Data Integration Via Analysis of Subspaces (DIVAS)

Jack Prothero, Meilei Jiang, Jan Hannig,
Quoc Tran-Dinh, Andrew Ackerman, J.S. Marron

August 7, 2023

Abstract

Modern data collection in many data paradigms, including bioinformatics, often incorporates multiple traits derived from different data types (i.e. platforms). We call this data multi-block, multi-view, or multi-omics data. The emergent field of data integration develops and applies new methods for studying multi-block data and identifying how different data types relate and differ. One major frontier in contemporary data integration research is methodology that can identify partially-shared structure between sub-collections of data types. This work presents a new approach: Data Integration Via Analysis of Subspaces (DIVAS). DIVAS combines new insights in angular subspace perturbation theory with recent developments in matrix signal processing and convex-concave optimization into one algorithm for exploring partially-shared structure. Based on principal angles between subspaces, DIVAS provides built-in inference on the results of the analysis, and is effective even in high-dimension-low-sample-size (HDLSS) situations.

1 Introduction

Modern experiments are increasingly likely to produce complex data derived from multiple sources. One common example is a single group of n *objects* (i.e. cases, observations, patients) being observed across K different *views* or *data blocks*, each with their own sets of $d_k,\ k = 1, \ldots, K$, *traits* (i.e. variables, features, descriptors) and measurement methodologies. We call data collected and organized this way *multi-block data*. Simple data analysis approaches would concatenate multiple data blocks into a single data matrix of n data objects and $d_1 + \cdots + d_K$ traits. However, these approaches ignore the often important relationships between the views and obscure insights on which information comes from which data block. We specifically aim to address this challenge by considering the data blocks as separate units and searching for *shared structure* between them. Shared structure can be defined in several ways. Here and in some other contemporary approaches seen in Section 1.1 it means common modes of variation modeled by low-rank matrices. More mathematical details of this modeling choice can be found in Section 2. Our proposed method, *Data Integration Via Analysis of Subspaces* (DIVAS), incorporates state-of-the-art advances in matrix perturbation theory and optimization to provide insights about both shared and partially-shared joint structure between several data blocks.

DIVAS gives a novel approach for finding structure in a multi-block data set based on searching for shared subspaces between different collections of data blocks. The data blocks are intrinsically linked by having the same *trait space* $\mathbb{R}^n$, so we primarily search for shared subspaces within the trait space. In contrast to other approaches, angles form the foundation of our analysis of the relationships between these subspaces. In particular, *principal angles* are the measure of choice of proximity between subspaces. These angles are applied in a rigorous framework of inference that provides relevant statistical significance determinations for the chosen subspaces.

Subspaces of trait space have corresponding induced subspaces of *object space* $\mathbb{R}^{d_k}$ for each data block. Combined together, these subspaces can be decomposed into *modes of variation*. A mode of variation is a rank 1 matrix formed from the outer product of two vectors: one in object space and one in trait space. In the terminology of Principal Components Analysis (PCA) these would be a loadings (direction) vector and a scores vector, respectively. Considering subspaces in terms of modes of variation is particularly useful for visualization. The scores vectors demonstrate relationships between the data objects, and the loadings vectors provide information about which

traits are driving the variation. The ultimate result of a DIVAS exploration of a data set is a set of modes of variation for each data block associated with each block collection.

The rest of the paper proceeds as follows. The remainder of Section 1 continues with a background information on data integration in general. Section 2 details DIVAS methodology and demonstrates the method's performance on a synthetic data set. Section 3 provides a prototypical application of DIVAS in cancer genomics. Section 4 contains a case study on twentieth century mortality using DIVAS. Section 5 summarizes some brief conclusions. The Appendices A, B, C, and D provide reference materials on random matrix theory, principal angle analysis, residual matrix estimation, and details on the optimization problem solved by DIVAS, respectively. MATLAB code is available for download at `https://github.com/jbprothero/DIVAS2021`.

1.1 Data Integration Literature

A time-honored multi-block data analysis method is Canonical Correlations Analysis (CCA), proposed by (Hotelling, 1936). Given two blocks of data $\mathbf{X}$ and $\mathbf{Y}$, CCA seeks to maximize the Pearson correlation between vectors from the span of each data block in trait space. The fundamental ideas of CCA have been thoroughly extended to more general settings. Several authors propose different generalizations of CCA for locating highly correlated structure between three or more data blocks, including (Horst, 1961; Kettenring, 1971; Nielsen, 2002). Some others have experimented with kernel methods for CCA, as in (Akaho, 2007) and (Cai and Huang, 2017).

In the context of machine learning, methods like CCA are termed *multi-view methods*. Multi-view learning broadly includes methods like co-training for semi-supervised learning (Blum and Mitchell, 1998), SVM-2K (Farquhar et al., 2005), subspace learning (White et al., 2012), and other multi-view extensions of paradigms like active learning and ensemble learning. See Sun (2013), Xu et al. (2013), and Li et al. (2019) for more details on these extensions.

Any CCA-based method is ultimately focused on finding jointly shared structure between each available block or view of the data. Oftentimes, given low-rank approximations of each data block we are also interested in a full factorization of the signal present in each block into a joint component shared between all blocks and an individual component unique to each block. One algorithm that produces such a factorization is Joint and Individual Variation Explained (JIVE) (Lock et al., 2013).

After initially choosing a signal rank for each data matrix using a permutation testing approach, the algorithm seeks to minimize residual energy by alternating between determining joint structure and individual structure. Broadly the algorithm accomplishes its goal, but the optimization problem can proceed slowly and there is no underlying inferential justification for the chosen boundary between joint and individual structure.

A later generation of JIVE, dubbed Angle-based JIVE (AJIVE) (Feng et al., 2018), was proposed to address the above shortcomings. Selection of joint structure happens in a quick, single step based on principal angle analysis and the delineation between joint and individual structure is based on a bound on the angles between original and perturbed subspaces found in Wedin (1972). Initial rank selection, however, is performed ad hoc in a separate initial step, and as described in Feng et al. (2018) the perturbation angle bounds used can become extremely conservative under rank mis-specification. Additionally, neither JIVE nor AJIVE consider partially-shared joint structure between subsets of blocks.

Decomposition of data blocks into partially-shared joint structure components is one of the primary frontiers in contemporary data integration research. Two approaches to the problem, (Gaynanova and Li, 2019) and (Zhao et al., 2016), both model partially-shared information via structured sparsity in a basis matrix for the concatenated data blocks. Gaynanova and Li (2019) determine sparse structure via bi-cross-validation, while Zhao et al. (2016) determine sparse structure via a collection of Bayesian priors. A third approach is found in the recent unpublished manuscript (Yi et al., 2022). Their method models partially-shared information via subspace intersections and a hierarchical matrix nuclear norm regularization scheme. This formulation eschews factorization into scores and loadings subspaces entirely in order to work with a convex optimization problem, to capture potentially non-orthogonal partially-shared structure across block collections, and to ensure identifiability. DIVAS maintains identifiability even among potentially non-orthogonal partially-shared structures via a sequential search through each collection of data blocks. DIVAS also involves a more challenging non-convex optimization problem based on factorized matrices, but in doing so achieves relevant statistical significance measurements for the resulting shared and partially-shared subspaces.

Another recent pursuit in data integration research is complete incorporation of all the information from the data blocks. In most cases the data blocks share the same number of data objects,

so integrative analysis often takes place primarily in trait space, with corresponding information about the contributions of certain traits to shared structure being determined subsequently. In many cases, information about the contributing traits is just as pertinent as the shared structure itself, including if data blocks are bi-dimensionally linked as in (Lock et al., 2020) or bi-dimensionally matched as in (Yuan and Gaynanova, 2021). This is often the case in bioinformatics where the traits represent measurements of particular genes and the primary goal is to identify genes or other biological factors that contribute to patterns observed across the data blocks. The above papers each propose their own method for incorporating trait information in the analysis in the situations where the data blocks are appropriately linked. Another methodology found in (Shu and Qu, 2021) attempts to incorporate trait information for more general multi-block situations, but it relies on a computationally-taxing row-matching algorithm as part of its procedure and the method cannot parse partially-shared joint structure. Our method utilizes subspace perturbation theory that applies along either dimension of the data blocks and in any scenario with mutli-block data. This makes trait information easy to incorporate throughout the algorithm that locates partially-shared structure.

2 Methodology

Let $\mathbf{X}_1, \ldots, \mathbf{X}_K$ be data blocks each containing the same set of n data objects and distinct sets of $d_1, \ldots, d_K$ traits. In matrix calculations we follow the bioinformatics convention where matrix columns are data objects and matrix rows are traits (i.e. the matrix $\mathbf{X}_k$ has d_k rows and n columns). In our data model, shown in (1), each data block is assumed to be a low-rank signal matrix $\mathbf{A}_k$ plus a full-rank noise matrix $\mathbf{E}_k$:

$$\mathbf{X}_k = \mathbf{A}_k + \mathbf{E}_k. \tag{1}$$

We assume each entry of $\mathbf{E}_k$ is independent with identical variance σ^2. For inferential purposes, we also assume *rotational invariance* between signal and noise data matrices. The mathematical details of this assumption are provided in Section 2.1.3 where they are maximally relevant.

Further discussion of the modeling assumptions for DIVAS requires notation for describing different collections of data blocks in detail. Consider the power set $2^{\{1,\ldots,K\}}$ as a set of index sets, where each element $\mathbf{i} \in 2^{\{1,\ldots,K\}}$ represents a particular collection of data block indices. Each $\mathbf{i}$

indexes a shared structure among the data blocks $\mathbf{X}_k$ with $k \in \mathbf{i}$. Denote by $|\mathbf{i}|$ the cardinality of $\mathbf{i}$.

In order to define structure partially-shared across the data blocks, we decompose the signal matrices $\mathbf{A}_k,\ k = 1 \ldots, K$ into a sum of low-rank signal matrices, each of which corresponds to the joint structure shared between the collection of blocks indicated by the index $\mathbf{i}$:

$$\mathbf{A}_k = \sum_{\mathbf{i}|k\in\mathbf{i}} \mathbf{L}_{\mathbf{i},k}\mathbf{V}_\mathbf{i}^\top. \tag{2}$$

Here the sum extends over all index sets $\mathbf{i} \in 2^{\{1,\ldots,K\}}$ that satisfy $k \in \mathbf{i}$. The $n \times r_\mathbf{i}$ *scores matrices* $\mathbf{V}_\mathbf{i}$ model the shared structure in the trait space between the data blocks $\mathbf{X}_k$ with $k \in \mathbf{i}$. The $d_k \times r_\mathbf{i}$ *loadings matrices* $\mathbf{L}_{\mathbf{i},k}$ contain the induced object space structure in each block $\mathbf{X}_k$ with $k \in \mathbf{i}$. In order to ensure identifiability of the decomposition (2), the factorized signal matrices $\mathbf{L}_{\mathbf{i},k}$ and $\mathbf{V}_\mathbf{i}$ are required to satisfy the following conditions: (Here and for the rest of the manuscript the notation $[\mathbf{V}_\mathbf{i}]_{\mathbf{i}|\mathbf{i}\in S}$ denotes horizontal matrix concatenation $\left[\mathbf{V}_{\mathbf{i}_1} \cdots \mathbf{V}_{\mathbf{i}_{|S|}}\right]$ of all matrices $\mathbf{V}_\mathbf{i}$ with $\mathbf{i} \in S$.)

Conditions 1. *Identifiability conditions for decomposition (2):*

1. *The columns of each $\mathbf{V}_\mathbf{i}$ are orthonormal.*
2. *For two different block index sets $\mathbf{i} \neq \mathbf{j}$, if $\mathbf{i} \subset \mathbf{j}$ or $\mathbf{j} \subset \mathbf{i}$, then the subspaces spanned by the columns of $\mathbf{V}_\mathbf{i}$ and $\mathbf{V}_\mathbf{j}$ in the trait space are orthogonal.*
3. *The matrix $[\mathbf{V}_\mathbf{i}]_{\mathbf{i}|\mathbf{i}\in 2^{\{1,\ldots,K\}}}$, concatenated over all $\mathbf{i} \in 2^{\{1,\ldots,K\}}$, has rank equal to its number of columns.*
4. *For all k, the matrix $[\mathbf{L}_{\mathbf{i},k}]_{\mathbf{i}\,|\,k\in\mathbf{i}}$, concatenated over all $\mathbf{i} \in 2^{\{1,\ldots,K\}}$ so that $k \in \mathbf{i}$, has rank equal to its number of columns.*

Note that the columns of the loadings matrices $\mathbf{L}_{\mathbf{i},k}$ are not required to be orthogonal and may have arbitrary magnitude in order to encode scale information. The $d_k \times n$ matrix $\mathbf{A}_{\mathbf{i},k} = \mathbf{L}_{\mathbf{i},k}\mathbf{V}_\mathbf{i}^\top$ has rank $r_\mathbf{i}$, the number of columns of $\mathbf{V}_\mathbf{i}$. The matrix $\mathbf{A}_{\mathbf{i},k}$ will be called the *partially shared joint structure* between blocks in $\mathbf{i}$. When $\mathbf{i}$ is a singleton we also call it *individual structure*, and when $\mathbf{i} = \{1, \ldots, K\}$ we also call it *fully joint structure*.

Next we prove existence and uniqueness of the joint structure decomposition (2) in the absence of noise:

Theorem 1. *For a set of signal matrices* $\mathbf{A}_1, \ldots, \mathbf{A}_K$*, there exists a set of matrices* $\mathbf{L}_{\mathbf{i},k}, \mathbf{V}_\mathbf{i}$ *satisfying* (2) *and identifiability Condition 1. The joint structure matrices* $\mathbf{A}_{\mathbf{i},k} = \mathbf{L}_{\mathbf{i},k}\mathbf{V}_\mathbf{i}^\top$ *are uniquely determined for all* $k = 1, \ldots, K$ *and all* $\mathbf{i} \in 2^{\{1,\ldots,K\}}$ *so that* $k \in \mathbf{i}$.

Proof. We proceed with a constructive proof by induction. Denote by $\mathcal{V}_\mathbf{i}$ the intersection of the subspaces spanned by the columns of transposed signal matrices $\mathbf{A}_k^\top$ with $k \in \mathbf{i}$ in trait space.

Step 1: Consider the index set $\mathbf{i} = \{1, \ldots, K\}$. Choose $\mathbf{V}_{\{1,\ldots,K\}}$ such that its columns form an orthonormal basis for $\mathcal{V}_{\{1,\ldots,K\}}$. Clearly the relevant parts of Condition 1 are satisfied. Also notice that any vector orthogonal to $\mathbf{V}_{\{1,\ldots,K\}}$ is in at most $K-1$ subspaces spanned by the columns of $\mathbf{A}_k^\top$.

Step 2: Let us assume that we have defined $\mathbf{V}_\mathbf{i}$ for all $K \geq |\mathbf{i}| \geq q$ that satisfy Condition 1, and, for any $|\mathbf{i}| = q$, any vector orthogonal to $[\mathbf{V}_\mathbf{j}]_{\mathbf{j}\supset\mathbf{i}}$ is included in at most $q-1$ subspaces spanned by the columns of $\mathbf{A}_k^\top$, $k \in \mathbf{i}$. For any $\mathbf{i}$ with $|\mathbf{i}| = q-1$ select $\mathbf{V}_\mathbf{i}$ such that its columns form an orthonormal basis for $\mathcal{V}_\mathbf{i} \cap [\mathbf{V}_\mathbf{j}]^\perp_{\mathbf{j}\supset\mathbf{i}}$, the part of the space $\mathcal{V}_\mathbf{i}$ orthogonal to all $\mathbf{V}_\mathbf{j}$, $\mathbf{j} \supset \mathbf{i}$. Each $\mathbf{V}_\mathbf{i}$ chosen this way satisfies parts 1 and 2 of Condition 1 by construction. Condition 3 is satisfied, because if there was a rank deficiency in a concatenated matrix $[\mathbf{V}_\mathbf{i}]_{\mathbf{i}\,||\,\mathbf{i}|\geq q-1}$ there would be two indices $\mathbf{i}, \mathbf{j}$ such that $\mathbf{i} \neq \mathbf{j}$ and the spans of the matrices $\mathbf{V}_\mathbf{i}$ and $\mathbf{V}_\mathbf{j}$ share some vector in common. However this vector would already be included in $\mathbf{V}_{\mathbf{i}\cup\mathbf{j}}$ which is a contradiction. Finally, for any $|\mathbf{i}| = q-1$, any vector orthogonal to $[\mathbf{V}_\mathbf{j}]_{\mathbf{j}\supset\mathbf{i}}$ is included in at most $q-2$ subspaces spanned by the columns of $\mathbf{A}_k^\top$, $k \in \mathbf{i}$. This completes the inductive construction of the collection of $\mathbf{V}_\mathbf{i}$.

For each k, notice that the column span of $[\mathbf{V}_\mathbf{i}]_{\mathbf{i}|k\in\mathbf{i}}$ is $\mathcal{V}_{\{k\}}$, the space spanned by columns of $\mathbf{A}_k^\top$. Thanks to part 3 of Condition 1, the matrices $\mathbf{L}_{\mathbf{i},k}$ are chosen as the unique solution of the equation formed using the concatenated matrices $[\mathbf{L}_{\mathbf{i},k}]_{\mathbf{i}|k\in\mathbf{i}} \cdot [\mathbf{V}_\mathbf{i}]^\top_{\mathbf{i}|k\in\mathbf{i}} = \mathbf{A}_k$. These $\mathbf{L}_{\mathbf{i},k}$ satisfy part 4 of Condition 1 by construction.

Now assume there exists some other collection $\mathbf{A}_k = \sum_{\mathbf{i}|k\in\mathbf{i}} \tilde{\mathbf{L}}_{\mathbf{i},k}\tilde{\mathbf{V}}_\mathbf{i}^\top$ also satisfying Condition 1. Following similar arguments as above we see that the columns spaces of $\tilde{\mathbf{V}}_\mathbf{i}$ and $\mathbf{V}_\mathbf{i}$ are the same, and therefore there exists an orthonormal $r_\mathbf{i} \times r_\mathbf{i}$ matrix $\mathbf{Q}_\mathbf{i}$, so that $\tilde{\mathbf{V}}_\mathbf{i} = \mathbf{V}_\mathbf{i}\mathbf{Q}_\mathbf{i}$. Consequently $\tilde{\mathbf{L}}_{\mathbf{i},k} = \mathbf{L}_{\mathbf{i},k}\mathbf{Q}_\mathbf{i}$ and $\tilde{\mathbf{L}}_{\mathbf{i},k}\tilde{\mathbf{V}}_\mathbf{i}^\top = \mathbf{L}_{\mathbf{i},k}\mathbf{V}_\mathbf{i}^\top$. □

The fact that the matrices $\mathbf{L}_{\mathbf{i},k}$ and $\mathbf{V}_\mathbf{i}$ are only determined up to basis rotation is a natural

result of DIVAS being a subspace based method and not focused on matrices. In particular, the most important information contained in $\mathbf{L_{i,}}_k$ and $\mathbf{V_i}$ is the subspaces their columns span in object space and trait space respectively. This allows DIVAS to efficiently handle near equal singular values that could cause problems for matrix based approaches. For interpretive purposes, it can be helpful to choose a particular informative basis for the shared subspaces and examine modes of variation of the data along those basis directions. These modes of variation of the data may be formed by outer-multiplying corresponding columns of suitably rotated $\mathbf{L_{i,}}_k$ and $\mathbf{V_i}$. In Section 2.3, we discuss a procedure to choose such informative basis rotations $\mathbf{Q_i}$ using SVD and projection.

