Hierarchical nuclear norm penalization for multi-view data integration

Sangyoon Yi\textsuperscript{1} | Raymond Ka Wai Wong\textsuperscript{2} | Irina Gaynanova\textsuperscript{2}

\textsuperscript{1}Department of Statistics, Oklahoma State University, Stillwater, Oklahoma, USA
\textsuperscript{2}Department of Statistics, Texas A&M University, College Station, Texas, USA

**Abstract**

The prevalence of data collected on the same set of samples from multiple sources (i.e., multi-view data) has prompted significant development of data integration methods based on low-rank matrix factorizations. These methods decompose signal matrices from each view into the sum of shared and individual structures, which are further used for dimension reduction, exploratory analyses, and quantifying associations across views. However, existing methods have limitations in modeling partially-shared structures due to either too restrictive models, or restrictive identifiability conditions. To address these challenges, we propose a new formulation for signal structures that include partially-shared signals based on grouping the views into so-called hierarchical levels with identifiable guarantees under suitable conditions. The proposed hierarchy leads us to introduce a new penalty, hierarchical nuclear norm (HNN), for signal estimation. In contrast to existing methods, HNN penalization avoids scores and loadings factorization of the signals and leads to a convex optimization problem, which we solve using a dual forward–backward algorithm. We propose a simple refitting procedure to adjust the penalization bias and develop an adapted version of bi-cross-validation for selecting tuning parameters. Extensive simulation studies and analysis of the genotype-tissue expression data demonstrate the advantages of our method over existing alternatives.

**KEYWORDS**

bi-cross-validation, data fusion, low-rank matrix, multi-source data, optimization, rank selection

---

1 | INTRODUCTION

Technological advances in biomedical fields led to prevalence of multi-source data collected from the same set of samples, often called multi-view data. Our motivating example is the Genotype-Tissue Expression (GTEx) project (The GTEx Consortium, 2020) that collects gene-expression data on the same individuals across multiple tissues (views). The gene expression is often tissue-specific. For instance, the p53 gene is critical for cell regulation playing prominent role in the cancer development (Tanikawa et al., 2017), however p53 tissue-specific expression makes it difficult to develop targeted therapies (Vlatkovic et al., 2011). It is thus crucial to identify patterns in gene expression that are shared and unique across multiple tissues. Traditional methods for signal extraction...
tend to be applied either separately to each view or to all views combined, thus failing to separate shared and unique parts of the signal.

Specifically, let $X_d \in \mathbb{R}^{n \times p_d}$ be the observed data matrix for the $d$th view with $n$ matched samples and $p_d$ measurements for $d = 1, \ldots, D$. We consider an additive error model

$$X_d = M_d + E_d, \quad d = 1, \ldots, D, \quad (1)$$

where $M_d$ is the signal matrix and $E_d$ is the noise matrix. Successful extraction of the signals in the model (1) requires additional structural assumptions. For single-view case, $M_d$ is commonly assumed to be low-rank (Candès and Recht, 2009; Candès et al., 2013; Srebro & Jaakkola, 2003). A direct application of this approach for multi-view data leads to low-rank assumption on either (i) each $M_d$ or (ii) the concatenated signal $M = [M_1, \ldots, M_D]$.

The main limitation of (i) is that it not only ignores any structural relationships between the signals due to the matched samples, but also separately estimates signals without borrowing strength across views. The main limitation of (ii) is that it assumes a joint structure across all $M_d$, resulting in all estimated $\tilde{M}_d$ having the same rank. Therefore, it does not take into account possible heterogeneity across the views. In multi-view context, it is of interest to explore structural assumptions on $M_d$ that allow both joint (thus accounting for matched samples) and individual (thus permitting differences across views) parts of the signal matrices.

The simultaneous exploration of the joint and individual signals has been a prominent line of research for multi-view data (Feng et al., 2018; Gaynanova & Li, 2019; Lock et al., 2013; Yang & Michailidis, 2016; Zhou et al., 2016). The joint structure is a shared pattern across all views and defined as the intersection of the column spaces of $M_d$ by JIVE (Joint and Individual Variation Explained, Lock et al., 2013), COBE (Common Orthogonal Basis Extraction, Zhou et al., 2016), and AJIVE (Angle-based JIVE, Feng et al., 2018), while the individual structure is unique to a particular view adjusted after the joint structure, see Example S1 in Web Appendix A.1 for illustration. Despite these advances, many existing approaches (Feng et al., 2018; Lock et al., 2013; Yang & Michailidis, 2016; Zhou et al., 2016) lack explicit definition of partially-shared structures when $D > 2$, for example, a signal structure shared by muscle and blood tissues, but not the skin tissue in the motivating GTEx example. This implies that the corresponding methods cannot take advantage of partially-shared structures for signal estimation, leading to rank mis-identification and worse signal estimation performance compared with methods that account for such structures (Gaynanova & Li, 2019).

While several methods for identification of partially-shared structures have been proposed, they have limitations from modeling perspective. The penalized matrix factorization approaches by Jia et al. (2010) and Van Deun et al. (2011) lack explicit formulation of underlying models, making it difficult to assess identifiability and provide interpretation. SLIDE (Gaynanova & Li, 2019) requires orthogonality across individual and partially-shared signals for model identifiability. This implies that the SLIDE model is not always unique, causing issues in estimation as well as interpretation.

In addition to modeling issues (lack of partially-shared structures or lack of identifiability), existing methods (Gaynanova & Li, 2019; Jia et al., 2010; Lock et al., 2013; Van Deun et al., 2011; Yang & Michailidis, 2016; Zhou et al., 2016) estimate the signals $M_d$ based on explicit score-and-loading factorizations, leading to non-convex optimization problems. The convergence to the global optimum is not guaranteed and the obtained solution depends on the starting point. Thus, in practice, the obtained solution may be sub-optimal. The recently proposed BIDIFAC (Park & Lock, 2020) and BIDIFAC+ (Lock et al., 2022) are exceptions, as those methods provide a convex formulation via the use of nuclear norm penalization. Both are designed for bidimensionally-matched data, which includes multi-view data as a special case. However, BIDIFAC does not consider partially-shared structures. Furthermore, both BIDIFAC and BIDIFAC+ suffer from bias associated with nuclear norm penalization (Chen et al., 2013; Josse & Sardy, 2016), affecting their signal estimation performance.

In this work, we address the limitations of existing models by both formulating an explicit model for partially-shared structures and avoiding identifiability issues. We also address the limitations of associated estimation approaches by avoiding scores and loadings factorization, and correcting for possible estimation bias. We achieve these goals as follows. First, we propose a recursive definition of signal structures based on the new concept of the hierarchical level that are shown to be identifiable under suitable conditions. For $D = 3$ views, the hierarchical levels are: (i) all three views (Level 1); (ii) any two views (Level 2); (iii) individual views (Level 3). The key idea is to work with the column spaces of $M_d$, and recursively consider column spaces at each level after accounting for previous levels. Our formulation includes JIVE, COBE, and AJIVE models as a special case and avoids identifiability issues of SLIDE. Second, guided by our hierarchical levels, we propose a new penalty, hierarchical nuclear norm (HNN), for direct estimation of $M_d$ without scores and loadings factorization. To our knowledge, HNN is novel in the literature, and its hierarchical construction allows to restrict the ranks of the signals at each level. By varying the corresponding tuning parameters, the solution path encompasses a wide range of models with different patterns of shared signals. HNN is
shows how many existing methods (Feng et al., 2017). We address this challenge by adapting the dual block-coordinate forward–backward algorithm (Abboud et al., 2017). Third, we address penalization bias by a new type of refitting. As the proposed definition of partially-shared structures relies on the column spaces, we use HNN to estimate the column spaces, and then refit $M_d$ under the column space constraint, thus preserving the estimated shared and individual patterns. Finally, we take advantage of SURE criterion (Candès and Recht, 2009) to reduce the number of tuning parameters and use the adapted version of the bi-cross validation (BCV) (Owen and Perry, 2009) for model selection. In numerical studies, the resulting HNN method outperforms the competitors in both signal estimation and structure identification.

