Angle-based joint and individual variation explained

Qing Feng, Meilei Jiang, Jan Hannig, J.S. Marron

PII: S0047-259X(17)30204-X
DOI: https://doi.org/10.1016/j.jmva.2018.03.008
Reference: YJMVA 4340

To appear in: *Journal of Multivariate Analysis*

Received date : 7 April 2017

Angle-based joint and individual variation explained

Qing Feng, Meilei Jiang[*], Jan Hannig, J. S. Marron

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract

Integrative analysis of disparate data blocks measured on a common set of experimental subjects is a major challenge in modern data analysis. This data structure naturally motivates the simultaneous exploration of the joint and individual variation within each data block resulting in new insights. For instance, there is a strong desire to integrate the multiple genomic data sets in The Cancer Genome Atlas to characterize the common and also the unique aspects of cancer genetics and cell biology for each source. In this paper we introduce Angle-Based Joint and Individual Variation Explained capturing both joint and individual variation within each data block. This is a major improvement over earlier approaches to this challenge in terms of a new conceptual understanding, much better adaption to data heterogeneity and a fast linear algebra computation. Important mathematical contributions are the use of score subspaces as the principal descriptors of variation structure and the use of perturbation theory as the guide for variation segmentation. This leads to an exploratory data analysis method which is insensitive to the heterogeneity among data blocks and does not require separate normalization. An application to cancer data reveals different behaviors of each type of signal in characterizing tumor subtypes. An application to a mortality data set reveals interesting historical lessons. Software and data are available at GitHub `https://github.com/MeileiJiang/AJIVE_Project`.

Keywords: Data integration, Heterogeneity, Perturbation theory, Principal angle, Singular value decomposition

1. Introduction

A major challenge in modern data analysis is data integration, combining diverse information from disparate data sets measured on a common set of experimental subjects. Simultaneous variation decomposition has been useful in many practical applications. For example, Kühnle [14], Lock and Dunson [19], and Mo et al. [24] performed integrative clustering on multiple sources to reveal novel and consistent cancer subtypes based on understanding of joint and individual variation. The Cancer Genome Atlas (TCGA) [25] provides a prototypical example for this problem. TCGA contains disparate data types generated from high-throughput technologies. Integration of these is fundamental for studying cancer on a molecular level. Other types of application include analysis of multi-source metabolomic data [15], extraction of commuting patterns in railway networks [10], recognition of brain-computer interface [49], etc.

A unified and insightful understanding of the set of data blocks is expected from simultaneously exploring the joint variation representing the inter-block associations and the individual variation specific to each block. Lock et al. [20] formulated this challenge into a matrix decomposition problem. Each data block is decomposed into three matrices modeling different types of variation, including a low-rank approximation of the joint variation across the blocks, low-rank approximations of the individual variation for each data block, and residual noise. Definitions and constraints were proposed for the joint and individual variation together with a method named JIVE; see `https://genome.unc.edu/jive/` and O'Connell and Lock [27] for Matlab and R implementations of JIVE, respectively.

JIVE was a promising framework for studying multiple data matrices. However, Lock et al. [20] algorithm and its implementation was iterative (thus slow) and performed rank selection based on a permutation test. It had no

[*] Corresponding author.

Email addresses: `qing.feng1014@gmail.com` (Qing Feng), `jiangm@live.unc.edu` (Meilei Jiang), `jan.hannig@unc.edu` (Jan Hannig), `marron@unc.edu` (J. S. Marron) guarantee of achieving a solution that satisfied the definitions of JIVE, especially in the case of some correlation between individual components. The example in Figure B.16 in Appendix B shows that this can be a serious issue. An important related algorithm named COBE was developed by Zhou et al. [50]. COBE considers a JIVE-type decomposition as a quadratic optimization problem with restrictions to ensure identifiability. While COBE removed many of the shortcomings of the original JIVE, it was still iterative and often required longer computation time than the Lock et al. [20] algorithm. Neither Zhou et al. [50] nor Lock et al. [20] provided any theoretical basis for selection of a thresholding parameter used for separation of the joint and individual components.

A novel solution, *Angle-based Joint and Individual Variation Explained (AJIVE)*, is proposed here for addressing this matrix decomposition problem. It provides an efficient *angle-based algorithm* ensuring an identifiable decomposition and also an insightful new interpretation of extracted variation structure. The key insight is the use of row spaces, i.e., a focus on scores, as the principal descriptor of the joint and individual variation, assuming columns are the n data objects, e.g., vectors of measurements on patients. This focuses the methodology on variation patterns across data objects, which gives straightforward definitions of the components and thus provides identifiability. These variation patterns are captured by the *score subspaces* of $\mathbb{R}^n$. Segmentation of joint and individual variation is based on studying the relationship between these score subspaces and using perturbation theory to quantify noise effects [36].

The main idea of AJIVE is illustrated in the flowchart of Figure 1. AJIVE works in three steps. First we find a low-rank approximation of each data block (shown as the far left color blocks in the flowchart) using SVD. This is depicted (using blocks with colored dashed line boundaries) on the left side of Figure 1 with the black arrows signifying thresholded SVD. Next, in the middle of the figure, SVD of the concatenated bases of row spaces from the first step (the gray blocks with colored boundaries) gives a joint row space (the gray box next to the circle), using a mathematically rigorous threshold derived using perturbation theory in Section 2.3. This SVD is a natural extension of Principal Angle Analysis, which is also closely related to the multi-block extension of Canonical Correlation Analysis [26] as well as to the flag means of the row spaces [5]; see Section 4.2 for details. Finally, the joint and individual space approximations are found using projection of the joint row space and its orthogonal complements on the data blocks as shown as colored boundary gray squares on the right with the three joint components at the top and the individual components at the bottom.

Using score subspaces to describe variation contained in a matrix not only empowers the interpretation of analysis but also improves understanding of the problem and the efficiency of the algorithm. An identifiable decomposition can now be obtained with all definitions and constraints satisfied even in situations when individual spaces are somewhat correlated. Moreover, the need to select a tuning parameter used to distinguish joint and individual variation is eliminated based on theoretical justification using perturbation theory. A consequence is an algorithm which uses a fast built-in singular value decomposition to replace lengthy iterative algorithms. For the example in Section 1.1, implemented in Matlab, the computational time of AJIVE (10.8 seconds) is about 11 times faster than the old JIVE (121 seconds) and 39 times faster than COBE (422 seconds). The computational advantages of AJIVE get even more pronounced on data sets with higher dimensionality and more complex heterogeneity such as the TCGA data analyzed in Section 3.1. For a very successful application of AJIVE on integrating fMRI imaging and behavioral data, see Yu et al. [48].

Other methods that aim to study joint variation patterns and/or individual variation patterns have also been developed. Westerhuis et al. [42] discuss two types of methods. One main type extends traditional Principal Component Analysis (PCA), including Consensus PCA and Hierarchical PCA first introduced by Wold et al. [45, 46]. An overview of extended PCA methods is discussed in Smilde et al. [35]. Abdi et al. [1] discuss a multiple block extension of PCA called multiple factor analysis. This type of method computes the block scores, block loadings, global loadings and global scores.

The other main type of method is extensions of Partial Least Squares (PLS) [44] or Canonical Correlation Analysis (CCA) [9] that seek associated patterns between the two data blocks by maximizing covariance/correlation. For example, Wold et al. [46] introduced multi-block PLS and hierarchical PLS (HPLS) and Trygg and Wold [37] proposed *O2-PLS* to better reconstruct joint signals by removing structured individual variation. A multi-block extension can be found in Löfstedt et al. [21].

Yang and Michailidis [47] provide a very nice integrative joint and individual component analysis based on non-negative matrix factorization. Ray et al. [31] do integrative analysis using factorial models in the Bayesian setting. Schouteden et al. [33, 34] propose a method called DISCO-SCA that is a low-rank approximation with rotation to sparsity of the concatenated data matrices.

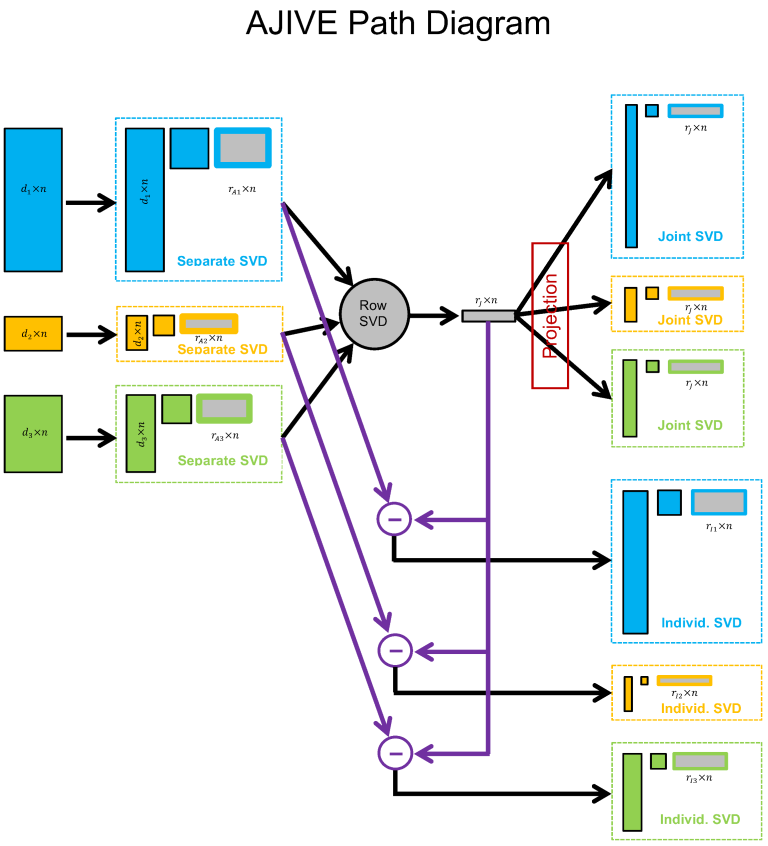

Figure 1: Flow chart demonstrating the main steps of AJIVE. First low-rank approximation of each data block is obtained on the right. Then in the middle joint structure between the low-rank approximations is extracted using SVD of the stacked row basis matrices. Finally, on the right, the joint components (upper) are obtained by projection of each data block onto the joint basis (middle) and the individual components (lower) come from orthonormal basis subtraction.

A connection between extended PCA and extended PLS methods is discussed in Hanafi et al. [6]. Both types of methods provide an integrative analysis by taking the inter-block associations into account. These papers recommend use of normalization to address potential scale heterogeneity, including normalizing by the Frobenius norm, or the largest singular value of each data block, etc. However, there are no consistent criteria for normalization and some of these methods have convergence problems. An important point is that none of these approaches provide simultaneous decomposition highlighting joint and individual modes of variation with the goal of contrasting these to reveal new insights.