Throughout the description of methodology we will use the following synthetic three-block data example to illustrate each step of DIVAS. Each block includes a different set of traits associated with 400 observations. To mimic challenging data situations with large disparities in trait set sizes, Block 1 has 200 traits, Block 2 has 400 traits, and Block 3 has 10000 traits. Figure 1 displays this synthetic data set using matrix *heatmaps*. Heatmaps are a graphical display of matrix entry magnitude using color. Negative entries are shown in shades of blue and positive entries are shown in red, with color saturation indicating the magnitude of each entry. This means that entries close to zero are shown with a low-saturation white color. The color scaling ranges of each heatmap are shown in the color bar below each individual plot.

Each row demonstrates the formation of one of the data blocks via the data model in (1). The left-most column of heatmaps shows the observed data blocks $\mathbf{X}_k$, and the right-most column of heatmaps shows the noise matrices $\mathbf{E}_k$, which in this example are i.i.d. Gaussian matrices. The middle columns display the various rank 1 components that sum to each block's signal matrix $\mathbf{A}_k$ as in (2). Each signal matrix is comprised of a fully-shared component (second column) and partially-shared components between two of the three data blocks (third, fourth, and fifth columns). As required by our model, these fully-shared and partially-shared components are constructed such that the trait space subspace of the fully-shared component is orthogonal to the corresponding trait space subspaces of each partially-shared component. However, the partially-shared subspaces are not mutually orthogonal. In fact, each pair of partially-shared trait subspaces each has a principal angle of 60 degrees between them in the trait space, $\mathbb{R}^{400}$. Adding the matrices in the middle columns across each row in the manner of (2) combines the signal components into rank 3 signal matrices. Adding the noise matrices in the manner of (1) then produces the observed data blocks (first column).

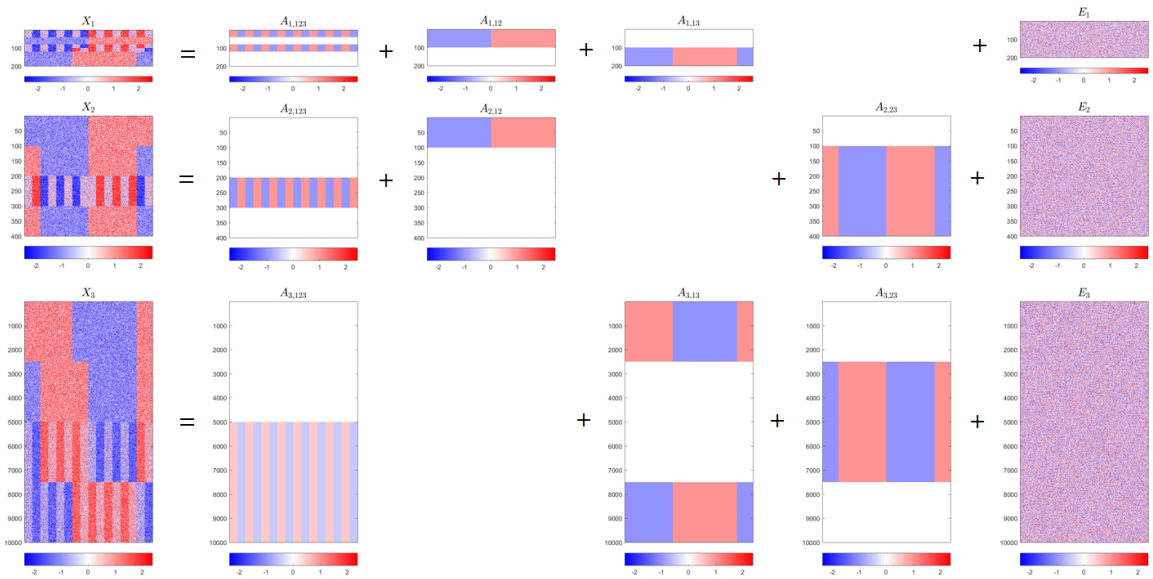

Figure 1: Heatmap view of $K = 3$ synthetic example construction. Heatmaps share a common color scheme displayed in the color bars below each plot, with white representing 0 magnitude. The three blocks in the first column are the observed data, and are formed by adding up the other matrices in their respective row. The three blocks in the second column show the rank 1 fully-shared structure common to each block. The next three columns show the rank 1 partially-shared structure common to each subset of two blocks. The final column shows the noise matrices for each data block.

The procedure of DIVAS takes the observed data blocks as input and outputs a full breakdown of shared and partially-shared structure between them over three steps. The first step, described in Section 2.1, extracts and estimates the dimension, magnitude, and direction of each block's signal subspace. The second step, described in Section 2.2, combines the information from each block to locate shared directions between subspaces. The third step, described in Section 2.3, uses those shared directions to form estimates $\hat{\mathbf{A}}_{\mathbf{i},k}$ of the partially shared joint structure matrices for each data block, e.g., estimates of the middle columns of Figure 1. Section 2.4 describes the visual display of DIVAS output and corresponding diagnostic measurements.

2.1 Signal Subspace Extraction

The first step of DIVAS is to estimate the subspaces spanned in both object space and trait space by the signal matrix for each data block. As the properties and analyses that take place in this

subsection apply to each data block independently, we suppress the subscript k indexing data blocks when referring to data-block-specific quantities for simplicity of notation. The indexing subscript will return in the remaining subsections when information from different data blocks is combined together. When we observe $\mathbf{X}$, we are observing data that's been perturbed in both magnitude (singular values) and orientation (basis vectors) from the signal $\mathbf{A}$. The signal magnitude is readily recoverable from the data magnitude via the signal extraction procedure described subsequently in Section 2.1.1. The signal orientation is itself more challenging to estimate from the data orientation; there's no reason to favor one direction of rotation over another under the rotational invariance assumptions on the noise matrix $\mathbf{E}$. However, the key to DIVAS is to quantify a range of feasible signal orientations given the observed data orientation using bounds on principal angles. These techniques are described in Sections 2.1.2 and 2.1.3. The results of this step for the synthetic data set presented in Figure 1 are explored in Section 2.1.4

2.1.1 Signal Subspaces

Shabalin and Nobel (2013) demonstrated in Proposition 5 that if the noise in (1) is orthogonally invariant, any procedure for extracting signal from a data matrix $\mathbf{X}$ need only consider the singular values of the data matrix. Motivated by this fact, we perform SVD on $\mathbf{X}$ to find $\mathbf{X} = \overline{\mathbf{U}}\overline{\mathbf{D}}\overline{\mathbf{V}}^\top$. The columns of the matrices $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ are orthonormal bases for the subspaces spanned in object space and trait space of $\mathbf{X}$ respectively, and the diagonal entries of $\overline{\mathbf{D}}$ are the singular values of $\mathbf{X}$. Denote these singular values as $\bar{\nu}_1, \ldots, \bar{\nu}_{d \wedge n}$. Estimations of the signal matrix $\hat{\mathbf{A}}$ typically take the form of a decomposition in terms of rank 1 matrices/approximations that combine to form the estimated object space and trait space subspaces. The vectors $\bar{\mathbf{u}}_i$ and $\bar{\mathbf{v}}_i$ denote the ith columns of the matrices $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$ respectively, and $\eta(\bullet)$ is a function from $\mathbb{R}^+$ to $\mathbb{R}^+$ for shrinking the singular values:

$$\hat{\mathbf{A}} = \sum_{i=1}^{d \wedge n} \eta(\bar{\nu}_i) \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top. \tag{3}$$

Common choices for η include *soft thresholding*: $\eta_{soft}(\nu) = (\nu - c) \vee 0$, and *hard thresholding*: $\eta_{hard}(\nu) = \nu \mathbb{I}_{\{\nu \geq c\}}$, for some constant c, and where $\mathbb{I}_{\{\bullet\}}$ represents an indicator function. In either case, any singular value smaller than c is set to 0. This means both procedures have dimension-reducing effects on the estimated $\hat{\mathbf{A}}$, and only subspaces $\bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top$ associated with nonzero transformed singular values contribute to the estimate. Gavish and Donoho (2014) outline optimal choices for c

for both soft and hard thresholding in terms of the *aspect ratio* $\beta = \frac{d \wedge n}{d \vee n}$ and the standard deviation σ of the noise matrix $\mathbf{E}$.

In the additive noise data matrix model (1), the presence of noise inflates the singular values associated with the signal component of the data matrix. Hard thresholding does not account for this phenomenon at all and soft thresholding often overcorrects by applying the same amount of shrinkage to each nonzero singular value. Shabalin and Nobel (2013) and Gavish and Donoho (2017) each propose optimal thresholding functions based on the Marchenko-Pastur distribution (see Appendix A) under a variety of matrix norms. We use the operator-norm-optimal function η^* from (Gavish and Donoho, 2017) for DIVAS:

$$\eta^*(\nu) = \begin{cases} \frac{1}{\sqrt{2}}\sqrt{\nu^2 - \beta - 1 + \sqrt{(\nu^2 - \beta - 1)^2 - 4\beta}}, & \nu \geq 1 + \sqrt{\beta}; \\ 0, & \nu < 1 + \sqrt{\beta}. \end{cases} \tag{4}$$

Figure 2 demonstrates how this shrinkage function (4), blue solid line, compromises between soft and hard thresholding for singular values with different magnitudes for a matrix with aspect ratio $\beta = 1$ and noise standard deviation $\sigma = 1$. Small values are thresholded according to optimal soft thresholding (magenta dot-dash line), but the shrinkage function approaches optimal hard thresholding (black dashed line) for larger values.

Equation (4) assumes noise standard deviation $\sigma = 1$. To use the shrinkage function in general settings, we must appropriately scale the singular values before and after shrinkage according to some estimate of the standard deviation of the noise $\hat{\sigma}$. Shabalin and Nobel (2013) use a grid search over several candidate values for $\hat{\sigma}$ to find a value that minimizes the Kolmogorov-Smirnov distance between the non-signal singular values and the appropriate Marchenko-Pastur distribution. Gavish and Donoho (2017) opt for the simple, robust, closed-form estimate $\hat{\sigma} = \frac{\nu_{median}}{\sqrt{MP(\beta)_{0.5}}}$, where ν_{median} denotes the median singular value of $\mathbf{X}$ and $MP(\beta)_{0.5}$ denotes the median of the Marchenko-Pastur distribution with parameter β (see Section A). We use the latter method for DIVAS noise standard deviation estimation.

Combining the previous equations, our estimate for the signal matrix $\hat{\mathbf{A}}$ for a given data block $\mathbf{X}$ is:

$$\hat{\mathbf{A}} = \sum_{i=1}^{d \wedge n} \hat{\sigma}\eta^*(\bar{\nu}_i/\hat{\sigma})\bar{\mathbf{u}}_i\bar{\mathbf{v}}_i^\top. \tag{5}$$

Let $\hat{\nu_i} = \hat{\sigma}\eta^*(\bar{\nu}_i/\hat{\sigma})$ be the ith shrunken singular value of $\mathbf{X}$. Let $\hat{r}$ be the number of nonzero

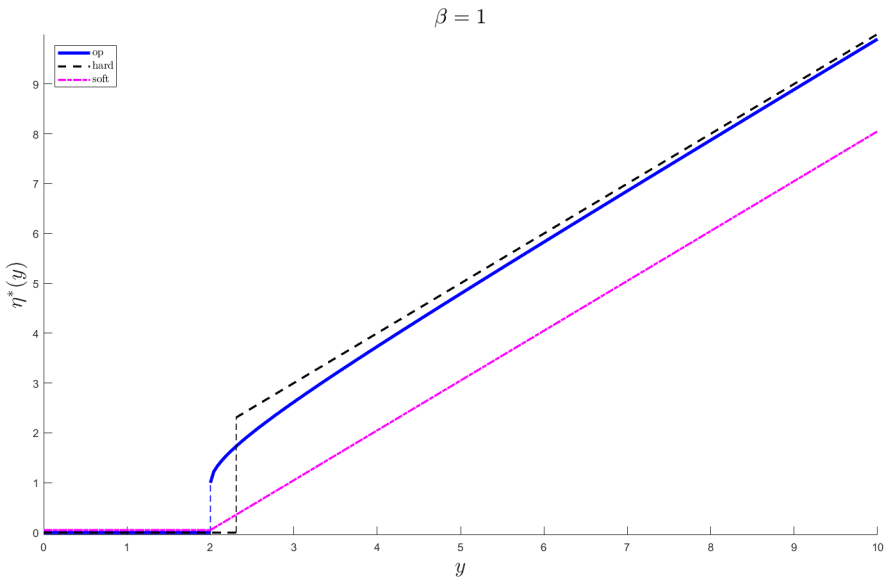

Figure 2: Functions for hard thresholding, soft thresholding, and optimal shrinkage under operator norm loss for a square matrix ($\beta = 1$). The optimal shrinkage function compromises between the other two approaches. Figure produced with code from (Gavish and Donoho, 2017).

shrunken singular values, and therefore the estimated rank of $\mathbf{A}$. Let $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ be matrices containing the first $\hat{r}$ columns of $\overline{\mathbf{U}}$ and $\overline{\mathbf{V}}$, respectively. Using this notation and defining the matrix $\hat{\mathbf{D}}$ as the $\hat{r} \times \hat{r}$ diagonal matrix with diagonal entries equal to $\hat{\nu}_1, \ldots, \hat{\nu}_{\hat{r}}$, we can also write $\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}^\top$. Note that $\hat{\mathbf{U}}$ is therefore an orthonormal basis for the subspace spanned in object space of $\hat{\mathbf{A}}$ and $\hat{\mathbf{V}}$ is an orthonormal basis for the subspace spanned in trait space of $\hat{\mathbf{A}}$.

2.1.2 Angle Perturbation Theory

The foundation of DIVAS is determining whether candidate directions $\mathbf{v}^\star \in \mathbb{R}^n$ lie in the trait space span $\mathbf{TS}(\mathbf{A})$ of the signal matrix $\mathbf{A}$. If $\mathbf{A}$ was observable, this would simply amount to checking whether the angle θ between $\mathbf{v}^\star$ and $\mathbf{TS}(\mathbf{A})$ was 0. Since $\mathbf{A}$ and θ are unobservable, we aim to estimate θ based on the observable estimated low-rank signal matrix $\hat{\mathbf{A}}$ and an estimate of the noise variation $\hat{\mathbf{E}}$ developed in Appendix C. Specifically, we want to choose a *perturbation angle bound* $\hat{\phi}$ that defines a cone-shaped significance region around $\mathbf{TS}(\hat{\mathbf{A}})$ which contains $\mathbf{TS}(\mathbf{A})$ with high probability. Directions $\mathbf{v}^\star$ lying within that significance region would then be potential basis directions for $\mathbf{TS}(\mathbf{A})$. Directions $\mathbf{v}^\star$ that lie within the significance regions of multiple data blocks

would then be potential basis directions for partially-shared joint structure between those blocks. To arrive at such a data-block-wise uniform perturbation angle bound $\hat{\phi}$, we first look to bound the range of possible values for θ for one given candidate direction $\mathbf{v}^\star$.

The first step is construction of the range of values for θ based on the relationships between the projections of $\mathbf{v}^\star$ onto various subspaces. This is illustrated using a simple low-dimensional example in Figure 3. In particular, $\mathbf{v}^\star$ (green vector in both panels of Figure 3) is projected onto each of $\mathbf{TS}(\mathbf{A})$ (translucent purple plane) and $\mathbf{TS}(\hat{\mathbf{A}})$ (solid gold plane), with those projections denoted $\mathbf{v}^\star_{proj}$ (red solid lines) and $\hat{\mathbf{v}}^\star_{proj}$ (blue solid lines) respectively. Computations of these projections are based on the orthonormal basis matrices $\mathbf{V}$ for $\mathbf{TS}(\mathbf{A})$ and $\hat{\mathbf{V}}$ for $\mathbf{TS}(\hat{\mathbf{A}})$. Let $\hat{\theta}$ be the angle between $\mathbf{v}^\star$ and $\mathbf{TS}(\hat{\mathbf{A}})$. We can write expressions for θ and $\hat{\theta}$ in terms of the above quantities as follows:

$$\theta = \arccos\left(\frac{\langle \mathbf{v}^\star, \mathbf{v}^\star_{proj}\rangle}{\| \mathbf{v}^\star \| \| \mathbf{v}^\star_{proj} \|}\right); \quad \mathbf{v}^\star_{proj} = \mathbf{V}\mathbf{V}^\top \mathbf{v}^\star.$$

$$\hat{\theta} = \arccos\left(\frac{\langle \mathbf{v}^\star, \hat{\mathbf{v}}^\star_{proj}\rangle}{\| \mathbf{v}^\star \| \| \hat{\mathbf{v}}^\star_{proj} \|}\right); \quad \hat{\mathbf{v}}^\star_{proj} = \hat{\mathbf{V}}\hat{\mathbf{V}}^\top \mathbf{v}^\star.$$

We can construct bounds involving θ and $\hat{\theta}$ by considering further projections between $\mathbf{TS}(\mathbf{A})$ and $\mathbf{TS}(\hat{\mathbf{A}})$. The total angle traversed by projecting $\mathbf{v}^\star$ to $\mathbf{TS}(\mathbf{A})$ and then projecting that result, $\mathbf{v}^\star_{proj}$, onto $\mathbf{TS}(\hat{\mathbf{A}})$ (red dashed line in left panel of Figure 3) is at least as large as $\hat{\theta}$, the angle between $\mathbf{v}^\star$ and $\mathbf{TS}(\hat{\mathbf{A}})$. Define $\theta_1^\star$ as the angle between $\mathbf{v}^\star_{proj}$ and $\mathbf{TS}(\hat{\mathbf{A}})$. By the triangle inequality, $\hat{\theta} \leq \theta + \theta_1^\star$. Via an analogous projection of $\hat{\mathbf{v}}^\star_{proj}$ onto $\mathbf{TS}(\mathbf{A})$ (blue dashed line in right panel of Figure 3), define $\theta_2^\star$ as the angle between $\hat{\mathbf{v}}^\star_{proj}$ and $\mathbf{TS}(\mathbf{A})$. Then θ, the angle between $\mathbf{v}^\star$ and $\mathbf{TS}(\mathbf{A})$, is no larger than the total angle traversed by projecting $\mathbf{v}^\star$ onto $\mathbf{TS}(\hat{\mathbf{A}})$ and then projecting that result, $\hat{\mathbf{v}}^\star_{proj}$, onto $\mathbf{TS}(\mathbf{A})$: $\theta \leq \hat{\theta} + \theta_2^\star$.

