Biometrics 000, 000–000 DOI: 000
000 0000

Multiple Augmented Reduced Rank Regression for Pan-Cancer Analysis

Jiuzhou Wang, Eric F. Lock *

Division of Biostatistics, University of Minnesota, Minneapolis, MN, USA,

**email:* elock@umn.edu

Summary:

Statistical approaches that successfully combine multiple datasets are more powerful, efficient, and scientifically informative than separate analyses. To address variation architectures correctly and comprehensively for high-dimensional data across multiple sample sets (i.e., cohorts), we propose multiple augmented reduced rank regression (maRRR), a flexible matrix regression and factorization method to concurrently learn both covariate-driven and auxiliary structured variation. We consider a structured nuclear norm objective that is motivated by random matrix theory, in which the regression or factorization terms may be shared or specific to any number of cohorts. Our framework subsumes several existing methods, such as reduced rank regression and unsupervised matrix factorization, and includes a promising novel approach to regression and factorization of a single dataset (aRRR) as a special case. Simulations demonstrate substantial gains in power from combining multiple datasets, and from parsimoniously accounting for all structured variation. We apply maRRR to gene expression data from multiple cancer types (i.e., pan-cancer) from TCGA, with somatic mutations as covariates. The method performs well with respect to prediction and imputation of held-out data, and provides new insights into mutation-driven and auxiliary variation that is shared or specific to certain cancer types.

Key words: cancer genomics, data integration, low rank matrix factorization, missing data imputation, nuclear norm, reduced rank regression.

1. Introduction

The dramatic proliferation of omics data in biomedicine and genomics research has allowed for increasingly comprehensive investigations that span multiple distinct sample sets and multiple molecular facets. Statistical approaches that successfully combine multiple datasets within a single analytical framework are more powerful, efficient, and scientifically informative than separate analyses. This has spurred several recent advances in methodology for high-dimensional data integration, however, there remain unmet needs especially for the multi-cohort context (data on the same features from different groups of samples). For example, the Cancer Genome Atlas (TCGA) program has collected over 10,000 tumor samples from individuals from 33 cohorts corresponding to different cancer types, and detecting signals across cohorts on multiple genomic levels is of interest (Hutter and Zenklusen, 2018).

Several unsupervised multi-matrix factorization methods provide low-rank representations of underlying structure. The singular value decomposition (SVD), principle component analysis (PCA) and other well-known approaches allow a rank r approximation of a single matrix $\mathbf{X}_{m\times n} \approx \mathbf{U}_{m\times r}\mathbf{V}^T_{n\times r}, r < \min(m, n)$. Loadings $\mathbf{U}$ and scores $\mathbf{V}$ explain variation in the rows or columns, respectively. The joint and individual variation explained (JIVE) method extends PCA to multiple datasets with shared columns $\{\mathbf{X}_1, \ldots, \mathbf{X}_J\}$ via $\mathbf{X}_i \approx \mathbf{U}_i\mathbf{V}^T + \mathbf{W}_i\mathbf{V}_i^T$. Here the joint scores $\mathbf{V}$ capture shared structure among the datasets, and the individual scores $\mathbf{V}_i$ capture structure speficic to dataset i. Numerous relevant approaches, such as AJIVE (Feng et al., 2018) and SLIDE (Gaynanova and Li, 2019), have been proposed to factorize multiple data from other perspectives. Moreover, BIDIFAC+ (Lock et al., 2022) enables a more flexible way to identify multiple shared and specific modules of variation, which may be partially shared over row subset or column subsets. However, these unsupervised methods suffer from neglecting covariate information. Other supervised techniques (Safo et al., 2022; Wang and Safo, 2021; Zhang and Gaynanova, 2021) identify structures across multiple datasets that is relevant to predicting an outcome, but they do not capture both covariate-driven and auxiliary structures.

To impose low-rank covariate effects, different types of penalties have been introduced in the the multivariate least square regression framework. Reduced rank regression (RRR) (Izenman, 1975) is a popular approach to predict $\mathbf{X} : p \times n$ from $\mathbf{Y} : q \times n$ via least squares in which the coefficients have low-rank, $\mathbf{X} \approx \mathbf{B}$ with $\text{rank}(\mathbf{B}) < \min(p, q)$. Rank penalized (RSC) (Bunea et al., 2011) and nuclear-norm penalized (NNP) least square criteria (Yuan et al., 2007) are widely used alternatives with penalties that enforce low-rank coefficients. Combining RRR with adaptive NNP (Chen et al., 2013) shows a better performance than RSC. Integrative RRR (Li et al., 2019) extends the estimation to multiple covariate sets all at once. Nonetheless, those regression methods have two limitations: (1) they do not allow for potentially unique covariate-driven signals across multiple sample cohorts and (2) they do not account for additional low-rank structure unrelated to the covariates.

Missing values, especially in genomics data, are prevalent due to cost limitations or other technical issues. The data usually has three types of missing: entry missing, column missing and row missing. To impute missing values with global structures, matrix factorization based approaches, such as SVDImpute (Troyanskaya et al., 2001) and SoftImpute (Mazumder et al., 2010) are popular since they are effective and straightforward to implement. The aforementioned methods can be potentially used to impute missing data as well. However, they will suffer from the same limitations as we described above.

Unifying reduced rank regression and unsupervised low-rank factorization using the nuclear norm penalty, we develop the multiple augmented reduced rank regression (maRRR) method for multi-cohort data that enables a very flexible approach for the simultaneous identification of covariate-driven effects and auxiliary structured variation. These covariate effects and augmented structures may be shared across any cohorts via a general objective function.

This novel low-rank regression and factorization method can be used to impute various types of missing data, accurately capture the relationship between covariates and high-dimensional outcomes, and explore covariate-related and covariate-unrelated patterns of variation that are shared across or specific to different cohorts.

Due to space limitations for this submission, proofs are omitted from the main article and are available in Appendix A.

2. Proposed Model

Let $\mathbf{X}_j : p \times n_j$ denote data matrices with accompanying covariates $\mathbf{Y}_j : q \times n_j$ for j sample cohorts $j = 1, ..., J$. Concatenations across the cohorts are denoted by $\cdot$, e.g., $\mathbf{X}_\cdot = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_J]$ and $\mathbf{Y}_\cdot = [\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_J]$. For our application, we consider gene expression data $\mathbf{X}_\cdot$ and somatic mutations $\mathbf{Y}_\cdot$ for several patients across $J = 30$ cancer types. We are interested in estimating low-rank coefficient matrices $\mathbf{B}_k : p \times q, k = 1, ..., K$ and auxiliary variation structures $\mathbf{S}_j^{(l)} : p \times n_j, l = 1, ..., L, j = 1, ..., J$. Acknowledging the errors $\mathbf{E}_j : p \times n_j, j = 1, ..., J$ for each cohort, the full model can be written as:

$$\mathbf{X}_\cdot = \sum_{k=1}^{K} \mathbf{B}_k \mathbf{Y}_\cdot^{(k)} + \sum_{l=1}^{L} \mathbf{S}_\cdot^{(l)} + \mathbf{E}_\cdot \tag{1}$$

where

$$\mathbf{Y}_\cdot^{(k)} = [\mathbf{Y}_1^{(k)}, \mathbf{Y}_2^{(k)}, ..., \mathbf{Y}_J^{(k)}], \mathbf{S}_\cdot^{(l)} = [\mathbf{S}_1^{(l)}, \mathbf{S}_2^{(l)}, ..., \mathbf{S}_J^{(l)}], \mathbf{E}_\cdot^{(l)} = [\mathbf{E}_1^{(l)}, \mathbf{E}_2^{(l)}, ..., \mathbf{E}_J^{(l)}].$$

The presence of each $\mathbf{Y}_j^{(k)}$ or $\mathbf{S}_j^{(l)}$ across the cohorts are determined by binary indicator matrices $\mathbf{C}_Y : J \times K$ and $\mathbf{C}_S : J \times L$ respectively:

$$\mathbf{Y}_j^{(k)} = \begin{cases} 0_{q \times n_j} & \text{if } \mathbf{C}_Y[j,k] = 0 \\ \mathbf{Y}_j & \text{if } \mathbf{C}_Y[j,k] = 1 \end{cases}, \qquad \mathbf{S}_j^{(l)} = \begin{cases} 0_{p \times n_j} & \text{if } \mathbf{C}_S[j,l] = 0 \\ \mathbf{U}_S^{(l)} \mathbf{V}_{Sj}^{(l)T} & \text{if } \mathbf{C}_S[j,l] = 1. \end{cases}$$

We refer to each $\mathbf{B}_k Y^{(k)}$ and $\mathbf{S}^{(l)}$ as a module. Each $\mathbf{S}^{(l)}$ gives a low-rank module that explains covariate-unrelated structured variability within the cancer types identified by

$\mathbf{C}_S[:,l]$. Each $\mathbf{B}_k\mathbf{Y}^{(k)}$ gives another low-rank module for covariate-driven structure for the cancer type identified by $\mathbf{C}_Y[:,l]$. Each module is assumed to be low-rank, meaning it can be factorized as the product of a small number of row and column vectors, $\mathbf{B}_k = \mathbf{U}_B^{(k)}\mathbf{V}_B^{(k)T}$ and $\mathbf{S}_k = \mathbf{U}_S^{(l)}\mathbf{V}_S^{(l)T}$.