1.1. Toy example

We give a toy example to provide a clear view of multiple challenges brought by potentially very disparate data blocks. This toy example has two data blocks, X (100×100) and Y ($10{,}000 \times 100$), with patterns corresponding to joint and individual structures. Such data set sizes are reasonable in modern genetic studies, as seen in Section 3.1. Figure 2 shows colormap views of these matrices, with the value of each matrix entry colored according to the color bar at the bottom of each subplot. The data have been simulated so expected row means are 0. Therefore mean centering is not necessary in this case. A careful look at the color bar scalings shows the values are almost four orders of magnitude larger for the top matrices. Each column of these matrices is regarded as a common data object and each row is considered as one feature. The number of features is also very different as labeled in the y-axis. Each of the two raw data matrices, X and Y in the left panel of Figure 2, is the sum of joint, individual and noise components shown in the other panels.

The joint variation for both blocks, second column of panels, presents a contrast between the left and right halves of the data matrix, thus having the same rank-1 score subspace. If for example the left half columns were male and right half were female, this joint variation component could be interpreted as a contrast of gender groups which exists in both data blocks for those features where color appears.

The X individual variation, third column of panels, partitions the columns into two other groups of size 50 that are arranged so the row space is orthogonal to that of the joint score subspace. The individual signal for Y contains two variation components, each driven by half of the features. The first component, displayed in the first 5000 rows, partitions the columns into three groups. The other component is driven by the bottom half of the features and partitions the columns into two groups, both with row spaces orthogonal to the joint. Note that these two individual score subspaces for X and Y are different but not orthogonal. The smallest angle between the individual subspaces is 45°.

This example presents several challenging aspects, which also appear in real data sets such as TCGA, as studied in Section 3.1. One is that both the values and the number of the features are orders of magnitude different between X and Y. Another important challenge is that because the individual spaces are not orthogonal, the individual signals are correlated. Correctly handling these in an integrated manner is a major improvement of AJIVE over earlier methods. In particular, normalization is no longer an issue because AJIVE only uses the low-rank initial *scores* (represented as the gray boxes in the SVD shown on the left of Figure 1), while signal power appears in the central subblocks and the features only in the left subblocks. Appropriate handing of potential correlation among individual components is done using perturbation theory in Section 2.

The noise matrices, the right panels of Figure 2, are standard Gaussian random matrices (scaled by 5000 for X) which generates a noisy context for both data blocks and thus a challenge for analysis, as shown in the left panels of Figure 2.

Simply concatenating X and Y on columns and performing a singular value decomposition on this concatenated matrix completely fails to give a meaningful joint analysis. PLS and CCA might be used to address the magnitude difference in this example and capture the signal components. However, they target common relationships between two data matrices and therefore are unable to simultaneously extract and distinguish the two types of variation. Moreover, because of its sensitivity to the strength of the signal PLS misclassifies correlated individual components as joint components. The original JIVE of Lock et al. [20] also fails on this toy example. Details on all of these can be found in Appendix B.

In this toy example, the selection of the initial low-rank parameters $r_X = 2$ and $r_Y = 3$ is unambiguous. The left panel of Figure 3 shows this AJIVE-approximation well captures the signal variations within both X and Y. What's more, our method correctly distinguishes the types of variation showing its robustness against heterogeneity across

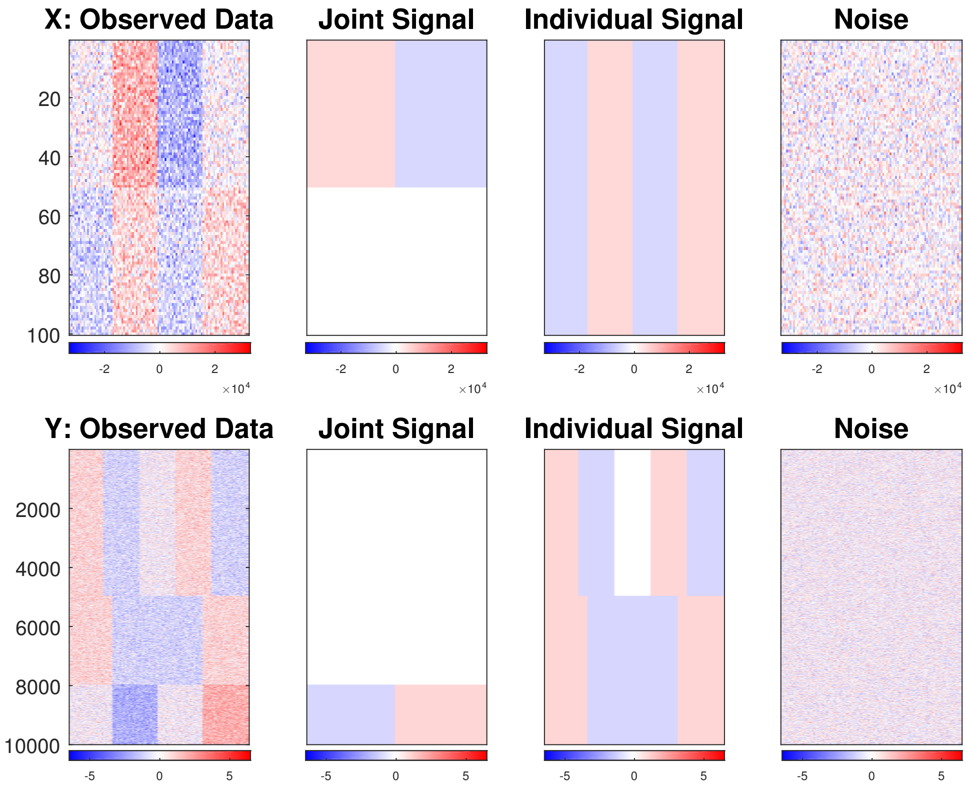

Figure 2: Data blocks X (top) and Y (bottom) in the toy example. The left panels present the observed data matrices which are a sum of the signal and noise matrices depicted in the remaining panels. Scale is indicated by color bars at the bottom of each sub-plot. These structures are challenging to capture using conventional methods due to very different orders of magnitude and numbers of features.

data blocks and correlation between individual data blocks. The approximations of both joint and individual signal are depicted in the remaining panels. A careful study of the impact of initial rank misspecification on the AJIVE results for this toy example is in Sections 2.2 and 2.3.

The rest of this paper is organized as follows. Section 2 describes the population model and mathematical details of the estimation approach. Results of application to a TCGA breast cancer data set and a mortality data set are presented in Section 3. Relationships between the proposed AJIVE and other methods from an optimization point of view are discussed in Section 4. The AJIVE Matlab software, the related Matlab scripts and associated datasets, which can be used to reproduce all the results in this paper, are available at the GitHub repository `https://github.com/MeileiJiang/AJIVE_Project`.

2. Proposed method

In this section the details of the new proposed AJIVE are presented. The population model is proposed in Sections 2.1. The estimation approaches are given in Section 2.2, 2.3, and 2.4.

2.1. Population model

Matrices $X_1, \dots, X_K$ each of size $d_k \times n$ are a set of data blocks for study, e.g., the colored blocks on the left of Figure 1. The columns are regarded as data objects, one vector of measurements for each experimental subject, while rows are considered as features. All X_ks therefore have the same number of columns and perhaps a different number of rows.

Each X_k is modeled as low-rank true underlying signals A_k perturbed by additive noise matrices E_k. Each low-rank signal A_k is the sum of two matrices containing joint and individual variation, denoted as J_k and I_k respectively for each block, viz.

$$\begin{bmatrix} X_1 \\ \vdots \\ X_K \end{bmatrix} = \begin{bmatrix} A_1 \\ \vdots \\ A_K \end{bmatrix} + \begin{bmatrix} E_1 \\ \vdots \\ E_K \end{bmatrix} = \begin{bmatrix} J_1 \\ \vdots \\ J_K \end{bmatrix} + \begin{bmatrix} I_1 \\ \vdots \\ I_K \end{bmatrix} + \begin{bmatrix} E_1 \\ \vdots \\ E_K \end{bmatrix}.$$

Our approach focuses on the vectors in the row space of our matrices. In this context these vectors are often called *score vectors* and the row space of the matrix is often called *score subspace* ($\subset \mathbb{R}^n$). Therefore, the row spaces of the matrices capturing joint variation, i.e., joint matrices, are defined as sharing a common score subspace denoted as row(J), viz.

$$\text{row}(J_1) = \cdots = \text{row}(J_K) = \text{row}(J).$$

The individual matrices are individual in the sense that they are orthogonal to the joint space, i.e., $\text{row}(I_k) \perp \text{row}(J)$ for all $k \in \{1, \dots, K\}$, and the intersection of their score subspaces is the zero vector space, i.e.,

$$\bigcap_{k=1}^{K} \text{row}(I_k) = \{\vec{0}\}.$$

This means that there is no non-trivial common row pattern in every individual score subspaces across blocks.

To ensure an identifiable variation decomposition we assume $\text{row}(J) \subset \text{row}(A_k)$, which also implies $\text{row}(I_k) \subset \text{row}(A_k)$, for all $k \in \{1, \dots, K\}$. Note that orthogonality between individual matrices $\{I_1, \dots, I_K\}$ is *not* assumed as it is not required for the model to be uniquely determined.

Under these assumptions, the model is identifiable in the following sense.