The above discussion of angles between subspaces summarizes the proof of the following theorem. More details can be found in the Ph.D. dissertation of Jiang (2018).

Theorem 2. *Let $\mathbf{X} = \mathbf{A} + \mathbf{E}$ be a $d \times n$ data matrix which is a sum of a signal matrix $\mathbf{A}$ and a noise matrix $\mathbf{E}$ under the assumptions of* (1)*. Given θ, $\hat{\theta}$, $\theta_1^\star$, and $\theta_2^\star$ defined as above, and using $(\bullet)_+ = \max(\bullet, 0)$, we have:*

$$(\hat{\theta} - \theta_1^\star)_+ \leq \theta \leq \hat{\theta} + \theta_2^\star. \tag{6}$$

Both inequalities in (6) will be used to rule out directions $\mathbf{v}^\star$ as a candidate for the joint spaces.

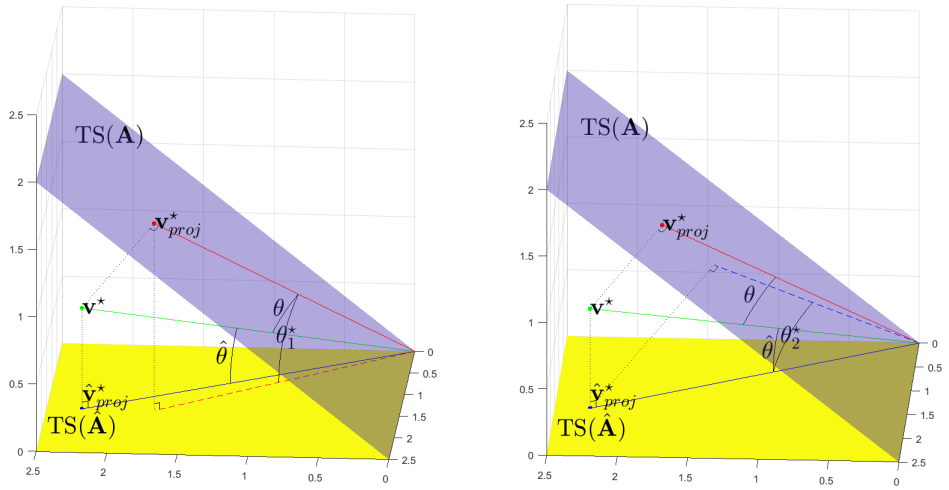

Figure 3: Locations of θ, $\hat{\theta}$, $\theta_1^\star$, and $\theta_2^\star$ in a low-dimensional example. Each panel demonstrates a different angle bound. Left: $\hat{\theta} \leq \theta + \theta_1^\star$. Right: $\theta \leq \hat{\theta} + \theta_2^\star$.

These exclusions will be based on two distinct statistical arguments: a novel rotational bootstrap used for both bounds, and a distribution of angles between random directions used only for the upper bound.

If the angle to the estimated signal subspace $\hat{\theta}$ for a given candidate direction $\mathbf{v}^\star$ is less than $\theta_1^\star$, then the lower bound in (6) for the angle to the true signal subspace is 0, indicating that this direction can't be ruled out as lying in the true signal subspace. The angle $\theta_1^\star$ is behaving much like the desired perturbation angle bound, but as noted in Jiang (2018), $\theta_1^\star$ is not directly estimable for a given direction. However, $\theta_1^\star$ is uniformly bounded from above for all $\mathbf{v}^\star$ by the maximum principal angle ϕ between $\mathbf{TS}(\mathbf{A})$ and $\mathbf{TS}(\hat{\mathbf{A}})$ (see Appendix B). Our chosen perturbation angle bound will therefore be a statistical estimate $\hat{\phi}$ of that maximum principal angle ϕ. This estimation is performed via a rotational bootstrap as described in Section 2.1.3.

Unlike $\theta_1^\star$, the angle $\theta_2^\star$ in the upper bound of (6) can be estimated for each $\mathbf{v}^\star$ using

$$\theta_2^\star = \arccos\left(\frac{\| \mathbf{V}^\top \hat{\mathbf{V}} \hat{\mathbf{V}}^\top \mathbf{v}^\star \|}{\| \hat{\mathbf{V}} \hat{\mathbf{V}}^\top \mathbf{v}^\star \|}\right). \tag{7}$$

The only unknown quantity in this formula is the matrix $\mathbf{V}^\top \hat{\mathbf{V}}$, and we generate samples from the estimated distribution of this matrix as part of the rotational bootstrap in Section 2.1.3. Therefore by recording those samples and using them in (7) we can generate from a bootstrap distribution of

$\theta_2^\star$ and choose a high percentile denoted as $\hat{\theta}_2^\star$.

The main use of this estimate is determining whether $\mathbf{v}^\star$ can be distinguished from an arbitrarily chosen direction based on the estimated upper bound $\hat{\theta} + \hat{\theta}_2^\star$. The rotational invariance property assumed of the signal and noise in (1) implies a natural null distribution for comparison. In particular, we choose the distribution of angles between a fixed arbitrary $\hat{r}$-dimensional subspace of $\mathbb{R}^n$ (recall that $\hat{r}$ is the estimated signal rank) and unit vectors chosen uniformly at random. We pick a *random direction angle bound* θ_0 as a low percentile of that null distribution. If $\hat{\theta} + \hat{\theta}_2^\star$ lies above θ_0 for some direction $\mathbf{v}^\star$, then that direction cannot be distinguished from an arbitrarily chosen direction, which provides statistical evidence that $\mathbf{v}^\star$ is far from $\mathbf{TS}(\mathbf{A})$.

The above derivations of a perturbation angle bound and other angle-based inference have taken place entirely in trait space. Analogous derivations can be carried out in object space, and the estimation of perturbation angle bounds in both spaces can take place simultaneously during the rotational bootstrap. When considering candidate directions $\mathbf{v}^\star$, we should additionally rule out directions whose corresponding basis directions in object space do not obey the object space perturbation angle bounds. Therefore both space's angle bounds play key roles in the optimization problem for locating joint structure between data blocks. This leads to more precise estimates of joint subspaces compared to methods like AJIVE (Feng et al., 2018) that consider only trait space in their algorithms.

2.1.3 Rotational Bootstrap

We estimate perturbation angle bounds for the object space and trait space of a data block using a novel *rotational bootstrap*. This technique is designed to take advantage of the assumed rotational invariance property and aims to estimate the distribution of principal angles between object space and trait space subspaces of $\mathbf{X}$ and $\mathbf{A}$ through random generation of replicate signal subspaces.

Recall (1), $\mathbf{X} = \mathbf{A} + \mathbf{E}$, and as in Section 2.1.2 the compact SVD of the rank r signal matrix is $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ and the compact SVD of the rank r approximation to the data $\mathbf{X}$ is $\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}^\top$. Next consider a random replication $\mathbf{X}^\diamond = \mathbf{U}^\diamond\mathbf{D}\mathbf{V}^{\diamond\top} + \mathbf{E}^\diamond$, where $\mathbf{E}^\diamond$ has the same distribution as $\mathbf{E}$, and $\mathbf{U}^\diamond$ and $\mathbf{V}^\diamond$ are random $d \times r$ and $n \times r$ orthonormal matrices, respectively. The corresponding compact SVD of the rank r approximation to $\mathbf{X}^\diamond$ is $\hat{\mathbf{A}}^\diamond = \hat{\mathbf{U}}^\diamond\hat{\mathbf{D}}^\diamond\hat{\mathbf{V}}^{\diamond\top}$.

Assumption 1. *Data matrix model is* rotationally invariant *if the matrices* $\hat{\mathbf{D}}, \mathbf{V}^\top\hat{\mathbf{V}}, \hat{\mathbf{V}}^\top\mathbf{V}, \mathbf{U}^\top\hat{\mathbf{U}}, \hat{\mathbf{U}}^\top\mathbf{U}$ *have the same distribution as the corresponding* $\hat{\mathbf{D}}^\diamond, \mathbf{V}^{\diamond\top}\hat{\mathbf{V}}^\diamond, \hat{\mathbf{V}}^{\diamond\top}\mathbf{V}^\diamond, \mathbf{U}^{\diamond\top}\hat{\mathbf{U}}^\diamond, \hat{\mathbf{U}}^{\diamond\top}\mathbf{U}^\diamond$.

As discussed in Appendix B, these matrices determine the principle angle structure between the spaced spanned by the low rank signal and its estimate in both trait and object spaces. Theorem 7 of Jiang (2018) shows that if the noise distribution is rotationally invariant, e.g., having i.i.d. centered Gaussian entries, then the model satisfies Assumption 1. Alternatively, the model will be rotationally invariant, if the signal matrix $\mathbf{A}$ is considered random following a rotationally invariant prior distribution.

The continuity of these distributions in the singular values $\mathbf{D}$ suggests use of a *parametric bootstrap* estimator of these quantities based on the estimated $\hat{r} \times \hat{r}$ singular value matrix $\hat{\mathbf{D}}$. In particular, we form a bootstrap replication of a signal matrix $\mathbf{A}^\circ = \mathbf{U}^\circ\hat{\mathbf{D}}\mathbf{V}^{\circ\top}$, where $\mathbf{U}^\circ$ and $\mathbf{V}^\circ$ are random $d \times \hat{r}$ and $n \times \hat{r}$ orthonormal matrices, respectively. Using this randomly rotated estimated signal matrix along with an estimate $\hat{\mathbf{E}}$ of the noise matrix $\mathbf{E}$, we form a bootstrap replication of the data matrix $\mathbf{X}^\circ = \mathbf{A}^\circ + \hat{\mathbf{E}}$. If the rotational invariance assumption is satisfied, and the estimated $\hat{\mathbf{D}}$ is close to $\mathbf{D}$ this construction produces replicate signal and data matrices with principal angle structure drawn from a similar distribution as the unobserved principal angle structure between the true signal and data matrices. An important, and perhaps surprising point is that the naïve noise matrix estimate $\hat{\mathbf{E}} = \mathbf{X} - \hat{\mathbf{A}}$ is not appropriate for use in this construction. This is because $\mathbf{X} - \hat{\mathbf{A}}$ has insufficient energy in the directions associated with $\hat{\mathbf{A}}$, and therefore has eigenvalues that don't follow the Marchenko-Pastur distribution in the manner expected for a noise matrix under our assumptions. Our proposed estimator, labeled $\hat{\mathbf{E}}_{impute}$, is shown in (12) and corrects for the insufficient energy through imputation via Marchenko-Pastur random variates. See Appendix A for details on the Marchenko-Pastur distribution and Appendix C for full details on the poor performance of $\mathbf{X} - \hat{\mathbf{A}}$ and the motivation of $\hat{\mathbf{E}}_{impute}$.

As mentioned in Section 2.1.2, we will use estimates of the maximum principal angles between the subspaces spanned in object space and trait space by $\mathbf{X}$ and $\mathbf{A}$ as perturbation angle bounds. Through repeated replications of the randomly rotated signal and data matrices described in the previous paragraph, we generate bootstrap samples estimating the *distribution* of principal angles between subspaces spanned by $\mathbf{X}$ and $\mathbf{A}$ in both object and trait space. With sufficiently many replications (we use $M = 400$) we can choose high quantiles (e.g. 0.95) of the empirical distributions

of maximum principal angles as statistical perturbation angle bounds. Recall that $\hat{\phi}$ is the trait space perturbation angle bound, and denote the corresponding object space perturbation angle bound as $\hat{\psi}$.

The procedure described in Section 2.1.1 discriminates noise fairly well, but as our algorithm is based on angles we find that additional angle-based rank selection is necessary for good performance. Therefore as part of the rotational bootstrap we *filter* the signal subspaces according to the random direction angle bound θ_0 defined in Section 2.1.2. In particular we choose a filtered rank $\check{r}$ such that the estimated maximum principal angles between true and estimated signal don't exceed $\xi\theta_0$, where $\xi \in (0, 0.5]$ is a tuning parameter. In our analyses we explored a grid of values for ξ ranging from 0.3 to 0.5 and found that a value between 0.35 and 0.4 often captured an appropriate amount of signal for our data sets. Therefore the case studies in Sections 3 and 4 choose $\xi = 1 - \frac{2}{1+\sqrt{5}} \approx 0.382$, a value based on the golden ratio. This hyperparameter can be tuned up or down to include more or less information from the estimated signal matrix in the analysis.

The limitation of $\xi \leq 0.5$ follows from the statistical inference framework laid out in Section 2.1.2. If the lower bound $(\hat{\theta} - \hat{\phi})_+$ is 0 and the upper bound $\hat{\theta} + \hat{\theta}_2^\star$ is simultaneously greater than θ_0 for a given candidate direction $\mathbf{v}^\star$, the inference procedure says there is evidence that $\mathbf{v}^\star$ is both significantly close to the true signal subspace and indistinguishable from an arbitrary direction. This inference outcome is completely non-informative. In this case, both $\hat{\theta}$ and $\hat{\theta}_2^\star$ must be bounded from above by the maximum principal angle between estimated and true signal subspaces. Therefore this non-informative inference outcome is avoided by filtering the estimated signal subspace until the rotational-bootstrap-estimated maximum principal angle is at most $0.5 \cdot \theta_0$.

The above description of the rotational bootstrap algorithm is formulated in Algorithm 1. For each block k, we have as inputs to the algorithm the estimated signal matrix $\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}^\top$ from (5) in Section 2.1, the estimated residual matrix $\hat{\mathbf{E}}_{impute}$ from (12) in Appendix C, the random direction angle bound θ_0 discussed at the end of Section 2.1.2, and the hyperparameters ξ and α. α is the desired confidence level for the perturbation angle bounds and the default is 0.95. At the end of the algorithm, we have as outputs estimates of the trait space and object space perturbation angle bounds $\hat{\psi}$ and $\hat{\phi}$ respectively for each block k.

During each replication, random subspaces are generated from i.i.d. standard Gaussian matrices with the same centering operations used on the data. Note that orthogonalization of an i.i.d.

random matrix in this fashion is identical to sampling from a rotationally uniform distribution of subspaces, according to Theorem 2.2.1 from Chikuse (2012). The inner **for** loop records maximum principal angles at each possible filtered rank from 1 to $\hat{r}$. After the outer **for** loop concludes, the algorithm chooses a filtered rank $\check{r}$ to align with the chosen value of ξ. In the case where the filtered ranks in object space and trait space differ, the smaller of the two is chosen to ensure compliance in both spaces. The trait and object space perturbation angle bounds $\hat{\phi}$ and $\hat{\psi}$ are chosen as the $1-\alpha$ percentile of the empirical distributions of angles at the filtered rank $\check{r}$. Once the filtered rank is selected, we filter the columns of the estimated basis matrices for the signal object and trait space subspaces to correspond with the reduced rank. Let $\check{\mathbf{U}} = \overline{\mathbf{U}}_{1:\check{r}}$ and $\check{\mathbf{V}} = \overline{\mathbf{V}}_{1:\check{r}}$ be the final estimates of the signal object and trait space bases respectively.

2.1.4 Signal Extraction for Synthetic Data

The results of signal space extraction on the synthetic data example from Figure 1 are shown in Figure 4. Each heatmap shows the estimated signal matrix $\hat{\mathbf{A}}$ for the respective data block. The denoising of each data block appears to have been successful when comparing the visual impression of these heatmaps to the original data matrices shown in the first column of Figure 1. The signal rank is correctly chosen as 3 for all three blocks. In this case, angle-based rank filtering didn't further reduce the rank beyond the value determined from eigenvalue-based rank selection.

Since we know the true signal object and trait subspaces for the synthetic data example, we can compare the estimated perturbation angle bounds to the actual angles between estimated and true signal subspaces to check the performance of the bounds. Tables 1 and 2 display the perturbation angle bounds and angles between the true and estimated signal subspaces in trait space and object space respectively. The calculated bounds exceed the true angles in all cases, so the true basis directions all lie within the cones of feasibility defined by the bounds. Since the bounds are calculated as uniform 95% bounds, we'd expect to not cover the truth about 1 in 20 times, and the performance of the bounds in this case aligns with that expectation.

Each synthetic data block has a similar signal-to-noise ratio, so the observed differences in perturbation angle bounds in the second column of each table are primarily explained by differences in matrix dimension. Recall that all four data blocks have 400 data objects, X1 has 200 traits, X2 has 400 traits, and X3 has 10000 traits. The trait space perturbation angle bounds decrease

Algorithm 1 Rotational Bootstrap

$objectAngles \leftarrow 90 * \mathbf{1}^{M \times \hat{r}}$; $traitAngles \leftarrow 90 * \mathbf{1}^{M \times \hat{r}}$

for all $m \in \{1, \ldots, M\}$ **do**

$\mathbf{U}^{\circ} \leftarrow \text{rand}^{d \times \hat{r}}$; $\mathbf{V}^{\circ} \leftarrow \text{rand}^{n \times \hat{r}}$

▷ Replications must be orthogonal to constant function direction in appropriate spaces.

if $\mathbf{X}$ is trait-centered **then**

$\mathbf{U}^{\circ} \leftarrow \left(\mathbf{I}^{d \times d} - \frac{1}{d}\mathbf{1}^{d \times d}\right)\mathbf{U}^{\circ}$

end if

if $\mathbf{X}$ is object-centered **then**

$\mathbf{V}^{\circ} \leftarrow \left(\mathbf{I}^{n \times n} - \frac{1}{n}\mathbf{1}^{n \times n}\right)\mathbf{V}^{\circ}$

end if

$\mathbf{U}^{\circ} \leftarrow \text{orth}(\mathbf{U}^{\circ})$; $\mathbf{V}^{\circ} \leftarrow \text{orth}(\mathbf{V}^{\circ})$; $\mathbf{A}^{\circ} \leftarrow \mathbf{U}^{\circ}\hat{\mathbf{D}}\mathbf{V}^{\circ}$; $\mathbf{X}^{\circ} \leftarrow \mathbf{A}^{\circ} + \hat{\mathbf{E}}$

$[\overline{\mathbf{U}}^{\circ}, \sim, \overline{\mathbf{V}}^{\circ}] \leftarrow \text{SVD}(\mathbf{X}^{\circ})$.

for all $j \in \{1, \ldots, \hat{r}\}$ **do**

▷ Smallest singular value equal to cosine of largest angle.

