 372 2481–2498. [PubMed: 26061751]

Network TR et al. (2012). Comprehensive molecular portraits of human breast tumors. Nature 490 61. [PubMed: 23000897]

Network TR et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature 511 543. [PubMed: 25079552]

Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP et al. (2010). Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 17 510–522. [PubMed: 20399149]

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 75 582–592. MR3999181 10.1111/biom.13010 [PubMed: 30516272]

Park JY and Lock EF (2020). Integrative factorization of bidimensionally linked matrices. Biometrics 76 61–74. MR4098544 10.1111/biom.13141 [PubMed: 31444786]

Rudelson M and Vershynin R (2010). Non-asymptotic theory of random matrices: Extreme singular values. In Proceedings of the International Congress of Mathematicians. Volume III 1576–1602. Hindustan Book Agency, New Delhi. MR2827856

Shabalin AA and Nobel AB (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. J. Multivariate Anal 118 67–76. MR3054091 10.1016/jjmva.2013.03.005

Shen R, Wang S and Mo Q (2013). Sparse integrative clustering of multiple omics data sets. Ann. Appl. Stat 7 269–294. MR3086419 10.1214/12-AOAS578 [PubMed: 24587839]

Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, GOLUB T et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17 98–110. 10.1016/j.ccr.2009.12.020 [PubMed: 20129251]

Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM et al. (2013). The cancer genome atlas pan-cancer analysis project. Nat. Genet 45 1113–1120. [PubMed: 24071849]

Yang Z and Michailidis G (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. Bioinformatics 32 1–8. 10.1093/bioinformatics/btv544 [PubMed: 26377073]

Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhang C-Z, Wala J et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet 45 1134–1140. [PubMed: 24071852]

Zhang S, Wang W, Ford J, Makedon F and Pearlman J (2005). Using singular value decomposition approximation for collaborative filtering. In Seventh IEEE International Conference on E-Commerce Technology (CEC '05) 257–264. IEEE.