Lemma 1. *Given a set* $\{A_1, \dots, A_K\}$ *of matrices, there are unique sets* $\{J_1, \dots, J_K\}$ *and* $\{I_1, \dots, I_K\}$ *of matrices so that*

(i) $A_k = J_k + I_k$*, for all* $k \in \{1, \dots, K\}$*;*

(ii) $\text{row}(J_k) = \text{row}(J) \subset \text{row}(A_k)$*, for all* $k \in \{1, \dots, K\}$*;*

(iii) $\text{row}(J) \perp \text{row}(I_k)$*, for all* $k \in \{1, \dots, K\}$*;*

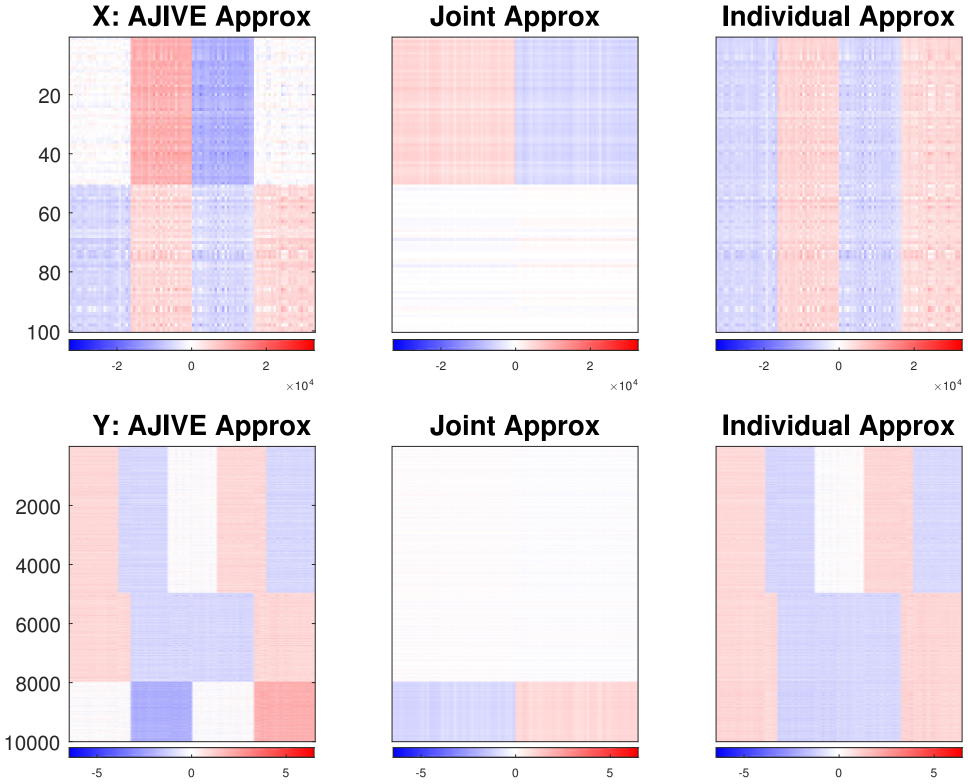

Figure 3: AJIVE approximation of the data blocks *X* and *Y* in the toy example are shown in the first column, with the joint and individual signal matrices depicted in the remaining columns. Both quite diverse types of variations are well captured for each data block by AJIVE, in contrast to other usual methods as seen in Appendix B.

(iv) $\bigcap_{k=1}^{K} \text{row}(I_k) = \{\vec{0}\}$.

The proof is provided in Appendix A. Lemma 1 is very similar to Theorem 1.1 in Lock et al. [20]. The main difference is that the rank conditions are replaced by conditions on row spaces. In our view, this provides a clearer mathematical framework and more precise understanding of the different types of variation.

The additive noise matrices, $E_1, \ldots, E_K$, are assumed to follow an isotropic error model where the energy of projection is invariant to direction in both row and column spaces. Important examples include the multivariate standard normal distribution and the matrix multivariate Student t distribution [13]. All singular values of each noise matrix are assumed to be smaller than the smallest singular values of each signal to give identifiability. This assumption on the noise distribution here is weaker than the classical iid Gaussian random matrix, and only comes into play when determining the number of joint components.

The estimation algorithm, which segments the data into joint and individual components in the presence of noise, has three main steps, as follows.

Step 1: Signal Space Initial Extraction. Low-rank approximation of each data block, as shown on the left in Figure 1. A novel approach together with careful assessment of accuracy using matrix perturbation theory from linear algebra [36], is provided in Sections 2.2 and 2.3.

Step 2: Score Space Segmentation. Initial determination of joint and individual components, as shown in the center of Figure 1. Our approach to this is based on an extension of Principal Angle Analysis, and an inferential based graphical diagnostic tool. The two block case is discussed in Section 2.3.1, with the multi-block case appearing in Section 2.3.2.

Step 3: Final Decomposition and Outputs. Check segmented components still meet initial thresholds in Step 1, and reproject for appropriate outputs, as shown in the right of Figure 1. Details of this are in Section 2.4.

2.2. *Step 1: Signal space initial extraction*

Even though the signal components $A_1, \ldots, A_K$ are low-rank, the data matrices $X_1, \ldots, X_K$ are usually of full rank due to the presence of noise. SVD works as a signal extraction device in this step, keeping components with singular values greater than selected thresholds individually for each data block, as discussed in Section 2.2.1. The accuracy of this SVD approximation will be carefully estimated in Section 2.2.2, and will play an essential role in segmenting the joint space in Step 2.

2.2.1. *Initial low-rank approximation*

Each signal block A_k is estimated using SVD of X_k. Given a threshold t_k, the estimator $\tilde{A}_k$ (represented in Figure 1 as the boxes with dashed colored boundaries on the left) is defined by setting all singular values below t_k to 0. The resulting rank $\tilde{r}_k$ of $\tilde{A}_k$ is an initial estimator of the signal rank r_k. The reduced-rank decompositions of the $\tilde{A}_k$s are

$$\tilde{A}_k = \tilde{U}_k \tilde{\Sigma}_k \tilde{V}_k^\top, \tag{1}$$

where $\tilde{U}_k$ contains the left singular vectors that correspond to the largest $\tilde{r}_k$ singular values respectively for each data block. The initial estimate of the signal score space, denoted as $\text{row}(\tilde{A}_k)$, is spanned by the right singular vectors in $\tilde{V}_k$ (shown as gray boxes with colored boundaries on the left of Figure 1).

When selecting these thresholds, one needs to be aware of a bias/variance like trade-off. Setting the threshold too high will provide an accurate estimation of the parts of the joint space that are included in the low-rank approximation. The downside is that significant portions of the joint signal might be thresholded out. This could be viewed as a low-variance high-bias situation. If the threshold is set low, then it is likely that the joint signal is included in all of the blocks. However, the precision of the segmentation in the next step can deteriorate to the point that individual components, or even worse, noise components, can be selected in the joint space. This can be viewed as the low-bias high-variance situation.

Most off-the-shelf automatic procedures for low-rank matrix approximation have as their stated goal signal reconstruction and prediction, which based on our experience tends toward thresholds that are too small, i.e., input ranks

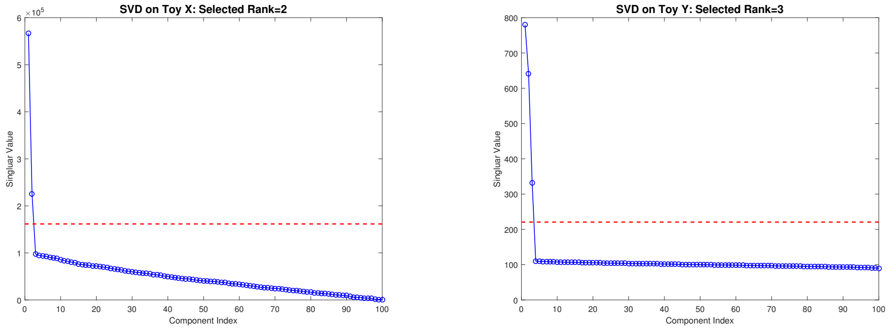

Figure 4: Scree plots for the toy data sets X (left) and Y (right). Both plots display the singular values associated with a component in descending order versus the index of the component. The components with singular values above the dashed red threshold line are regarded as the initial signal components in the first step of AJIVE.

that are too large. This is sensible as adding a little bit more noise usually helps prediction but it has bad effects on signal segmentation. We therefore recommend taking a multi-scale perspective and trying several threshold choices, e.g., by considering several relatively big jumps in a scree plot. A useful inferential graphical device to assist with this choice is developed in Section 2.3.

Figure 4 shows the scree plots of each data block for the toy example in Section 1. The left scree plot for X clearly indicates a selection of rank $\tilde{r}_1 = 2$ and the right scree plot for Y points to rank $\tilde{r}_2 = 3$; in both cases those components stand out while the rest of the singular values decay slowly showing no clear jump.

2.2.2. Approximation accuracy estimation

A major challenge is segmentation of the joint and individual variation in the presence of noise which individually perturbs each signal. A first step towards addressing this is a careful study of how well A_k is approximated by $\tilde{A}_k$ using the *Generalized* $\sin\theta$ *Theorem* [40].

Pseudometric between subspaces. To apply the Generalized $\sin\theta$ Theorem, we use the following pseudometric as a notion of distance between theoretical and perturbed subspaces. Recall that $\text{row}(A_k)$, $\text{row}(\tilde{A}_k)$ are respectively the r_k- and $\tilde{r}_k$-dimensional score subspaces of $\mathbb{R}^n$ for the matrix A_k and its approximation $\tilde{A}_k$. The corresponding projection matrices are P_{A_k} and $P_{\tilde{A}_k}$, respectively. A pseudometric between the two subspaces can be defined as the difference of the projection matrices under the operator L^2 norm, i.e., $\rho\{\text{row}(A_k), \text{row}(\tilde{A}_k)\} = \|P_{A_k} - P_{\tilde{A}_k}\|$ [36]. When $r_k = \tilde{r}_k$, this pseudometric is also a distance between the two subspaces.

An insightful understanding of this pseudometric $\rho\{\text{row}(A_k), \text{row}(\tilde{A}_k)\}$ comes from a principal angle analysis [9, 11] of the subspaces $\text{row}(A_k)$ and $\text{row}(\tilde{A}_k)$. Denote the principal angles between $\text{row}(A_k)$ and $\text{row}(\tilde{A}_k)$ as

$$\Theta\{\text{row}(A_k), \text{row}(\tilde{A}_k)\} = \{\theta_{k,1}, \ldots, \theta_{k,r_k \wedge \tilde{r}_k}\} \tag{2}$$

with $\theta_{k,1} \leq \cdots \leq \theta_{k,r_k \wedge \tilde{r}_k}$. The pseudometric $\rho\{\mathrm{row}(A_k), \mathrm{row}(\tilde{A}_k)\}$ is equal to the sine of the maximal principal angle, i.e., $\sin\theta_{k,r_k \wedge \tilde{r}_k}$. Thus the largest principal angle between two subspaces measures their closeness, i.e., distance.

The pseudometric $\rho\{\mathrm{row}(A_k), \mathrm{row}(\tilde{A}_k)\}$ can be also written as

$$\rho\{\mathrm{row}(A_k), \mathrm{row}(\tilde{A}_k)\} = \|(I - P_{A_k})P_{\tilde{A}_k}\| = \|(I - P_{\tilde{A}_k})P_{A_k}\|,$$

which gives another useful understanding of this definition. It measures the relative deviation of the signal variation from the theoretical subspace. Accordingly, the similarity/closeness between the subspaces and its perturbation can be written as $\|P_{A_k}P_{\tilde{A}_k}\|$ and is equal to the cosine of the maximal principal angle defined above, i.e., $\cos\theta_{k,r_k \wedge \tilde{r}_k}$. Hence, $\sin^2\theta_{k,r_k \wedge \tilde{r}_k}$ indicates the proportion of signal deviation and $\cos^2\theta_{k,r_k \wedge \tilde{r}_k}$ tells the proportion of remaining signal in the theoretical subspace.