3. Objective Function

To estimate model (1) and impose low-rank structures, we minimize the following least square criterion with a structured nuclear norm penalty:

$$\min_{\{\mathbf{B}_k\}_{k=1}^K,\{\mathbf{S}_\cdot^{(l)}\}_{l=1}^L} \{\frac{1}{2}||\mathbf{X}_\cdot - \sum_{k=1}^K \mathbf{B}_k\mathbf{Y}_\cdot^{(k)} - \sum_{l=1}^L \mathbf{S}_\cdot^{(l)}||_F^2 + \sum_{k=1}^K \lambda_B^{(k)}||\mathbf{B}_k||_* + \sum_{l=1}^L \lambda_S^{(l)}||\mathbf{S}_\cdot^{(l)}||_*\} \tag{2}$$

Here $||\cdot||_*$ denotes the nuclear norm, i.e., the sum of the singular values of the matrix, a convex penalty which encourages a low-rank solution.

There are three special cases of the general objective functions worth noticing. The first two are novel and the last is existing, listed as follows:

(1) Augmented reduced rank regression (aRRR), our proposed approach minimizing $K = L = J = 1$. The reduced rank regression model is "augmented" to account for auxiliary structured variation $\mathbf{S}$ simultaneously.

(2) Multi-cohort reduced rank regression (mRRR), our proposed approach minimizing $L = 0$, i.e. no auxiliary terms $\mathbf{S}$. The reduced rank regression is extended to recover multiple (shared or individual) covariate effects at once.

(3) Optimizing this objective with no covariate-driven structure ($K = 0$) corresponds to the BIDIFAC+ method (Lock et al., 2022) with horizontal structures only.

Mazumder et al. (2010) and others have noted the equivalence of the nuclear norm penalty and an additive L_2 penalty on the terms in the low-rank factorization, and this leads to an alternative form of our objective (2),

$$\min_{\{\mathbf{U}_B^{(k)},\mathbf{V}_B^{(k)}\}_{k=1}^K,\{\mathbf{U}_S^{(l)},\mathbf{V}_S^{(l)}\}_{l=1}^L} \frac{1}{2}\{||\mathbf{X}_{\cdot} - \sum_{k=1}^{K}\mathbf{U}_B^{(k)}\mathbf{V}_B^{(k)T}\mathbf{Y}_{\cdot}^{(k)} - \sum_{l=1}^{L}\mathbf{U}_S^{(l)}\mathbf{V}_S^{(l)T}||_F^2 + \sum_{k=1}^{K}\lambda_B^{(k)}(||\mathbf{U}_B^{(k)}||_F^2 + ||\mathbf{V}_B^{(k)}||_F^2) + \sum_{l=1}^{L}\lambda_S^{(l)}(||\mathbf{U}_S^{(l)}||_F^2 + ||\mathbf{V}_S^{(l)}||_F^2)\} \quad (3)$$

where we only need to set a general upper bound for the estimated rank of each $\mathbf{B}$ and $\mathbf{S}$, i.e. $r_{B,upper}$ and $r_{S,upper}$, as the number of columns of each $\mathbf{U}$ and $\mathbf{V}$. The actual ranks of the solution may be smaller, as it will correspond to that for the structured nuclear norm penalty (2).

Theorem 1. *If both (2) and (3) have the same penalty terms $\lambda_B^{(k)} > 0, k = 1, ..., K$ and $\lambda_S^{(l)} > 0, l = 1, ..., L$, the solutions to the objective functions coincide.*

In what follows we describe a random matrix theory approach to automatically select the nuclear norm penalty weights λ for the different modules.

4. Theoretical results

First we describe conditions on the penalty parameters needed to avoid degenerate cases in which certain modules are guaranteed to be zero in the solution (regardless of the data $\mathbf{X}_{\cdot}$ and $\mathbf{Y}_{\cdot}$) in Proposition 1.

Proposition 1. *The following conditions are needed for non-zero estimation in $\mathbf{B}, \mathbf{S}$:*

(1) Let $\mathcal{I}_k \subset \{1, ..., k-1, k+1, ..., K\}$ be any subset of $\mathbf{Y}$ modules for which the non-zero blocks of $\{\mathbf{Y}_{\cdot}^{(i)}\}_{i\in\mathcal{I}_k}$ cover exactly those of $\mathbf{Y}_k$, i.e. $\sum_{i\in\mathcal{I}_k}\mathbf{C}_Y[\cdot, i] = c_y \cdot \mathbf{C}_y[\cdot, k]$ for some positive integer c_y, then $\lambda_B^{(k)} < \frac{1}{c_y}\sum_{i\in\mathcal{I}_k}\lambda_B^{(i)}$. Note there may be multiple $\mathcal{I}_k$'s in one objective function.

(2) Let $\mathcal{I}_k \subset \{1, 2, ..., L\}$ be any subset of $\mathbf{S}$ modules for which the non-zero blocks of $\{\mathbf{S}_{\cdot}^{(i)}\}_{i\in\mathcal{I}_k}$ cover exactly those of $\mathbf{Y}_k$, i.e. $\sum_{i\in\mathcal{I}_k}\mathbf{C}_S[\cdot, i] = c_{sy} \cdot \mathbf{C}_Y[\cdot, k]$ for some positive integer c_{sy}, then $\lambda_B^{(k)} < \frac{1}{c_{sy}}\sum_{i\in\mathcal{I}_k}\lambda_S^{(i)}||\mathbf{Y}_{\cdot}^{(k)}||_*$. Note there may be multiple $\mathcal{I}_k$'s in one objective function.

(3) For $l \neq l'$, if there exists a module $\mathbf{S}^{(l')}_{\cdot}$ is contained in another module $\mathbf{S}^{(l)}_{\cdot}$, i.e. $\mathbf{C}_S[j,l] \geqslant \mathbf{C}_S[j,l'], \forall j$, then $\lambda_S^{(l')} < \lambda_S^{(l)}$.

(4) Let $\mathcal{I}_l \subset \{1,...,l-1,l+1,...,L\}$ be any subset of $\mathbf{S}$ modules that the non-zero blocks of $\{\mathbf{S}^{(i)}_{\cdot}\}_{i\in\mathcal{I}_l}$ cover exactly those of $\mathbf{S}^{(l)}_{\cdot}$, i.e. $\sum_{i\in\mathcal{I}_l}\mathbf{C}_S[\cdot,i] = c_s \cdot \mathbf{C}_S[\cdot,l]$ for some positive integer c_s, then $\lambda_S^{(l)} < \frac{1}{c_s}\sum_{i\in\mathcal{I}_l}\lambda_S^{(i)}$. Note there may be multiple $\mathcal{I}_l$'s in one objective function.

To motivate a random matrix theory approach to select the tuning parameters, we present two results establishing the connection between the nuclear norm penalty and singular value thresholding. Lemma 1 is a well-known result for the unsupervised case (Cai et al., 2010), and in Proposition 2 we extend it to the regression context.

Lemma 1. *(Cai et al., 2010) Let $\mathbf{UDV}^T$ be the SVD of a matrix $\mathbf{X}$. The solution to*

$$\min_{\mathbf{S}}\{\frac{1}{2}||\mathbf{X}-\mathbf{S}||_F^2 + \lambda||\mathbf{S}||_*\}$$

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

Proposition 2. *Let $\mathbf{Y}$ be a semi-orthogonal matrix such that $\mathbf{YY}^T = \mathbf{I}$ and $\mathbf{UDV}^T$ be the SVD of a matrix $\mathbf{XY}^T$. The solution to both of the following objectives:*

$$\min_{\mathbf{B}}\{\frac{1}{2}||\mathbf{X}-\mathbf{BY}||_F^2 + \lambda||\mathbf{B}||_*\} \;\; \text{and} \;\; \min_{\mathbf{B}}\{\frac{1}{2}||\mathbf{X}-\mathbf{BY}||_F^2 + \lambda||\mathbf{BY}||_*\},$$

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

While the relative merits of penalizing $||\mathbf{B}||_*$ or $||\mathbf{BY}||_*$ has been debated (Yuan et al., 2007; Chen et al., 2013), Proposition 2 shows they are identical if $\mathbf{Y}$ is semi-orthogonal. In practice, we orthogonalize the columns of $\mathbf{Y}$ prior to estimation. However, this requires that the number of features in $\mathbf{Y}$ is less than the sample size (e.g., $q < n$); if $q \geqslant n$ then $\mathbf{Y}$ will be orthogonal and the solution degenerates to the unsupervised case, which we establish in Proposition 3.