$[\sim, \vec{\nu}_{object}, \sim] \leftarrow \text{SVD}\left(\mathbf{U}^{\circ T}\overline{\mathbf{U}}^{\circ}_{1:j}\right)$; $[\sim, \vec{\nu}_{trait}, \sim] \leftarrow \text{SVD}\left(\mathbf{V}^{\circ T}\overline{\mathbf{V}}^{\circ}_{1:j}\right)$

$objectAngles[m, j] \leftarrow \arccos(\min(\vec{\nu}_{object}))$; $traitAngles[m, j] \leftarrow \arccos(\min(\vec{\nu}_{trait}))$

end for

end for

$objectAnglesSort \leftarrow \text{sort}(objectAngles, col, asc)$; $traitAnglesSort \leftarrow \text{sort}(traitAngles, col, asc)$

$\check{r} \leftarrow \min\left(\sum_{j=1}^{\hat{r}} \mathbb{I}_{\{objectAnglesSort[\alpha M, j] < \xi\theta_0\}}, \sum_{j=1}^{\hat{r}} \mathbb{I}_{\{traitAnglesSort[\alpha M, j] < \xi\theta_0\}}\right)$

$\hat{\psi} \leftarrow objectAnglesSort[\alpha M, \check{r}]$; $\hat{\phi} \leftarrow traitAnglesSort[\alpha M, \check{r}]$

as the number of traits in the data blocks increase since we have a more precise idea of where the true signal subspace is with more trait vectors in the same trait space $\mathbb{R}^{400}$. The object space perturbation angle bounds increase as the number of traits in the data blocks increase since we have a more precise idea of where the true signal subspace is in object space with 400 object vectors in $\mathbb{R}^{200}$ than we do with 400 object vectors in $\mathbb{R}^{10000}$.

2.2 Joint Subspace Estimation

We formally introduce the optimization problem for locating shared structure in DIVAS. The

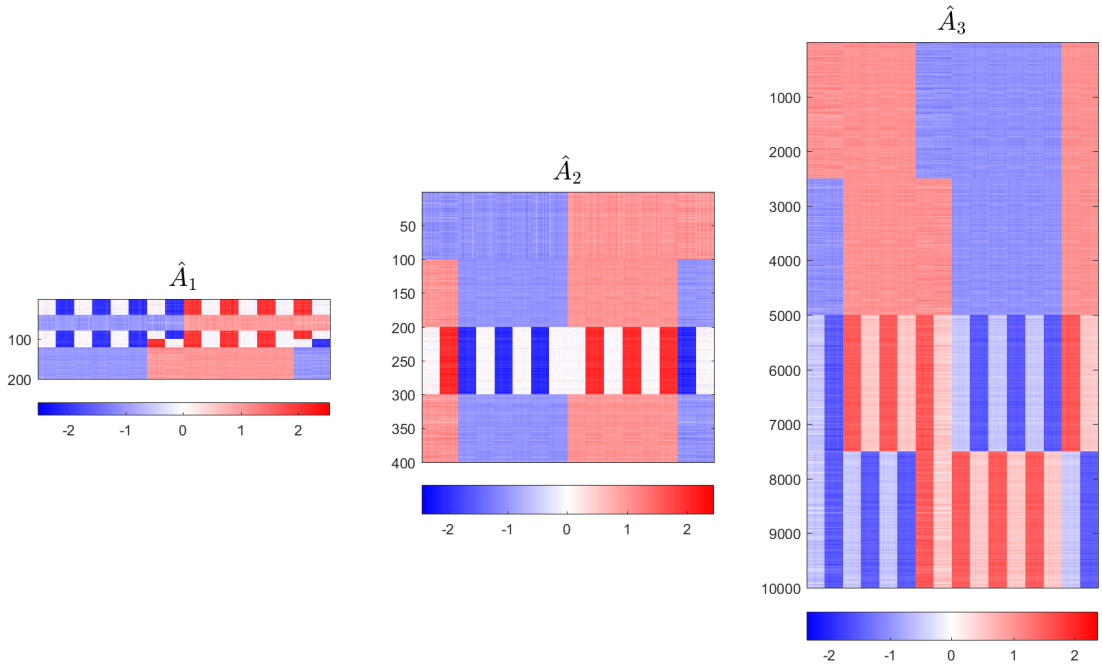

Figure 4: Estimated signal matrices for each block in the toy example defined in Figure 1. The heatmaps show good recovery of the original signal patterns. The trait and object spaces of each estimated matrix are rank 3.

conceptual constraints and objective function are shown in (8) and the full numerical algorithm is deferred to Appendix D with the main subproblem being a convex optimization problem (16).

For any given collection of blocks $\mathbf{i}$, the corresponding joint subspace should be near each of the included blocks in some sense. In DIVAS, proximity is evaluated in terms of angles between candidate directions and subspaces. In particular, during each phase of joint subspace estimation we minimize the angle between candidate directions $\mathbf{v}^\star$ and the estimated trait space subspaces of included blocks as our objective. This objective is expressed in terms of angle cosines in line 1 of (8). To ensure that a candidate direction lies in the true signal subspace of an included block $\mathbf{X}_k, k \in \mathbf{i}$ with high significance, the trait space angle between a candidate direction and the subspace spanned by the columns of $\check{\mathbf{V}}_k$ should be at most the trait space angle perturbation bound $\hat{\phi}_k$. Additionally, the object space angle between $\mathbf{X}_k\mathbf{v}^\star$ and the subspace spanned by the columns of $\check{\mathbf{U}}_k$ should be at most the object space angle perturbation bound $\hat{\psi}_k$. Finally, the angle between a candidate direction and an excluded block should be at least the trait space

Data Block	Trait Space Angle Bound	Angle to 1,2,3 Truth	Angle to 1,2 Truth	Angle to 1,3 Truth	Angle to 2,3 Truth
1	11.7	9.2	8.5	6.1	
2	8.6	6.9	5.6		4.0
3	2.8	2.5		1.0	1.0

Table 1: Table of angles between estimated signal trait spaces and true signal trait spaces. All angles are within the calculated perturbation angle bounds.

Data Block	Object Space Angle Bound	Angle to 1,2,3 Truth	Angle to 1,2 Truth	Angle to 1,3 Truth	Angle to 2,3 Truth
1	8.6	4.5	4.9	4.7	
2	8.6	5.8	6.6		4.0
3	13.1	7.9		4.6	4.7

Table 2: Table of angles between estimated signal object spaces and true signal object spaces. All angles are within the calculated perturbation angle bounds.

angle perturbation bound $\hat{\phi}_k$. These requirements are expressed as constraints for the optimization problem in lines 2-6 in (8), with subscripts T and O indicating angles in trait space and object space respectively. Crucially, our dimensionally flexible subspace-based angle perturbation approach to signal extraction allows object space information to be incorporated very naturally into the joint subspace estimation algorithm. This innovation enhances the significance and interpretability of loadings vectors found using DIVAS.

Following the proof of Theorem 1, we determine each block collection's potential joint structure in turn, starting with larger block collections and ending with singleton block collections. Within a joint structure search for a given block collection $\mathbf{i}$, joint subspace basis directions are found one at a time via successive solves of (8). If no new feasible direction is found, the search among the current block collection ends and the search among the next block collection begins. Candidate directions in trait space for a particular block collection must also obey orthogonality constraints expressed in parts 1 and 2 of Condition 1. These conditions are concisely expressed in the constraint in line 7 of (8). The Gothic script symbol $\mathfrak{V}_\mathbf{i}$ is used to denote the current estimated trait space basis for the

shared structure for block collection $\mathbf{i}$, i.e., an estimate of the $\mathbf{V_i}$ from (2). Note that the search for joint structure between two block collections of the same size is embarassingly parallelizable, as the orthogonality constraint will only include joint structure found for strictly larger block collections. The order in which block collections of equal size are searched does not affect DIVAS output.

$$
\begin{aligned}
\min_{\mathbf{v}^\star} \quad & -\sum_{k\in\mathbf{i}} \cos^2 \hat{\theta}_{Tk} \\
s.t. \quad & \hat{\theta}_{Tk} = \angle(\mathbf{v}^\star, \check{\mathbf{V}}_k) && \forall k \\
& \hat{\theta}_{Ok} = \angle(\mathbf{X}_k\mathbf{v}^\star, \check{\mathbf{U}}_k) && \forall k \\
& \hat{\theta}_{Tk} \leq \hat{\phi}_k && \forall k \in \mathbf{i} \\
& \hat{\theta}_{Tk} > \hat{\phi}_k && \forall k \in \mathbf{i}^c \\
& \hat{\theta}_{Ok} \leq \hat{\psi}_k && \forall k \in \mathbf{i} \\
& \mathbf{v}^\star \perp \mathfrak{V}_\mathbf{j} && \forall \mathbf{j} \supseteq \mathbf{i}
\end{aligned}
\tag{8}
$$

Note that as basis directions are found for block collection $\mathbf{i}$, $\mathfrak{V}_\mathbf{i}$ grows larger, and so the constraint in line 7 becomes more stringent as more basis directions are located. Also as directions are located the angle constraints change. Each orthonormal basis for estimated trait space signal $\check{\mathbf{V}}_k$ is shrunk to only include directions in the null space of $[\mathfrak{V}_\mathbf{j}]_{\mathbf{j}\supseteq\mathbf{i}}$. This basis shrinking improves computation time and assists in the choice of basis directions satisfying our assumptions.

In practice, this problem is solved via an iterative procedure called convex-concave procedure as described in (Ismailova and Lu, 2016). This algorithm is also called DC (Difference of two Convex functions) algorithm in the literature. Full details can be found in Appendix D.

2.3 Signal Reconstruction

Once we have located all possible joint structure, the remaining task is to reconstruct the signal matrix components for each data block. Recall that in Section 2.2 we denote the estimated orthonormal basis for the joint structure among blocks in collection $\mathbf{i}$ as $\mathfrak{V}_\mathbf{i}$. For a given data block k, we first horizontally concatenate all joint structure basis matrices found involving block k into one matrix $[\mathfrak{V}_\mathbf{i}]_{\mathbf{i}|k\in\mathbf{i}}$. Then we form a linear regression problem to find the corresponding loadings vectors for block k associated with the common scores vectors for block collection $\mathbf{i}$ in a similar

fashion as the proof of Theorem 1. In particular $[\mathfrak{L}_{\mathbf{i},k}]_{\mathbf{i}|k\in\mathbf{i}}$ is chosen as the least-squares solution of the regression problem $\min_{\mathfrak{L}} \|\mathbf{X}_k - \mathfrak{L} \cdot [\mathfrak{V}_\mathbf{i}]_{\mathbf{i}|k\in\mathbf{i}}^\top\|_2^2$. This solution is unique when $[\mathfrak{V}_\mathbf{i}]_{\mathbf{i}|k\in\mathbf{i}}$ is full-rank. The columns of $[\mathfrak{L}_{\mathbf{i},k}]_{\mathbf{i}|k\in\mathbf{i}}$ can then be partitioned into loadings matrices $\mathfrak{L}_{\mathbf{i},k}$, each of which is associated with joint structure for one block collection $\mathbf{i}$ with $k \in \mathbf{i}$, i.e., $\mathfrak{L}_{\mathbf{i},k}$ is an estimate of $\mathbf{L}_{\mathbf{i},k}$ from (2). The estimated partially shared joint structure between between data blocks in $\mathbf{i}$ is then simply $\hat{\mathbf{A}}_{\mathbf{i},k} = \mathfrak{L}_{\mathbf{i},k}\mathfrak{V}_\mathbf{i}^\top$.

As a brief remark, DIVAS may in certain cases select shared structure such that the rank of $[\mathfrak{V}_\mathbf{i}]_{\mathbf{i}|k\in\mathbf{i}}$ is larger than $\check{r}_k$. Since the subspaces spanned by $\mathfrak{V}_\mathbf{i}$ and $\mathfrak{V}_\mathbf{j}$ need not be orthogonal unless $\mathbf{i} \subseteq \mathbf{j}$ or $\mathbf{j} \subseteq \mathbf{i}$, more than $\check{r}_k$ total basis directions may be selected within the cone of feasibility for block k.

Additional insight comes from further decomposition of $\hat{\mathbf{A}}_{\mathbf{i},k}$ into a sum of rank 1 modes of variation. Each mode comes from an outer product of corresponding columns of $\mathfrak{L}_{\mathbf{i},k}$ and $\mathfrak{V}_\mathbf{i}$. Since $\mathfrak{L}_{\mathbf{i},k}$ and $\mathfrak{V}_\mathbf{i}$ are only determined up to basis rotation, we first select a rotation matrix $\mathbf{Q}_\mathbf{i}$, and then examine modes of variation formed with the matrices $\mathfrak{L}_{\mathbf{i},k}\mathbf{Q}_\mathbf{i}$ and $\mathfrak{V}_\mathbf{i}\mathbf{Q}_\mathbf{i}$. In particular, we take an SVD of the projection of the stacked data matrix $[\mathbf{X}_k^\top]_{k\in\mathbf{i}}^\top$ onto the subspace spanned by $\mathfrak{V}_\mathbf{i}$ in trait space, and choose $\mathbf{Q}_\mathbf{i}$ as the matrix of right singular vectors from that calculation. This re-rotation can be thought of as sorting the modes of variation within the shared subspace in order of importance.

Since DIVAS is based on angles, additional insight into the modes of variation is derived from angles between the loadings (columns of $\mathfrak{L}_{\mathbf{i},k}\mathbf{Q}_\mathbf{i}$) and scores (columns of $\mathfrak{V}_\mathbf{i}\mathbf{Q}_\mathbf{i}$) and the object and trait spaces spanned by the estimated low rank matrix $\hat{\mathbf{A}}_k$, respectively. In particular, for each estimated score vector and loadings vector, the angle to each estimated signal matrix $\hat{\theta}_k$ and the upper bound on the angle to the true signal $\hat{\theta}_k + \theta_2^\star$ for each direction in both trait space and object space are computed. See Section 2.1.2 for a definition of $\theta_2^\star$. To calculate the upper bound for one of the vectors, we choose the 95th percentile of an empirical distribution of $\theta_2^\star$ generated using (7) and the cached matrices from the rotational bootstrap (See Section 2.1.3). Scores for all blocks are in a shared trait space, and therefore in trait space we calculate these angles not only for included ($k \in \mathbf{i}$) but also for excluded ($k \notin \mathbf{i}$) blocks for each block collection $\mathbf{i}$. The angle to the included block is expected to be small and the angle to the excluded block is expected to be large, though not necessarily 90 degrees. If some score vector has an upper bound below the random direction

bound θ_0 for an excluded block, then the corresponding mode of variation is correlated with that excluded block even though it is not joint with that block. The object spaces are block specific, and therefore for loadings we calculate angles to the included ($k \in \mathbf{i}$) data blocks only. These diagnostic angles form the crux of the overall diagnostic displays that we describe next.

2.4 DIVAS Diagnostic Graphics

We compile all angle-based diagnostics for DIVAS into comprehensive displays. These displays can be seen in Figures 5 and 6 for the synthetic data set shown in Figure 1, in Figure 7 for breast cancer omics data, and in Figures 10 and 11 for 20th century mortality data. We explain the interpretation of these displays in this section using the synthetic data example.

Figure 5 shows the diagnostic angles for the joint scores vectors found for the synthetic data example from Figure 1, and Figure 6 shows those same diagnostic angles for the joint loadings vectors. Each row of boxes corresponds to a data block and the various block collections appear in the columns. Boxes for included blocks in a given column are colored-in while boxes for excluded blocks are white. The number in each colored box specifies the rank of the estimated joint subspace between the blocks included in that column. Block collections where no partially shared joint structure was found are labeled with a 0 and grayed out. The last column labeled *Ranks* contains the dimensions of key subspaces for each data block $k = 1, \ldots, K$. The *final* rank is the dimension of the subspace spanned by all structure involving that data block, i.e. the rank of $[\mathfrak{V}_{\mathbf{i}}]_{\mathbf{i}|k\in\mathbf{i}}$. The *filtered* rank is the dimension of the estimated signal subspace in both object and trait space for that data block, i.e. $\check{r}_k$. The *maximum* rank is the largest possible dimension spanned by structure involving that data block, i.e. $d_k \wedge n$. These three ranks will usually appear in ascending order, but as discussed in Section 2.3, the final rank is sometimes larger than the filtered rank.

To explain the interpretation of the comprehensive information in DIVAS diagnostic displays, we first focus on the top-left corner of Figure 5. Each box is a scatter plot, with the horizontal axis indicating basis direction index and the vertical axis indicating angle from 0° at the bottom to 90° at the top. Within a box of this figure, each candidate direction found for that column's joint structure is represented by two points: $\times$ and $\bullet$. The $\times$ represents the angle $\hat{\theta}_k$ between the candidate direction and the corresponding estimated signal matrix for data block k, and the $\bullet$ represents the upper bound $\hat{\theta}_k + \hat{\theta}_2^\star$ on θ_k, the angle between the direction and the true subspace.

The dashed line represents the perturbation angle bound $\hat{\phi}_k$ (for trait space) or $\hat{\psi}_k$ (for object space) and the dot-dash line represents the random direction angle bound $\theta_{0,k}$. The numerical values of those angle bounds are given to the right of each group of columns. As per the inferential framework laid out in Section 2.1.2, a $\times$ below the dashed line indicates strong evidence that the direction can't be ruled out as joint structure for that data block, and a $\bullet$ above the dot-dash line indicates strong statistical evidence that the direction can't be distinguished from an arbitrarily chosen direction with respect to that data block. Due to the rank filtering procedure that takes place during the rotational bootstrap, no direction has both a $\times$ below the dashed line and a $\bullet$ above the dot-dash line.

Based on the placements of $\times$ and $\bullet$ in each colored box in Figures 5 and 6, we have strong evidence that each piece of estimated joint structure located by DIVAS is statistically significant in both trait space and object space, respectively. Specifically, all $\times$ are below the perturbation angle bound dashed line within their respective boxes. We also gain additional insight about the angular relationships between the two-way joint subspaces via the angles to the excluded blocks in Figure 5. Each $\bullet$ lies below the random direction angle bound dot-dash line in columns 3-5 of the display, indicating strong evidence that the chosen joint subspaces are distinguishable from arbitrary directions with respect to the excluded block in each column. In fact, the true joint subspaces were constructed to have pairwise principal angles of 60° between them, so this statistical rejection of arbitrariness is not surprising.