**Notation:** For a matrix $A = (a_{ij}) \in \mathbb{R}^{n \times p}$, let $\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{p} a_{ij}^2}$ the Frobenius norm, and $\|A\|_n = \sum_{i=1}^{\min(n,p)} \sigma_i(A)$ be the nuclear norm, where $\sigma_i(A)$ is the $i$th largest singular value of $A$. For the column space $C(A)$, let $P_{C(A)}$ and $P_{C(A)}^\perp$ be the projection matrices onto $C(A)$ and its orthogonal complement $C(A)^\perp$, respectively. Let $0$ and $I$ be the zero and identity matrix, respectively, with dimensions inferred from the context. Let $A_{k} = [A_k, A_I] \in \mathbb{R}^{n \times (p_k+p_I)}$ be formed by concatenating $A_k \in \mathbb{R}^{n \times p_k}$ and $A_I \in \mathbb{R}^{n \times p_I}$ column-wisely. Let $C_1 + \cdots + C_k = \text{span}\{C_1, \ldots, C_k\}$ be subspace spanned by the union of sets of basis vectors of subspaces $C_1, \ldots, C_k \subset \mathbb{R}^n$.

**2 | METHODOLOGY**

**2.1 | Motivating example**

Here, we provide a toy example that illustrates the definition of joint and individual structures of JIVE (Lock et al., 2013), COBE (Zhou et al., 2016), and AJIVE (Feng et al., 2018), and how this definition does not account for partial sharing. We also illustrate how it is possible to have non-orthogonality between different structures (individual, joint, partially-shared), thus violating identifiability conditions for SLIDE (Gaynanova & Li, 2019).

**Example 1.** Consider $D = 3$ with the signal matrices:

$$M_1 = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad M_3 = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & -1 \\ 1 & -1 & 0 \\ -1 & 0 & 2 \\ 1 & 0 & 1 \end{bmatrix}.$$ 

JIVE, COBE, and AJIVE define the joint structure as the intersection of the column spaces $C(M_1) \cap C(M_2) \cap C(M_3) = \text{span}\{[1 \ -1 \ 1 \ -1 \ 1]^T\}$, with the individual structures being the remaining signal:

- Individual view 1: $C(P_{\cap_d \cup}^1 C(M_d) M_1) = \text{span}\{[1 \ 1 \ 0 \ 0 \ 0]^T, [1 \ 0 \ -1 \ 0 \ 0]^T\}$
- Individual view 2: $C(P_{\cap_d \cup}^2 C(M_d) M_2) = \text{span}\{[1 \ 1 \ 0 \ 0 \ 0]^T, [0 \ 0 \ 1 \ 0 \ -1]^T\}$
- Individual view 3: $C(P_{\cap_d \cup}^3 C(M_d) M_3) = \text{span}\{[1 \ 0 \ -1 \ 0 \ 0]^T, [0 \ -1 \ 0 \ 2 \ 1]^T\}$

However, these individual structures are not truly unique to each view: while their overall intersection is zero, they have non-trivial pairwise intersections. Concretely, it follows that (i) $\text{span}\{[1 \ 1 \ 0 \ 0 \ 0]^T\}$ is shared by the first and second views (but not the third view); (ii) $\text{span}\{[1 \ 0 \ -1 \ 0 \ 0]^T\}$ is shared by the first and third views (but not the second view). Likewise, in many applications, it is reasonable to expect that certain signals are shared across a few views rather than all views. We refer to such signals as partially-shared.

Example 1 shows how many existing methods (Feng et al., 2018; Lock et al., 2013; Yang & Michailidis, 2016; Zhou et al., 2016) treat partially-shared signals as individual. To allow for partial-sharing, we can apply these methods separately to each of the $D \choose 2$ view pairs: if the estimated joint rank from any pair is larger than that from all $D$ views, it indicates the existence of a partially-shared structure between the corresponding pair. However, such approach significantly increases computational burden, and the resulting estimates across pairs are not guaranteed to be consistent with each other due to estimation errors.

To address these challenges, we propose to define partially-shared and individual structures in terms of column spaces of signal matrices. As existing models, we define the joint structure as an intersection of all column spaces. In contrast to these methods, we further decompose each $C(P_{\cap_d \cup}^1 C(M_d) M_d)$ to incorporate partially-shared structures. Our key idea is to consider remaining signal structures after accounting for the joint structure. More precisely, the partially-shared signal structures (rigorously defined in Definition 1) in Example 1 are
FIGURE 1  Illustration of Definition 1 applied to the signals in Example 1. The solid arrows (in yellow) indicate the corresponding intersection of column-spaces (joint structures). The red dash-dotted lines (in red) together with the dotted arrows (in blue) indicate the remainder terms after accounting for joint structures. Observe the missing $J(2, 3)$ (due to zero intersection or $R(1)_2$ and $R(1)_3$), and missing $R(2)_1$, $I(1)$ (due to $R(1)_1$ being fully captured by $J(1, 2)$ and $J(1, 3)$). [This figure appears in color in the electronic version of this paper, and any mention of color refers to that version].

specified as

\[
J(1, 2) = \text{span}\{ [1, 1, 0, 0, 0]^\top \}, \\
J(1, 3) = \text{span}\{ [1, 0, -1, 0, 0]^\top \}, \quad J(2, 3) = \{0\}.
\]

By considering the remaining signals, we redefine the individual structures in Example 1 as

(proposed) individual view 1: $I(1) = \{0\}$,
(proposed) individual view 2:

\[
I(2) = \text{span}\{ [0, 0, 1, 0, -1]^\top \},
\]
(proposed) individual view 3:

\[
I(3) = \text{span}\{ [0, -1, 0, 2, 1]^\top \}.
\]

In comparison to COBE, JIVE, and AJIVE, our approach models the individual structures that are indeed view-specific after taking into account both partially-shared structures in Example 1. Our signal structure modeling is discussed with more details in Section 2.2.

At the same time, SLIDE (Gaynanova & Li, 2019) also allows partial-sharing but requires orthogonality between different types of structures to guarantee uniqueness. In Example 1, the individual and partially-shared structures are not orthogonal to each other, that is, $J(1, 2) \not\perp J(1, 3)$, $J(1, 2) \not\perp I(3)$, $J(1, 3) \not\perp I(2)$, and $I(2) \not\perp I(3)$. This implies that SLIDE’s identifiability conditions are violated, and the resulting parameterization is not unique, leading to difficulties in estimation and interpretation (see Example S2 in Web Appendix A.2).