Wedin bound. For a signal matrix A_k and its perturbation $X_k = A_k + E_k$, the generalized $\sin\theta$ theorem provides a bound for the distance between the rank $\tilde{r}_k$ $(\leq r_k)$ singular subspaces of A_k and X_k. This bound quantifies how the theoretical singular subspaces are affected by noise.

Theorem 1 (Wedin, 1972)**.** *Let A_k be a signal matrix with rank r_k. Letting $A_{k,1} = U_{k,1}\Sigma_{k,1}V_{k,1}^\top$ denote the rank $\tilde{r}_k$ SVD of A_k, where $\tilde{r}_k \leq r_k$, write $A_k = A_{k,1} + A_{k,0}$. For the perturbation $X_k = A_k + E_k$, a corresponding decomposition can be made as $X_k = \tilde{A}_{k,1} + \tilde{E}_k$, where $\tilde{A}_{k,1} = \tilde{U}_{k,1}\tilde{\Sigma}_{k,1}\tilde{V}_{k,1}^\top$ is the rank $\tilde{r}_k$ SVD of X_k. Assume that there exists an $\alpha \geq 0$ and a $\delta > 0$ such that for $\sigma_{\min}(\tilde{A}_{k,1})$ and $\sigma_{\max}(A_{k,0})$ denoting appropriate minimum and maximum singular values*

$$\sigma_{\min}(\tilde{A}_{k,1}) \geq \alpha + \delta \quad \textit{and} \quad \sigma_{\max}(A_{k,0}) \leq \alpha.$$

Then the distance between the row spaces of $\tilde{A}_{k,1}$ and $A_{k,1}$ is bounded by

$$\rho\{\mathrm{row}(\tilde{A}_{k,1}), \mathrm{row}(A_{k,1})\} \leq \frac{\max\left(\|E_k\tilde{V}_{k,1}\|, \|E_k^\top\tilde{U}_{k,1}\|\right)}{\delta} \wedge 1.$$

In practice we do not observe $A_{k,0}$ thus δ cannot be estimated in general. A special case of interest for AJIVE is $\tilde{r}_k = r_k$, in which case $A_{k,0} = 0, A_k = A_{k,1}$. The following is an adaptation of the generalized $\sin\theta$ theorem to this case.

Corollary 1 (Bound for correctly specified rank)**.** *For each $k \in \{1, \ldots, K\}$, the signal matrix A_k is perturbed by additive noise E_k. Let $\theta_{k,\tilde{r}_k}$ be the largest principal angle for the subspace of signal A_k and its approximation $\tilde{A}_k$, where $\tilde{r}_k = r_k$. Denote the SVD of $\tilde{A}_k$ as $\tilde{U}_k\tilde{\Sigma}_k\tilde{V}_k^\top$. The distance between the subspaces of A_k and $\tilde{A}_k$, $\rho\{\mathrm{row}(A_k), \mathrm{row}(\tilde{A}_k)\}$, i.e., sine of $\theta_{k,\tilde{r}_k}$, is bounded above by*

$$\rho\{\mathrm{row}(A_k), \mathrm{row}(\tilde{A}_k)\} = \sin\theta_{k,\tilde{r}_k} \leq \frac{\max(\|E_k\tilde{V}_k\|, \|E_k^\top\tilde{U}_k\|)}{\sigma_{\min}(\tilde{A}_k)} \wedge 1. \tag{3}$$

In this case the bound is driven by the maximal value of noise energy in the column and row spaces and by the estimated smallest signal singular value. This is consistent with the intuition that a deviation distance, i.e., a largest principal angle, is small when the signal is strong and perturbations are weak.

In general, it can be very challenging to correctly estimate the true rank of A_k. If the true rank r_k is not correctly specified, then different applications of the Wedin bound are useful. In particular, when $A_{k,0}$ is not 0, i.e., $\tilde{r}_k < r_k$, insights come from replacing E_k by $E_k + A_{k,0}$ in the Wedin bound.

Corollary 2 (Bound for under-specified rank)**.** *For each $k \in \{1, \ldots, K\}$, the signal matrix A_k with rank r_k is perturbed by additive noise E_k. Let $\tilde{A}_k = \tilde{U}_k\tilde{\Sigma}_k\tilde{V}_k^\top$ be the rank $\tilde{r}_k$ SVD approximation of A_k from the perturbed matrix, where $\tilde{r}_k < r_k$. Denote $A_k = A_{k,1} + A_{k,0}$, where $A_{k,1}$ is the rank $\tilde{r}_k$ SVD of A. Then the distance between* $\mathrm{row}(A_{k,1})$ *and* $\mathrm{row}(\tilde{A}_k)$ *is bounded above by*

$$\rho\{\mathrm{row}(A_{k,1}), \mathrm{row}(\tilde{A}_k)\} \leq \frac{\max\left(\|(E_k + A_{k,0})\tilde{V}_k\|, \|(E_k + A_{k,0})^\top\tilde{U}_k\|\right)}{\sigma_{\min}(\tilde{A}_k)} \wedge 1.$$

For the other type of initial rank misspecification, $\tilde{r}_k > r_k$, we augment A_k with appropriate noise components to be able to use the Wedin bound.

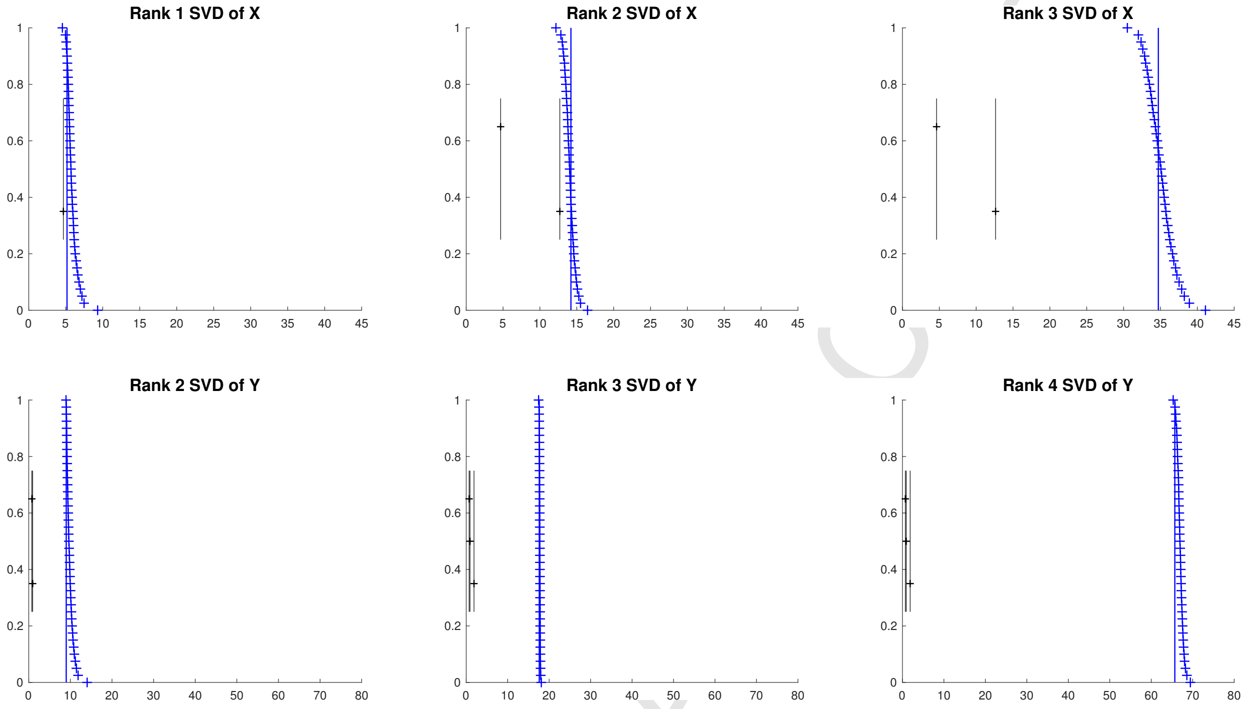

Figure 5: Principal angle plots between each singular subspace of the signal matrix $A_{k,1}$ and its estimator $\tilde{A}_k$ for the toy dataset. Graphics for X are on the upper row, with Y on the lower row. The left, middle and right columns are the under-specified, correctly specified and over-specified signal matrix rank cases respectively. Each x-axis represents the angle. The y-axis shows the values of the survival function of the resampled distribution, which are shown as blue plus signs in the figure. The vertical blue solid line is the theoretical Wedin bound, showing this bound is well estimated. The vertical black solid line segments represent the principal angles $\theta_{k,1}, \ldots, \theta_{k,r_k \wedge \tilde{r}_k}$ between $\mathrm{row}(A_{k,1})$ and $\mathrm{row}(\tilde{A}_k)$. The distance between the black and blue lines reveals when the Wedin bound is tight.

Corollary 3 (Bound for over-specified rank)**.** *For each $k \in \{1, \ldots, K\}$, the signal matrix $A_k = U_k \Sigma_k V_k^\top$ with rank r_k is perturbed by additive noise E_k. Let $\tilde{A}_k = \tilde{U}_k \tilde{\Sigma}_k \tilde{V}_k^\top$ be the rank $\tilde{r}_k$ SVD of X_k, where $\tilde{r}_k > r_k$. Let E_0 be the rank $\tilde{r}_k - r_k$ SVD of $(I - U_k U_k^\top) E_k (I - V_k V_k^\top)$. Then the pseudometric between* $\mathrm{row}(A_k)$ *and* $\mathrm{row}(\tilde{A}_k)$ *is bounded above by*

$$\rho\{\mathrm{row}(A_k), \mathrm{row}(\tilde{A}_k)\} \leq \frac{\max\left(\|(E_k - E_0)\tilde{V}_k\|, \|(E_k - E_0)^\top \tilde{U}_k\|\right)}{\sigma_{\min}(\tilde{A}_k)} \wedge 1.$$

The bounds in Corollaries 1–3 provide many useful insights. However, these bounds still cannot be used directly since we do not observe the error matrices $E_1, \ldots, E_K$. A re-sampling based estimator of the Wedin bounds is provided in the next paragraph. As seen in Figure 5, this estimator appropriately adapts to each of the above three cases. Moreover, Figure 5 also indicate that the Wedin bound for over-specified rank is usually very conservative.