Proposition 3. *For any orthogonal matrix $\mathbf{Y}$, if the optimization problems $\min_{\mathbf{B}}\{\frac{1}{2}||\mathbf{X}-$*

$\mathbf{BY}||_F^2 + \lambda||\mathbf{B}||_*\}$ *and* $\min_{\mathbf{S}}\{\frac{1}{2}||\mathbf{X} - \mathbf{S}||_F^2 + \lambda||\mathbf{S}||_*\}$ *have their optimal solutions as* $\widehat{\mathbf{B}}$ *and* $\widehat{\mathbf{S}}$ *respectively, then* $\widehat{\mathbf{S}} = \widehat{\mathbf{B}}\mathbf{Y}$.

The following propositions describe the distribution of the singular values of a random matrix under general assumptions, which can then be used to motivate tuning parameters.

Proposition 4. *Let* λ_{max} *be the largest singular value of a matrix* $\mathbf{E} : m \times n$ *of independent Guassian entries with mean 0, variance* σ^2 *and finite fourth moment. As* $m, n \to \infty$, *we have* $E(\lambda_{max}) \leqslant \sigma(\sqrt{m} + \sqrt{n})$.

Proposition 5. *Let* $\mathbf{Y}_{q\times n}$ *is semi-orthogonal such that* $\mathbf{YY}^T = \mathbf{I}$. *For integers* $m, q \geqslant 1$ *defined in a way that* $\frac{m}{q} \to c > 0$ *as* $q \to \infty$, *Let* $\mathbf{X}_{m\times n}, \mathbf{B}_{m\times q}, \mathbf{E}_{m\times n}$ *be three matrices such that* $\mathbf{X} = \mathbf{BY}_{q\times n} + \frac{1}{\sqrt{q}}\mathbf{E}$, *where entries of* $\mathbf{E}$ *are independent Guassian with mean 0 and variance* σ^2. *Assume* $rank(\mathbf{B}) = r$. *Denote the singular values of* $\mathbf{B}$ *and* $\mathbf{XY}^T$ *are* $\sigma_1(\mathbf{B}) \geqslant ... \geqslant \sigma_r(\mathbf{B}) > 0$ *and* $\sigma_1(\mathbf{XY}^T) \geqslant ... \geqslant \sigma_r(\mathbf{XY}^T) > 0$ *respectively. As* $n \to \infty$, *for* $1 \leqslant j \leqslant r$,

$$\sigma_j(\mathbf{XY}^T) \xrightarrow{P} \begin{cases} s(\sigma_j(\mathbf{B})) > 1 + \sqrt{c}, & \textit{if } \sigma_j(\mathbf{B}) > \sqrt[4]{c} \\ 1 + \sqrt{c}, & \textit{if } \sigma_j(\mathbf{B}) \leqslant \sqrt[4]{c} \end{cases}$$

where $s(\cdot)$ *is a known function. In particular, when* $\mathbf{Y}$ *is an identity matrix* ($q = n$) *and* $\mathbf{X} = \mathbf{B} + \frac{1}{\sqrt{n}}\mathbf{E}$, *it follows that*

$$\sigma_j(\mathbf{X}) \xrightarrow{P} \begin{cases} s(\sigma_j(\mathbf{B})) > 1 + \sqrt{c}, & \textit{if } \sigma_j(\mathbf{B}) > \sqrt[4]{c} \\ 1 + \sqrt{c}, & \textit{if } \sigma_j(\mathbf{B}) \leqslant \sqrt[4]{c}. \end{cases}$$

Proposition 4 comes directly from (Rudelson and Vershynin, 2010), and Proposition 5 is closely related to the result in (Shabalin and Nobel, 2013). Consider the reasonable penalty for $\mathbf{S}$ in Lemma 1, i.e. $\mathbf{X}_{m\times n} = \mathbf{S}_{m\times n} + \mathbf{E}_{m\times n}$. A set of reasonable tuning parameters will 1) detect the low-rank signals and 2) not capture components that are solely due to noise. Considering Propositions 4 and 5, setting $\lambda = \sigma(\sqrt{m} + \sqrt{n})$ is reasonable because it only keeps the signals (top r components) whose singular values are expected to be greater than those of independent random noise. Consider the reasonable penalty for $\mathbf{B}$ in Proposition 2, i.e. $\mathbf{X}_{m\times n} = \mathbf{B}_{m\times q}\mathbf{Y}_{q\times n} + \mathbf{E}_{m\times n}$. Following a similar argument, we set $\lambda = \sigma(\sqrt{m} + \sqrt{q})$.

In our content we normalize real data $\mathbf{X}_{\cdot}$ to have $\sigma = 1$ for residual error. In order to distinguish true signals $\{\mathbf{B}_k\}_{k=1}^{K}, \{\mathbf{S}_{\cdot}^{(l)}\}_{l=1}^{L}$ from Guassian noises in the objective (2), we sequentially optimize $\mathbf{B}_k$ or $\mathbf{S}_{\cdot}^{(l)}$ conditional on all other modules with $\lambda_B^{(k)} = \sqrt{p} + \sqrt{q}$ for any module $\mathbf{B}_k, k = 1, ..., K$ and $\lambda_S^{(l)} = \sqrt{p} + \sqrt{\sum_{j=1}^{J} n_j \mathbf{C}_S[j, l]}$ for any module $\mathbf{S}_{\cdot}^{(l)}, l = 1, ..., L$. Note that propositions 3 and 5 require orthogonality in $\mathbf{Y}$, which generally does not hold in practice. However, we can still make use of these results by orthogonalizing the data and then transforming back to their original scale after analysis.

5. Estimation

5.1 *Scaling*

In practice we scale $\mathbf{X}$ and orthogonalize $\mathbf{Y}$ prior to optimization. We first center each row of $\mathbf{X}_{\cdot}$ to have mean 0. In order to satisfy the standard normal noise requirement, we estimate the error variance for $\mathbf{X}_{\cdot}$ by using the median absolute deviation estimator from (Gavish and Donoho, 2017). The estimated variance of $\mathbf{X}_{\cdot}$ is denoted as $\hat{\sigma}^2$. Then, we use $\mathbf{X}_{\cdot}/\hat{\sigma}$ as the final data matrix for optimization. By this, we have the residual variance approximately 1. The details of scaling for $\mathbf{Y}$ are left out in Appendix B.

5.2 *Optimization*

We estimate all regression coefficients $\mathbf{B}$ and extra variation sources $\mathbf{S}$ simultaneously by iterativily solving objective function (3). The introduction of $\mathbf{U}$ and $\mathbf{V}$ makes the optimization algorithm more efficient since the objective function has a closed-form gradient. Given all other estimates, we update every single $\mathbf{U}_B^{(k)}, \mathbf{V}_B^{(k)}, \mathbf{U}_S^{(l)}, \mathbf{V}_S^{(l)}$ by setting its corresponding gradient to be zero. The detailed optimization algorithm is as follows (Algorithm 1):

Algorithm 1 Alternating Least Square with Matrix Decomposition

```
Input: Covariates $\mathbf{Y}$ and corresponding multivariate outcomes $\mathbf{X}$; penalizing terms $\lambda_B, \lambda_S$
Output: $\mathbf{B}, \mathbf{S}$
Initialization Assign initialized numbers for each entry of $\{\mathbf{U}_B^{(k)}, \mathbf{V}_B^{(k)}\}_{k=1}^{K}, \{\mathbf{U}_S^{(l)}, \mathbf{V}_S^{(l)}\}_{l=1}^{L}$
while convergence criterion does not meet do
    for $k = 1, ..., K$ do
        Compute the residual matrix $X_{\cdot}^{(k)} = \mathbf{X}_{\cdot} - \sum_{k'=1,k'\neq k}^{K} \mathbf{U}_B^{(k')}\mathbf{V}_B^{(k')T}\mathbf{Y}_{\cdot}^{(k')} - \sum_{l=1}^{L} \mathbf{U}_S^{(l)}\mathbf{V}_S^{(l)T}$
        Update $\mathbf{U}_B^{(k)} = \mathbf{X}_{\cdot}^{(k)}\mathbf{Y}_{\cdot}^{(k)T}\mathbf{V}_B^{(k)}(\mathbf{V}_B^{(k)T}\mathbf{Y}_{\cdot}^{(k)}\mathbf{Y}_{\cdot}^{(k)T}\mathbf{V}_B^{(k)} + \lambda_B^{(k)}\mathbf{I}_{r_B})^{-1}$
        Update $vec(\mathbf{V}_B^{(k)}) = [(\mathbf{U}_B^{(k)T}\mathbf{U}_B^{(k)})\bigotimes(\mathbf{Y}_{\cdot}^{(k)}\mathbf{Y}_{\cdot}^{(k)T}) + \lambda_B^{(k)}\mathbf{I}_{q*r_B}]^{-1}vec[\mathbf{Y}_{\cdot}^{(k)}(\mathbf{X}_{\cdot}^{(k)T})\mathbf{U}_B^{(k)}]$
        Transform $vec(\mathbf{V}_B^{(k)})$ to $\mathbf{V}_B^{(k)}$
    end for
    for $l = 1, .., L$ do
        Compute the residual matrix $\mathbf{X}_{\cdot}^{(l)} = \mathbf{X}_{\cdot} - \sum_{k=1}^{K} \mathbf{U}_B^{(k)}\mathbf{V}_B^{(k)T}\mathbf{Y}_{\cdot}^{(k)} - \sum_{l'=1,l'\neq l}^{L} \mathbf{U}_S^{(l')}\mathbf{V}_S^{(l')T}$
        Set $\mathbf{X}_j^{(l)} = 0$ where $\mathbf{C}_s[j, l] = 0$ for $j = 1, ..., J$
        Update $\mathbf{U}_S^{(l)} = \mathbf{X}_{\cdot}^{(l)}\mathbf{V}_S^{(l)}(\mathbf{V}_S^{(l)T}\mathbf{V}_S^{(l)} + \lambda_S^{(l)}\mathbf{I}_{r_S})^{-1}$
        Update $\mathbf{V}_S^{(l)} = \mathbf{X}_{\cdot}^{(l)T}\mathbf{U}_S^{(l)}(\mathbf{U}_S^{(l)T}\mathbf{U}_S^{(l)} + \lambda_S^{(l)}\mathbf{I}_{r_S})^{-1}$
    end for
end while
Set $\mathbf{B}_k = \mathbf{U}_B^{(k)}\mathbf{V}_B^{(k)T}$ for all $k = 1, .., K$, and $\mathbf{S}_{\cdot}^{(l)} = \mathbf{U}_S^{(l)}\mathbf{V}_S^{(l)T}$ for all $l = 1, .., L$
```