### 2.2 Proposed definition of hierarchical signal structure

Motivated by Example 1, we propose to define joint, partially-joint, and individual structures via hierarchical levels. When $D = 3$, we consider three hierarchical levels:

\[
\begin{align*}
\text{Level 1} & : \{(1, 2, 3)\}, \\
\text{Level 2} & : \{(1, 2), (1, 3), (2, 3)\}, \\
\text{Level 3} & : \{(1, 2, 3)\}.
\end{align*}
\]

(2)

where the integers indicate the views. Levels 1–3 correspond to the joint structure, partially-shared structures between $\binom{3}{2}$ distinct view pairs, and the three individual structures, respectively. The structure at each level is defined recursively based on the previous level, starting with the intersection of all column-spaces $C(M_d)$ at Level 1. Figure 1 provides an illustration of the proposed definition as applied to Example 1, which we formalize below.

At Level 1, the joint structure is defined as $J(1, 2, 3) = \bigcap_{d=1}^{3} C(M_d)$. We use $R(1)_d = C(P_{J(1,2,3) \setminus M_d})$ to represent the remaining signal in each view after accounting for the joint structure. At Level 2, we use the signal remaining from Level 1 to define the joint structures across each view pair. Formally, $J(l, m) = R(1)_l \cap R(1)_m$ for $(l, m) =$
(1, 2), (1, 3), (2, 3). For example, J(1, 2) is shared across views 1 and 2, is not present in view 3, and by construction is orthogonal to the globally-shared J(1, 2, 3). We use 
\( R_d^{(2)} \) to define the remaining signal in each view after accounting for structures in Levels 1 and 2. That is, 
\[ R_1^{(2)} = C(\mathbf{P}_1^{(2)} \mathbf{p}_j(1,2,3) \mathbf{M}_1), \quad R_2^{(2)} = C(\mathbf{P}_2^{(2)} \mathbf{p}_j(1,2,3) \mathbf{M}_2) \quad \text{and} \quad R_3^{(2)} = C(\mathbf{P}_3^{(2)} \mathbf{p}_j(1,2,3) \mathbf{M}_3) \]
where 
\[ Q_1^{(2)} = J(1, 2) + J(1, 3), \quad Q_2^{(2)} = J(1, 2) + J(2, 3) + J(1, 3) + J(2, 3). \]
At Level 3, only singletons remain, with individual structure for view \( d \) being the remainder \( I(d) = R_d^{(2)} \) since it is already adjusted for both joint and partially-shared structures at the previous two levels.

The above construction of signal structures decomposes each \( C(\mathbf{M}_d) \) as
\[
C(\mathbf{M}_1) = J(1, 2, 3) + \text{span}[J(1, 2), J(1, 3)] + I(1), \\
C(\mathbf{M}_2) = J(1, 2, 3) + \text{span}[J(1, 2), J(2, 3)] + I(2), \\
C(\mathbf{M}_3) = J(1, 2, 3) + \text{span}[J(1, 3), J(2, 3)] + I(3).
\]
In the column-space representation, JIVE, COBE, and AJIVE correspond to the same decomposition \( C(\mathbf{M}_d) = J(1, 2, 3) + I_{\text{JIVE}}(d) \), where \( I_{\text{JIVE}}(d) = C(\mathbf{p}_j(1,2,3) \mathbf{M}_d) \) is the individual structure for \( d = 1, 2, 3 \). In contrast, Equation (3) further decomposes \( I_{\text{JIVE}}(d) \), making partially-shared structures explicit. Theorem 1 summarizes identifiability conditions for the proposed decomposition (3). The proof is in Web Appendix A.3.

**Theorem 1.** Given \( C(\mathbf{M}_1), C(\mathbf{M}_2), \) and \( C(\mathbf{M}_3) \) in \( \mathbb{R}^n \), the signal subspaces \( J(1, 2, 3), J(1, 2), J(1, 3), J(2, 3), I(1), I(2) \) and \( I(3) \) defined above (i.e., Definition 1 when \( D = 3 \)) exist and are the only spaces that jointly satisfy the following conditions:

(a) **(Decomposition of \( C(\mathbf{M}_d) \) into signal structures)**

\[
C(\mathbf{M}_d) = J(1, 2, 3) + Q_d^{(2)} + I(d) \quad \text{for} \quad d = 1, 2, 3,
\]

where \( Q_1^{(2)} = J(1, 2) + J(1, 3) \), \( Q_2^{(2)} = J(1, 2) + J(2, 3) \), and \( Q_3^{(2)} = J(1, 3) + J(2, 3) \).

(b) **(Orthogonality between signal structures)**

\[
J(1, 2, 3) \perp I(d) \quad \text{for} \quad d = 1, 2, 3,
\]

\[
J(1, 2, 3) \perp J(k, l), \quad J(k, l) \perp I(k), \quad J(k, l) \perp I(l)
\]

\[
\text{for} \quad (k, l) = (1, 2), (1, 3), (2, 3);
\]

(c) **(Linear independence)**

\[
\cap_{d=1}^3 \{ Q_d^{(2)} + I(d) \} = \{ 0 \};
\]

(d) \( \{ Q_k^{(2)} + I(k) \} \cap \{ Q_l^{(2)} + I(l) \} = J(k, l) \)

for \( (k, l) = (1, 2), (1, 3), (2, 3) \).

The above (a)–(d) are the identifiability conditions for our proposed hierarchical signal structures. Condition (a) states that the column space of the signal matrix of each view is spanned by the union of the basis vectors from each type of the signal structure. Condition (b) specifies that the signal structures are orthogonal across hierarchical levels (joint, partially-joint, individual). Since the signal structures at the same level are allowed to be non-orthogonal, condition (b) is weaker than identifiability conditions of SLIDE (Gaynanova & Li, 2019). Condition (c) guarantees that each partially-shared structure is solely spanned by the remaining signals after accounting for \( J(1, 2, 3) \).

For general \( D \), there are in total \( D \) hierarchical levels. For \( 1 \leq k \leq D \), the \( (D-k+1) \) level corresponds to the structures shared across \( k \) views, and there are \( \binom{D}{k} \) distinct view combinations \( (i_1, \ldots, i_k) \). The key idea is to recursively decompose the remaining signal from the previous level into the shared structures specified at the current level and the remaining signal. Below, we provide the full definition of joint, partially-joint, and individual structures.

**Definition 1.** At Level 1 = \{1, ..., D\}, the joint structure and the remaining signals are defined as \( J(2013) \).

\[
\mathbf{M}_d^{(2)} = \prod_{l=k}^{D} \mathbf{P}_l^{(D-l+1)} \mathbf{M}_d
\]

At Level \( D \), the individual structure is defined as \( I(d) = R_d^{(D-1)} \).
The main advantage of Definition 1 is that each type of structure is defined in terms of column space, instead of relying on matrix representations (via decomposition and factorization) that existing methods (Feng et al., 2018; Gaynanova & Li, 2019; Lock et al., 2013; Yang & Michailidis, 2016; Zhou et al., 2016) use. However, our approach can still be used to obtain such a representation by extracting the basis vectors of each signal subspace and considering a corresponding projection, see Web Appendix A.4. The decomposition for general $D > 3$ case is identifiable under conditions similar to Theorem 1 (Web Appendix A.3).