Estimation and evaluation of the Wedin bound. As mentioned above, the perturbation bounds of each $\theta_{k, r_k \wedge \tilde{r}_k}$ require the estimation of terms $\|E_k \tilde{V}_k\|$, $\|E_k^\top \tilde{U}_k\|$ for $k \in \{1, 2\}$. These terms are measurements of energies of the noise matrices projected onto the signal column and row spaces. Since an isotropic error model is assumed, the *distributions* of energy of the noise matrices in arbitrary fixed directions are equal. Thus, if we sample random subspaces of dimension $\tilde{r}_k$, that are orthogonal to the estimated signal $\tilde{A}_k$, and use the observed residual $\tilde{E}_k = X_k - \tilde{A}_k$, this should provide a good estimator of the distribution of the unobserved terms $\|E_k \tilde{V}_k\|$, $\|E_k^\top \tilde{U}_k\|$.

In particular, consider the estimation of the term $\|E_k \tilde{V}_k\|$. We draw a random subspace of dimension $\tilde{r}_k$ that is orthogonal to $\tilde{V}_k$, denoted as $V_k^\star$. The observed data block X_k is projected onto the subspace spanned by $V_k^\star$, written as $X_k V_k^\star$. The distribution (with respect to the $V_k^\star$ variation) of the operator L^2 norm $\|X_k V_k^\star\| = \|\tilde{E}_k V_k^\star\|$ approximates the

Table 1: Coverages of the prediction intervals of the true angle between the signal row($A_{k,1}$) and its estimator row($\tilde{A}_k$) for the matrix X in the toy example. Rows are nominal levels. Columns are ranks of approximation (where 2 is the correct rank). The simulation based on 10000 realizations of X shows good performance for this square matrix.

	1	**2**	3
50%	91.9%	**63.6%**	100.0%
90%	100.0%	**89.6%**	100.0%
95%	100.0%	**93.7%**	100.0%
99%	100.0%	**98.0%**	100.0%

distribution of the unknown $\|E_k \tilde{V}_k\|$ because both measure noise energy in essentially random directions. Similarly the estimation of $\|E_k^\top \tilde{U}_k\|$ is approximated by $\|X_k^\top U_k^\star\|$, where $U_k^\star$ is a random $\tilde{r}_k$-dimensional subspace orthogonal to $\tilde{U}_k$. These distributions are used to estimate the Wedin bound by generating 1000 replications of $\|X_k V_k^\star\|$ and $\|X_k^\top U_k^\star\|$, and plugging these into (3). The quantiles of the resulting distributions are used as prediction intervals for the unknown theoretical Wedin bound. Note this random subspace sampling scheme provides a distribution with smaller variance than simply sampling from the remaining singular values of X_k, i.e., using 1000 subspaces each generated by a random sample of $\tilde{r}_k$ remaining singular vectors.

There are two criteria for evaluating the effectiveness of the estimator. First is how well the resampled distributions approximate the underlying theoretical Wedin bounds. This is addressed in Figure 5, which is based on the toy example in Section 1.1. For each of the matrices X and Y (top and bottom rows), the under, correctly, and over specified signal rank cases (Corollaries 2, 1 and 3, respectively) are carefully investigated. In each case the theoretical Wedin bound (calculated using the true underlying quantities, that are only known in a simulation study) are shown as vertical blue lines. Our resampling approach provides an estimated distribution, the survival function of which is shown using blue plus signs. This indicates remarkably effective estimation of the Wedin bound in all cases.

The second more important criterion is how well the prediction interval covers the actual principal angles between row(A_k) and row($\tilde{A}_k$). These angles are shown as vertical black line segments in Figure 5. For the square matrix X, in the under and correctly specified case (top, left, and center), the Wedin bound seems relatively tight. In all other cases, the Wedin bound is conservative.

Figure 5 shows one realization of the noise in the toy example. A corresponding simulation study is summarized in Table 1. For this we generated 10,000 independent copies of the data sets X (100×100, true signal rank $r_1 = 2$) and Y ($10{,}000 \times 100$, true signal rank $r_2 = 3$). Then for several low-rank approximations (columns of Table 1) we calculated the estimate of the angle between the true signal and the low-rank approximation. Table 1 reports the percentage of the times the corresponding quantile of the resampled estimate is bigger than the true angle for the matrix X. When the rank is correctly specified, i.e., $\tilde{r}_1 = r_1 = 2$, we see that the performance for the square matrix X is satisfactory as the empirical percentages are close to the nominal values. When the rank is misspecified, the empirical upper bound is conservative. Corresponding empirical percentages for the high dimension, low sample size data set Y are all 100%, and thus are not shown. This is caused by the fact that Wedin bound can be very conservative if the matrix is far from square. As seen in Figure 6 this can cause identification of spurious joint components. This motivates our development of a diagnostic plot in Section 2.3. Recent works of Cai and Zhang [3] and O'Rourke et al. [28] may provide potential approaches for improvement of the Wedin bound.

2.3. Step 2: Score space segmentation

2.3.1. Two-block case

For a clear introduction to the basic idea of score space segmentation into joint and individual components, the two-block special case ($K = 2$) is first studied. The goal is to use the low-rank approximations $\tilde{A}_k$ from Eq. (1) to obtain estimates of the common joint and individual score subspaces. Due to the presence of noise, the components of row($\tilde{A}_1$) and row($\tilde{A}_2$) corresponding to the underlying joint space, no longer are the same, but should have a relatively small angle. Similarly, the components corresponding to the underlying individual spaces are expected to have a relatively large angle. This motivates the use of principal angle analysis to separate the joint from the individual components.

Principal angle analysis. One of the ways of computing the principal angles between $\text{row}(\tilde{A}_1)$ and $\text{row}(\tilde{A}_2)$ is to perform SVD on a concatenation of their right singular vector matrices [23], i.e.,

$$M \triangleq \begin{bmatrix} \tilde{V}_1^\top \\ \tilde{V}_2^\top \end{bmatrix} = U_M \Sigma_M V_M^\top, \tag{4}$$

where the singular values, $\sigma_{M,i}$, on the diagonal of Σ_M, determine the principal angles, $\Phi\{\text{row}(\tilde{A}_1), \text{row}(\tilde{A}_2)\} = \{\phi_1, \ldots, \phi_{\tilde{r}_1 \wedge \tilde{r}_2}\}$, where, for each $i \in \{1, \ldots, \tilde{r}_1 \wedge \tilde{r}_2\}$,

$$\phi_i = \arccos\{(\sigma_{M,i})^2 - 1\}. \tag{5}$$

This SVD decomposition can be understood as a tool that finds pairs of directions in the two subspaces $\text{row}(\tilde{A}_1)$ and $\text{row}(\tilde{A}_2)$ of minimum angle, sorted in increasing order. These angles are shown as vertical black line segments in our main diagnostic graphic introduced in Figure 6. The first $\tilde{r}_J$ column vectors in V_M will form the orthonormal basis of the estimated joint space, $\text{row}(J) \subseteq \mathbb{R}^n$. A deeper investigation of the relationship between V_M and the canonical correlation vectors in U_M appears in Section 4.2. Next we determine which angles are small enough to be labeled as joint components, i.e., the selection of $\tilde{r}_J$.

Random direction bound. In order to investigate which principal angles correspond to random directions, we need to estimate the distribution of principal angles generated by random subspaces. This distribution only depends on the initial input ranks of each data block, $\tilde{r}_k$, and the dimension of the row spaces, n. We obtain this distribution by simulation. In particular, $\tilde{V}_1$ and $\tilde{V}_2$ are replaced in (5) by random subspaces, i.e., each is right multiplied by an independent random orthonormal matrix. The distribution of the smallest principal angle, corresponding to the largest singular value, indicates angles potentially driven by pure noise. We recommend the 5th percentile of the angle distribution as cutoff in practice. Principal angles larger than this are not included in the joint component, which provide 95% confidence that the selected joint space does not have pure noise components. This cutoff is prominently shown in Figure 6 as the vertical dot-dashed red line. The cumulative distribution function of the underlying simulated distribution is shown as red circles.

When the individual spaces are not orthogonal, a sharper threshold based on the Wedin bounds is available.

Threshold based on the Wedin bound. The following lemma provides a bound on the largest allowable principal angle of the joint part of the initial estimated spaces.

Lemma 2. *Let ϕ be the largest principal angle between two subspaces that are each a perturbation of the common row space within $\text{row}(\tilde{A}_1)$ and $\text{row}(\tilde{A}_2)$. That angle is bounded by $\sin\phi \le \sin(\theta_{1,\tilde{r}_1 \wedge r_1} + \theta_{2,\tilde{r}_2 \wedge r_2})$ in which $\theta_{1,\tilde{r}_1 \wedge r_1}$ and $\theta_{2,\tilde{r}_2 \wedge r_2}$ are the angles given in Eq.* (2).

The proof is provided in Appendix A. As with the theoretical Wedin bound, the unknown $\theta_{1,\tilde{r}_1 \wedge r_1}$ and $\theta_{2,\tilde{r}_2 \wedge r_2}$ are replaced by distribution estimators of the Wedin bounds. The survival function of the distribution estimator of this upper bound on ϕ is shown in Figure 6 using blue plus signs. The vertical dashed blue line is the 95th percentile of this distribution, giving 95% confidence that angles larger do not correspond to joint components of the lower rank approximations in Step 1. The joint rank $\tilde{r}_J$ is selected to be the number of principal angles, ϕ_i in (5), that are smaller than both the 5th percentile of the random direction distribution and the 95th percentile of the resampled Wedin bound distribution.

Figure 6 illustrates how this diagnostic graphic provides many insights that are useful for initial rank selection. This considers several candidates of initial ranks. Recall for Section 1.1, this toy example has one joint component, one individual X component, and two individual Y components. The row subspaces of their individual components are not orthogonal and the true principal angle (only known in simulation study) is 45°. Furthermore, PCA of Y reveals that 79.6% of the joint component appears in the third principal component.

The upper left panel of Figure 6 shows the under specified rank case of $\tilde{r}_1 = \tilde{r}_2 = 2$. The principal angles (black lines) are larger than the Wedin bound (blue dashed line), so we conclude neither is joint variation. This is sensible since the true joint signal is mostly contained in the 3rd Y component. However, both are smaller than the random direction bound (red dashed line), so we conclude each indicates presence of correlated individual spaces.