The symbol $\bigotimes$ means Kronecker product. Note Algorithm 1 does not require the columns of $\mathbf{Y}_{\cdot}^{(k)}$ to be orthogonal. Moreover, when $\mathbf{Y}_{\cdot}^{(k)}$ is semi-orthogonal, in light of Lemma 1 and Proposition 2, we develop soft-singular value estimators without matrix decomposition as (Algorithm 2):

For both algorithms, we use the same convergence criteria to decide whether to stop the optimization process:

$$\sum_{k=1}^{K} ||\widehat{\mathbf{B}}_k - \widetilde{\mathbf{B}}_k||_F^2 + \sum_{l=1}^{L} ||\widehat{\mathbf{S}}_{\cdot}^{(l)} - \widetilde{\mathbf{S}}_{\cdot}^{(l)}||_F^2 < \epsilon$$

where $\widehat{}$ denotes the estimation in the current epoch and $\widetilde{}$ denotes the estimation in the previous epoch. It is also reasonable to use convergence of the loss function as the criteria.

In practice Algorithm 1 and Algorithm 2 have different strengths and weaknesses. Theoretically, Algorithm 2 can be used only when we orthogonalize the original $\mathbf{Y}$, because otherwise soft-thresholding to update $\mathbf{B}$ is not possible. In general, we find that the algorithms require

Algorithm 2 Alternating Least Square with Soft-threshold Estimators

```
Input: Orthogonal covariates $\mathbf{Y}$ and corresponding multivariate outcomes $\mathbf{X}$; penalizing
    terms $\lambda_B, \lambda_S$
Output: $\mathbf{B}, \mathbf{S}$
Initialization Assign initialized numbers for each entry of $\{\mathbf{B}_k\}_{k=1}^K, \{\mathbf{S}_\cdot^{(l)}\}_{l=1}^L$
while convergence criterion does not meet do
    for $k = 1, .., K$ do
        Compute the residual matrix $\mathbf{X}_\cdot^{(k)} = \mathbf{X}_\cdot - \sum_{k'=1,k'\neq k}^K \mathbf{B}_{k'}\mathbf{Y}_\cdot^{(k')} - \sum_{l=1}^L \mathbf{S}_\cdot^{(l)}$
        Compute the SVD of $\mathbf{X}_\cdot^{(k)}\mathbf{Y}_\cdot^{(k)T}$, i.e. $\mathbf{X}_\cdot^{(k)}\mathbf{Y}_\cdot^{(k)T} = L_B^{(k)}\mathbf{D}_B^{(k)}\mathbf{R}_B^{(k)}$
        Update $\mathbf{B}_k = L_B^{(k)}\widehat{\mathbf{D}}_B^{(k)}\mathbf{R}_B^{(k)}$ where $\widehat{\mathbf{D}}_B^{(k)}$ is a diagonal matrix with $\widehat{\mathbf{D}}_B^{(k)}[r,r] =$
    $max(\mathbf{D}_B^{(k)}[r,r] - \lambda_B^{(k)}, 0)$ for $r = 1, 2, ...$ on its diagonal entries and zero otherwise
    end for
    for $l = 1, .., L$ do
        Compute the residual matrix $\mathbf{X}_\cdot^{(l)} = \mathbf{X}_\cdot - \sum_{k=1}^K \mathbf{B}_{k'}\mathbf{Y}_\cdot^{(k')} - \sum_{l=1,l'\neq l}^L \mathbf{S}_\cdot^{(l')}$
        Set $\mathbf{X}_j^{(l)} = 0$ where $\mathbf{C}_s[j,l] = 0$ for $j = 1, ..., J$
        Compute the SVD of $\mathbf{X}_\cdot^{(l)}$, i.e. $\mathbf{X}_\cdot^{(l)} = L_S^{(l)}\mathbf{D}_S^{(l)}\mathbf{R}_S^{(l)}$
        Update $\mathbf{S}_\cdot^{(l)} = L_S^{(l)}\widehat{\mathbf{D}}_S^{(l)}\mathbf{R}_S^{(l)}$ where $\widehat{\mathbf{D}}_S^{(l)}$ is a diagonal matrix with $\widehat{\mathbf{D}}_S^{(l)}[r,r] =$
    $max(\mathbf{D}_S^{(l)}[r,r] - \lambda_S^{(l)}, 0)$ for $r = 1, 2, ...$ on its diagonal entries and zero otherwise
    end for
end while
```

similar computation time to achieve the same convergence criterion: Algorithm 1 tends to require less time if the true ranks (and accompanying maximum ranks specified for the algoithm) are small, while Algorithm 2 is quicker and consumes less computational resources when the true rank and maximum ranks specific for Algorithm 1 are large.

5.3 *Missing data imputation*

One of the main uses of our proposed method is to impute various types of missing data. Based on the assumption that the abundance of existing entries provides sufficient information to uncover the global structures (both covariate and auxiliary effects) and therefore, to estimate the calues of absent entries. Denote the set of indexes of all missing entries as M. Our iterative imputation process is as follows:

(1) Initialize $\widetilde{\mathbf{X}}_\cdot$ by $\widetilde{\mathbf{X}}_\cdot[m,n] = \begin{cases} 0 & \text{if } [m,n] \in M \\ \mathbf{X}_\cdot[m,n] & \text{if } [m,n] \notin M. \end{cases}$

(2) Estimate $\{\mathbf{B}_k\}_{k=1}^K, \{\mathbf{S}_\cdot^{(l)}\}_{l=1}^L$ by Algorithm 1 or Algorithm 2 with current $\widetilde{\mathbf{X}}_\cdot$;

(3) Update $\widetilde{\mathbf{X}}_\cdot$ by setting $\widetilde{\mathbf{X}}_\cdot[m,n] = (\sum_{k=1}^{K} \mathbf{B}_k \mathbf{Y}_\cdot^{(k)} + \sum_{l=1}^{L} \mathbf{S}_\cdot^{(l)})[m,n]$ for all $[m,n] \in M$;

(4) Back to (2) unless convergence. The final $\widetilde{\mathbf{X}}_\cdot$ is the imputation result.

This can be considered a modified EM-algorithm, and is similar to the approach used for softImpute (Mazumder et al., 2010) for nuclear-norm penalized imputation of a single matrix with no covariates.

6. Simulations

6.1 *Recovery of true structure for special cases*

Here, we present simulations as proof-of-concept for two novel scenarios within our approach: (i) simultaneous modeling of covariate effects and auxiliary low-rank variation and (ii) simultaneous modeling of shared or specific covariate effects across multiple cohorts.

For (i), we consider a single data matrix $\mathbf{X} : 100 \times 100$ and single set of covariates $\mathbf{Y} : 10 \times 100$ and generate data via $\mathbf{X} = \mathbf{BY} + \mathbf{S} + \mathbf{E}$, where $\mathbf{BY}$ is covariate-driven variation, $\mathbf{S}$ is auxiliary structured variation, and $\mathbf{E}$ is error. The coefficient array $\mathbf{B}$ has rank R_y via $\mathbf{B} = a\mathbf{U}_B\mathbf{V}_B^T$ where $\mathbf{U}_B : 100 \times R_y$ and $\mathbf{V}_B : 10 \times R_y$, and $\mathbf{S}$ has rank 5 via $\mathbf{S} = b\mathbf{U}_S\mathbf{V}_S$ where $\mathbf{U}_S : 100 \times 5$ and $\mathbf{V}_S : 5 \times 100$. The entries of $\mathbf{E}$, $\mathbf{Y}$, $\mathbf{U}_B$, $\mathbf{V}_B$, $\mathbf{U}_S$ and $\mathbf{V}_S$ are all generated independently from a standard normal distribution. We consider $R_y = 1$ or $R_y = 5$, and consider three conditions with different signal strength for each term by adjusting a and b: sd($\mathbf{BY}$) $= 0.5$ and sd($\mathbf{S}$) $= 5$ ($||\mathbf{BY}||/||\mathbf{S}|| = 0.1$, sd($\mathbf{BY}$) $=$ sd($\mathbf{S}$) $= 1$ ($||\mathbf{BY}||/||\mathbf{S}|| = 1$), and sd($\mathbf{BY}$) $= 5$ and sd($\mathbf{S}$) $= 0.5$ ($||\mathbf{BY}||/||\mathbf{S}|| = 10$). For each set of conditions, we estimate $\mathbf{B}$ and $\mathbf{S}$ using three approaches:

(1) Augmented reduced rank regression (aRRR), our proposed approach as decsribed in Section 3.