### 2.3 Hierarchical rank constraints

Based on Definition 1, one possibility is to use matrix factorization for signal estimation, that is, to decompose each $\mathbf{M}_d$ into a sum of several matrices that represent each signal structure. The scores and loadings factorization of each structure can be used for estimation, which is pursued by many existing methods (Feng et al., 2018; Gaynanova & Li, 2019; Lock et al., 2013). In our framework, however, it is unclear how to adopt such factorization while preserving the potential non-orthogonality. Another possibility is to consider signal estimation as optimization problem with explicit subspace constraints (Yuan & Gaynanova, 2022). However, in the presence of multiple types of structures, this leads to a highly non-trivial non-convex, manifold optimization. Here, we propose a new strategy for direct signal estimation that avoids both matrix factorization, and subspace constraints, leading to a convex optimization problem. Our key idea is to introduce an HNN penalty, motivated by the rank constraints.

Specifically, we propose to take advantage of the constraint on the ranks of the matrices at each hierarchical levels. As shown in Section 2.1, the existing models’ definition of the joint structure (Feng et al., 2018; Lock et al., 2013; Zhou et al., 2016) relies on non-trivial intersection of column spaces of signal matrices, which implies rank constraint for $D = 3$: $\text{rank}([\mathbf{M}_1 \, \mathbf{M}_2 \, \mathbf{M}_3]) < \text{rank}([\mathbf{M}_1]) + \text{rank}([\mathbf{M}_2]) + \text{rank}([\mathbf{M}_3])$. When estimating $\mathbf{M}_d$ from noisy $\mathbf{X}_d$ in model (1), these methods effectively impose the low-rank constraints with some $r_d \ll \min(n, p_d)$ and $r_{123} \ll \min(n, \sum_{d=1}^3 p_d)$ so that

$$\text{rank}([\mathbf{M}_1 \, \mathbf{M}_2 \, \mathbf{M}_3]) \leq r_{123}, \quad \text{rank}([\mathbf{M}_d]) \leq r_d, \quad d = 1, 2, 3, \quad (4)$$

Choosing ranks such that $r_{123} < \sum_{d=1}^3 r_d$ leads to non-trivial intersections in column spaces.

Using the intuition from Equation (4), we introduce the hierarchical rank constraints which are imposed on the ranks of all matrices at all hierarchical levels. Considering Levels 1–3 in Figure 2 for $D = 3$ (see Web Appendix B.1 for $D > 3$), the constraints become

| Level 1 | M1 | M2 | M3 |
|--------|----|----|----|
| Level 2 | M1 | M2 |
|         | M1 | M3 | M3 |
| Level 3 | M1 | M2 | M3 |

**FIGURE 2** A diagram showing the matrices in each hierarchical level when $D = 3$. 

\[
\begin{align*}
\text{rank}([\mathbf{M}_1 \, \mathbf{M}_2 \, \mathbf{M}_3]) &\leq r_{123}, \quad \text{rank}([\mathbf{M}_1]) \leq r_{12}, \\
\text{rank}([\mathbf{M}_1 \, \mathbf{M}_3]) &\leq r_{13}, \quad \text{rank}([\mathbf{M}_2 \, \mathbf{M}_3]) \leq r_{23}, \\
\text{rank}(\mathbf{M}_1) &\leq r_1, \quad \text{rank}(\mathbf{M}_2) \leq r_2, \quad \text{rank}(\mathbf{M}_3) \leq r_3. \quad (5)
\end{align*}
\]

Various combinations of ranks in Equation (5) lead to intersection of column spaces with varying dimensions, allowing a wide range of signal structures. However, direct use of Equation (5) is difficult as: (i) rank constraints lead to a non-convex, NP-hard optimization problem (Candès and Recht, 2009; Mazumder et al., 2010); (ii) the total number of potential rank combinations for a given $r_{123}$ is $(r_{123} + 2^3 - 1)$ (Gaynanova & Li, 2019), making it practically hard to explore all models. Instead, in Section 2.4 we propose to replace each constraint in Equation (5) with corresponding nuclear norm penalty and only consider the set of models within the solution path.

### 2.4 Hierarchical nuclear norm penalization

Let each $\mathbf{X}_d$ be column-centered and scaled so that $\|\mathbf{X}_d\|_F = 1$ in line with common preprocessing of multiview data to accommodate potentially different scaling across views (Gaynanova & Li, 2019; Lock et al., 2013; Smilde et al., 2003). Nuclear norm penalization (Bach, 2008; Candès and Recht, 2009; Negahban & Wainwright, 2011) is commonly used for the estimation of the low-rank signal $\mathbf{M}_d$ in single-view settings since nuclear norm is a convex relaxation to rank leading to the convex optimization problem

\[
\begin{align*}
\min_{\mathbf{M}_d} &\left\{ \frac{1}{2} \|\mathbf{X}_d - \mathbf{M}_d\|_F^2 + \lambda_d \|\mathbf{M}_d\|_* \right\}, \quad (6)
\end{align*}
\]

where $\lambda_d \geq 0$ is a tuning parameter. Parallel to $\ell_1$-norm penalization for sparse vector estimation, the nuclear norm penalty in Equation (6) encourages sparse singular values for low-rank signal estimation. The solution to Equation (6) can be written in closed-form via the proximal operator of the nuclear norm (Polson et al., 2015):
Algorithm 1: Dual block-coordinate forward–backward algorithm for Equation (9)

Input: \( X = [X_1, X_2, X_3] \), \( \lambda_{123}, \lambda_{112}, \lambda_{132}, \lambda_{213}, \lambda_{213}, \lambda_{312} \geq 0 \), \( \epsilon > 0 \), \( \gamma \in (0, 2) \), \( t = 1 \).

Initialize with \( M^{(0)} = [M^{(0)}_1, M^{(0)}_2, M^{(0)}_3] = X \), \( D^{(0)} = \mathbf{0} \), \( D_d^{(0)} = \mathbf{0} \) for \( d = 1, 2, 3 \) and \( D_d^{(0)} = 0 \) for \( (k, l) = (1, 2), (1, 3), (2, 3) \).

repeat

\[
D_d^{(t)} = D_d^{(t-1)} + \gamma M_d^{(t-1)} - S \left( D_d^{(t-1)} + \gamma M_d^{(t-1)}, \lambda_d \right) \quad \text{for } d = 1, 2, 3
\]

\[
D_{kl}^{(t)} = D_{kl}^{(t-1)} + \gamma M_{kl}^{(t-1)} - S \left( D_{kl}^{(t-1)} + \gamma M_{kl}^{(t-1)}, \lambda_{kl} \right) \quad \text{for } M_{kl}^{(t-1)} = [M^{(t-1)}_k, M^{(t-1)}_l]
\]

\[
D^{(t)} = D^{(t-1)} + \gamma M^{(t-1)} - S \left( D^{(t-1)} + \gamma M^{(t-1)}, \lambda \right)
\]

\[
M^{(t)} = M^{(t-1)} - \left[ (D^{(t)}_1 - D^{(t-1)}_1) \cdot 0_{n \times p_1} \right] - \left[ (D^{(t)}_{13} - D^{(t-1)}_{13}) \cdot 0_{n \times p_2} \right] - \left[ (D^{(t)}_{13} - D^{(t-1)}_{13}) \cdot 0_{n \times p_2} \right]
\]

for the submatrices \( D_{13,1} \in \mathbb{R}^{n \times p_1} \), \( D_{13,2} \in \mathbb{R}^{n \times p_2} \) of \( D^{(t)} = [D^{(t)}_{13,1}, D^{(t)}_{13,2}] \)

\[
l = l + 1
\]

until \( \|M^{(t)} - M^{(t-1)}\|_F < \epsilon \).