There are some situations where DIVAS basis directions appear statistically significant from an angular perspective but depend on a very small number of observations or traits. Useful insight comes from summarizing the contributions of each observation and/or trait to the shared structure. In the case of observations, we quantify their involvement with a summary statistic called the *Effective Number of Cases* (ENC) based on ideas in importance sampling from (Kish, 1965). Let v_j for $j \in \{1, \ldots, n\}$ be the entries of a chosen direction $\mathbf{v}^\star$. Note that the entries are scaled so $\mathbf{v}^\star$ has norm 1 (i.e. $\sum_{j=1}^{n} v_j^2 = 1$). In this case, the ENC is:

$$ENC = \frac{1}{\sum_{j=1}^{n} v_j^4}$$

If one entry v_j is ± 1 while the rest are 0, meaning a single observation determines the direction, then the ENC evaluates to 1. If all entries v_j have the same magnitude $\pm\frac{1}{\sqrt{n}}$, meaning all observations have equal influence on the direction, then the ENC evaluates to n. Any chosen direction will fall

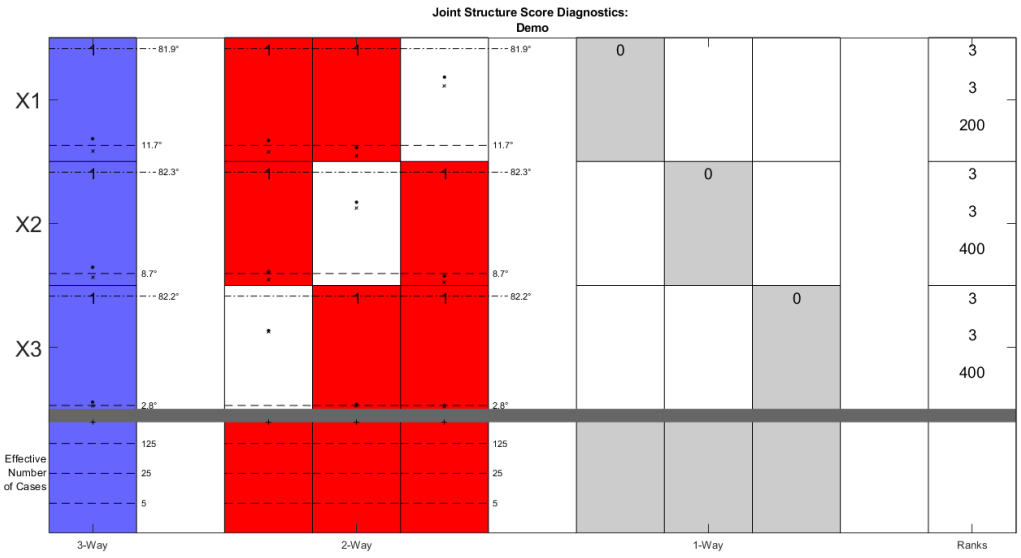

Figure 5: Summary of joint structure diagnostics for the trait spaces of the synthetic data example. All joint structure located is statistically significant, and angles to excluded blocks confirm underlying angular relationships between two-way shared subspaces. Effective Numbers of Cases (ENC) values in last row align well with the true score vectors.

somewhere between those two extremes.

The last row of Figure 5 shows the ENC for each joint scores direction found for the synthetic data example. Each box is again a scatter plot, with the horizontal axis indicating basis direction index and the vertical axis indicating the ENC from 1 to n on a logarithmic scale. Each ENC value is shown with a +. All the values for this example are very close to $n = 400$, indicating near equal contribution from each observation in all the scores vectors. This aligns with expectations since the entries of the true shared scores directions all have equal magnitude.

In the case of summarizing individual trait contributions, we use an analogous metric that takes into account the differing magnitudes of loadings vectors within data blocks and the different dimensions of loadings vectors between data blocks. To differentiate the two metrics we call this one *Effective Contribution of Traits* (ECT). ECT performs the same operation as ENC, except it uses the entries l_m for $m \in \{1, \ldots, d_k\}$ of candidate loadings directions $\mathbf{l}_k^\star$. Furthermore, it scales the result by both the magnitude $||\mathbf{l}_k^\star|| = \sum_{m=1}^{d_k} l_m^2$ of the candidate direction and by the number

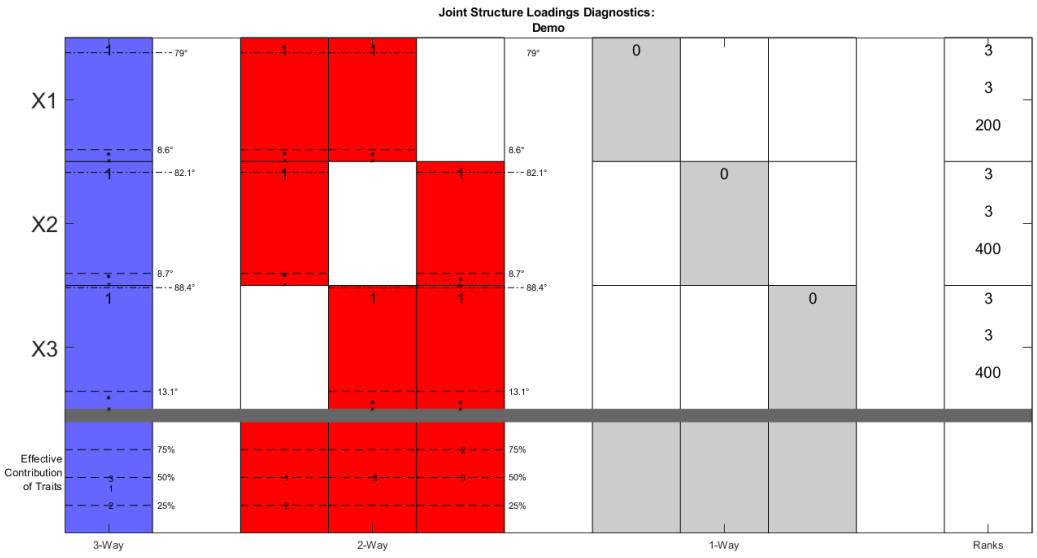

Figure 6: Summary of joint structure diagnostics for the object spaces of the synthetic example. All joint structure located is statistically significant. Effective Contributions of Traits (ECT) in last row align with expectations per the proportion of colored rows in each heatmap of Figure 1.

of traits d_k to allow for comparisons between data blocks:

$$ECT = \frac{1}{d_k} \frac{\left(\sum_{m=1}^{d_k} l_m^2\right)^2}{\sum_{m=1}^{d_k} l_m^4}.$$

The last row of Figure 6 shows the ECT for each loadings direction found for the synthetic data example. Within each box, the horizontal axis indicates basis direction index and the vertical axis indicates ECT percentage ranging from 0% to 100%. Each block has its own loadings direction for each basis direction, so each block's ECT is shown with a number corresponding with that block's index in the analysis. In all cases, the contribution percentages align quite well with the percentage of traits involved in each piece of true joint structure as per Figure 1. For example, half of the traits in X3 have the characteristic pinstripe pattern of the fully joint structure in Figure 1, and the "3" in the bottom-left box of Figure 6 sits right around 50%.

3 Case Study: Cancer Genomics

One of the examples that motivates the development of DIVAS is a four-block data set containing different views of omics data from $n = 616$ breast cancer patients from The Cancer Genome Atlas (TCGA) (Network et al., 2012). We have a gene expression (GE) data block containing 16615 gene traits, a gene copy number (CN) data block with 24174 traits, a protein expression (RPPA) data block containing 187 protein traits, and a 0-1 mutation detection (Mut) block containing traits for 128 genes. Each patient is labeled as one of four breast cancer subtypes: Luminal A, Luminal B, Basal, or Her2-enriched.

We wish to obtain the entire hierarchy of joint structure among the four data blocks. Of particular interest are the four-way joint structure, three-way partially-shared joint structures, and partially-shared structure involving proteins or mutations. The biological understanding of the protein production pathway suggests that after accounting for the gene expression and gene copy number there should be no additional variation shared between mutations and proteins. Once all joint structure is cataloged, further conclusions can be drawn from loadings of the joint modes of variation. For example, the loadings from the gene expression data block would reveal which genes are involved in a certain cancer subtype if one of the joint modes of variation discriminates that subtype from the others. Note that DIVAS is an unsupervised method that does not make use of the class labels. Development of a supervised version of DIVAS remains an interesting open problem.

Figure 7 shows the DIVAS decomposition and angle diagnostics for the joint scores vectors for this data set. Detailed descriptions of the information plotted in the figure can be found in Section 2.4. DIVAS finds a single shared component between all four data blocks, a seven-dimensional subspace shared between all data blocks besides mutation, and lots of shared structure between pairs of data blocks not involving mutation. This result aligns well with biological expectations, particularly the large amount of structure shared uniquely between gene expression and copy number.

Figure 8 displays a scores scatter plot matrix of directions from the four-way joint and three-way joint components of the data. In these plots, each point corresponds to a single data object. Each cancer subtype is shown with a different color and point symbol: basal with red triangles,

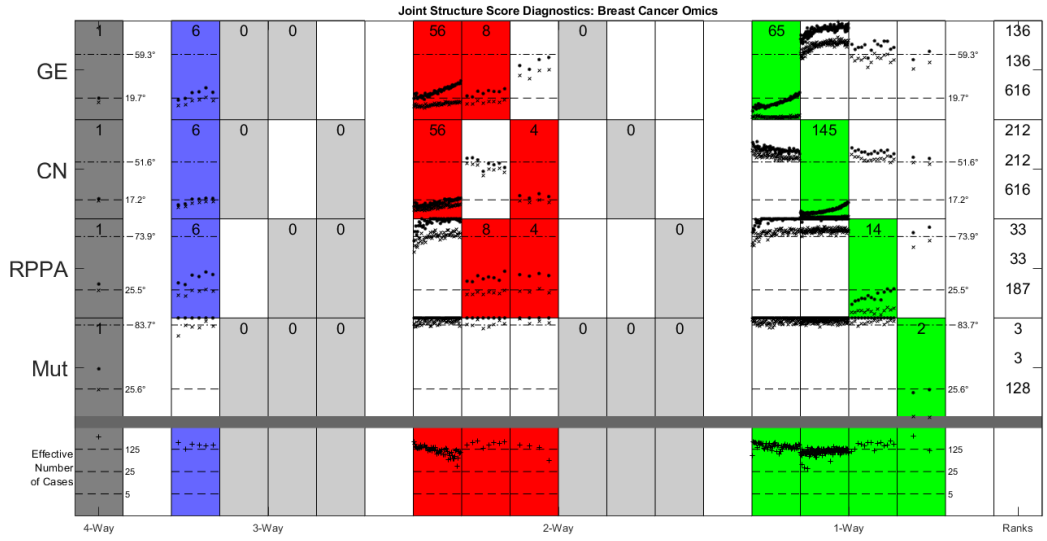

Figure 7: Joint structure breakdown and diagnostics for the breast cancer omics score vectors.

luminal A with blue asterisks, luminal B with cyan x'es, and Her2-enriched with magenta pluses. On-diagonal plots show the coefficients of projection of data objects onto the score vector indicated. The vertical axis provides a jitter for ease of visual interpretation. Solid curves in on-diagonal plots are kernel density estimates, with the black curve including all data objects and the colored curves corresponding with each respective subtype. Off-diagonal plots show scatter plots of coefficients of projection of the data objects onto the score vectors in that plot's respective axes labels. More information about these plots may be found in Chapter 1 of Marron and Dryden (2021). We chose to include the first two directions in the basis for the three-way joint subspace along with the four-way joint direction for ease of visual interpretation. The four-way joint component separates basal cases from other cases. This is typically the first component found in any analysis of the modes of variation in breast cancer patients, as basal cell cancers have a very different gene expression profile than the other subtypes. The two-dimensional plot of the scores along the first two directions in the three-way joint subspace separates Her2 cases from Basal cases primarily, and from Luminal A cases secondarily. This indicates the potential for identifying useful genes, proteins, and copy number regions that drive the variation in this three-way joint subspace that distinguish Her2 cases from other breast cancer subtypes.

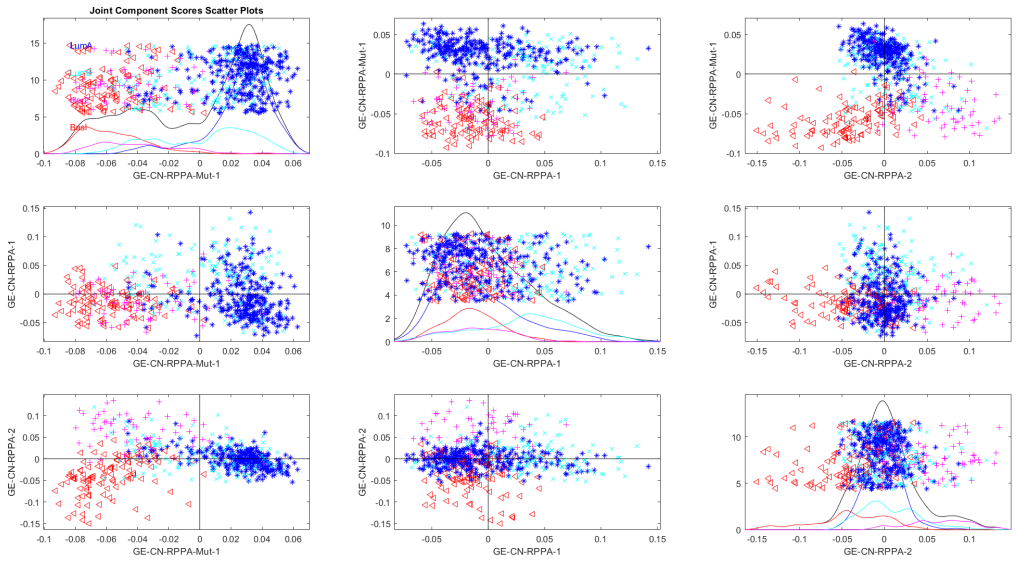

Figure 8: Scores scatterplot matrix of four-way joint direction and first two three-way joint directions. The four-way joint subspace (top left) distinguishes basal (red triangles) from other subtypes and the three-way joint subspace (bottom right) distinguishes Her2 (magenta plusses) from other subtypes.

4 Case Study: Twentieth Century Mortality

Marron and Alonso (2014) consider a data matrix containing mortality rates (proportion of the population of a given age that died in a year) of Spanish males from 1908 to 2002. We expand on this initial analysis by incorporating three additional data blocks: one for Spanish women and two more for Swiss men and women, and by increasing the end of the time frame to 2018. We are interested in how mortality rates changed over the course of this time period as a function of age. Hence, we will treat each year as a data object and the mortality rates of each age as a trait. We consider ages 12 to 90 to avoid zero counts for particularly high and low ages. Data was downloaded on April 8, 2021 from the Human Mortality Database (Wilmoth and Shkolnikov, 2021).

To appropriately handle the multiple orders of magnitude present in mortality proportions, we transform each entry of the data blocks with a *logit* function $f(x) = \log\left(\frac{x}{1-x}\right)$. After the logit transformation, each data block was *double-centered*: the mean vectors in both object space and trait space were removed from each data block. As per the discussion in Prothero et al. (2021),

this type of centering is effective when all the data blocks share a common mode of variation in the trait mean direction. The object and trait means for each data block are displayed as curves in Figure 9. Each curve is colored according to year using a rainbow color scheme starting at magenta and blue, through orange and red. In the object mean panels (top row), we see the overall mortality profile across ages for each country and gender. Males exhibit a slightly higher increase in mortality upon entering adulthood than females in both countries due to increased risk-taking behaviors at that age (Barbara Blatt Kalben F.S.A., 2000; Patton et al., 2009). We also observe systematic anomalies in the mortality rates for older Spanish individuals that are not present in the Swiss data. As discussed in Marron and Alonso (2014), these anomalies are manifestations of an age-rounding effect, and reflect major early differences in demographic record keeping practices between the two countries. The distinct rainbow sequence in each trait mean panel (bottom row) shows steady overall decreases in average mortality rate over time. The worst year of the 20th century flu pandemic, 1918, appears prominently at the top of each trait mean panel in violet. Spanish data block panels (especially males) have out-of-sequence light blue lines in their plots due to a civil war in the late 1930s.

Figure 10 shows the scores diagnostic graphic for the DIVAS decomposition of the mortality data, and Figure 11 shows the angle diagnostics for the corresponding loadings vectors (see Section 2.3). Since all four data blocks have identical trait and object dimensions and relatively similar variation, all the perturbation angles are also similar to each other. DIVAS finds a two-dimensional four-way shared component, a one-dimensional three-way component shared between each block besides Swiss females, and a six-dimensional shared component between Spanish males and females. Intuitive reasons for these findings are explained below via discussion of the modes of variation.

We further investigate the joint structure by visualizing the joint modes of variation about the mean in curve plots throughout Figures 12-14. Figure 12 shows such a visualization for the four-way joint structure. In this and subsequent mode of variation figures, each row of panels corresponds to a different basis direction and each column of panels corresponds to a data block. The final plot in each row shows the entries of the common score vector corresponding to that mode of variation. Each row of panels in this figure contains one trend in mortality that was found to be common across both countries and genders. The first mode is a contrast between older and younger individuals that manifests as a change in slope over time. In particular, while mortality rates decreased for all ages over the 20th century as per the trend seen in the trait mean, this mode of variation shows that

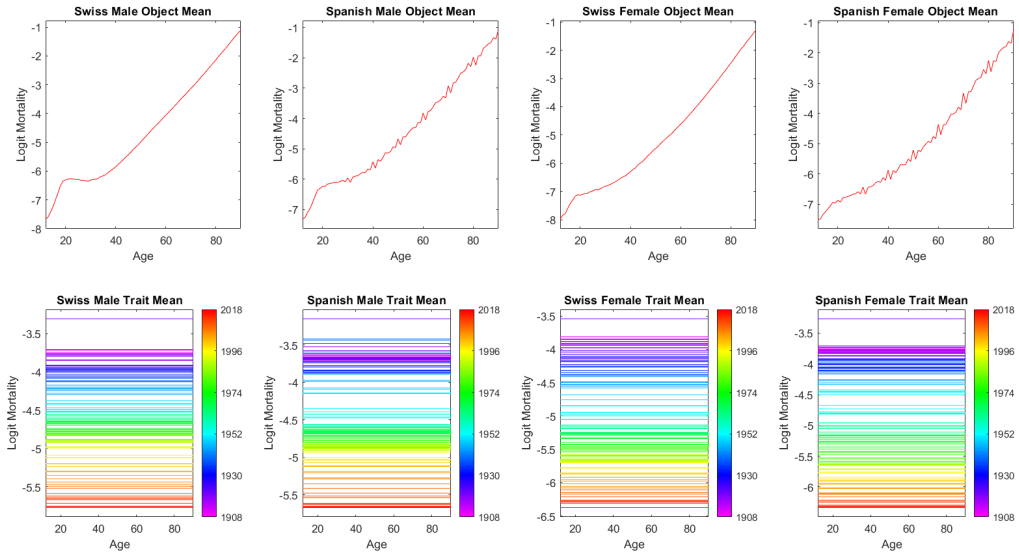

Figure 9: Object means (top row) and trait means (bottom row) for each data block. Both male data blocks have a more dramatic increase in mortality for young adults than the female data blocks. Both Spanish data blocks display effects of record-keeping round-offs absent from Swiss data blocks. All trait means capture the overall improvement in mortality rate over time across the population.

decrease was more pronounced for younger individuals. The second mode is primarily a contrast between younger adults and middle-age adults that takes place between the 1970s and 1990s, with somewhat different age groupings across blocks. This contrast is the well-documented automotive safety effect described in Marron and Alonso (2014). The wide proliferation of cars in the mid-20th century without modern safety guidelines in this time frame led to a notable increase in automobile fatalities concentrated in younger individuals. As automotive safety improved across Europe in the 1980s and 1990s, this source of excess mortality dissipated.