(2) Two-stage least squares, in which the coefficients $\mathbf{B}$ are determined by ordinary least squares

regression and $\mathbf{S}$ is determined by an SVD approximation with the true rank ($R = 5$) on the residuals $\mathbf{X} - \hat{\mathbf{B}}\mathbf{Y}$.

(3) Two-stage nuclear norm (NN), in which the coefficients $\mathbf{B}$ are determined by an NN-penalized reduced rank regression and $\mathbf{S}$ is determined by a NN-penalized matrix approximation to the residuals $\mathbf{X} - \hat{\mathbf{B}}\mathbf{Y}$.

For each method, we compute the relative mean squared error (MSE) for $\mathbf{B}$ and $\mathbf{S}$, e.g., $||\mathbf{B} - \hat{\mathbf{B}}||_F^2/||\mathbf{B}||_F^2$. Average relative MSEs for each condition, over 100 replications, are shown in Table 1A. This demonstrates clear advantages of a nuclear norm penalty on $\mathbf{B}$, and the dramatic advantage of aRRR when the auxiliary signal $\mathbf{S}$ is strong. The latter point is critical, because molecular data typically have a large amount of structured variation that is driven by coordinated biological processes or other latent effects; it is common for such variation to be stronger than the signal of interest (i.e., $\mathbf{BY}$), but it is not systematically adjusted for in practice.

For scenario (ii), we generate data $\{\mathbf{X}_j : 100 \times 100, \mathbf{Y}_j : 10 \times 100\}$ via $\mathbf{X}_j = (\mathbf{B} + \mathbf{B}_i)\mathbf{Y} + \mathbf{E}$ for two cohorts $j \in \{1, 2\}$. Here, $\mathbf{B}_j$ are covariate effects specific to cohort j and $\mathbf{B}$ are shared effects. The coefficient arrays are generated via $\mathbf{B} = a\mathbf{U}_B\mathbf{V}_B^T$, $\mathbf{B}_1 = b\mathbf{U}_{B_1}\mathbf{V}_{B_1}^T$, and $\mathbf{B}_2 = b\mathbf{U}_{B_2}\mathbf{V}_{B_2}^T$ where $\{\mathbf{U}_B, \mathbf{U}_{B_1}, \mathbf{U}_{B_2}\}$ are each $100 \times R_y$ and $\{\mathbf{V}, \mathbf{V}_{B_1}, \mathbf{V}_{B_2}\}$ are each $R_y \times 10$. The entries of $\{\mathbf{E}, \mathbf{Y}, \mathbf{U}_B, \mathbf{U}_{B_1}, \mathbf{U}_{B_2}, \mathbf{V}_B, \mathbf{V}_{B_1}, \mathbf{V}_{B_2}\}$ are each generated independently from a standard normal distribution. We consider $R_y = 1$ or 5, and three conditions with different signal strength for each term by adjusting a and b: $a = 2$ and $b = 0.2$ ($||\mathbf{B}||/||\mathbf{B}_i|| = 10$), $a = b = 1$ (($||\mathbf{B}||/||\mathbf{B}_i|| = 1$), and $a = 0.2$ and $b = 2$ ($||\mathbf{B}||/||\mathbf{B}_i|| = 0.1$). For each set of conditions, we estimate $\mathbf{B}$, $\mathbf{B}_1$ and $\mathbf{B}_2$ for $J = 2$ via maRRR with no auxiliary terms $\mathbf{S}$, termed multi-cohort reduced rank regression (mRRR). Table 1B shows average relative MSEs of $\mathbf{B}$ and the $\mathbf{B}_i$'s for mRRR in comparison to two-stage approaches analogous to those described previously. The mRRR approach can effectively recover shared and cohort specific effects, with dramatic improvement over ad-hoc multi-step approaches.

6.2 *Missing data imputation*

The simulation is composed of 1) complete data generation; 2) missingness assignment; 3) imputation analysis. In reality the true main signals may come from covariate effects or auxiliary structures and can be individual-level or shared across multiple cohorts. So we consider four fundamental scenarios: (a) large $\mathbf{B}$, main signals from one global auxiliary structure which is shared by all cohorts; (b) large $\mathbf{S}$, main signals from one global covariate effect which is shared by all cohorts; (c) large $\mathbf{B}_i$, main signals from individual covariate effect of each cohort; (d) large $\mathbf{S}_i$, main signals from individual auxiliary structure in each individual cohort. In order to mimic the real situation, the number of samples and dimensions of the data is set to be the same as the TCGA data analyzed in Section 7. That is, $\mathbf{X}$ consists of 1000 features and 6581 samples from 30 study cohorts and $\mathbf{Y}$ consists of 50 predictors. Therefore, the ground truth can be written as $\mathbf{X}_j = \mathbf{B}\mathbf{Y}_j + \mathbf{B}_j\mathbf{Y}_j + \mathbf{S}_{shared,j} + \mathbf{S}_j + \mathbf{E}_j, j = 1, ..., 30$. In each simulation, the standard deviation for the main signals is set to be $\sqrt{10}$ while that of the remaining signals and random errors are set to be 1. The complete data generation process is as follows:

(1) Every entry in each $\mathbf{Y}_j, j = 1, ..., J = 30$ is drawn independently from a standard normal distribution. By default, the variance of each feature is 1. Construct one global shared $\mathbf{Y}_{\cdot}^{(1)} = [\mathbf{Y}_1, ..., \mathbf{Y}_{30}]$ and 30 individual $\mathbf{Y}_{\cdot}^{(k)} = [\mathbf{0}, ..., \mathbf{Y}_{k-1}, ..., \mathbf{0}], k = 2, ..., 31$ (only the $k - 1$th submatrix is non-zero).

(2) For a number of $K = 31$ modules of $\mathbf{Y}_{\cdot}^{(k)}$ involved, generate $\mathbf{B}_k = \mathbf{U}_B^{(k)}\mathbf{V}_B^{(k)T}/sd(\mathbf{U}_B^{(k)}\mathbf{V}_B^{(k)T}\mathbf{Y}_{\cdot}^{(k)}) * \sqrt{n_k/n}$, where n_k is the number of samples in module k. Each entry of $\mathbf{U}_B^{(k)} : p \times r, \mathbf{V}_B^{(k)} : q \times r$ comes from standard Normal distribution.

(3) For a number of $L = 31$ modules of $\mathbf{S}_{\cdot}^{(l)}$ involved, draw one global score matrix $\mathbf{V}_S^{(1)} : n \times r$

and 30 individual score matrices $\mathbf{V}_S^{(l)} = [\mathbf{0}, ..., \mathbf{V}_{l-1}, ..., \mathbf{0}], l = 2, ..., 31$, where each entry of $\mathbf{V}_1, ..., \mathbf{V}_{30}$ comes from standard Normal. Generate $\mathbf{S}_\cdot^{(l)} = \mathbf{U}_S^{(l)}\mathbf{V}_S^{(l)T}/sd(\mathbf{U}_S^{(l)}\mathbf{V}_S^{(l)T}) * \sqrt{n_l/n}$, where each entry of $\mathbf{U}_S^{(l)} : p \times r$ comes from standard Normal distribution and n_l is the number of samples in module l.

(4) Draw each entry of $\mathbf{E}_\cdot$ from a standard normal distribution.

(5) Generate $\mathbf{X}_\cdot = a * \mathbf{B}_1\mathbf{Y}_\cdot^{(1)} + b * \mathbf{S}_\cdot^{(1)} + c * \sum_{k=2}^{31} \mathbf{B}_k\mathbf{Y}_\cdot^{(k)} + d * \sum_{l=2}^{31} \mathbf{S}_\cdot^{(l)} + \mathbf{E}_\cdot$. The letters a, b, c, d are constant for signal size. For instance, scenario (a), $a = \sqrt{10}$ and the remaining equal 1.