Return: \( \bar{M} = [\bar{M}_1, \bar{M}_2, \bar{M}_3] = M^{(t)} \)

with the solution path covering a wide range of models with different configurations of hierarchical ranks. The choice of the tuning parameters will be discussed in Section 2.6.

Compared to existing approaches based on matrix factorizations that lead to non-convex optimization, our formulation (9) is convex. We specialize the dual block-coordinate forward–backward algorithm (Abboud et al., 2017) to solve Equation (9), which is guaranteed to converge to global optimum for our setting for any step size \( \gamma \in (0, 2) \). The main idea is that the proximal operator of the HNN can be evaluated efficiently because the norm is a sum of convex functions with linear operators. Algorithm 1 presents the pseudo-code, where each penalty term has an associated dual variable \( D \), and the algorithm alternates the updates of dual and primal variables. We set the step size \( \gamma = 0.5 \) and the numerical tolerance \( \epsilon = 10^{-5} \) in all our numerical studies. Web Appendix B.2 provides detailed derivations.

2.5 Refitting procedure

Like many penalized estimation methods, it is well-known that the nuclear norm penalization causes shrinkage bias (Chen et al., 2013; Josse & Sardy, 2016). In particular, since it penalizes both small and large non-zero singular values by equal amount, the so-called overshrinkage phenomenon could occur, where non-zero singular values are shrunken too much. This causes inaccurate estimation of signal as the non-zero singular values are likely to be estimated smaller than the original values.
Since our HNN penalty consists of the nuclear norms of multiple matrices, our approach is also affected by such overshrinkage phenomenon. Thus, we adopt a simple refitting step to “un-shrink” our estimate. The key idea is to extract estimated low-rank column space from preliminary \( \hat{M}_d \) from Algorithm 1, and then construct \( \hat{M}_d^{\text{refit}} \) with the given column space without penalization of singular values. This idea is formalized in Algorithm 2. Since the refitting step corresponds to solving the least-squares problem (10) with a closed-form solution, it does not require a significant amount of computation. As illustrated in Figure S1 of Web Appendix C.1, while the singular values of each \( \hat{M}_d \) are smaller than the true values, our refitting procedure is effective in correcting the bias.

### 2.6 Selection of tuning parameters

The tuning parameters in Equation (9) need to be carefully selected to ensure good performance. If the nuclear norms are over-penalized, there is a degradation in signal estimation performance due to underestimated ranks (which the above refitting procedure cannot correct). On the other hand, if the tuning parameters are too small, the ranks are overestimated, and the estimated signals contain noise. One clear selection difficulty is a large number of total parameters. Even a modest \( D = 3 \) case requires seven parameters in Equation (9). Tuning all these parameters freely via cross-validation or a model selection criterion is computationally prohibitive.

We propose to reduce the total number of tuning parameters from \( 2^D - 1 \) to only \( D \) by taking advantage of the Stein’s unbiased risk estimates (SURE) (Candès et al., 2013; Josse & Sardy, 2016). The SURE formulation for single-view nuclear-norm minimization allows us to derive a closed form of optimal \( \lambda_d^{\text{SURE}} \) to use in \( S(X_d, \lambda_d) \) (7) as a minimizer of SURE criterion, see Candès and Recht (2009) for the closed-form of \( \lambda_d^{\text{SURE}} \) and derivation details. While SURE formulation for HNN minimization is intractable, we propose to take advantage of closed-form \( \lambda_d^{\text{SURE}} \) in the single matrix case to determine the relative weights that should be assigned to each matrix within each hierarchical level. Specifically, let \( \lambda_d^{\text{SURE}} \) and \( \lambda_{kl}^{\text{SURE}} \) be the SURE tuning parameters for \( S(X_d, \lambda_d) \) and \( S(X_{kl}, \lambda_{kl}) \), respectively, where \( X_{kl} = [X_k X_l] \). We propose to reparameterize penalty in Equation (9) as

\[
\tau \sum_{d=1}^{3} \omega_d^{\text{SURE}} ||M_d||_s + \chi \sum_{k,l=1,k<l}^{3} \omega_{kl}^{\text{SURE}} ||M_{kl}||_s + \lambda ||M||_s
\]

(11)

where \( \tau, \chi > 0 \) control the overall shrinkage amount for the individual and pairwise matrices, respectively, and the relative weights are chosen as

\[
\omega_d^{\text{SURE}} = \frac{\lambda_d^{\text{SURE}}}{\lambda_1^{\text{SURE}} + \lambda_2^{\text{SURE}} + \lambda_3^{\text{SURE}}}, \quad \omega_{kl}^{\text{SURE}} = \frac{\lambda_{kl}^{\text{SURE}}}{\lambda_{12}^{\text{SURE}} + \lambda_{13}^{\text{SURE}} + \lambda_{23}^{\text{SURE}}}
\]

(12)

That is, we form a convex combination of the nuclear norms within each hierarchical level in the penalty (11). The approach for \( D > 3 \) is similar. By design, \( \tau, \chi, \lambda \) control the amount of penalization across different hierarchical levels, while the weights (12) induce (potentially) differential shrinkage within the same level based on SURE. Since the weights are closed form, only three parameters \( (\tau, \chi, \lambda) \) remain to be tuned.

To choose the remaining \( D \) tuning parameters, we adapt the BCV (Owen and Perry, 2009), which is also used in SLIDE (Gaynanova & Li, 2019). The original BCV is designed for rank selection in the single-view case, and can be viewed as matrix extension of \( k \)-fold cross-validation. The \( (j \times k) \)-fold BCV splits the rows and the columns of \( X \) into \( j \) and \( k \) blocks, respectively. The sub-matrix corresponding to each block is hold out as test data, with the remaining submatrices used to fit the model of specified rank and make predictions on the test. To illustrate BCV in our case, we focus on the \( 2 \times 2 \) folds.
**Algorithm 3** Proposed 2 × 2 BCV procedure for evaluation of prediction errors

Input: X with random split as (13) and \( \{(\tau_i, \kappa_i, \lambda_i)\}_{l=1}^{L} \)

for \( j, k = 1, 2 \) do

Calculate \( Z_{d}^{j,−k} \) by column-centering and scaling \( X_{d}^{j,−k} \) so that \( \|Z_{d}^{j,−k}\|_F = 1 \)

for \( l = 1, 2, \ldots, L \) do

(i) Obtain \( \hat{M}_{d}^{\text{refit}}_{j,−k,l} \), \( \hat{M}_{d,−j,k,l}^{\text{refit}} \), \( \hat{M}_{d,−j,−k,l}^{\text{refit}} \) by back-scaling and back-centering the resulting matrix from Algorithm 2 with \( Z_{d}^{j,−k} \) and \( (\tau_i, \kappa_i, \lambda_i) \).