Figure 13 shows the single mode of variation found as three-way shared structure between Swiss men, Spanish men, and Spanish women. Before analyzing this data, we primarily expected to see four-way shared structure and pairwise shared structure across blocks with gender or country in common; this three-way shared structure deviates from that expectation. This mode of variation indicates a contrast between the mortality rates of young adults and the rest of the population during the late 1980s and early 1990s. Considering that the automotive safety effect already

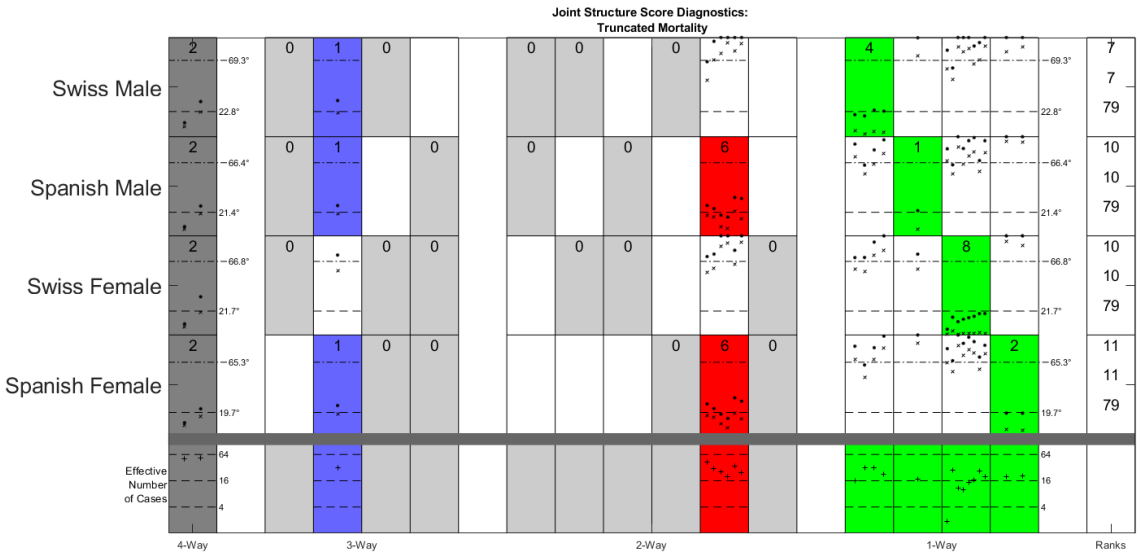

Figure 10: Joint structure breakdown and diagnostics for the mortality data score vectors. Perhaps surprising is some amount of three-way partially shared joint structure. Spanish men and women have complex two-way partially shared joint structure due to age rounding.

appeared in the fully joint component, we suspect this mode of variation is capturing a different phenomenon. Our hypothesis given the time frame and groups affected is that this component is capturing increased mortality from HIV/AIDS in the late 20th century. The contrast focusing on young males in both countries is the main driver of this hypothesis; the corresponding effect in Spanish women seems to be concentrated in older individuals so some additional effect might be entering the mode of variation in that block. This motivates further mortality research.

Figure 14 shows the six modes of variation found as two-way joint structure shared between the two Spanish blocks. Component 1 captures excess mortality of young people, and especially young men, during the Spanish Civil War. Component 2 is a contrast within young adults that we do not fully understand. The remaining four modes of variation seem to be harmonic components generated by the age rounding effect discussed in Marron and Alonso (2014).

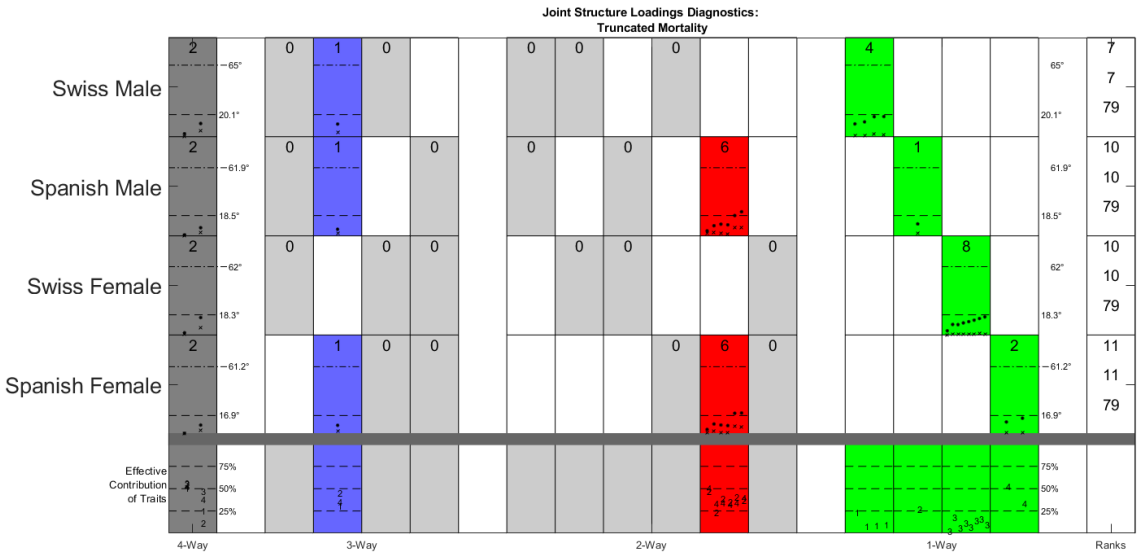

Figure 11: Joint structure breakdown and diagnostics for the mortality data loadings vectors.

5 Conclusions

This paper proposes DIVAS, a novel exploratory data analysis method for statistical data integration that allows for partially-shared structure between several distinct data blocks. The main contributions of DIVAS are twofold. First, we develop a rigorous, angle-based framework of statistical inference for diagnostically evaluating estimated shared structure. Second, we consider integration across both dimensions of the data blocks simultaneously and produce more thorough and higher-fidelity results as a consequence.

Future work on DIVAS could proceed in at several different directions. Methodologically, there remains room for additional refinement of noise estimation throughout the first step of DIVAS, both in the noise variance estimator and the residual matrix estimator. Structurally, DIVAS is fundamentally a linear, unsupervised statistical model. Generalizations and expansions that extend DIVAS to tackle supervised learning and nonlinear relationships between data blocks are promising future directions. Practically, the driving force behind development of the method has always been appropriately complex data like the breast cancer omics data set that demands such methodological sophistication. Therefore we expect further improvements in DIVAS development will be found during analysis of ever more demanding data.

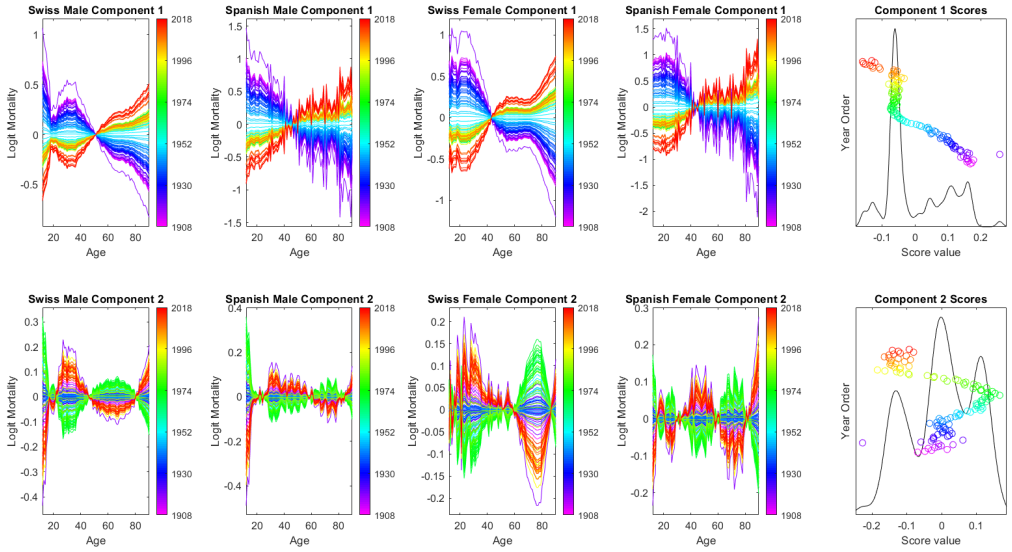

Figure 12: Mode of variation curve plots of the two four-way joint components. Component 1 shows stronger improvement in mortality for younger people, and component 2 shows the rise and fall of automobile fatalities.

6 Acknowledgement

Jan Hannig's research was supported in part by the National Science Foundation under Grant No. DMS-1916115, 2113404, and 2210337. J.S. Marron's research was partially supported by the National Science Foundation Grant No. DMS-2113404.

References

Akaho, S. (2007). A kernel method for canonical correlation analysis. *arXiv preprint*.

Barbara Blatt Kalben F.S.A., E.A., M. (2000). Why men die younger. *North American Actuarial Journal*, 4(4):83–111.

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computation Learning Theory*.

Cai, J. and Huang, X. (2017). Robust kernel canonical correlation analysis with applications to information retrieval. *Engineering Applications of Artificial Intelligence*, 64:33–42.

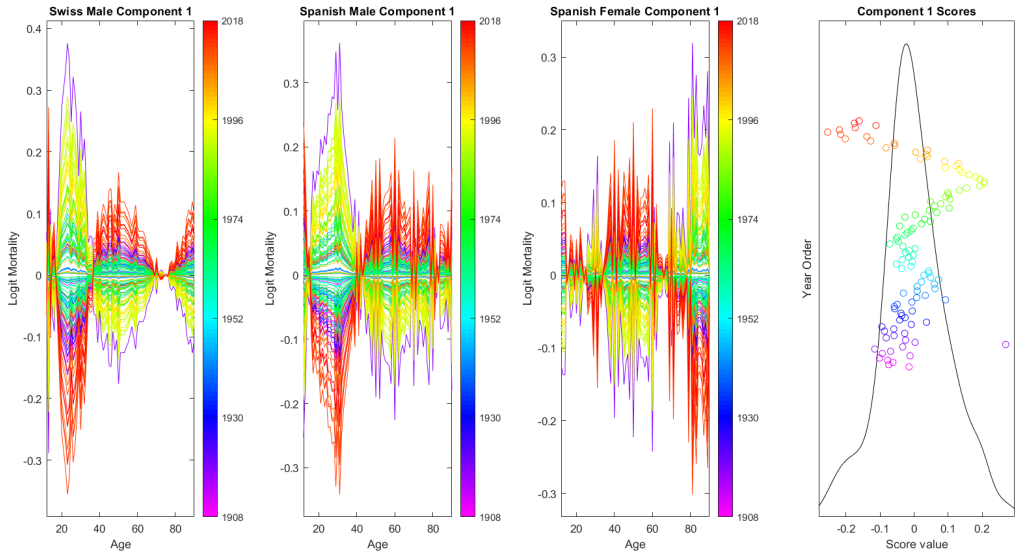

Figure 13: Mode of variation curve plots of three-way joint component between Swiss men, Spanish men, and Spanish women. Mode of variation contrasts mortality in young adults with others around 1990. Potentially related to the emergence of HIV/AIDS.

Chikuse, Y. (2012). *Statistics on Special Manifolds*. Lecture Notes in Statistics. Springer New York.

Farquhar, J. D., Hardoon, D. R., Meng, H., and Shawe-Taylor, J. (2005). Two-view learning: Svm-2k, theory and practive. In *Advances in Neural Information Processing Systems*.

Feng, Q., Jiang, M., Hannig, J., and Marron, J. S. (2018). Angle-based joint and individual variation explained. *J. Multivariate Anal.*, 166:241–265.

Gavish, M. and Donoho, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*.

Gavish, M. and Donoho, D. L. (2017). Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*.

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

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In Blondel, V., Boyd, S., and Kimura, H., editors, *Recent Advances in Learning and Control*,

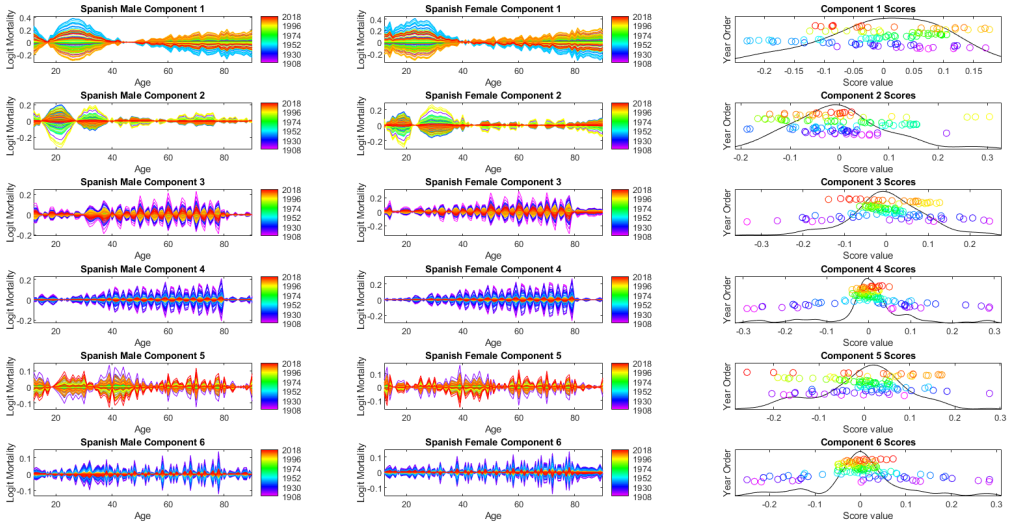

Figure 14: Mode of variation curve plots of the two-way joint components between Spanish data blocks. First component driven by the Spanish Civil War in the 1930s. Last four components are primarily driven by high-frequency-harmonic age-rounding record-keeping anomalies.

Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited. `http://stanford.edu/~boyd/graph_dcp.html`.

Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

Horst, P. (1961). Relations amongm sets of measures. *Psychometrika*, 26:129–149.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3-4):321–377.

Ismailova, D. and Lu, W.-S. (2016). Penalty convex-concave procedure for source localization problem. In *2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*.

Jiang, M. (2018). *Statistical Learning of Integrative Analysis*. PhD thesis, University of North Carolina at Chapel Hill.

Kettenring, J. (1971). Canonical analysis of several sets of variables. *Biometrika*.

Kish, L. (1965). *Survey Sampling*. J. Wiley.

Li, Y., Yang, M., and Zhang, Z. (2019). A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883.

Lock, E., Hoadley, K., Marron, J., and Nobel, A. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics*.

Lock, E. F., Park, J. Y., and Hoadley, K. A. (2020). Bidimensional linked matrix factorization for pan-omics pan-cancer analysis.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*.

Marron, J. and Dryden, I. (2021). *Object Oriented Data Analysis*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press.

Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biom. J.*, 56(5):732–753.

Miao, J. and Ben-Israel, A. (1992). On principal angles between subspaces in rn. *Linear Algebra and its Applications*, 171:81–98.

Network, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61.

Nielsen, A. (2002). Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Transactions on Image Processing*, 11(3):293–305.

Patton, G. C., Coffey, C., Sawyer, S. M., Viner, R. M., Haller, D. M., Bose, K., Vos, T., Ferguson, J., and Mathers, C. D. (2009). Global patterns of mortality in young people: a systematic analysis of population health data. *The Lancet*, 374(9693):881–892.

Prothero, J. B. (2021). *Data Integration Via Analysis of Subspaces*. PhD thesis, University of North Carolina at Chapel Hill.

Prothero, J. B., Hannig, J., and Marron, J. S. (2021). New perspectives on centering. *arXiv preprint*.

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

Shu, H. and Qu, Z. (2021). CDPA: Common and distinctive pattern analysis between high-dimensional datasets.

Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications*.

Tran-Dinh, Q. and Diehl, M. (2009). Sequential Convex Programming Methods for Solving Nonlinear Optimization Problems with DC constraints. Tech. report, ESAT/SCD and OPTEC, KU Leuven, Belgium.

Wedin, P.-A. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*.

White, M., Yu, Y., Zhang, X., and Schuurmans, D. (2012). Convex multi-view subspace learning. In *25th Internation Conference on Neural Information Processing Systems*.

Wilmoth, J. R. and Shkolnikov, V. (2000-2021). Human mortality database. *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)*.

Xu, C., Tao, D., and Xu, C. (2013). A survey on multi-view learning. *arXiv preprint*.

Yi, S., Wong, R. K. W., and Gaynanova, I. (2022). Hierarchical nuclear norm penalization for multi-view data.

Yuan, D. and Gaynanova, I. (2021). Double-matched matrix decomposition for multi-view data.

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2016). Bayesian group factor analysis with structured sparsity. *Journal of Machine Learning Research*, 17(196):1–47.

Zhua, P. and Knyazev, A. V. (2012). Principal angles between subspaces and their tangents. *MITSUBISHI ELECTRIC RESEARCH LABORATORIES*.

A Review of Random Matrix Theory

Our chosen signal extraction procedure uses random matrix theory ideas. The classical result from Marchenko and Pastur (1967) on the distribution of the eigenvalues of random matrices underpins all of these ideas; we restate that result below.

Let $\mathbf{E}$ be a $d \times n$ random matrix. The entries of $\mathbf{E}$ are independent and identically distributed (i.i.d.) with mean 0, finite variance σ^2, and finite fourth moment. Form the $d \times d$ estimator of the covariance matrix $\mathbf{\Sigma}_n = \frac{1}{n}\mathbf{E}\mathbf{E}^\top$ and let $\lambda_1, \ldots, \lambda_d$ denote the eigenvalues of $\mathbf{\Sigma}_n$. Consider the empirical measure $\mu_d(A) = \frac{1}{d}\#\{\lambda_j \in A\}, A \subset \mathbb{R}$ representing the empirical distribution of the eigenvalues of $\mathbf{\Sigma}_n$ as random variables themselves. Define an *indicator fuction* $\mathbf{1}_{\{K\}}$ for a given condition K as a function that returns 1 when condition K is satisfied and returns 0 otherwise.