In order to mimic the missingness in reality, we conduct simulations in which four types of missingness are considered: (i) missing entries, (ii) missing columns, (iii) missing rows, and (iv) a balanced mix of these three types of missingness. Missingness is set to be 5% of the original data for the assumption that adequate information is provided for revealing global structures. All missing indexes are randomly selected. Denote $\widetilde{\mathbf{X}}_\cdot$ as the estimate for true observation $\mathbf{X}_\cdot$, based on non-missing entries. Similarly, we define the relative squared error (RSE) for missing data imputation:

$$RSE = \frac{\sum_{(m,n)\in M}(\mathbf{X}_\cdot[(m,n)] - \widetilde{\mathbf{X}}_\cdot[(m,n)])^2}{\sum_{(m,n)\in M}(\mathbf{X}_\cdot[(m,n)])^2}$$

We compare our method (maRRR with true 31 modules) with the following approaches:

(1) BIDIFAC+ with 31 modules, i.e. only auxiliary variation structure $\mathbf{S}$;

(2) mRRR with 31 modules, i.e. only covariate-related structure $\mathbf{BY}$;

(3) aRRR with only one module for all cancer types' cohorts together;

(4) aRRR separately on each cancer type's cohort;

(5) nuclear norm regression (without $\mathbf{S}$) of $\mathbf{X}_\cdot$ on $\mathbf{Y}_\cdot$, i.e. mRRR with one all-shared module;

(6) nuclear norm regression (without $\mathbf{S}$) of $\mathbf{X}_j$ on $\mathbf{Y}_j$ separate for each cancer type, i.e. mRRR with 30 individual modules;

(7) nuclear norm approximation (without $\mathbf{BY}$) for all cancer types together

(8) nuclear norm approximation (without $\mathbf{BY}$) for each cancer types separately.

Based on the simulation results shown in Table 2, in terms of the average performance, our proposed method maRRR has the lowest RSE. In particular, maRRR has a very close RSE to the best result in the case of missing columns or row under large individual covariate signals, and it performs the best in all other cases. BIDIFAC+ performs slightly worse than our proposed method while there are only missing entries or rows, but it cannot utilize any covariate information to impute in the case of missing columns. The models only considering covariate effects (mRRR and nuclear norm regression) cannot predict accurately when the auxiliary variation is large, no matter global or individual. On the contrary, the models without considering covariates (BIDIFAC+, nuclear norm approximation) can perform reasonably well in the cases of missing entries or rows since the variation from covariates may be counted into that of auxiliary structures. The special case of our proposed method, aRRR (for one cohort only), though worse than maRRR, has lower MSE than many other existing methods. Thus, we can implement aRRR for missing imputation for faster usage, as long as the requirement for precision is not extremely strict and decomposing shared or individual structures across cohorts is not of interest.

6.3 *Computation*

The proposed method is computational effective. For Algorithm 1, the time cost is 45 seconds per epoch, and for Algorithm 2 the time cost is 50 seconds per epoch for a 1000×6581 matrix, on average. We report a full table of computation time in the supplementary materials for a full comparison for all used methods.

7. Real Data Analysis

7.1 *Data description*

We consider gene expression data from the TCGA Pan-Cancer Project (Hoadley et al., 2018). We used data for 6581 tumor samples from 30 cohorts (cancer types). We filter 1000 gene expressions that have the highest standard deviation after centering, as outcomes ($\mathbf{X}$). We orthogonalize 50 somatic mutations as covariates ($\mathbf{Y}$).

7.2 *Decomposition results*

We first apply the optimization dynamic modules mentioned in (Lock et al., 2022) to uncover 50 low-rank modules. 50 is chosen as the upper bound since the variance explained by more modules than 50 are relatively inconsequential.

We order the modules by total variance explained by descending by maRRR estimates, i.e. $||\widehat{\mathbf{B}}_i\mathbf{Y}_i + \widehat{\mathbf{S}}_i||_2^2, i = 1, ..., 50$. The ordered result is shown in Table 3. The top 3 modules by total variance explained are those with one or two cancers: BRCA (breast carcinoma), THCA (thyroid carcinoma) and a combination of GBM and LGG (both neurological cancers), respectively. In general, auxiliary structures explained more variation than mutation-related structures, but their relative contribution varied widely across modules. For example, Modules 6 and 7 have fairly comparable amount of variation explained by both mutation-related and -unrelated parts. Other modules have negligible mutation-driven variation, such as Module 12 which is shared by all but one cancer type. The large amount of low-rank variation unrelated to covariates demonstrates the importance of accounting for this auxiliary structure. Moreover, the large amount of covariate-related and -unrelated variation that is specific to one or a small number of cancer types demonstrates the importance of accounting for individual and partially-shared structures.

Principal components plots of the first two modules (BRCA, THCA) are shown in Figure 1. Figure 1A displays mutation-related variation for the BRCA module, and samples are distinguished by whether they have a mutation in the TP53 gene or not. This makes sense, as TP53 is known to play a critical role in genomic activity for cancer and it is the most frequently mutated gene in breast cancer (Olivier et al., 2010). From Figure 1B,we see that the mutation-unrelated structure is driven by the 5 intrinsic BRCA subtypes (Cancer Genome Atlas Research Network, 2012): Normal-like, Luminal A (LumA), Luminal B (LumB), HER2-enriched (HER2), and Basal-enriched (Basal) tumors. This makes sense, as the BRCA subtypes are known to be genomically distinct, but (as is apparent) do not have a direct correspondence to TP53 or other common mutations. From Figure 1C, we observe that BRAF and non-BRAF groups are elegantly separated on the first principle component of mutation-driven variation for THCA, which explains much more variation than other components. This is consistent with research showing that THCA patients with the BRAF mutation common and define a genomically and clinically distinct subgroup (Dolezal et al., 2021).

Figure 2 illustrate the third module, which was shared by GBM and LGG. Figures 2A and 2B show the first two principal components for mutation-driven and auxiliary structure, respectively. Note that plots show substantial variation in both the GBM and LGG samples, indicating that the structure is shared among the two cancer types. The TP53 mutation drives separation of the mutation-driven structure, which makes sense as TP53 status is closely related to GBM and LGG aggressiveness (Ham et al., 2019). This was also observed in BRCA, however, from the scatterplot in Figure 2C we see that the TP53 regression coefficients from module 3 (GBM&LGG) has little correlation to those from module 1 (BRCA). This demonstrates that while TP53 is an important somatic mutation related to different types of cancer, its effect on gene expression can differ dramatically depending on the cancer type. In general, we find that the same somatic mutation plays different roles for different cohorts and modules. This embodies the necessity of the flexible modeling for different (combinations of) cohorts. In fact, one interesting finding of our analysis is that the effect of somatic mutations on gene expression are almost not entirely shared across different types of cancer. For example, for the module that is shared across almost all cancers (module 12) the mutation-driven component is negligible. This observation is further supported by our analysis in Section 7.3.

7.3 *Missing data imputation*

Similar to Section 6.3, we compare our proposed maRRR with other relevant methods under four types of missingness for these data. Beyond the aforementioned eight methods in Section 6.2, we added (9) linear least squares regression to predict $\mathbf{X}$ from $\mathbf{Y}$ for all cancer types together and (10) linear least-squares regression for each cancer types separately. Note that maRRR, BIDIFAC+ and mRRR are based on the detected 50 modules. Results are provided in Table 4. We first describe the results for the three different types of missingness separately. In the scenario of missing entries, both maRRR and BIDIFAC+ have the lowest RSE, with similar values; both methods allow for an efficient decomposition of joint and individual structures. In the case of missing columns (some samples' entire outcomes are missing), the methods that do not consider mutations and only consider $\mathbf{S}$ (BIDIFAC+, NN approx) have no predictive power, which is expected because the mutation data is needed to inform predictions if no gene expression is available for a sample. Here the methods that consider both $\mathbf{B}$ and $\mathbf{S}$ (maRRR, aRRR) are suboptimal compared to those methods that only consider mutation-driven variation ($\mathbf{BY}$), indicating that for these data allowing for auxiliary variation does not improve column-wise predictions. Moreover, methods that allow for separate mutation-driven structure across the cohorts perform substantially better, which is consistent with the fact that there were more individual modules in our analysis and mutation-driven variation was generally not shared (Table 3). In the event of missing rows (each cohort misses random features), methods that consider individual structure only do not perform well, as they cannot leverage shared structure when a gene is entirely missing within a cohort. On the contrary, methods with only one all-shared module (NN approx and aRRR) perform well. Here, maRR also performs reasonable well, as including several individual modules does not limit its performance. In this case methods considering covariate effects only (mRRR, NN reg) do not perform well, as they tend toward estimates of zero (i.e., no predictions) to minimize squared error loss. NN approx is slightly better than aRRR., perhaps because it does not consider covariate-driven variation.

Under the circumstances of a balanced mix of missingness for different conditions, maRRR has the best average recovering ability. This is largely because it is the most robust and flexible. Other comparable methods (BIDIFAC+, aRRR, NN approx) will have limited peformnance for at least one form of missingness. In reality, maRRR will be the most suitable for imputation since missingness is unpredictable and complex.

8. Discussion

Two strengths of our proposed maRRR approach are its flexibility and versatility. The method is flexible because it accounts for many different types of signals (covariate-driven or not, shared or unshared) at once, without requiring prior assumptions on the size or rank of these signals. Our method is versatile because it is capable of performing many tasks at once: e.g., dimension reduction, prediction and missing data imputation. These advantages are well-illustrated by our pan-cancer application, in which adequate amounts of variation are explained by different components and the patterns detected are both insightful and consistent with existing scientific research on cancer.