(ii) Evaluate the average BCV error

\[
\text{BCVErr}_{j,k,l} = \frac{1}{3} \sum_{d=1}^{3} \frac{\|X_{d}^{j,k} - X_{d}^{j,−k}(\hat{M}_{d,−j,k,l}^{\text{refit}}) + X_{d}^{−j,k}\|^2_2}{\|X_{d}^{j,k}\|^2_2}.
\]

where the superscript + denotes the pseudo-inverse of the corresponding matrix.

end for

end for

Return: \( \{\text{BCVErr}_{j,k,l} : j, k = 1, 2\}_{l=1}^{L} \)

---

where \( X_{d}^{j,k} \) denotes the sub-matrix of \( X_{d} \) corresponding to the \( j \)th row fold and \( k \)th column fold. Algorithm 3 summarizes the corresponding BCV procedure. The main difference with original BCV is that the splits are distributed equally across the views. Suppose we hold out \( X_{d}^{j,k} \) for all \( d = 1, 2, 3 \). Given the tuning parameters \( (\tau_i, \kappa_i, \lambda_i) \), we obtain the estimates \( \hat{M}_{d,−j,−k,l}^{\text{refit}} \) based on the submatrix of \( X \) that shares no rows or columns with held-out \( X_{d,−j,−k}^{(d)} \) (this submatrix is denoted by \( \hat{X}_{d,−j,−k}^{(d)} \) in Algorithm 3).

We then use both \( X_{d}^{j,−k} \) (sharing only the rows with \( X_{d}^{j,k} \)) and \( X_{d}^{−j,k} \) (sharing only the columns with \( X_{d}^{j,k} \)) to evaluate the prediction error BCVErr\(_{j,k,l}\) (14) for \( X_{d}^{j,k} \) as in Owen and Perry (2009). Given the grid of tuning parameters \( \{(\tau_i, \kappa_i, \lambda_i)\}_{l=1}^{L} \), we calculate the average BCV error \( \text{Avg-Err}_l = 4^{-1} \sum_{j,k,l=1}^{D} \text{BCVErr}_{j,k,l} \) for combination \( l \), and choose optimal \( (\tau^*, \kappa^*, \lambda^*) \) based on 1 standard error rule (Hastie et al., 2009), see Web Appendix C.2 for more details.

### 3 | SIMULATION STUDIES

We conduct simulations to evaluate the performance of our approach. We label our approach as HNN and select its tuning parameters as described in Section 2.6. To investigate the effect of our tuning parameter selection, we also consider our estimator with the smallest error (15) and label it as HNN_best. For comparison, we further implement the other competing methods as follows: (a) JIVE (Lock et al., 2013) with the default permutation method for rank selection and true ranks (labeled as JIVE_true); (b) AJIVE (Feng et al., 2018); (c) SLIDE (Gaynanova & Li, 2019); (d) BIDIFAC (Park & Lock, 2020) (e) BIDIFAC+ (Lock et al., 2022). Additional implementation details are discussed in Web Appendix D.

We simulate each data matrix \( X_{d} \) via the model \( X_{d} = M_{d} + E_{d} \), where each element of the noise matrix \( E_{d} \) is generated independently from \( N(0, \sigma_{d}^{2}) \). The noise level \( \sigma_{d}^{2} \) is set to satisfy the chosen signal-to-noise ratio (SNR), where \( \text{SNR} = ||M_{d}||_F^2/||E_{d}||_F^2 = ||M_{d}||_F^2/(\sigma_{d}^2 n p_d) \). We vary the number of views \( (D = 2 \text{ or } 3) \), and consider two cases for each \( D \): (i) orthogonal (all joint, partially-shared, and individual structures are orthogonal); (ii) non-orthogonal (some structures are non-orthogonal). Web Appendix D provides additional data generation details. Given an estimated \( \hat{M} = [\hat{M}_1, \ldots, \hat{M}_D] \), we compare all methods in terms of the estimated ranks and the scaled squared Frobenius norm error defined by

\[
L_F(M, \hat{M}) = \sum_{d=1}^{D} \frac{1}{\|M_d\|_F^2} \|M_d - \hat{M}_d\|_F^2.
\]

Figures 3 shows the results for \( D = 2 \) setting over 100 replications. In the orthogonal case, AJIVE and SLIDE are the best performing methods, followed closely by HNN and BIDIFAC+. The good performance of SLIDE is expected in the orthogonal case as the identifiability conditions are satisfied. The good performance of AJIVE can be attributed to its accurate ranks estimation, as no partially-shared signals exist when \( D = 2 \). JIVE and BIDIFAC+ perform the worst. Comparing the ranks reveals that JIVE overestimates the rank of \( M \) while correctly estimating the ranks of individual \( M_d \), hence it under-estimates the rank of the joint structure. Since JIVE with true ranks (JIVE_true) performs...
as well as AJIVE and SLIDE, the results indicate that JIVE’s poor performance in this setting is due to rank mis-identification. As expected, BIDIFAC+ performs significantly worse than BIDIFAC+_refit due to the bias associated with the penalization in original BIDIFAC+ solution. In the non-orthogonal case, the performance of both AJIVE and SLIDE deteriorates, with HNN becoming the best performing method. The deterioration of SLIDE is expected as non-orthogonal individual structures violate SLIDE’s identifiability conditions (Gaynanova & Li, 2019). The deterioration of AJIVE could be attributed to rank mis-identification, as it significantly under-estimates the total rank of $\mathbf{M}$. The median estimated ranks of HNN always match the true ranks (in both orthogonal and non-orthogonal cases).

Figure 4 shows the results for $D = 3$ setting over 100 replications. Unlike $D = 2$ setting, HNN performs better than AJIVE in both orthogonal and non-orthogonal cases, and is only second to SLIDE in the orthogonal case (albeit close in both errors and estimated ranks). AJIVE’s worse performance is expected since it does not account for partially-shared structures. JIVE performs better or similar to AJIVE, but worse than HNN, SLIDE, and BIDIFAC+_refit. As expected, the refitted versions of BIDIFAC are better than the original ones, with BIDIFAC+ performing better than BIDIFAC. As in $D = 2$ case, SLIDE’s performance is significantly worse in non-orthogonal case due to identifiability issues. In the non-orthogonal case, the median estimated ranks for HNN are the closest to true ranks.
Surprisingly, BIDIFAC and BIDIFAC+ have significantly worse performance than other methods in our experiments. This is due to the bias caused by the penalization as the resulting singular values from both BIDIFAC and BIDIFAC+ are shrunk much too much relative to the true values. In Figures 3 and 4, both BIDIFAC_refit and BIDIFAC+_refit show much smaller Frobenius norm errors compared to the corresponding results without refitting.

Overall, the proposed HNN performs the best in terms of median estimation errors and median estimated ranks. As expected, the errors of oracle HNN_best are always better. However, the difference between the two medians is negligible compared to the difference with other methods. This supports the good performance of the proposed tuning parameter selection procedure. At the same time, HNN shows high variability across replications, especially in the non-orthogonal two-view case. As shown in Web Appendix D, this extra variability is due to the randomness in BCV splits. An occasional “poor” split may lead to rank under-estimation, subsequently leading to higher errors, whereas a different split on the same data has no such issue. In practice, we recommend running BCV with several split choices to assess the ranks that are selected most consistently across splits. In the orthogonal case, SLIDE has an advantage over HNN as it has similar or slightly higher accuracy, and lower variance. However, SLIDE performs significantly worse in non-orthogonal cases, which is further supported by additional results in Web Appendix E. Since in practice it is difficult to assess a priori whether orthogonality condition holds, HNN has an advantage over SLIDE.