Zhu H, Li G and Lock EF (2020). Generalized integrative principal component analysis for multi-type data with block-wise missing structure. Biostatistics 21 302–318. MR4133362 10.1093/biostatistics/kxy052 [PubMed: 30247540]
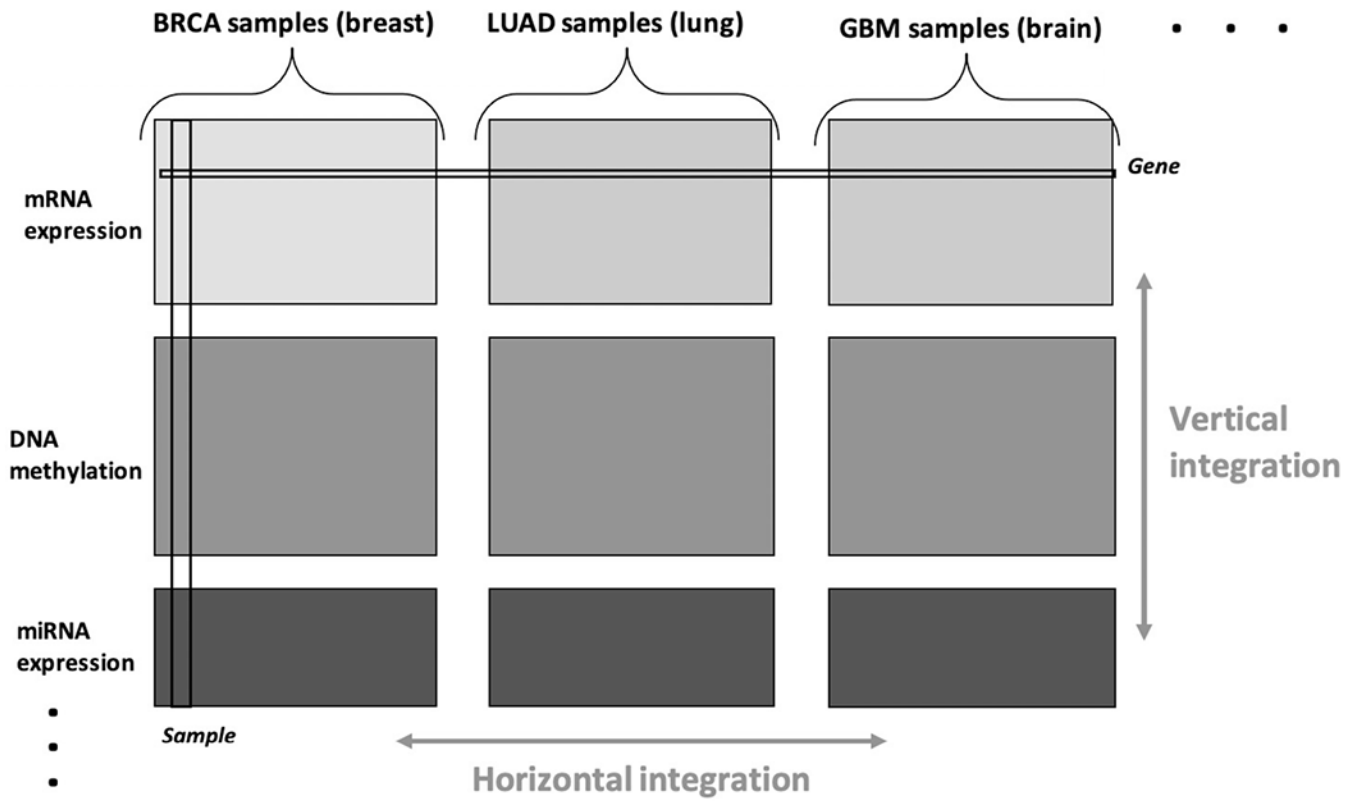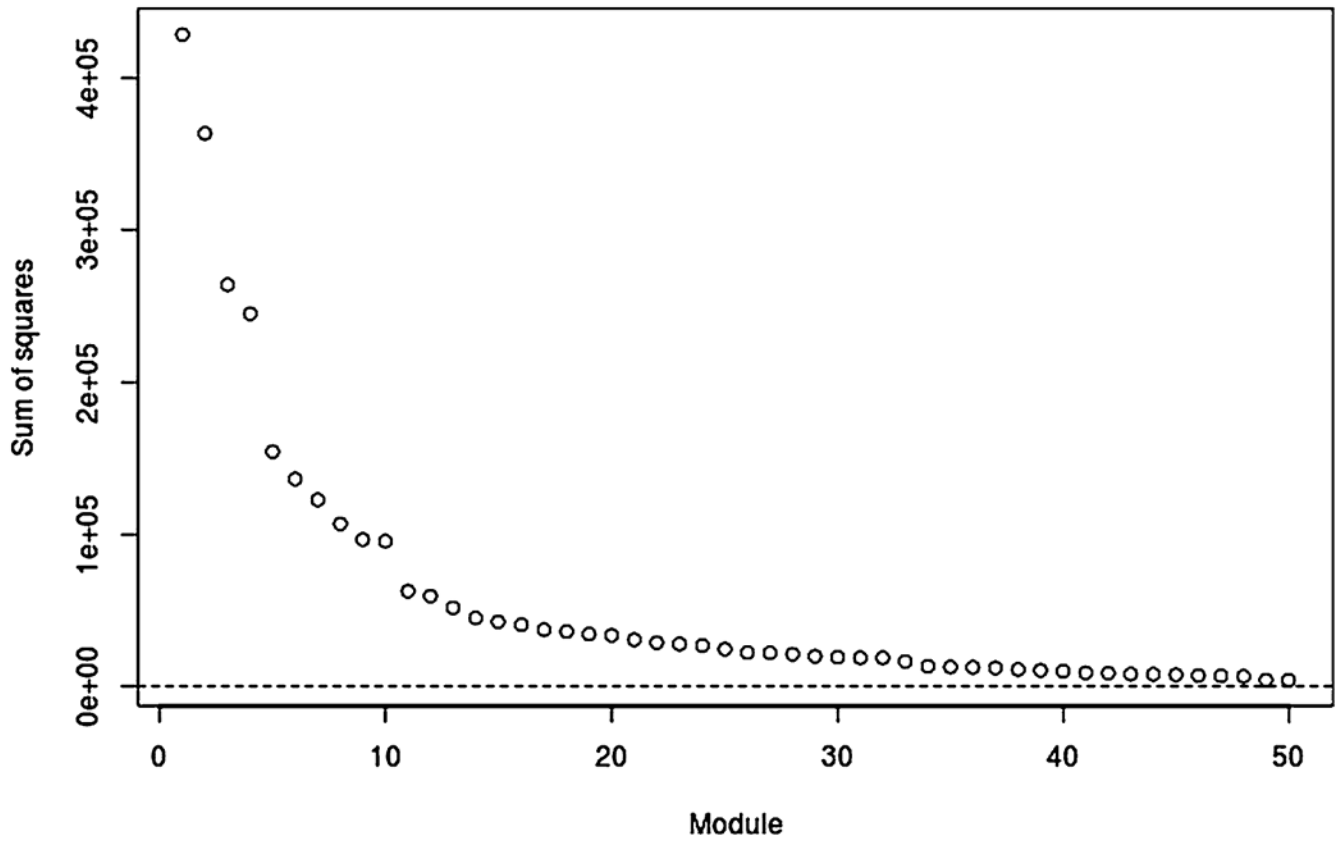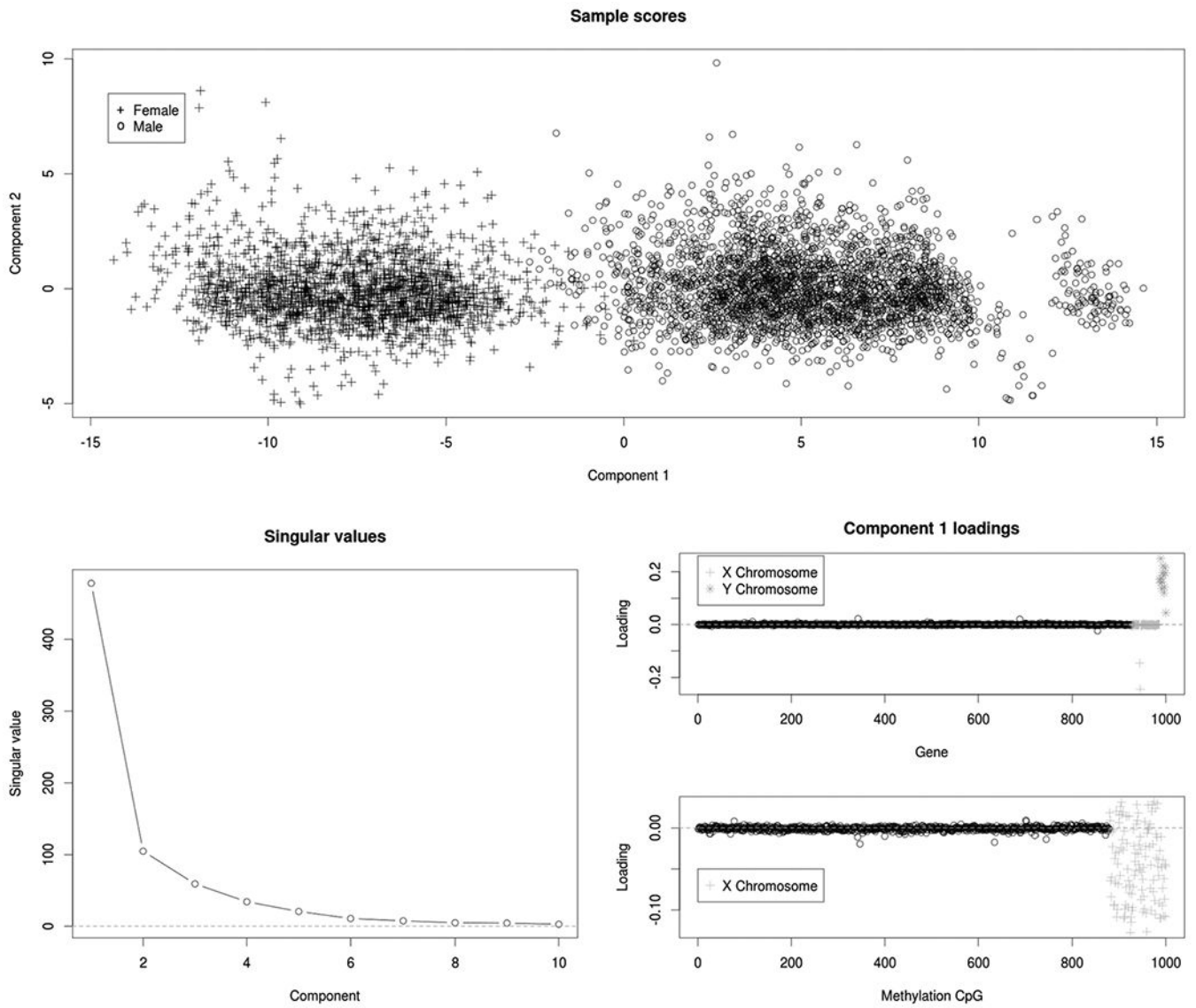
Fig. 1.
Bidimensional integration of pan-omics pan-cancer data.