Theorem 3 (Marchenko and Pastur 1967)**.** *If $d, n \to \infty$ such that $\frac{d}{n} \to \beta \in (0, +\infty)$, then μ_d converges weakly to the measure whose density is $\mu(\lambda)$:*

$$\mu(\lambda) = \begin{cases} h(\lambda)\mathbf{1}_{(1-\sqrt{\beta})^2 \leq \frac{\lambda}{\sigma^2} \leq (1+\sqrt{\beta})^2} & 0 < \beta \leq 1 \\ h(\lambda)\mathbf{1}_{(1-\sqrt{\beta})^2 \leq \frac{\lambda}{\sigma^2} \leq (1+\sqrt{\beta})^2} + \left(1 - \frac{1}{\beta}\right)\mathbf{1}_{\lambda=0} & \beta > 1 \end{cases} \tag{9}$$

where the function $h(\lambda)$ is defined below:

$$h(\lambda) = \frac{1}{2\pi}\frac{\sqrt{\left((1+\sqrt{\beta})^2 - \frac{\lambda}{\sigma^2}\right)\left(\frac{\lambda}{\sigma^2} - (1-\sqrt{\beta})^2\right)}}{\beta\lambda}. \tag{10}$$

If $d < n$, then $\beta < 1$ for $\mathbf{E}$ and $\mathbf{\Sigma}_n$ is rank d. In this case, since $\mathbf{\Sigma}_n$ is full-rank, all eigenvalues are nonzero, and asymptotically fall between $\sigma^2(1-\sqrt{\beta})^2$ and $\sigma^2(1+\sqrt{\beta})^2$. Alternatively, if $d > n$, then $\beta > 1$ for $\mathbf{E}$ and $\mathbf{\Sigma}_n$ is rank n. In this case $\mathbf{\Sigma}_n$ is not full rank so the eigenvalues $\lambda_{n+1} \ldots \lambda_d$ are all 0. In cases where $\beta > 1$ the Marchenko-Pastur density is therefore a mixture between a point mass of $1 - \frac{1}{\beta}$ at zero and a continuous portion bounded between $\sigma^2(1-\sqrt{\beta})^2$ and $\sigma^2(1+\sqrt{\beta})^2$ with total area $\frac{1}{\beta}$.

B Review of Principal Angle Analysis

The following is based on (Zhua and Knyazev, 2012) and (Miao and Ben-Israel, 1992). *Principal angle analysis* characterizes the relative positions of two subspaces $\mathcal{X}$ and $\mathcal{Y}$ in Euclidean space using canonical angles found via SVD. In particular, let $\mathbf{W}_\mathcal{X}$ and $\mathbf{W}_\mathcal{Y}$ be orthonormal basis matrices for $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then the singular value decomposition of $\mathbf{W}_\mathcal{X}^\top\mathbf{W}_\mathcal{Y}$ finds both the principal angles between $\mathcal{X}$ and $\mathcal{Y}$ and the corresponding *principal vectors*. Write the singular

value decomposition of $\mathbf{W}_{\mathcal{X}}^{\top}\mathbf{W}_{\mathcal{Y}}$ as $\mathbf{W}_{\mathcal{X}}^{\top}\mathbf{W}_{\mathcal{Y}} = \mathbf{U}\mathbf{D}\mathbf{V}^{\top}$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices containing the principal vectors of $\mathcal{X}$ and $\mathcal{Y}$ respectively, and $\mathbf{D}$ is a diagonal matrix. The inverse cosines of the nonzero entries of $\mathbf{D}$ give the principal angles between $\mathcal{X}$ and $\mathcal{Y}$, and in particular the angles between each pair of corresponding principal vectors. The jth pair of principal vectors have an angle between them equal to the jth principal angle.

This perspective also demonstrates the result of principal angle analysis when the dimensions of $\mathcal{X}$ and $\mathcal{Y}$ differ. Let the dimensions of $\mathcal{X}$ and $\mathcal{Y}$ be p and q respectively, with $p < q$. In this case some of the singular values will be zero as the matrix $\mathbf{W}_{\mathcal{X}}^{\top}\mathbf{W}_{\mathcal{Y}}$ is non-square, and the inverse cosine of zero is 90°. If $p < q$, the principal angles $\theta_{p+1}, \ldots, \theta_q$ are all 90°.

Principal angle analysis is also orthogonally invariant. In particular, the principal angles between $\mathcal{X}$ and $\mathcal{Y}$ will be identical to the principal angles between reoriented versions $\mathbf{O}\mathcal{X}$ and $\mathbf{O}\mathcal{Y}$, where $\mathbf{O}$ is an orthogonal matrix and $\mathbf{O}\mathcal{X} = \{\mathbf{O}\mathbf{x} | \mathbf{x} \in \mathcal{X}\}$. The matrices $\mathbf{O}\mathbf{W}_{\mathcal{X}}$ and $\mathbf{O}\mathbf{W}_{\mathcal{Y}}$ represent orthonormal bases for $\mathbf{O}\mathcal{X}$ and $\mathbf{O}\mathcal{Y}$, so the principal angle structure between the two rotated subspaces is found by taking a singular value decomposition of $\mathbf{W}_{\mathcal{X}}^{\top}\mathbf{O}^{\top}\mathbf{O}\mathbf{W}_{\mathcal{Y}}$, which is equivalent to that of $\mathbf{W}_{\mathcal{X}}^{\top}\mathbf{W}_{\mathcal{Y}}$.

C Noise Matrix Estimation

The residual $\hat{\mathbf{E}} = \mathbf{X} - \hat{\mathbf{A}}$ is a poor estimate of the non-signal component of the data, especially in the case of a non-square matrix. Heuristically, this is caused by the residual lying entirely in the subspace spanned by the data. We investigate the causes of this phenomenon and propose a solution.

For this investigation our synthetic data will be a 5000×500 matrix $\mathbf{X} = \mathbf{A} + \mathbf{E}$. The signal matrix $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{\top}$ is rank 50 with equally-spaced singular values from 0.1 to 5. $\mathbf{E}$ is a full-rank i.i.d. Gaussian matrix with variance $\frac{\sigma^2}{5000}$. This scaling of the noise variance by the number of traits is common in the matrix signal processing literature (Gavish and Donoho, 2014, 2017). It results in columns with expected norm σ and sets the noise at a level commensurate with the magnitude of the signal. With $\sigma = 1$, we expect most of the singular values to be easily recoverable while others are indistinguishable from the noise. We perform signal extraction as described in Section 2.1.1 on this matrix and subsequently examine estimators of $\mathbf{E}$ given $\hat{\mathbf{A}}$.

One way to check the efficacy of an estimator $\hat{\mathbf{E}}$ for $\mathbf{E}$ is to see how well its eigenvalues align with the Marchenko-Pastur distribution (see Appendix A). We can compare the observed values to theoretical quantiles using a *quantile-quantile* (Q-Q) plot. On the horizontal axis we plot the sorted observed eigenvalues for a noise matrix estimate $\hat{\mathbf{E}}$, and on the vertical axis we plot evenly-spaced quantiles of the Marchenko-Pastur distribution with parameter $\beta = \frac{d \wedge n}{d \vee n}$ (here d and n are the row and column dimensions of the matrix respectively). Typically, if the plotted points on a Q-Q plot roughly follow the 45° line, the conclusion is that the observed data aligns well with the theoretical distribution. To get a sense of how much variability to expect about the 45° line, we generate $M = 100$ i.i.d. Gaussian matrices and plot their eigenvalues as green lines underneath the magenta Q-Q points. These traces create a visually striking region of acceptable variability which can be used to judge the goodness of fit at a glance.

Figure 15 shows such a Q-Q plot for the eigenvalues of the naïve noise estimate $\hat{\mathbf{E}} = \mathbf{X} - \hat{\mathbf{A}}$ for the synthetic data matrix. The naïve estimated non-signal component tends to display perhaps unexpectedly low energy in directions associated with the estimated signal subspace. This phenomenon leads to Q-Q plots that are challenging to interpret. For this matrix the estimated signal rank is 44, and the bottom 44 eigenvalues of the estimated noise matrix completely deviate from the theoretical Marchenko-Pastur distribution. Importantly, this phenomenon (explained in detail below) occurs regardless of the chosen estimate for σ. The aberration in this graphic demonstrates the ineffectiveness of the naïve estimate. The rotational bootstrap procedure (see Section 2.1.3) central to DIVAS depends on effective estimation of the underlying noise matrix $\mathbf{E}$. This is accomplished via a correction to a portion of the singular values of $\hat{\mathbf{E}}$.

To explain this behavior and motivate our proposed correction, we consider our data model (1) in a special case where the signal is rank one, and the signal, noise and data are all vectors in $\mathbb{R}^2$, illustrated in Figure 16. The signal (green) and noise (red) vectors each lie in distinct one-dimensional subspaces. The two vectors are added together to form the data vector (blue). When we form our estimate of the signal $\hat{\mathbf{A}}$ (green-blue dashed) our shrinkage procedure gives us a good estimate of signal magnitude. However, it is challenging to recover directional information about $\mathbf{A}$ as the estimate $\hat{\mathbf{A}}$ lies in the same subspace as the data. When we next subtract $\hat{\mathbf{A}}$ from the data $\mathbf{X}$ to form the naïve estimate $\hat{\mathbf{E}}$ (red-blue dashed), the subtraction occurs entirely in the data subspace so we don't account for the angle between the initial signal and noise vectors at all. This leads to an underestimation of noise energy: the length of the estimated noise within the data

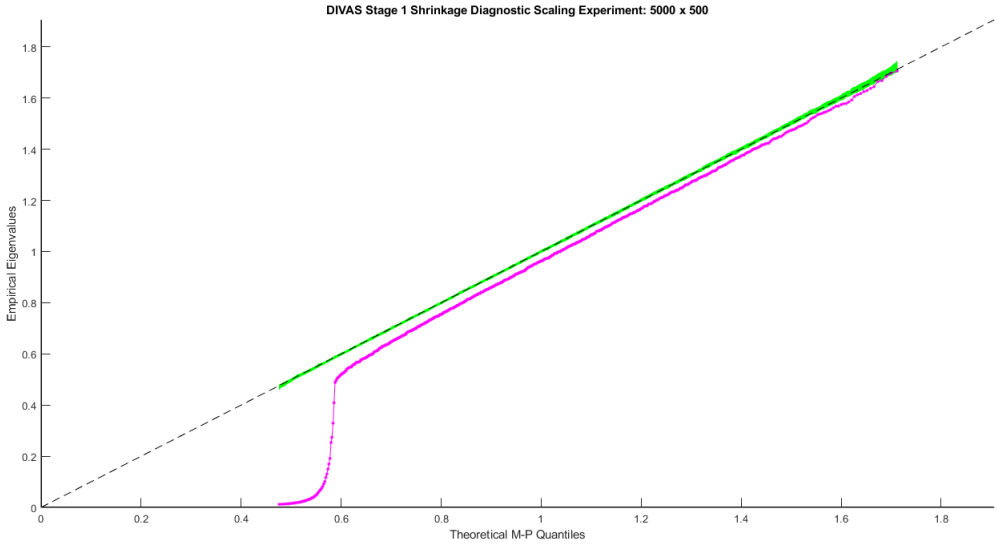

Figure 15: Q-Q plot for the eigenvalues of the naïve noise matrix estimate. The first $\hat{r}$ eigenvalues fall entirely outside the range determined by Theorem 3, signaling that the naive noise matrix estimate is flawed. These eigenvalues are also scaled using the original noise level estimate to retain some interpretability. Scaling using the apparent noise level in the estimated error matrix produces an even worse fit to the Marchenko-Pastur distribution because the apparent noise level is too low.

subspace (red-blue dashed) is distinctly shorter than the length of the original noise vector (red). This length discrepancy is the one-dimensional analog of the phenomenon shown in Figure 15 where many of the smallest eigenvalues are even smaller than expected.

Several potential corrections for this effect are proposed in Chapter 4 of Prothero (2021). We present the correction used in DIVAS here. Consider $\mathbf{X}$ as a sum of rank 1 approximations in the manner of (3): $\mathbf{X} = \sum_{i=1}^{d \wedge n} \bar{\nu}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top$. Once we estimate the signal singular values, we can split the energy in the associated singular vector directions into signal energy $\hat{\nu}$ and non-signal energy $\bar{\nu} - \hat{\nu}$:

$$\mathbf{X} = \sum_{i=1}^{\hat{r}} \hat{\nu}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top + \sum_{i=1}^{\hat{r}} (\bar{\nu}_i - \hat{\nu}_i) \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top + \sum_{i=\hat{r}+1}^{d \wedge n} \bar{\nu}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top \tag{11}$$

The Gavish-Donoho shrinkage function (4) gives us good estimates for the first $\hat{r}$ singular values $\hat{\nu}_{1:\hat{r}}$ while confirming many of the $\bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top$ subspaces and associated singular values as noise. However, by subtracting $\hat{\mathbf{A}} = \sum_{i=1}^{\hat{r}} \hat{\nu}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top$ from $\mathbf{X}$ we are overestimating the influence of the signal within the data subspace as the $\bar{\nu}_i - \hat{\nu}_i$ terms of $\hat{\mathbf{E}}$ have inordinately low energy in directions associated with the estimated signal.

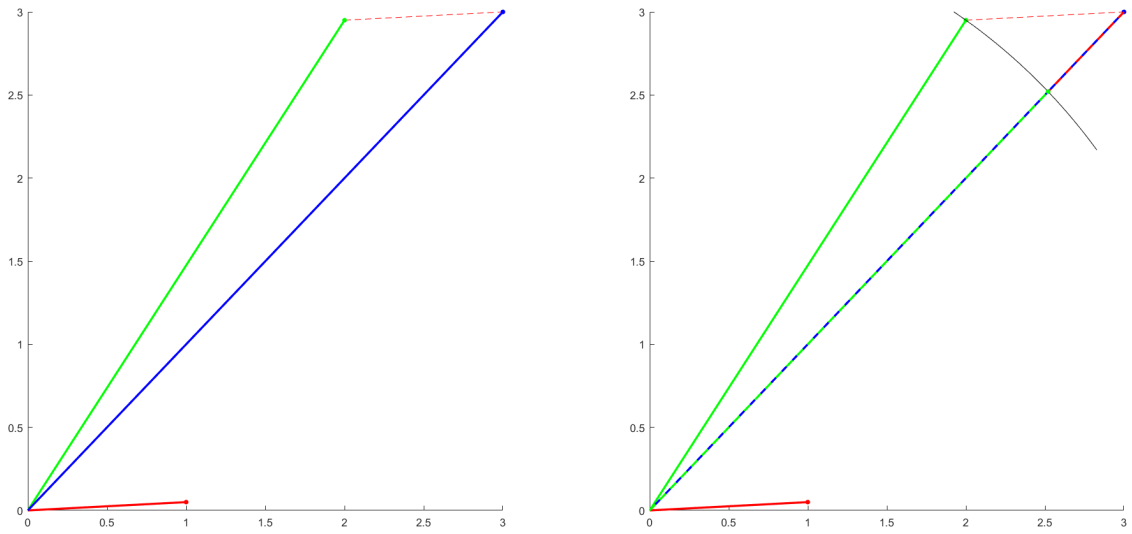

Figure 16: Example of noise energy underestimation for a rank-one signal subspace in $\mathbb{R}^2$. Left: Signal space (green), noise space (red), and data space (blue). Data vector formed by adding signal and noise vectors tip to tail. Right: Estimating $\hat{\mathbf{A}}$ (green-blue dashed) and $\hat{\mathbf{E}} = \mathbf{X} - \hat{\mathbf{A}}$ (red-blue dashed). When we remove energy equal to that of the signal space from the data space the leftover energy is noticeably smaller than the true noise energy. Note that the black arc indicates a rotation rather than a projection, so the green-blue dashed line has the same length as the green line.

The DIVAS solution to this energy deficiency is to replace each deficient singular value in $\hat{\mathbf{E}}$ with a Marchenko-Pastur random variate. Let $MP_q(\beta)$ be the qth percentile of the Marchenko-Pastur distribution with parameter β, let $U_{1:\hat{r}}$ be $\hat{r}$ i.i.d. standard uniform random variables, and let $\hat{\sigma}^2$ be an estimate of the noise variance. We form the *imputed* noise matrix estimate $\hat{\mathbf{E}}_{impute}$ as follows:

$$\hat{\mathbf{E}}_{impute} = \sum_{i=1}^{\hat{r}} \hat{\sigma} MP_{U_i}(\beta) \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top + \sum_{i=\hat{r}+1}^{d \wedge n} \bar{\nu}_i \bar{\mathbf{u}}_i \bar{\mathbf{v}}_i^\top \tag{12}$$

Figure 17 shows the original Q-Q plot from Figure 15 with the eigenvalues of $\hat{\mathbf{E}}_{impute}$ for the synthetic data matrix also included in black. After imputing the deficient singular values, the eigenvalues of the reconstructed noise matrix estimate follow the expected Marchenko-Pastur distribution quite closely; nearly all of them fall within the green acceptable variability envelope.

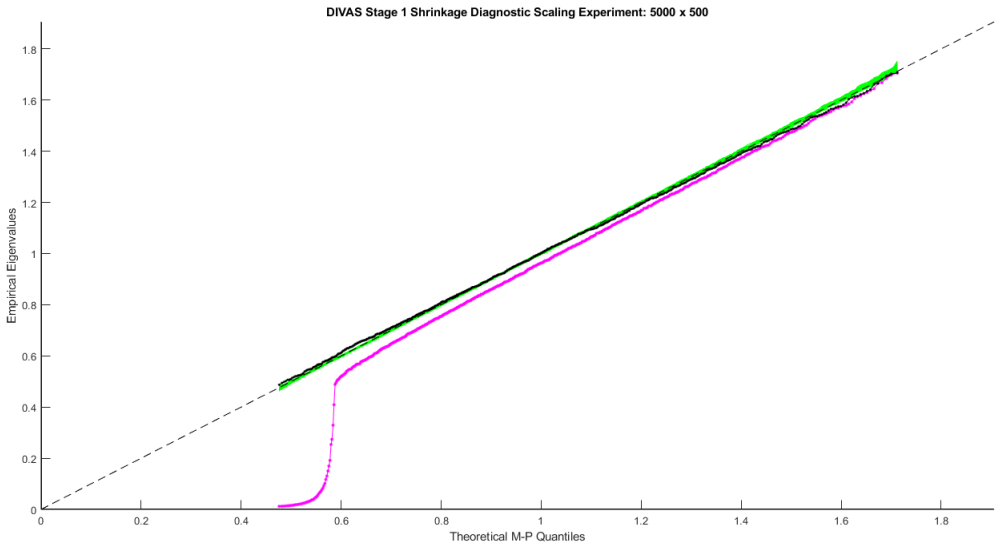

Figure 17: Q-Q plot for the eigenvalues of the naïve noise matrix estimate (magenta) and the imputed noise matrix estimate (black) from the synthetic data matrix. The corrected eigenvalues largely remain within the green acceptable variability envelope.

D Optimization Algorithm and Implementation

In this appendix we provide the details of our numerical algorithm to solve the optimization problem (8). First, we will explicitly rewrite (8) into a convex-concave optimization problem, also called a DC (difference of two convex functions) program. The detail primarily involves the steps to reformulate the problem (8) in terms of the convex-concave setting described in Ismailova and Lu (2016), and subsequently implements that convex-concave procedure for solving the resulting problem. The convex-concave procedure (or also called a DC algorithm) can be found in the literature including Ismailova and Lu (2016); Tran-Dinh and Diehl (2009). Since our problem has both the DC objective function and DC constraints, we can use the convergence analysis from (Tran-Dinh and Diehl, 2009) to guarantee the well-definedness of our algorithm.