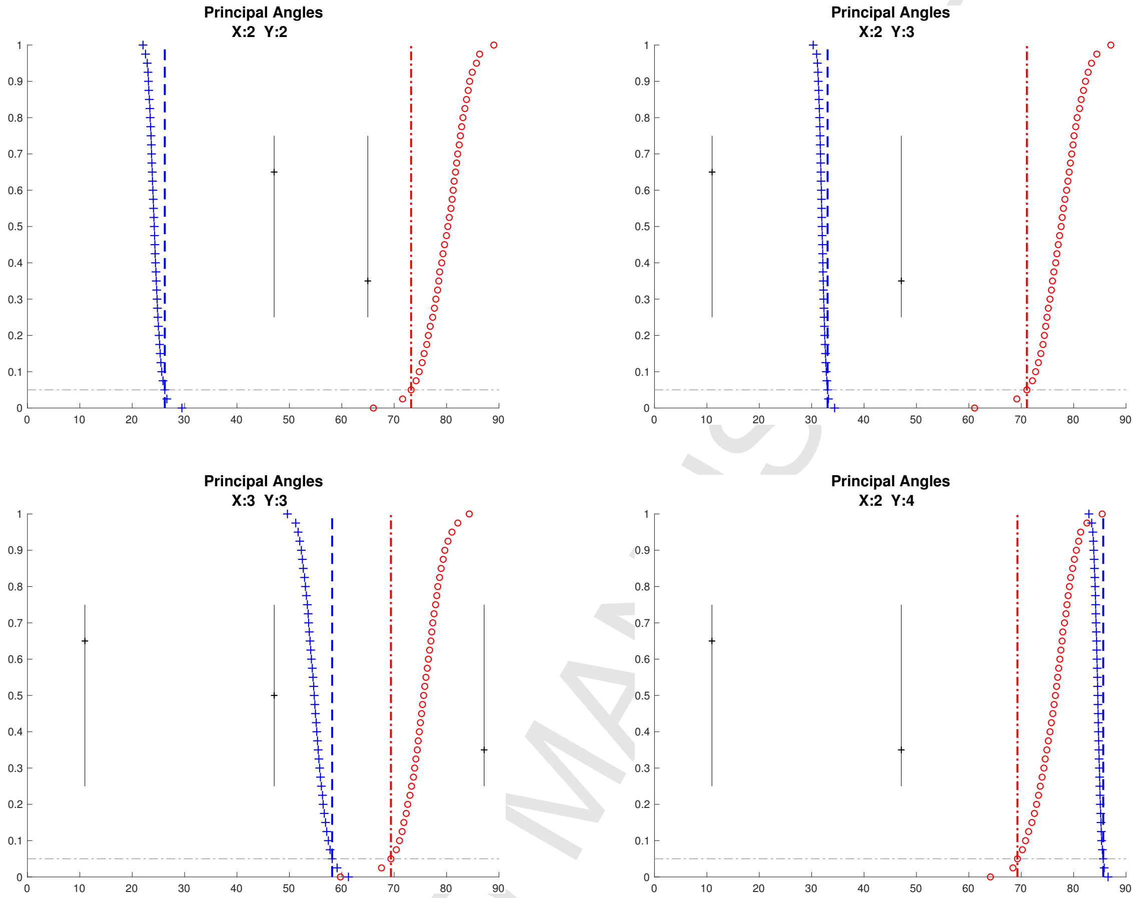

Figure 6: Principal angles and angle bounds used for segmentation in Step 2 of AJIVE for various input ranks. In each subfigure, the x-axis shows the angle and the y-axis shows the probabilities of the simulated distributions. The vertical black line segments are the values of the principal angles between row($\tilde{A}_1$) and row($\tilde{A}_2$), $\phi_1, \ldots, \phi_{\tilde{r}_1 \wedge \tilde{r}_2}$. The red circles show the values of the cumulative distribution function of the random direction distribution; the red dot-dashed line shows the 5th percentile of these angles. The blue plus signs show the values of the survival functions of the resampled Wedin bounds; the blue dashed line is the 95th percentile of the distribution. This figure contains several diagnostic plots, which provide guidance for rank selection. See Section 2.3.1 for details.

The correctly specified rank case of $\tilde{r}_1 = 2, \tilde{r}_2 = 3$ is studied in the upper right panel of Figure 6. Now the smallest angle is smaller than the blue Wedin bound, suggesting a joint component. The second principal angle is about 45°, which is the angle between the individual spaces. This is above the blue Wedin bound, so it is not joint structure.

The lower left panel considers the over specified initial rank of $\tilde{r}_1 = \tilde{r}_2 = 3$. The over specification results in a loosening of the blue Wedin bound, so that now we can no longer conclude 45° is not joint, i.e., $\tilde{r}_J = 2$ cannot be ruled out for this choice of ranks. Note that there is a third principal angle, larger than the red random direction bound, which thus cannot be distinguished from pure noise, which make sense because A_1 has only rank $r_1 = 2$.

A case where the Wedin bound is useless is shown in the lower right. Here the initial ranks are $\tilde{r}_1 = 2$ and $\tilde{r}_2 = 4$, which results in the blue Wedin bound being actually larger than the red random direction bound. In such cases, the Wedin bound inference is too conservative to be useful. While not always true, the fact that this can be caused by over specification gives a suggestion that the initial ranks may be too large. Further analysis of this is an interesting open problem.

2.3.2. *Multi-block case*

To generalize the above idea to more than two blocks, we focus on singular values rather than on principal angles in Eq. (5). In other words, instead of finding an upper bound on an angle, we will focus on a corresponding lower bound on the remaining energy as expressed by the sum of the squared singular values. Hence, an analogous SVD will be used for studying the closeness of multiple initial signal score subspace estimates.

For the vertical concatenation of right singular vector matrices

$$M \triangleq \begin{bmatrix} \tilde{V}_1^\top \\ \vdots \\ \tilde{V}_K^\top \end{bmatrix} = U_M \Sigma_M V_M^\top, \tag{6}$$

SVD sorts the directions within these K subspaces in increasing order of amount of deviation from the theoretical joint direction. The squared singular value $\sigma_{M,i}^2$ indicates the total amount of variation explained in the common direction $V_{M,i}^\top$ in the score subspace of $\mathbb{R}^n$. A large value of $\sigma_{M,i}^2$ (close to K) suggests that there is a set of K basis vectors within each subspace that are close to each other and thus are potential noisy versions of a common joint score vector. As in Section 2.3.1, the random direction bound and the Wedin bound for these singular values are used for segmentation of this joint and individual components in the multi-block case.

Random direction bound. The extension of the random direction bound in Section 2.3.1 is straightforward. The distribution of the largest squared singular value in (6) generated by random subspaces is also obtained by simulation. As in the two block case, each $\tilde{V}_k$ in M is replaced by an independent random subspace, i.e., right multiplied by an independent orthonormal matrix. The simulated distribution of the largest singular value of M indicates singular values potentially driven by pure noise. For the toy example, the values of the survival function of this distribution are shown as red circles in Figure 7, a singular value analog of Figure 6. The 5th percentile of this distribution, shown as the vertical red dot-dashed line in Figure 7, is used as the random direction bound for squared singular value, which provides 95% confidence that the squared singular values larger than this cutoff are not generated by random subspaces.

Threshold based on the Wedin bound. Next is the lower bound for segmentation of the joint space based on the Wedin bound.

Lemma 3. *Let $\theta_{k,\tilde{r}_k \wedge r_k}$ be the largest principal angle between the theoretical subspace* row(A_k) *and its estimation* row$(\tilde{A}_k)$ *for K data blocks from Eq.* (2). *The squared singular values* ($\sigma_{M,i}^2$) *corresponding to the estimates of the joint components satisfy*

$$\sigma_{M,i}^2 \geq K - \sum_{k=1}^{K} \sin^2 \theta_{k,\tilde{r}_k \wedge r_k} \geq K - \sum_{k=1}^{K} \left\{ \frac{\max(\|E_k \tilde{V}_k\|, \|E_k^\top \tilde{U}_k\|)}{\sigma_{\min}(\tilde{A}_k)} \wedge 1 \right\}^2 . \tag{7}$$

The proof is provided in Appendix A. This lower bound is independent of the variation magnitudes. This property makes AJIVE insensitive to scale heterogeneity across each block when extracting joint variation information.

As in Section 2.2.2, all the terms $\|E_k \tilde{V}_k\|$, $\|E_k^\top \tilde{U}_k\|$ are resampled to derive a distribution estimator for the lower bound in (7), which can provide a prediction interval as well. Figure 7 shows the values of the cumulative distribution function of this upper bound as blue plus signs for the toy example. As in the two-block case, if there are $\tilde{r}_J$ singular values larger than both this lower bound and the random direction bound, the first $\tilde{r}_J$ right singular vectors are used as the basis of the estimator of row(J).

In the two-block case, Figure 7 contains essentially the same information as Figure 6, thus the same insights are available. Since principal angles between multiple subspaces are not defined, Figure 7 provides appropriate generalization to the multi-block case, see Figure 8.

2.4. *Step 3: Final decomposition and outputs*

Based on the estimate of the joint row space, matrices containing joint variation in each data block can be reconstructed by projecting X_k onto this estimated space. Define the matrix $\tilde{V}_J$ as $[\vec{v}_{M,1}, \ldots, \vec{v}_{M,\tilde{r}_J}]$, where $\vec{v}_{M,i}$ is the