Fig. 2.
Total sum of squared entries in each of the 50 modules, ordered from largest to smallest.

Fig. 3.

For Module 4: Sample scores for the first two components (top), scree plot of singular values (bottom left), and loadings on genes and methylation CpGs (bottom right) for the first component. This module includes 22 cancer types with samples from both sexes, and it is dominated by molecular signals that distinguish males from females.
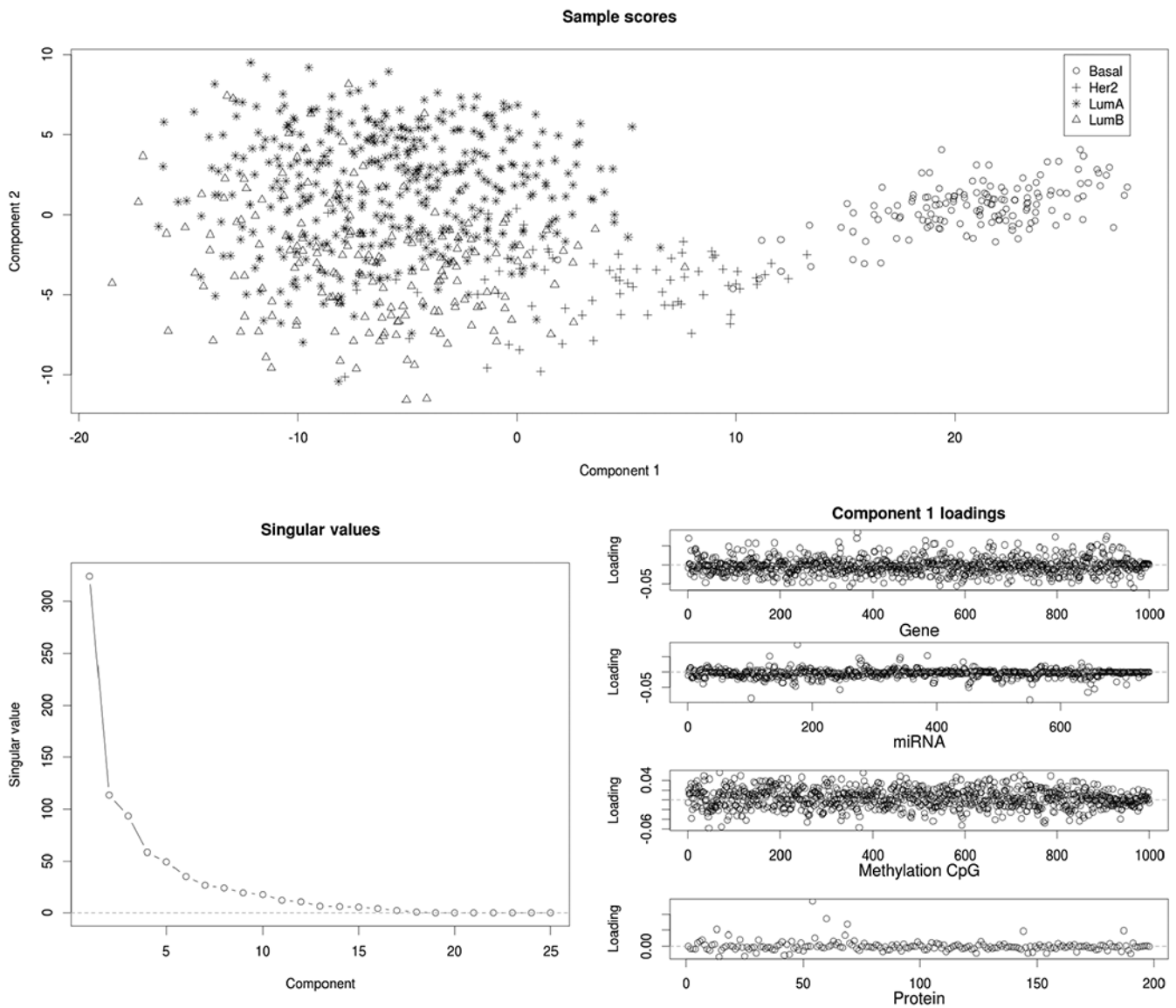
Fig. 4.