DC programming reformulation of (8)**.** To move towards to a DC programming reformulation of (8), we express the angles between candidate directions $\mathbf{v}^\star$ and various subspaces in terms of their squared cosines. For an arbitrary-magnitude $\mathbf{v}^\star$ and orthonormal basis matrix $\mathbf{V}$ for a subspace, if we define $\hat{\theta}_{\mathbf{V}} = \angle(\mathbf{v}^\star, \mathbf{V})$, then we have $\cos^2(\hat{\theta}_{\mathbf{V}}) = \frac{\mathbf{v}^{\star\top}\mathbf{V}\mathbf{V}^\top\mathbf{v}^\star}{\mathbf{v}^{\star\top}\mathbf{v}^\star}$. More specifically, by using the representation $\cos^2(\hat{\theta}_{Tk}) = \frac{\mathbf{v}^{\star\top}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\mathbf{v}^\star}{\mathbf{v}^{\star\top}\mathbf{v}^\star}$ and keeping in mind the orthonormal condition that

$\mathbf{v}^{\star\top}\mathbf{v}^\star = 1$, the objective function of (8) becomes $-\sum_{k\in\mathbf{i}}\cos^2(\hat{\theta}_{Tk}) = -\mathbf{v}^{\star\top}\left(\sum_{k\in\mathbf{i}}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\right)\mathbf{v}^\star$. Next, using the decreasing monotonicity of $\cos^2$ in $[0,\frac{\pi}{2}]$, the constraint $\hat{\theta}_{Tk} \le \hat{\phi}_k$ is equivalent to $\cos^2(\hat{\theta}_{Tk}) = \frac{\mathbf{v}^{\star\top}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\mathbf{v}^\star}{\mathbf{v}^{\star\top}\mathbf{v}^\star} \ge \cos^2(\hat{\phi}_k)$ for all $k \in \mathbf{i}$. Similarly, $\hat{\theta}_{Tk} \ge \hat{\phi}_k$ is equivalent to $\cos^2(\hat{\theta}_{Tk}) = \frac{\mathbf{v}^{\star\top}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\mathbf{v}^\star}{\mathbf{v}^{\star\top}\mathbf{v}^\star} \le \cos^2(\hat{\phi}_k)$ for all $k \in \mathbf{i}^c$. The constraint $\hat{\theta}_{Ok} \le \hat{\psi}_k$ is equivalent to $\cos^2(\hat{\theta}_{Ok}) = \frac{\mathbf{v}^{\star\top}\mathbf{X}_k^\top\check{\mathbf{U}}_k\check{\mathbf{U}}_k^\top\mathbf{X}_k\mathbf{v}^\star}{\mathbf{v}^{\star\top}\mathbf{X}_k^\top\mathbf{X}_k\mathbf{v}^\star} \ge \cos^2(\hat{\psi}_k)$ for all $k \in \mathbf{i}$. Finally, we multiply all these constraint reformulations by $\mathbf{v}^{\star\top}\mathbf{v}^\star$ to eliminate their denominator, and transform them into DC constraints. The orthonormal constraint $\mathbf{v}^{\star\top}\mathbf{v}^\star = 1$ is equivalent to $\mathbf{v}^{\star\top}\mathbf{v}^\star - 1 \le 0$ and $1 - \mathbf{v}^{\star\top}\mathbf{v}^\star \le 0$. Putting these transformations together, we can easily see that (8) is equivalent to the following problem:

$$\left\{\begin{array}{lllll}
\min_{\mathbf{v}^\star} & -\mathbf{v}^{\star\top}\left(\sum_{k\in\mathbf{i}}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\right)\mathbf{v}^\star & & & \\
s.t. & \hat{\theta}_{Tk} & = \angle(\mathbf{v}^\star, \check{\mathbf{V}}_k) & \forall k & \\
& \hat{\theta}_{Ok} & = \angle(\mathbf{X}_k\mathbf{v}^\star, \check{\mathbf{U}}_k) & \forall k & \\
& \cos^2(\hat{\phi}_k)\mathbf{v}^{\star\top}\mathbf{v}^\star - \mathbf{v}^{\star\top}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\mathbf{v}^\star & \le 0 & \forall k & \in \mathbf{i} \\
& \mathbf{v}^{\star\top}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\mathbf{v}^\star - \cos^2(\hat{\phi}_k)\mathbf{v}^{\star\top}\mathbf{v}^\star & \le 0 & \forall k & \in \mathbf{i}^c \\
& \cos^2(\hat{\psi}_k)\mathbf{v}^{\star\top}\mathbf{X}_k^\top\mathbf{X}_k\mathbf{v}^\star - \mathbf{v}^{\star\top}\mathbf{X}_k^\top\check{\mathbf{U}}_k\check{\mathbf{U}}_k^\top\mathbf{X}_k\mathbf{v}^\star & \le 0 & \forall k & \in \mathbf{i} \\
& \mathbf{v}^\star & \perp \mathfrak{V}_\mathbf{j} & \forall \mathbf{j} & \supseteq \mathbf{i} \\
& \mathbf{v}^{\star\top}\mathbf{v}^\star - 1 & \le 0 & & \\
& 1 - \mathbf{v}^{\star\top}\mathbf{v}^\star & \le 0 & &
\end{array}\right. \tag{13}$$

Clearly, the objective function of (13) is concave. In addition, the third, the fourth, the fifth, and the last constraints of (13) are DC constraints of the form $f(\mathbf{v}^\star) - g(\mathbf{v}^\star) \le 0$. Therefore, (13) is a DC program. This problem is feasible depending on the choice of $\hat{\phi}_k$ and $\hat{\psi}_k$. However, due to the orthonormal constraint $\mathbf{v}^{\star\top}\mathbf{v}^\star = 1$, the feasible set of problem (13) does not have nonempty interior. In this case, to guarantee the DC algorithm being well-defined, we will relax it by adding slack variables.

DC algorithm. Note that the objective function of (13) is though concave, it can be written into a DC function $f_0(\mathbf{v}^\star) - g_0(\mathbf{v}^\star)$, where $f_0 = 0$ and g_0 is a quadratic function. Assume that we have m_c DC constraints. Then, all the DC constraints can be written as $f_k(\mathbf{v}^\star) - g_k(\mathbf{v}^\star) \le 0$ for $k = 1, \cdots, m_c$. The other convex constraints are expressed as $\mathbf{v}^\star \in \mathcal{F}$, including the orthogonal constraints $\mathbf{v}^\star \perp \mathfrak{V}_\mathbf{j}$ for all $\mathbf{j} \supseteq \mathbf{i}$, which are in fact linear. However, to guarantee the feasibility of our DC program, we instead relax the DC constraints to obtain $f_k(\mathbf{v}^\star) - g_k(\mathbf{v}^\star) \le s_k$, where $s_k \ge 0$ are given slack variables. We also penalize the slack variables s_k into the objective function with a given penalty parameter $\tau > 0$ to better approximate feasible solutions of (13). Therefore, we can

write the relaxation form of (13) into the following DC program:

$$\begin{cases} \min\limits_{\mathbf{v}^\star, s_k} & f_0(\mathbf{v}^\star) - g_0(\mathbf{v}^\star) + \tau \sum_{k=1}^{m_c} s_k \\ \text{s.t.} & f_k(\mathbf{v}^\star) - g_k(\mathbf{v}^\star) \le s_k, \quad \forall i = 1, \cdots, m_c, \\ & \mathbf{v}^\star \in \mathcal{F}, \quad s_k \ge 0, \ (k = 1, \cdots, m_c). \end{cases} \tag{14}$$

Note that if $s_k = 0$ for $k = 1, \cdots, m_c$, then (14) reduces to (13). To solve (14), we apply a DC algorithm (see, e.g., Ismailova and Lu (2016); Tran-Dinh and Diehl (2009)), which can be roughly described as follows.

1. *Initialization:* At the iteration $t = 0$, find an initial point $\mathbf{v}_0$ of (14) (specified later).

2. *Iteration t.* At each iteration $t \ge 0$, given $\mathbf{v}_t$, linearize the concave parts of (14) to obtain the following convex optimization subproblem:

$$\begin{cases} \min\limits_{\mathbf{v}^\star, s_k} & f_0(\mathbf{v}^\star) - [g_0(\mathbf{v}_t) + \nabla g_0(\mathbf{v}_t)^\top (\mathbf{v} - \mathbf{v}_t)] + \tau \sum_{k=1}^{m_c} s_k \\ \text{s.t.} & f_k(\mathbf{v}^\star) - [g_k(\mathbf{v}_t) + \nabla g_k(\mathbf{v}_t)^\top (\mathbf{v} - \mathbf{v}_t)] \le s_k, \quad (k = 1, \cdots, m_c), \\ & \mathbf{v}^\star \in \mathcal{F}, \quad s_k \ge 0, \ (k = 1, \cdots, m_c). \end{cases} \tag{15}$$

   Solve (15) to obtain an optimal solution $\mathbf{v}_{t+1}$ and repeat the next iteration $t + 1$ with $\mathbf{v}_{t+1}$.

3. *Termination.* The algorithm is terminated if it does not significantly improve the objective values, or other criteria are met.

Note that as proven in (Tran-Dinh and Diehl, 2009), under mild conditions imposed on (14), our DC algorithm guarantees that the sequence $\{\mathbf{v}_t\}$ generated by our DC algorithm converges to a stationary point of (14) (i.e. the point satisfying the optimality condition of (14)). We do not repeat the convergence analysis of our DC procedure here, but refer to (Tran-Dinh and Diehl, 2009) for more details.

Detailed implementation. We now specify the detailed implementation of our DC procedure as follows. The first step is to choose an initial point $\mathbf{v}_0$ for our DC program (13). Without relaxation, choosing a feasible initial point for (13) is indeed challenging. Hence, we introduce slack variables s_k to the constraints to guarantee that our relaxed DC program is always feasible and thus our algorithm is well-defined and can proceed. For instance, one can directly choose an arbitrary $\mathbf{v}_0$ in $\mathcal{F}$ first, and then set $s_{k,0} = \max\{f_k(\mathbf{v}_0) - g_k(\mathbf{v}_0), 0\}$ for each k to obtain a feasible point of (13).

In our implementation, we choose as our initial point for the ith direction in the joint subspace of block collection $\mathbf{i}$ the ith right singular vector from the SVD of $[\mathbf{V}_k]_{k\in\mathbf{i}}^\top$, which is related to the joint structure found via the AJIVE algorithm (Feng et al., 2018). If necessary, the chosen initial point is also projected to obey any orthogonality constraints present at that point in the algorithm. As explained, the slack variables s_k introduced to allow for an infeasible initial condition also appear in the objective function. They are penalized with a weight τ (also called the penalty parameter) which changes on each iteration of the optimization problem as τ_t. Notably, the values of the quadratic forms involved in the object space constraints are often much larger than those for the trait space constraints as the object space constraints include the full-energy data matrices $\mathbf{X}_k$. Therefore, we downweight the slack penalty on those constraints by the leading singular value $\bar{\nu}_{1,k}$ of $\mathbf{X}_k$ so the optimization problem is not overly restricted by the object space constraints. For further computational efficiency, if the algorithm reaches a point where all the angle-constraint slack variables are zero, it will stop early and add to the current basis a normalized version of the current iteration's intermediate solution.

Next, if we specify the convex optimization subproblem (15) for (13), then it becomes

$$
\begin{aligned}
\min_{\mathbf{v}^\star} \quad & -2\mathbf{v}_0^\top\left(\sum_{k\in\mathbf{i}}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\right)\mathbf{v}^\star + \mathbf{v}_0^\top\left(\sum_{k\in\mathbf{i}}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\right)\mathbf{v}_0 + \tau_t\sum_{k=1}^{2K+2}s_k \\
s.t. \quad & \mathbf{v}^{\star\top}\mathbf{v}^\star - 2\frac{\mathbf{v}_0^\top\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\mathbf{v}^\star}{\cos^2\left(\hat{\phi}_k\right)} + \frac{\mathbf{v}_0^\top\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\mathbf{v}_0}{\cos^2\left(\hat{\phi}_k\right)} \le s_k \qquad \forall k\in\mathbf{i} \\
& \frac{\mathbf{v}^{\star\top}\check{\mathbf{V}}_k\check{\mathbf{V}}_k^\top\mathbf{v}^\star}{\cos^2\left(\hat{\phi}_k\right)} - 2\mathbf{v}_0^\top\mathbf{v}^\star + \mathbf{v}_0^\top\mathbf{v}_0 \le s_k \qquad \forall k\in\mathbf{i}^c \\
& \mathbf{v}^{\star\top}\mathbf{X}_k^\top\mathbf{X}_k\mathbf{v}^\star - 2\frac{\mathbf{v}_0^\top\mathbf{X}_k^\top\check{\mathbf{U}}_k\check{\mathbf{U}}_k^\top\mathbf{X}_k\mathbf{v}^\star}{\cos^2\left(\hat{\psi}_k\right)} + \frac{\mathbf{v}_0^\top\mathbf{X}_k^\top\check{\mathbf{U}}_k\check{\mathbf{U}}_k^\top\mathbf{X}_k\mathbf{v}_0}{\cos^2\left(\hat{\psi}_k\right)} \le s_{K+k}/\bar{\nu}_{1,k} \qquad \forall k\in\mathbf{i} \\
& 1 - 2\mathbf{v}_0^\top\mathbf{v}^\star + \mathbf{v}_0^\top\mathbf{v}_0 \le s_{2K+1} \\
& \mathbf{v}^{\star\top}\mathbf{v}^\star - 1 \le s_{2K+2} \\
& \mathfrak{V}_\mathbf{j}^\top\mathbf{v}^\star = \mathbf{0} \qquad \forall\mathbf{j}\supseteq\mathbf{i}.
\end{aligned}
\tag{16}
$$

This problem is in fact a convex optimization problem with linear objective function and convex quadratic and linear constraints, which can be efficiently solved by several convex optimization solvers, including interior-point methods. In our implementation, we use a MATLAB package associated with a default solver, called CVX from (Grant and Boyd, 2014, 2008) to model the subproblem (16). If we use CVX's default solver, SDPT3, then our algorithm runs in about 30

minutes on the mortality data example from Section 4 on the authors' laptop with the default solver. Larger data sets like the breast cancer genomics example can take considerably longer, but substantial speed-ups are possible with Mosek and other commercial solvers.

To terminate our algorithm, one can look at the objective value of (13) to see if it is actually improved through iterations. If after, e.g., five consecutive iterations, the objective values do not significantly improved, then we can terminate it. Alternatively, we can also look at the quality of the final solution to see if it is reasonable to terminate the algorithm or not.

Experiments. Figure 18 shows the iterative process of the optimization problem for the synthetic data example displayed in Figure 1. On each panel, the horizontal axis represents the number of iterations and the vertical axis represents the angles in degrees to the panel's respective estimated signal subspace in trait space. Each horizontal green line represents the trait space perturbation angle bound. The blue paths represent the angles to the low rank approximations of trait spaces at each iteration. The red dashed paths represent the angle to the true trait subspaces at each iteration, which are known since this is a synthetic data set. Note that in the right panel in the first row, the initial condition is infeasible for the X3 data block's angle perturbation bound, demonstrating how the algorithm can use the flexibility afforded by the slack variables to explore the space before choosing a final solution. Since the objective function is trying to minimize the total squared cosine of all included blocks, the angle with one data block may increase, as in the top middle panel, if it means the angle with other data blocks decreases.

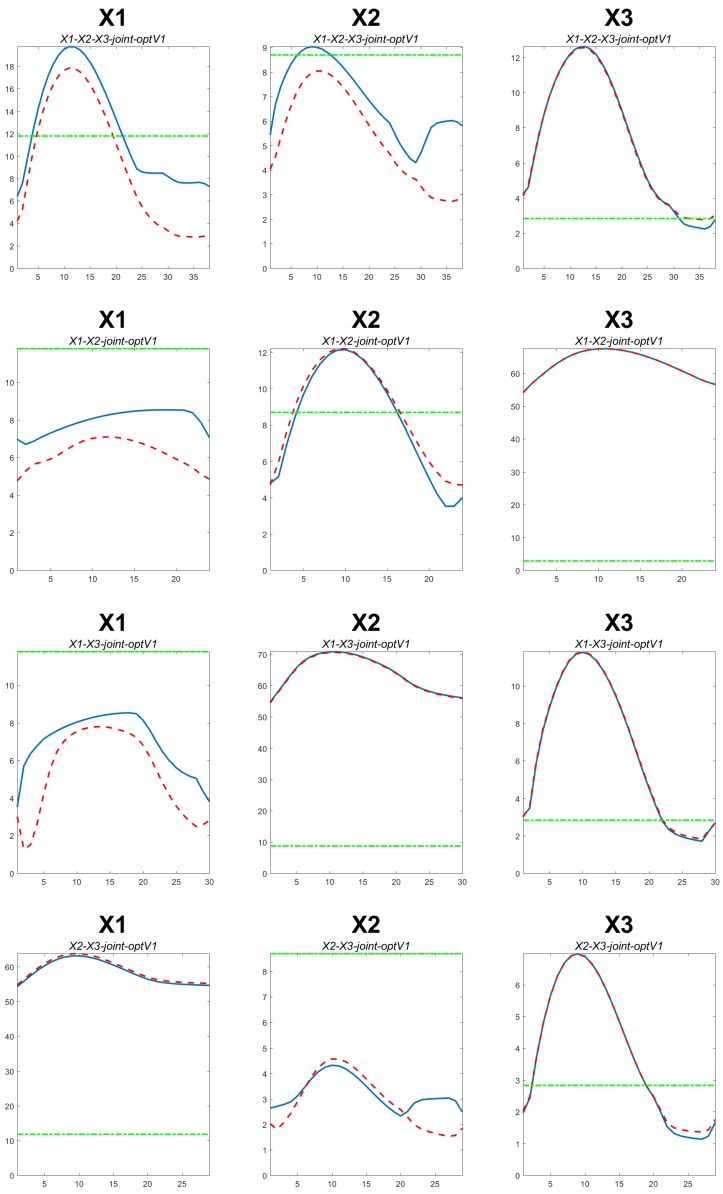

Figure 18: Iterative steps of the sequential optimizations locating joint structure between each possible combination of data blocks for the synthetic data from Figure 1. The horizontal axes represent iterations of the optimization problem and the vertical axes represent angles in degrees. Colored paths show progression of angles between the candidate direction and $\mathbf{TS}(\hat{\mathbf{A}}_k)$ (blue) and $\mathbf{TS}(\mathbf{A}_k)$ (red). Perturbation angle bounds $\hat{\phi}_k$ shown as green horizontal lines. From top to bottom: three-way joint, joint between blocks 1 and 2, joint between blocks 1 and 3, joint between blocks 2 and 3.