**FIGURE 4** Boxplots of scaled squared Frobenius norm errors (15) and rank estimates of the concatenated signals and each $\mathbf{M}_d$ over 100 independent replication for orthogonal scheme (left column) and non-orthogonal scheme (right column) when $D = 3$. The true ranks are indicated by the red dotted line. [This figure appears in color in the electronic version of this paper, and any mention of color refers to that version].
TABLE 1 The rank estimates of each view with the percentage of explained variation (Rows 1–3) and the estimated ranks of each signal structures (Rows 4–10).

|                | JIVE   | AJIVE  | SLIDE  | HNN    | BIDIFAC w/wo refit | BIDIFAC+ w/wo refit |
|----------------|--------|--------|--------|--------|--------------------|---------------------|
| Muscle         | 13 (66.3%) | 12 (66.9%) | 16 (69.2%) | 16 (71.4%) | 13 (64.4%/18.8%) | 16 (65.6%/19.0%)   |
| Blood          | 11 (72.4%) | 5 (68.0%)  | 12 (77.7%) | 17 (82.7%) | 11 (75.3%/30.3%) | 12 (75.7%/30.2%)   |
| Skin           | 15 (61.1%) | 11 (66.4%) | 19 (73.3%) | 19 (75.7%) | 14 (67.4%/16.4%) | 17 (68.4%/16.6%)   |
| Joint          | 2      | 1      | 3      | 1      | 5                  | 2                   |
| Muscle and blood | NA    | NA     | 0      | 3      | NA                | 2                   |
| Muscle and skin | NA    | NA     | 0      | 2      | NA                | 5                   |
| Blood and skin  | NA    | NA     | 1      | 2      | NA                | 2                   |
| Muscle only    | 11    | 11     | 13     | 10     | 8                 | 7                   |
| Blood only     | 9     | 4      | 8      | 11     | 6                 | 6                   |
| Skin only      | 13    | 10     | 15     | 14     | 9                 | 8                   |

Note: For both BIDIFAC and BIDIFAC+, we present the corresponding results with and without refitting.

4 ANALYSIS OF MULTI-TISSUE GENE EXPRESSION DATA

The p53 gene plays an important role in cell regulation with impact on aging processes and cancer (Seim et al., 2016; Tanikawa et al., 2017). However, the tissue-specific nature of p53 gene expressions makes it difficult to develop targeted therapies (Vlatkovic et al., 2011). Thus, it is crucial to identify patterns in p53 gene expression that are shared and unique across multiple tissues. To identify such patterns, we analyze gene expression data corresponding to p53 signaling pathway from three tissues (muscle, blood, and skin). The original data are available from The GTEx Consortium (2020), we use the preprocessed data by Li and Jung (2017) (see Web Appendix F for preprocessing details and how it affects the scaling of various methods). The resulting data contain measurements on 191 genes from 204 matched samples corresponding to three tissues: $X_1 \in \mathbb{R}^{204 \times 191}$ for muscle, $X_2 \in \mathbb{R}^{204 \times 191}$ for blood, and $X_3 \in \mathbb{R}^{204 \times 191}$ for skin. Our goal is to identify shared, partially-shared, and individual signal structures across three tissues, and compare the results across different methods.

In the top half of Table 1, we compare HNN with other methods on each view in terms of (i) the total estimated rank; (ii) the percentage of explained variation (the ratio of the squared Frobenius norm of the estimated signal to the squared Frobenius norm of data). JIVE, AJIVE, and BIDIFAC estimate smaller ranks compared to SLIDE, BIDIFAC+, and HNN. Furthermore, HNN has the highest total ranks for each view, with highest percentage of variance explained. While larger explained variance can be partially attributed to larger ranks, this is not the case for muscle and skin tissues. For muscle, HNN has higher percentage of variance explained than SLIDE and BIDIFAC+ _refit with the same rank of 16. For skin, HNN has higher percentage than SLIDE with the same rank of 19. To provide a more fair comparison, we truncate HNN estimates using truncated singular value decomposition (SVD) to match the ranks of other methods. These adjusted HNN estimates still explain more variation in each view than other methods (Figure S5 in Web Appendix F), while maintaining non-trivial shared patterns (unlike separate SVD). These comparisons suggest that HNN provides a better fit to the data among methods that account for shared structures.

The last seven rows in Table 1 show the ranks of the corresponding structures identified by each method. For all methods, individual structures have higher ranks than shared structures, which is consistent with known tissue-specificity of p53 gene expressions (Seim et al., 2016; Tanikawa et al., 2017). HNN identifies more partially-shared structures compared to SLIDE but less than BIDIFAC+. Since JIVE, AJIVE, and BIDIFAC are unable to identify partially-shared structures directly, we also apply these methods separately to each pair of views. If the joint rank estimated on a pair of views is larger than the corresponding joint ranks for all three views, there is evidence for partially-shared structures. Table 2 shows estimated ranks for JIVE, AJIVE, and BIDIFAC when applied to each pair of views. For muscle and skin, the joint ranks (3 for JIVE, 3 for AJIVE, and 6 for BIDIFAC) are larger than the corresponding joint ranks for all three views (2 for JIVE, 1 for AJIVE, and 5 for BIDIFAC), suggesting a missing partially-shared structure between the muscle and skin tissues. From Table 1, BIDIFAC+ and HNN are the only methods that identify partial-sharing between muscle and skin. To compare how partially-shared structures from HNN relate to other methods, in Figure S6 in Web Appendix F we illustrate cosines of principal angles between HNN partially-shared structures and the corresponding individual structures from the other approaches. For both muscle...
and blood and muscle and skin, we find that rank one subspace of corresponding partially-shared structure from HNN is identified as individual for muscle by JIVE and AJIVE (cosine value close to one). For blood and skin, the partially-shared structures of HNN are distinct, and are not identified as individual by any of the other methods.

The numerical results in Section 3 coupled with comparisons of HNN with other methods in terms of variance explained and identified ranks suggest that HNN provides new insights into specificity of p53 pathway gene expressions across tissues. The presence of partially-shared structures across each pair of tissues supports the existing research that each pair has unique similarity depending on the underlying characteristics used for comparison. Sonawane et al. (2017) clustered tissues into communities based on tissue-specific targeting patterns: both skin and blood tissues belong to the cell proliferation community, whereas muscle tissue belongs to a distinct extracellular structure community. The GTEx Consortium (2020) considered pairwise associations between tissues based on cis-eQTLs, expression quantitative trait loci located near the gene of origin. They found that skin and muscle have a high pairwise association, whereas the cis-eQTL-based association of each tissue with blood was considerably lower. Oliva et al. (2020) considered the impact of gender on tissue-expressions, with considerably higher number of sex-biased genes in skin tissue compared to muscle and blood tissues. Consequently, they found that muscle and blood tissues co-cluster based on the effect size of sex-biased genes, with skin tissue being significantly more distinct. While these prior analyses were not restricted to p53 signaling pathway, it is known that p53 pathway plays a crucial role in cell proliferation (Seim et al., 2016) and is enriched for sex-biased genes (Lopes-Ramos et al., 2020). The identified shared, partially-shared, and individual signal by HNN provide a principled data-driven foundation for further investigation of distinct biological characteristics affecting p53 expression.

5 | CONCLUSION

We introduce a formal model for joint, partially-shared, and individual signal structures in multi-view data based on hierarchical levels. We propose a new HNN penalization approach for signal estimation and a simple refitting step for bias adjustment. Simulation studies indicate superior performance of our method in terms of signal recovery and rank estimation, particularly when not all signal structures are orthogonal.