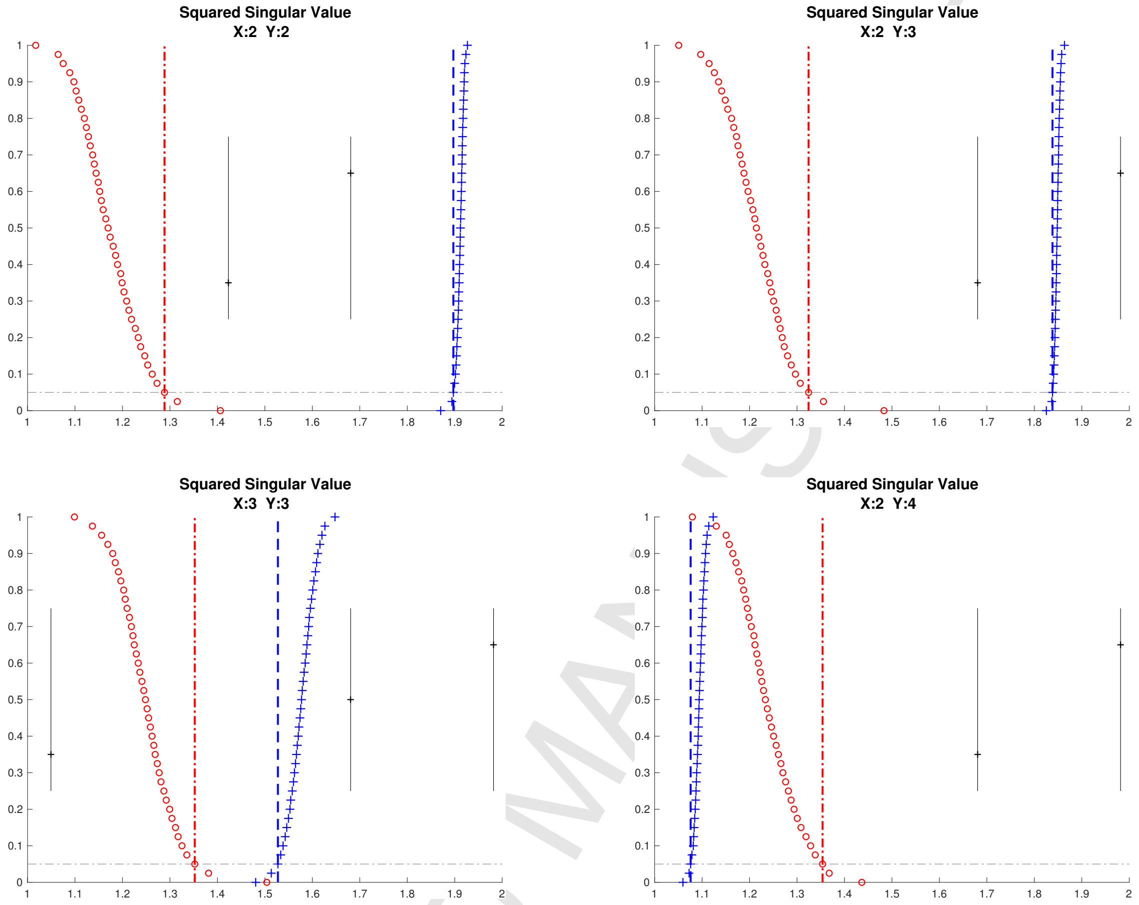

Figure 7: Squared singular values in (6) and bounds for Step 2 of AJIVE for various rank choices. The black vertical line segments shows the first $\tilde{r}_1 \wedge \tilde{r}_2$ squared singular values of M in equation (6). The values of the survival function of the random direction bounds are shown as the red circles and the red dot-dashed line is the 95th percentile of this distribution, which is the random direction bound. The values of the c.d.f of the Wedin bound are shown as the blue plus signs and the 5th percentile (blue dashed line) is used for a prediction interval for the Wedin bound. In the two-block case presented here this contains the essentially same information as in Figure 6. For the multi-block case it is the major diagnostic graphic.

ith column in the matrix V_M. To ensure that all components continue to satisfy the identifiability constraints from Section 2.2.1, we check that, for all the blocks, each $\|X_k\vec{v}_{M,i}\|$ is also above the corresponding threshold used in Step 1. If the constraint is not satisfied for any block, that component is removed from $\tilde{V}_J$. A real example of this happens in Section 3.1. An important point is that this removal can happen even when there is a common joint structure in all but a few blocks.

Denote $\hat{V}_J$ as the matrix $\tilde{V}_J$ after this removal and $\hat{r}_J$ as the final joint rank. The projection matrix onto the final estimated joint space row($\hat{J}$) is $P_J = \hat{V}_J\hat{V}_J^\top$, represented as the red rectangle in Figure 1. The estimate of the joint variation matrices in block $k \in \{1, \dots, K\}$ is $\hat{J}_k = X_kP_J$.

The row space of joint structure is orthogonal to the row spaces of each individual structure. Therefore, the original data blocks are projected to the orthogonal space of row($\hat{J}$). The projection matrix onto the orthogonal space of row($\hat{J}$) is $P_J^\perp = I - P_J$ and the projections of each data block are denoted as $X_k^\perp$ respectively for each block, i.e., $X_k^\perp = X_kP_J^\perp$. These projections are represented as the circled minus signs in Figure 1.

Finally we rethreshold this projection by performing SVD on $X_1^\perp, \dots, X_K^\perp$. The components with singular values larger than the first thresholds from Section 2.2.1 are kept as the individual components, denoted as $\hat{I}_1, \dots, \hat{I}_K$. The remaining components of each SVD are regarded as an estimate of the noise matrices.

By taking a direct sum of the estimated row spaces of each type of variation, denoted by $\oplus$, the estimated signal row spaces are

$$\mathrm{row}(\hat{A}_k) = \mathrm{row}(\hat{J}) \oplus \mathrm{row}(\hat{I}_k)$$

with rank $\hat{r}_k = \hat{r}_J + \hat{r}_{I_k}$ respectively for each $k \in \{1, \ldots, K\}$.

Due to this adjustment of directions of the joint components, these final estimates of signal row spaces may be different from those obtained in the initial signal extraction step. Note that even the estimates of rank $\hat{r}_k$ might also differ from the initial estimates $\tilde{r}_k$.

Given the variation decompositions of each AJIVE component, as shown on the right side of Figure 1, several types of post AJIVE representations are available for representing the joint and individual variation patterns. There are three important matrix representations of the information in the AJIVE joint output, i.e., the boxes on the right in Figure 1, with differing uses in post AJIVE analyses.

1. *Full matrix representation.* For applications where the original features are the main focus (such as finding driving genes), the full $d_k \times n$ matrix representations $\hat{J}_k$ and $\hat{I}_k$ with $k \in \{1, \ldots, K\}$ are most useful. Thus this AJIVE output is the product of all three blocks in each dashed box on the right side of Figure 1. Examples of these outputs are shown in the two right columns of Figure 3.

2. *Block specific representation.* For applications where the relationships between subjects are the main focus (such as discrimination between subtypes) large computational gains are available by using lower dimensional representations. These are based on SVDs as indicated in the right side of Figure 1, i.e., for each $k \in \{1, \ldots, K\}$,

$$\hat{J}_k = \hat{U}_J^k \hat{\Sigma}_J^k \hat{V}_J^{k\top}, \quad \hat{I}_k = \hat{U}_I^k \hat{\Sigma}_I^k \hat{V}_I^{k\top}. \tag{8}$$

The resulting AJIVE outputs include the joint and individual *Block Specific Score* (BSS) matrices $\hat{\Sigma}_J^k \hat{V}_J^{k\top}$ $(\hat{r}_J \times n)$, $\hat{\Sigma}_I^k \hat{V}_I^{k\top}$ $(\hat{r}_{I_k} \times n)$, respectively. This results in no loss of information when rotation invariant methods are used. The corresponding *Block Specific Loading* matrices are $\hat{U}_J^k$ $(d_k \times \hat{r}_J)$ and $\hat{U}_I^k$ $(d_k \times \hat{r}_{I_k})$.

3. *Common normalized representation.* Although $\mathrm{row}(\hat{V}_J^{k\top})$ in (8) are the same, the matrices are different. In particular, the rows in (8) can be completely different across k, because they are driven by the pattern of the singular values in each $\hat{\Sigma}_J^k$. In some applications, correspondence of components across data blocks is important. In this case the analysis should use a common basis of $\mathrm{row}(\hat{J})$, namely $\hat{V}_J^\top$ $(\hat{r}_J \times n)$, called the *Common Normalized Scores* (CNS). This is shown as the gray rectangular near the center of Figure 1. To get the corresponding loadings, we regress $\hat{J}_k$ on each score vector in $\hat{V}_J^\top$ (which is computed as $\hat{J}_k \hat{V}_J$) following by normalization. By doing this, there is no guarantee of orthogonality between CNS loading vectors. However, the loadings are linked across blocks by their common scores. For studying scale free individual spaces, use the *Individual Normalized Scores* (INS) $\hat{V}_I^{k\top}$ $(\hat{r}_{I_k} \times n)$. The individual loading matrices $\hat{U}_I^k$ are the same as the block specific individual loadings.

The relationship between Block Specific Representation and Common Normalized Representation is analogous to that of the traditional covariance, i.e., PLS, and correlation, i.e., CCA, modes of analysis. The default output in the AJIVE software is the Common Normalized Representation.

3. Data analysis

In this section, we apply AJIVE to two real data sets, TCGA breast cancer in Section 3.1 and Spanish mortality in Section 3.2.

3.1. TCGA Data

A prominent goal of modern cancer research, of which The Cancer Genome Atlas [25] is a major resource, is the combination of biological insights from multiple types of measurements made on common subjects.

TCGA provides prototypical data sets for the application of AJIVE. Here we study the 616 breast cancer tumor samples from Ciriello et al. [4], which had a common measurement set. For each tumor sample, there are measurements of 16615 gene expression features (GE), 24174 copy number variations features (CN), 187 reverse phase protein array features (RPPA) and 18256 mutation features (Mutation). These data sources have very different dimensions and scalings.

The tumor samples are classified into four molecular subtypes: Basal-like, HER2, Luminal A and Luminal B. An integrative analysis targets the association among the features of these four disparate data sources that jointly quantify the differences between tumor subtypes. In addition, identification of driving features for each source and subtype is obtained from studying loadings.

Scree plots were used to find a set of interesting candidates for the initial ranks selected in Step 1. Various combinations of them were investigated using the diagnostic graphic. Four interesting cases are shown in Figure 8. The upper left panel of Figure 8 is a case where the input ranks are too small, resulting in no joint components being identified, i.e., all the black lines are smaller than the dashed blue estimated Wedin bound. The upper right panel shows a case where only one joint component is identified. In addition to the joint component identified in the upper right panel, the lower left panel contains a second potential joint component close to the Wedin bound. The lower right panel shows a case where the Wedin bound becomes too small since the input ranks are too large. Many components are suggested as joint here, but these are dubious because the Wedin bound is smaller than the random direction bound. Between the two viable choices, in the upper right and the lower left, we investigate the latter in detail, as it best highlights important fine points of the AJIVE algorithm. In particular, we choose low-rank approximations of dimensions 20 (GE), 16 (CN), 15 (RPPA) and 27 (Mutation). However, detailed analysis of the upper right panel results in essentially the same final joint component. After selection of the threshold in Step 1, it took AJIVE 298 seconds (5.0 minutes, on Macbook Pro Mid 2012, 2.9 GHz) to finish Steps 2 and 3.

In the second AJIVE step, the one sided 95% prediction interval suggested selection of two joint components. However, the third step indicated dropping one joint component, because the norm of the projection of the mutation data on that direction, i.e., the second CNS, is below the threshold from Step 1. This result of one joint component was consistent with the expectation of cancer researchers, who believe the mutation component has only one interesting mode of variation. A careful study of all such projections shows that the other data types, i.e., GE, CN and RPPA, do have a common second joint component as discussed at the end of this section. The association between the CNS and genetic subtype differences is visualized in the left panel of Figure 9. The dots are a jitter plot of the patients, using colors and symbols to distinguish the subtypes (Blue for Basal-like, cyan for HER2, red for Luminal A and magenta for Luminal B). Each symbol is a data point whose horizontal coordinate is the value and vertical coordinate is the height based on data ordering. The curves are Gaussian kernel density estimates, i.e., smoothed histograms, which show the distribution of the subtypes.