We focus on multi-cohort integration rather than multi-view (data on the same subjects from different sources) integration, in part because shared or unshared covariate effects are straightforward to interpret across multiple cohorts. But one can still argue that in a multi-view (e.g., multi-omics) context each sample will have intrinsic underlying signals that will affect variables from different sources. Without loss of generality, this method can also be modified to analyze multi-view data as well. This is achieved by simply switching the way we integrate matrices: horizontally across shared rows or vertically across shared columns. An interesting future direction is to extend maRRR to the bidimensional integration context, where the data are both multi-cohort and multi-view.

References

Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. The Annals of Statistics **39,** 1282–1309.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on optimization **20,** 1956–1982.

Cancer Genome Atlas Research Network (2012). Comprehensive molecular portraits of human breast tumours. Nature **490,** 61–70.

Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. Biometrika **100,** 901–920.

Dolezal, J. M., Trzcinska, A., Liao, C.-Y., Kochanny, S., Blair, E., Agrawal, N., Keutgen, X. M., Angelos, P., Cipriani, N. A., and Pearson, A. T. (2021). Deep learning prediction of braf-ras gene expression signature identifies noninvasive follicular thyroid neoplasms with papillary-like nuclear features. Modern Pathology **34,** 862–874.

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

Gavish, M. and Donoho, D. L. (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 **75,** 1121–1132.

Ham, S. W., Jeon, H.-Y., Jin, X., Kim, E.-J., Kim, J.-K., Shin, Y. J., Lee, Y., Kim, S. H., Lee, S. Y., Seo, S., et al. (2019). Tp53 gain-of-function mutation promotes inflammation in glioblastoma. Cell Death & Differentiation **26,** 409–425.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell **173,** 291–304.

Hutter, C. and Zenklusen, J. C. (2018). The cancer genome atlas: creating lasting value beyond its data. Cell **173,** 283–285.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. Journal of multivariate analysis **5,** 248–264.

Li, G., Liu, X., and Chen, K. (2019). Integrative multi-view regression: Bridging group-sparse and low-rank models. Biometrics **75,** 593–602.

Lock, E. F., Park, J. Y., and Hoadley, K. A. (2022). Bidimensional linked matrix factorization for pan-omics pan-cancer analysis. The annals of applied statistics **16,** 193.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. The Journal of Machine Learning Research **11,** 2287–2322.

Olivier, M., Hollstein, M., and Hainaut, P. (2010). Tp53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harbor perspectives in biology **2,** a001008.

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. II–IV: Invited Lectures, pages 1576–1602. World Scientific.

Safo, S. E., Min, E. J., and Haine, L. (2022). Sparse linear discriminant analysis for multiview structured data. Biometrics **78,** 612–623.

Shabalin, A. A. and Nobel, A. B. (2013). Reconstruction of a low-rank matrix in the presence of gaussian noise. Journal of Multivariate Analysis **118,** 67–76.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. Bioinformatics **17,** 520–525.

Wang, J. and Safo, S. E. (2021). Deep ida: A deep learning method for integrative discriminant analysis of multi-view data with feature ranking–an application to covid-19 severity. ArXiv .

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **69,** 329–346.

Zhang, Y. and Gaynanova, I. (2021). Joint association and classification analysis of multi-view data. Biometrics .

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Table 4 about here.]

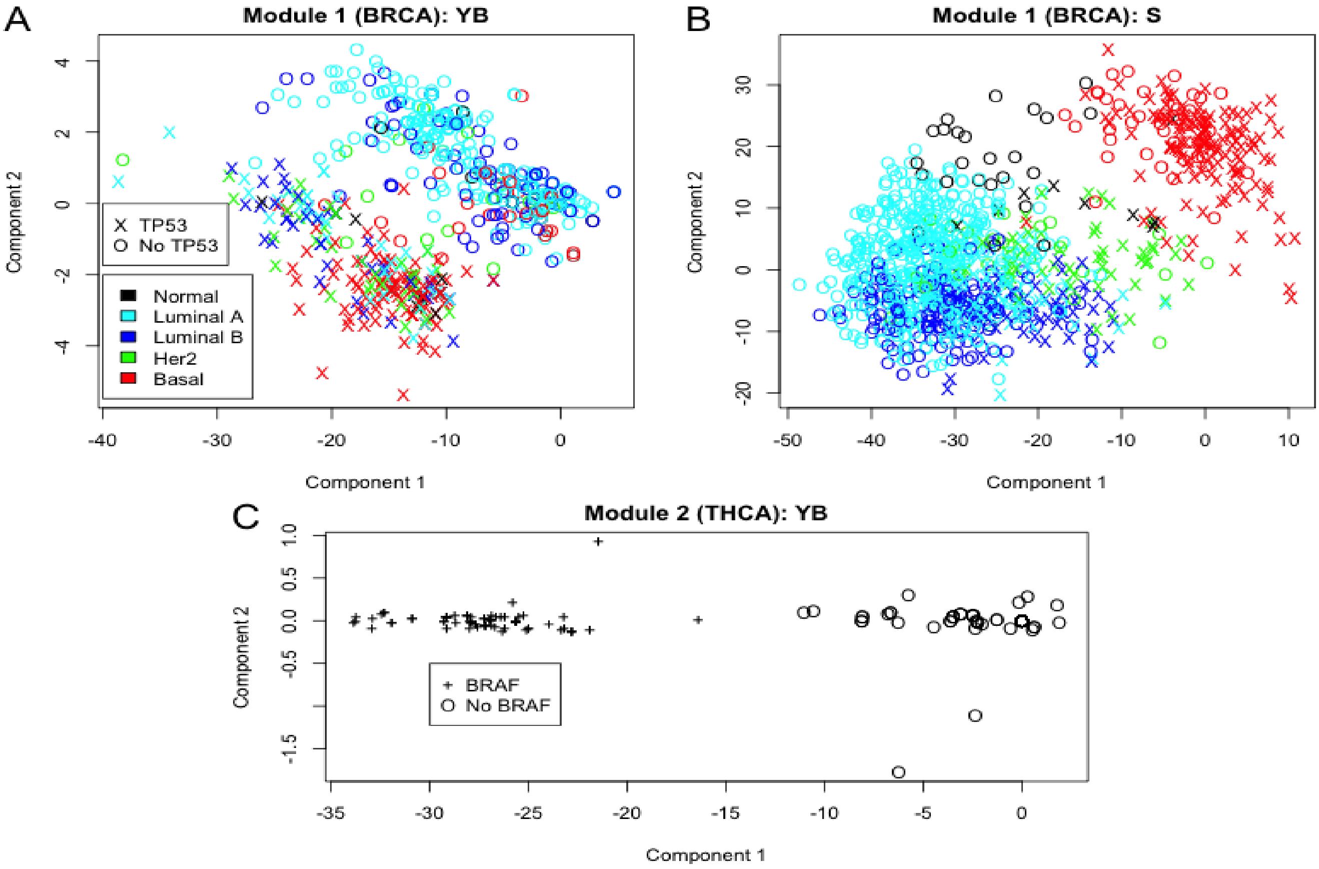

Figure 1. A: scores for the first two principle components of covariate-related variation (**BY**) from Module 1 (BRCA); B: scores for the first two principle components of covariate-unrelated auxiliary variation (**S**) from Module 1 (BRCA), with symbols and colors showing TP53 mutation and 5 subtypes of BRCA respectively. C: scores for the first two principle components of covariate-related variation (**BY**) from Module 2 (THCA), with symbols showing BRAF mutation