While the proposed BCV-based tuning parameter selection scheme performs well in our numerical studies, a large number of tuning parameters still needs to be chosen, leading to high computational burden. It would be of interest to explore a way of choosing tuning parameters that is analogous to the SURE approach in single-view nuclear norm penalization. This approach, however, is quite challenging because Algorithm 1 does not admit a closed-form solution. Given the tuning parameters, the proposed algorithm can be applied with any number of views $D$, but in practice we expect computational bottlenecks when $D > 5$. In such cases, we recommend dropping some hierarchical levels by setting corresponding tuning parameters in Equation (8) to zero. Such reduction should be dictated by the scientific question of interest and analyses goals. Theoretical analysis of our method is also an interesting future study, which would be possible thanks to considerable development of theories in single-view nuclear norm penalization (Bach, 2008; Negahban & Wainwright, 2011).

ACKNOWLEDGMENTS

This research was supported by NSF CCF-1934904. Irina Gaynanova was partially supported by NSF DMS-2044823.

DATA AVAILABILITY STATEMENT

The GTEx data (The GTEx Consortium, 2015) that supports the findings in this paper have been preprocessed by Li and Jung (2017) and are openly available at http://github.com/reagan0323/SIFA.

ORCID

Raymond Ka Wai Wong https://orcid.org/0000-0001-9342-3755
Irina Gaynanova https://orcid.org/0000-0002-4116-0268

REFERENCES

Abboud, F., Chouzenoux, E., Pesquet, J.C., Chenot, J.H. & Laborelli, L. (2017) Dual block-coordinate forward–backward algorithm with application to deconvolution and deinterlacing of video

| TABLE 2 | The estimated ranks of the structures from JIVE, AJIVE, and BIDIFAC with each pairs of the GTEx data. |
|---------|---------------------------------------------------------------|
|         | JIVE | Indiv 1 | Indiv 2 | AJIVE | Indiv 1 | Indiv 2 | BIDIFAC | Indiv 1 | Indiv 2 |
| Muscle and blood | Joint | 2 | 11 | 9 | 1 | 11 | 4 | 4 | 8 | 6 |
| Muscle and skin | Joint | 3 | 11 | 15 | 3 | 9 | 8 | 6 | 7 | 8 |
| Blood and skin | Joint | 2 | 8 | 13 | 1 | 4 | 10 | 3 | 6 | 9 |
sequences. Journal of Mathematical Imaging and Vision, 59(3), 415–431.

Bach, F.R. (2008) Consistency of trace norm minimization. Journal of Machine Learning Research, 9, 1019–1048.

Candès, E.J. & Recht, B. (2009) Exact matrix completion via convex optimization. Foundations of Computational Mathematics, 9(6), 717–772.

Candès, E.J., Sing-Long, C.A. & Trzasko, J.D. (2013) Unbiased risk estimates for singular value thresholding and spectral estimators. IEEE Transactions on Signal Processing, 61(19), 4643–4657.

Chen, K., Dong, H. & Chan, K.S. (2013) Reduced rank regression via adaptive nuclear norm penalization. Biometrika, 100(4), 901–920.

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

Gaynanova, I. & Li, G. (2019) Structural learning and integrative decomposition of multi-view data. Biometrics, 75(4), 1121–1132.

Hastie, T., Tibshirani, R., Friedman, J.H. & Friedman, J.H. (2009) The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer.

Jia, Y., Salzmann, M. & Darrell, T. (2010) Factorized latent spaces with structured sparsity. Advances in Neural Information Processing Systems, 23, 982–990.

Josse, J. & Sardy, S. (2016) Adaptive shrinkage of singular values. Statistics and Computing, 26(3), 715–724.

Li, G. & Jung, S. (2017) Incorporating covariates into integrated factor analysis of multi-view data. Biometrics, 73(4), 1433–1442.

Lock, E.F., Hoadley, K.A., Marron, J.S. & Nobel, A.B. (2013) Joint and individual variation explained (jive) for integrated analysis of multiple data types. Annals of Applied Statistics, 7(1), 523–542.

Lock, E.F., Park, J.Y. & Hoadley, K.A. (2022) Bidimensional linked matrix factorization for pan-omics pan-cancer analysis. Annals of Applied Statistics, 16(1), 193–215.

Lopes-Ramos, C.M., Quackenbush, J. & DeMeeo, D.L. (2020) Genome-wide sex and gender differences in cancer. Frontiers in Oncology, 10, 597788.

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

Negahban, S. & Wainwright, M.J. (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Annals of Statistics, 39(2), 1069–1097.

Oliva, M., Muñoz-Aguirre, M., Kim-Hellmuth, S., Wucher, V., Gewirtz, A.D., Cotter, D.J. et al. (2020) The impact of sex on gene expression across human tissues. Science, 369(6509), eaba3066.

Owen, A.B. & Perry, P.O. (2009) Bi-cross-validation of the SVD and the nonnegative matrix factorization. Annals of Applied Statistics, 3(2), 564–594.

Park, J.Y. & Lock, E.F. (2020) Integrative factorization of bidimensionally linked matrices. Biometrics, 76(1), 61–74.

Polson, N.G., Scott, J.G. & Willard, B.T. (2015) Proximal algorithms in statistics and machine learning. Statistical Science, 30(4), 559–581.

Seim, I., Ma, S. & Gladyshev, V.N. (2016) Gene expression signatures of human cell and tissue longevity. npj Aging and Mechanisms of Disease, 2(1), 1–8.

Smiilde, A.K., Westerhuis, J.A. & De Jong, S. (2003) A framework for sequential multiblock component methods. Journal of Chemometrics, 17(6), 323–337.

Sonawane, A.R., Platig, J., Fagny, M., Chen, C.Y., Paulson, J.N., Lopes-Ramos, C.M., DeMeeo, D.L., Quackenbush, J., Glass, K. & Kuijjer, M.L. (2017) Understanding tissue-specific gene regulation. Cell Reports, 21(4), 1077–1088.

Srebro, N. & Jaakkola, T. (2003) Weighted low-rank approximations. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 720–727.

Tanikawa, C., Zhang, Y.Z., Yamamoto, R., Tsuda, Y., Tanaka, M., Funauchi, Y. et al. (2017) The transcriptional landscape of p53 signalling pathway. EBioMedicine, 20, 109–119.

The GTEx Consortium (2015) Genotype-tissue expression data. Available at: http://www.gtexportal.org/. dbGaP Accession phs000424.v6.

The GTEx Consortium, (2020) The gtex consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318–1330.

Van Deun, K., Wilderjans, T.F., Van Den Berg, R.A., Antoniadis, A. & Vande Mechemel, I. (2011) A flexible framework for sparse simultaneous component based data integration. BMC Bioinformatics, 12(1), 1–17.

Vlatkovic, N., Crawford, K., Rubbi, C. & Boyd, M. (2011) Tissue-specific therapeutic targeting of p53 in cancer: one size does not fit all. Current Pharmaceutical Design, 17(6), 618–630.

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

Yuan, D. & Gaynanova, I. (2022) Double-matched matrix decomposition for multi-view data. Journal of Computational and Graphical Statistics, 31(4), 1114–1126.

Zhou, G., Cichocki, A., Zhang, Y. & Mandic, D.P. (2016) Group component analysis for multiblock data: common and individual feature extraction. IEEE Transactions on Neural Networks and Learning Systems, 27(11), 2426–2439.

SUPPORTING INFORMATION
Web Appendices, Figures, and Examples referenced in Sections 1–4, and R codes with corresponding analyses are available with this paper at the Biometrics website on Wiley Online Library. R codes are also available at http://github.com/sangyoonstat/HNN_paper.