For Module 6: Sample scores for the first two components (top), scree plot of singular values (bottom left), and loadings for all four omics platforms (bottom right). This module includes only breast (BRCA) tumor samples, and it is dominated by molecular signals that distinguish the PAM50 subtypes (Bsasal, Her2, LumA, and LumB).

Fig. 5.
Scores for the first two components of Module 12 (*LGG, methylation*), *with symbols and colors showing separation by IDH mutation status* (*wild-type, IDH mutant, and IDH mutant with codeletion*).

Table 1

TCGA acronyms for the 29 different cancer types considered

Acronym	Cancer type	Acronym	Cancer type
ACC	Adrenocortical carcinoma	BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma	CESC	Cervical carcinoma
CHOL	Cholangiocarcinoma	CORE	Colorectal adenocarcinoma
DLBC	Diffuse large B-cell lymphoma	ESCA	Esophageal carcinoma
HNSC	Head/neck squamous cell	KICH	Kidney chromophobe
KIRC	Kidney renal clear cell	KIRP	Kidney renal papillary cell
LGG	Brain lower grade glioma	LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma	LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma	OV	Ovarian cancer
PAAD	Pancreatic adenocarcinoma	PCPG	Pheochromocytoma and paraganglioma
PRAD	Prostate adenocarcinoma	SARC	Sarcoma
SKCM	Skin cutaneous melanoma	STAD	Stomach adenocarcinoma
TGCT	Testicular germ cell tumors	THCA	Thyroid carcinoma
THYM	Thymoma	UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma		

Table 2

Cancer types and sources for the first 15 modules, ordered by variation explained

Module	Rank	Cancer types	Omics sources
1	37	All cancers	mRNA miRNA Meth Protein
2	25	All cancers	miRNA
3	22	BLCA BRCA CESC CHOL CORE DLBC ESCA HNSC LIHC LUAD LUSC OV PAAD PRAD SKCM STAD TGCT UCEC UCS	Meth
4	10	ACC BLCA CHOL CORE DLBC ESCA HNSC KICH KIRC KIRP LGG LIHC LUAD LUSC MESO PAAD PCPG SARC SKCM STAD THCA THYM	mRNA Meth
5	24	All cancers	mRNA
6	17	BRCA	mRNA miRNA Meth Protein
7	15	LGG	mRNA miRNA Protein
8	20	All cancers *but* LGG	Protein
9	15	THCA	mRNA miRNA Protein
10	20	All cancers *but* LGG and TGCT	miRNA
11	15	CHOL KIRC KIRP LIHC	mRNA miRNA Meth Protein
12	34	LGG	Meth
13	20	BLCA CESC CORE ESCA HNSC LUSC SARC STAD	mRNA miRNA Meth Protein
14	8	KICH KIRC KIRP	mRNA miRNA Protein
15	21	BLCA BRCA CESC CHOL ESCA HNSC LUAD LUSC PAAD PRAD SKCM STAD TGCT UCEC UCS	mRNA miRNA

Table 3

Imputation RSE under different approaches and different types of missingness

Method	Observed	Entrywise	Row	Column	Both
BIDIFAC+	0.510	0.558	0.670	0.807	0.881
Soft-SVD (joint)	0.531	0.621	0.678	0.834	0.894
Soft-SVD (separate)	0.564	0.610	1.000	1.000	1.000
Hard-SVD (joint)	0.431	0.559	0.829	0.908	1.200
Hard-SVD (separate)	0.344	0.581	1.000	1.000	1.000