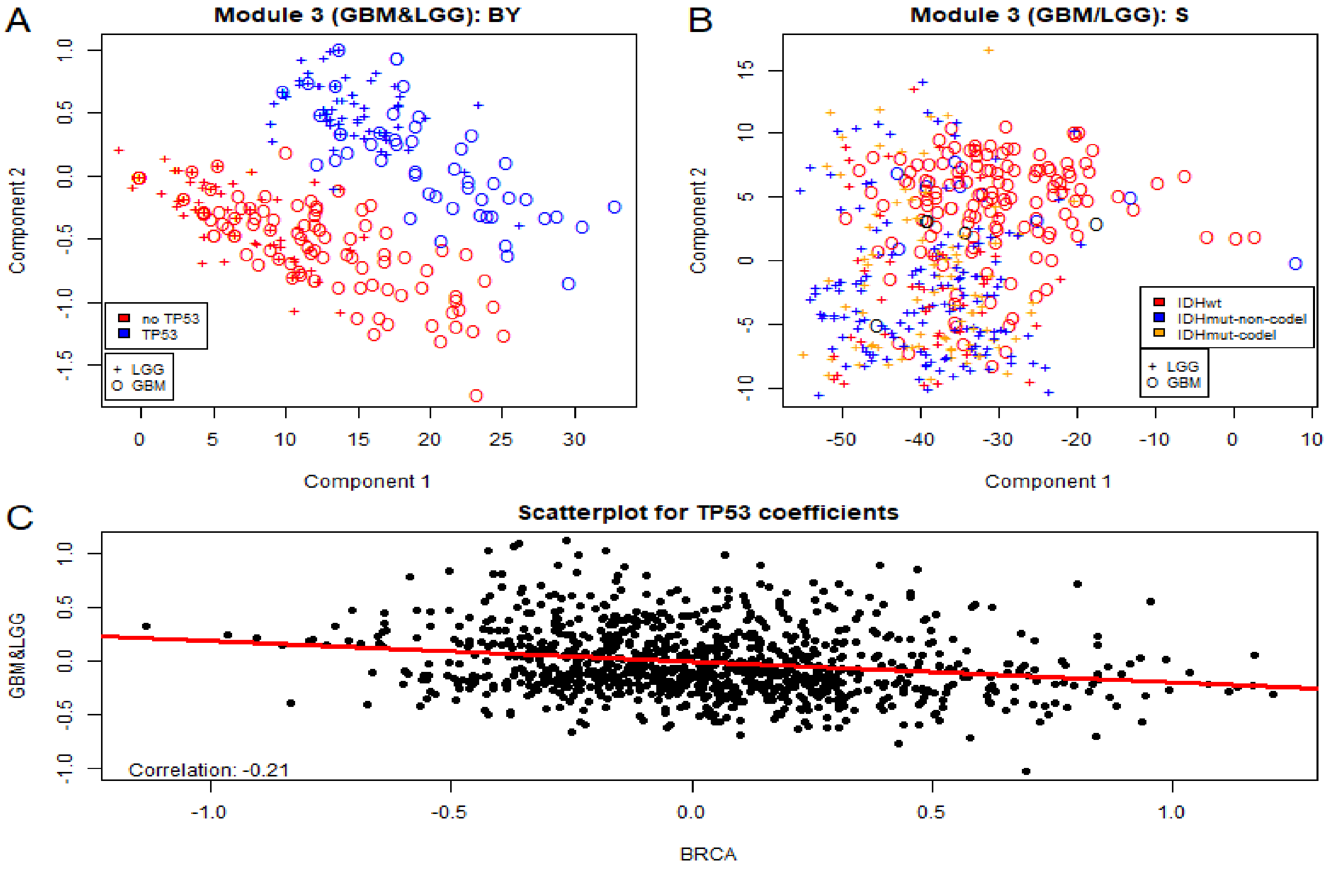

Figure 2. A: scores for the first two principle components of covariate-related variation (**BY**) from Module 3 (GBM&LGG); B: scores for the first two principle components of covariate-unrelated auxiliary variation (**S**) from Module 3 (GBM&LGG), with symbols and colors showing TP53 mutation and 3 IDH/codel subtypes of GBM&LGG respectively. C: scatterplot for the regression coefficients for TP53 in module 1 and module 3

Table 1

*Relative MSE for scenarios assessing aRRR (**A**) and mRRR (**B**).*

A		aRRR		Two-stage LS		Two-stage NN	
$\frac{\|\|\mathbf{BY}\|\|}{\|\|\mathbf{S}\|\|}$	R_y	**B**	**S**	**B**	**S**	**B**	**S**
10	1	0.01	0.60	0.01	0.60	0.01	0.62
1	1	0.04	0.23	0.23	0.21	0.06	0.26
0.1	1	0.19	0.01	11.49	0.10	7.67	0.08
10	5	0.01	0.63	0.01	0.59	0.01	0.64
1	5	0.14	0.24	0.22	0.19	0.14	0.26
0.1	5	0.41	0.01	11.85	0.11	8.07	0.09

B		mRRR		Two-stage LS		Two-stage NN	
$\frac{\|\|\mathbf{B}\|\|}{\|\|\mathbf{B}_i\|\|}$	R_y	**B**	$\mathbf{B}_i$	**B**	$\mathbf{B}_i$	**B**	$\mathbf{B}_i$
10	1	0.01	0.11	0.01	0.77	0.01	0.42
1	1	0.01	0.01	0.75	0.60	0.67	0.52
0.1	1	0.07	0.01	80.57	0.55	76.17	0.52
10	5	0.01	0.28	0.01	0.57	0.01	0.50
1	5	0.08	0.08	0.49	0.54	0.45	0.50
0.1	5	0.49	0.01	54.79	0.56	52.41	0.54

Table 2

Imputation relative squared error(RSE) under different methods and different types of missingness, simulated data is set to be large at only one type of modules. Missingness is set to be 5% of the original $\mathbf{X}$*. The number of epochs for each method is set as 30. Each result is a mean of 10 replications. The standard error is less than 0.01 for all of the means shown.*

large_B	Method	missing entries	missing columns	missing rows	mean
	maRRR	0.082	0.228	0.216	0.175
	BIDIFAC+	0.085	1	0.231	0.439
	mRRR	0.202	0.241	0.285	0.243
	aRRR, one all-shared	0.125	0.288	0.227	0.213
	aRRR, 30 separate	0.093	0.287	1.014	0.465
	NN reg, one all-shared	0.283	0.287	0.288	0.286
	NN reg, 30 separate	0.212	0.255	1	0.489
	NN approx, one all-shared	0.127	1	0.229	0.452
	NN approx, 30 separate	0.096	1	1	0.699
large_S	Method	missing entries	missing columns	missing rows	mean
	maRRR	0.082	0.877	0.218	0.392
	BIDIFAC+	0.085	1	0.225	0.437
	mRRR	0.759	1.066	0.927	0.917
	aRRR, one all-shared	0.125	0.929	0.228	0.427
	aRRR, 30 separate	0.093	0.884	1.004	0.66
	NN reg, one all-shared	0.928	0.931	0.933	0.93
	NN reg, 30 separate	0.783	1.072	1	0.952
	NN approx, one all-shared	0.127	1	0.23	0.452
	NN approx, 30 separate	0.096	1	1	0.699
large_Bi	Method	missing entries	missing columns	missing rows	mean
	maRRR	0.083	0.262	0.901	0.415
	BIDIFAC+	0.086	1	0.867	0.651
	mRRR	0.204	0.246	0.928	0.459
	aRRR, one all-shared	0.149	0.913	0.902	0.655
	aRRR, 30 separate	0.093	0.287	1.013	0.464
	NN reg, one all-shared	0.896	0.899	0.964	0.92
	NN reg, 30 separate	0.212	0.252	1	0.488
	NN approx, one all-shared	0.151	1	0.907	0.686
	NN approx, 30 separate	0.096	1	1	0.699
large_Si	Method	missing entries	missing columns	missing rows	mean
	maRRR	0.083	0.873	0.861	0.606
	BIDIFAC+	0.086	1	0.866	0.651
	mRRR	0.76	1.066	0.928	0.918
	aRRR, one all-shared	0.148	0.93	0.906	0.661
	aRRR, 30 separate	0.093	0.885	1.003	0.661
	NN reg, one all-shared	0.929	0.931	0.935	0.932
	NN reg, 30 separate	0.784	1.072	1	0.952
	NN approx, one all-shared	0.15	1	0.91	0.687
	NN approx, 30 separate	0.096	1	1	0.699

Table 3
Cancer types and sources for the first 15 modules, ordered by variation explained by maRRR. Variance of S/BY refers to total variance explained by S/BY.

Module	Variance of **BY**	Variance of **S**	Cancer types
1	129652.09	1085292.83	BRCA
2	167384.83	551269.19	THCA
3	67242.88	645839.92	GBM, LGG
4	49.52	730462.99	PRAD
5	7231.61	633034.84	LIHC
6	137934.04	316746.84	BLCA, CESC, ESCA, HNSC, KICH, LUSC
7	137765.92	298195.14	COAD, ESCA, PAAD, READ, STAD
8	20.61	396671.22	PCPG
9	53.04	347389.29	LAML
10	13530.50	231107.59	KIRC, KIRP
11	30138.96	192612.50	SKCM, UVM
12	0.14	264645.95	All cancers *but* LAML
13	63836.59	87452.29	COAD, READ
14	495.86	237614.82	TGCT
15	97.81	242978.25	THYM

Table 4

Imputation relative squared error(RSE) under different methods and different types of missingness. Missingness is set to be 5% *of the original* $\mathbf{X}$*. "one all shared" means data for 30 groups are stacked together to form one matrix to analyze; "30 separate" means each group has its only model. "Missing entries" refers to missingness is entrywise; "missing columns" means some samples' entire observation are missing; "missing rows" means each group has several features entirely missing. "N/A" means some specific method is not applicable.*

Methods	Missing entries	Missing columns	Missing rows	Average
maRRR	0.233	0.813	0.600	0.548
BIDIFAC+	0.233	0.999	0.613	0.615
mRRR	0.603	0.711	0.998	0.770
aRRR, one all-shared	0.261	0.930	0.487	0.559
aRRR, 30 separate	0.376	0.780	1.001	0.719
LS reg, one all-shared	0.908	0.906	0.899	0.904
LS reg, 30 separate	0.560	N/A	N/A	N/A
NN reg, one all-shared	0.912	0.913	1.032	0.953
NN reg, 30 separate	0.599	0.727	1.000	0.775
NN approx, one all-shared	0.273	1.000	0.454	0.576
NN approx, 30 separate	0.252	1.000	1.009	0.754