Table 4

Comparison of BIDIFAC+ and SLIDE signal decomposition RSE ($I = 3$ sources)

Scenario	Structure	BIDIFAC+		SLIDE	
		True R	**Estimated R**	**True ranks**	**Estimated ranks**
Rank = 1, s2n = 0.5	Global	0.130	0.130	0.120	0.120
	Pairwise	0.157	0.156	0.103	0.103
	Individual	0.197	0.197	0.118	0.118
Rank = 1, s2n = 1	Global	0.060	0.060	0.084	0.084
	Pairwise	0.068	0.068	0.053	0.053
	Individual	0.070	0.070	0.048	0.048
Rank = 1, s2n = 0	Global	0.010	0.010	0.035	3.65
	Pairwise	0.005	0.005	0.027	1.00
	Individual	0.008	0.008	0.037	0.689
Rank = 5, s2n = 0.5	Global	0.270	0.270	0.276	0.869
	Pairwise	0.268	0.268	0.263	0.460
	Individual	0.329	0.329	0.306	0.317
Rank = 5, s2n = 1	Global	0.123	0.123	0.232	1.320
	Pairwise	0.121	0.121	0.189	0.674
	Individual	0.148	0.148	0.241	0.485
Rank = 5, s2n = 10	Global	0.080	0.080	0.233	2.36
	Pairwise	0.060	0.060	0.189	0.917
	Individual	0.089	0.089	0.249	0.703

Table 5

Comparison of BIDFAC+ and SLIDE signal decomposition RSE ($I = 10$ sources)

Ranks	s2n	BIDIFAC+		SLIDE	
		True R	Estimated R	True ranks	Estimated ranks
1	0.5	0.150	0.150	0.116	0.116
1	1	0.076	0.078	0.105	0.105
1	10	0.032	0.025	0.060	0.060
5	0.5	0.297	0.320	0.402	0.685
5	1	0.177	0.189	0.324	0.603
5	10	0.167	0.245	0.347	0.347

Table 6

Imputation RSE under different approaches, for different types of missingness and joint and individual signal strengths (σ_j^2 and σ_i^2, respectively)

	Entry missing			Column missing			Block missing		
Joint signal (σ_j^2)	$\sqrt{10}$	**1**	**1**	$\sqrt{10}$	**1**	**1**	$\sqrt{10}$	**1**	**1**
Individual signal (σ_i^2)	**1**	**1**	$\sqrt{10}$	**1**	**1**	$\sqrt{10}$	**1**	**1**	$\sqrt{10}$
BIDIFAC+	0.01	0.06	0.01	0.11	0.56	0.93	0.16	0.57	0.94
BIDIFAC+ (random init)	0.01	0.06	0.01	0.11	0.56	0.93	0.16	0.57	0.94
BIDIFAC+ (true **R**, **C**)	0.01	0.06	0.01	0.10	0.55	0.92	0.10	0.52	0.93
Soft-SVD (joint)	0.03	0.15	0.04	0.11	0.58	0.94	0.10	0.53	0.93
Soft-SVD (separate)	0.02	0.09	0.02	1.00	1.00	1.00	1.00	1.00	1.00
Hard-SVD (joint)	0.01	0.06	0.01	0.99	2.02	0.99	1.26	2.15	1.04
Hard-SVD (separate)	0.01	0.03	0.01	1.00	1.00	1.00	1.00	1.00	1.00

Table 7

Relative squared error of the decomposition (*RSE*) *and relative overall signal recovery* (*ROSR*) *using BIDIFAC+ with known modules* ($\mathbf{R}$ *and* $\mathbf{C}$) *and estimated modules* ($\hat{\mathbf{R}}$ *and* $\hat{\mathbf{C}}$)

s2n	RSE(R, C)	RSE ($\tilde{\mathbf{R}}, \tilde{\mathbf{C}}$)	ROSR (R, C)	ROSR ($\tilde{\mathbf{R}}, \tilde{\mathbf{C}}$)
0.2	0.356	0.531	0.170	0.189
0.5	0.242	0.386	0.131	0.143
5	0.128	0.346	0.012	0.026