The clear separation among density estimates suggest that this joint variation component is strongly connected with the subtype difference between Luminal A versus the other subtypes. To quantify this subtype difference, a test is performed using the CNS of this joint component evaluated by the DiProPerm hypothesis test [41] based on 100 permutations. Strength of the evidence is usually measured by permutation *p*-values. However, in this context empirical *p*-values are frequently zero. Thus a more interpretable measure of strength of the evidence is the DiProPerm *z*-score. This is 26.54 for this CNS. An area under the receiver operating characteristic (ROC) curve (AUC) [7] of 0.878, is also obtained to reflect the classification accuracy. These numbers confirm the strong Luminal A property shared by these four data types.

A further understanding can be obtained by identifying the feature set of each data type which jointly works with the others in characterizing the Luminal A property. By studying the loading coefficients, important mutation features TP53,TTN and PIK3CA are identified which are well known features from previous studies. Similarly the strong role played by GATA3 in RPPA is well known, and is connected with the large GATA3 mutation loading. A less well known result of this analysis is the genes appearing with large GE loadings. Many of these were not flagged in earlier studies, which had focused on subgroup separation, instead of joint behavior.

As noted in the discussion of Step 2 above, all four data types have only one significant joint component. However, the individual components for all of GE, CN and RPPA seem to have 3-way joint components. This is investigated

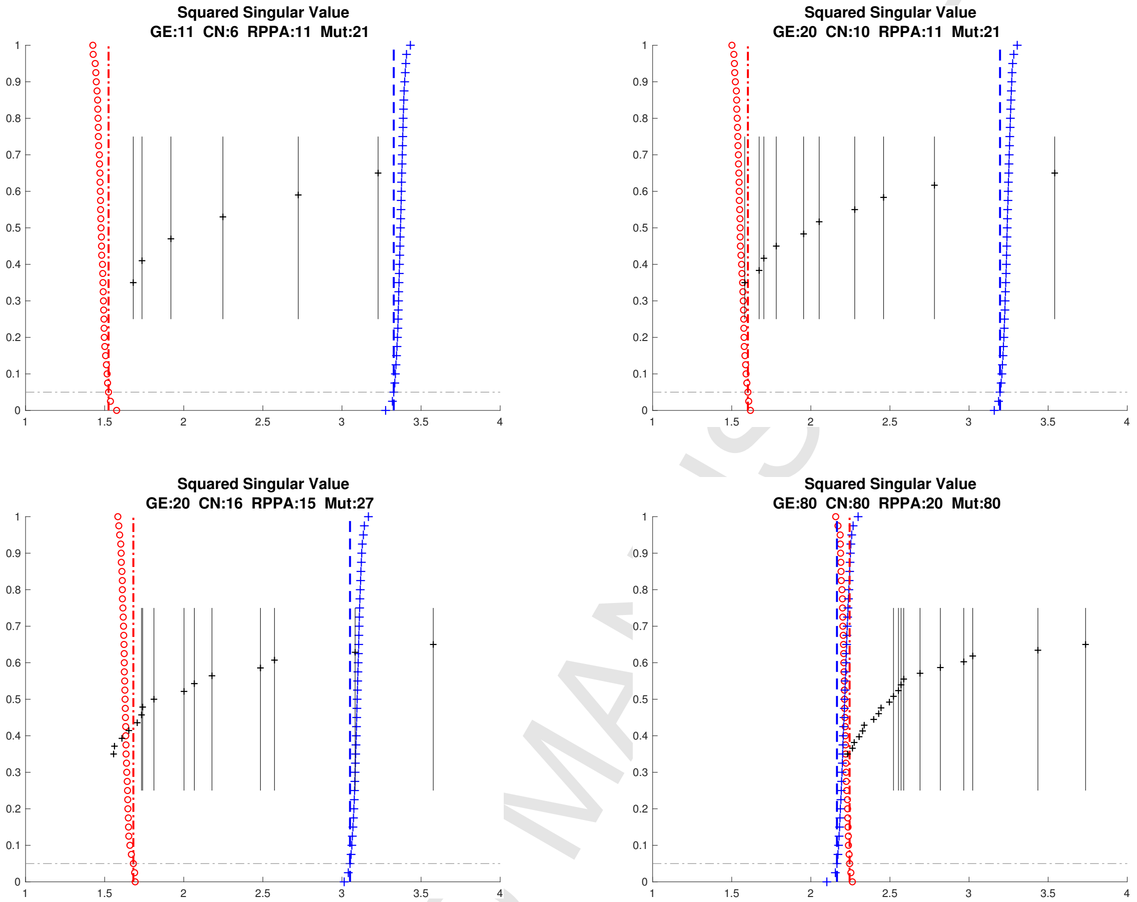

Figure 8: Squared singular value diagnostic graphics for TCGA dataset over various rank choices. Indicates that there are one joint component among four data blocks and one joint component among three data blocks.

by performing a second AJIVE analysis. In particular, we apply the second and third step to the three individual variation matrices from the initial analysis. Notice that all individual matrices are low-rank and thus the first step is not necessary. The AJIVE analysis results in one joint variation component which is displayed in the right panel of Figure 9. This joint variation component clearly shows the differences among Basal, HER2 and Luminal subtypes. In particular, a subtype difference between Basal-like versus the others is quantified using the DiProPerm z-score (31.60) and the AUC (0.996). Considering the fact that the AUC of the classification between Basal-like versus the others using all the original separate GE features is 0.999, this single joint component contains almost all the variation information for separating Basal-like from the others. This hierarchical application of AJIVE reveals an important joint component that is specific to GE, CN and RPPA but not to Mutation.

3.2. *Spanish mortality data*

A quite different data set from the Human Mortality Database is studied here, which consists of both Spanish males and females. For each gender data block, there is a matrix of *mortality*, defined as the number of people who died divided by the total, for a given age group and year. Because mortality varies by several orders of magnitude, the $\log_{10}$ mortality is studied here. Each row represents an age group from 0 to 95, and each column represents a year between 1908 and 2002. In order to associate the historical events with the variations of mortality, columns, i.e., mortality as a function of age, are considered as the common set of data objects of each gender block. Marron and

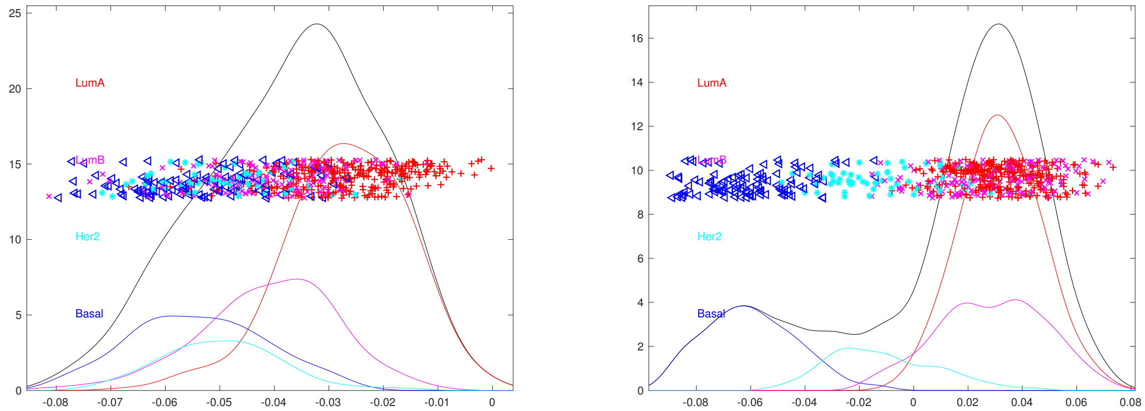

Figure 9: Left: Kernel density estimates of the CNS among GE, CN, RPPA and mutation. The clear separation among Luminal A versus Her2 and Basal indicates that these four data blocks share a very strong Luminal A property captured in this joint variation component; Right: The CNS from applying AJIVE to the individual matrices of GE, CN, and RPPA. The clear separation indicates that these contain a joint variation component that is consistent with the subtype difference between Basal versus the others.

Alonso [22] performed analysis on the male block and showed interesting interpretations related to Spanish history. Here we are looking for a deeper analysis which integrates both males and females by exploring joint and individual variation patterns.

AJIVE is applied to the two gender blocks centered by subtracting the mean of each age group. The principal angle diagnostic graphics introduced in Section 2.3 are provided for this mortality dataset over various rank choices in Figure 10 to guide the selection of initial ranks in Step 1. The upper left panel shows the case $\tilde{r}_1 = \tilde{r}_2 = 1$. The only principal angle is larger than the 95th percentile of the resampled Wedin bound and thus we conclude that no joint space is identified. The upper right panel shows the effect of increasing the initial rank choices to $\tilde{r}_1 = \tilde{r}_2 = 2$. In this case, the first principal angle becomes smaller. Because it is smaller than the Wedin bound it is identified as a joint component. The second principal angle is still larger than the Wedin bound. Thus we concluded that only one joint component is identified in this case. In the lower left panel we increase the input rank of male mortality to 3. The second principal angle becomes much smaller, in particular smaller than the Wedin bound, and thus is also labeled as joint component. This indicates that the third principal component of male mortality contains joint information. The lower right panel shows the case where $\tilde{r}_1 = 4, \tilde{r}_2 = 5$. In this case the two smallest principal angle are unchanged (and still joint). Two more principal angle appear. One is larger than the random direction bound, and thus cannot be distinguished from pure noise. The other is just inside the boundary of the much increased Wedin bound suggesting correlation among individual components. Based on these, the choice $\tilde{r}_1 = 3, \tilde{r}_2 = 2$ is used in the subsequent analysis.

The resulting AJIVE gives two joint components and one individual component for the male. Since the loading matrices provide important information on the effect of different age groups, block specific analysis together with loading matrices is most informative here.

Figure 11 shows a view of the first joint components for the males (left) and females (right) that is very different from the heat map views used in Section 1.1. While these components are matrices, additional insights come from plotting the rows of the matrices as curves over year (top) and the columns as curves over age (bottom). The curves over year (top) are colored using a heat color scheme, indexing age (black = 0 through red = 40 to yellow = 95 as shown in the vertical color bar on the bottom left). The curves over age (bottom) are colored using a rainbow color scheme (magenta = 1908 through green = 1960 to red = 2002, shown in the horizontal color bar in the top) and use the vertical axis as domain with horizontal axis as range to highlight the fact that these are column vectors. Additional visual cues to the matrix structure are the horizontal rainbow color bar in the top panel, showing that year indexes columns of the data matrix and the vertical heat color bar (bottom) showing that age indexes rows of the component matrix. Because this is a single component, i.e., a rank-one approximation of the data, each curve is a multiple of a

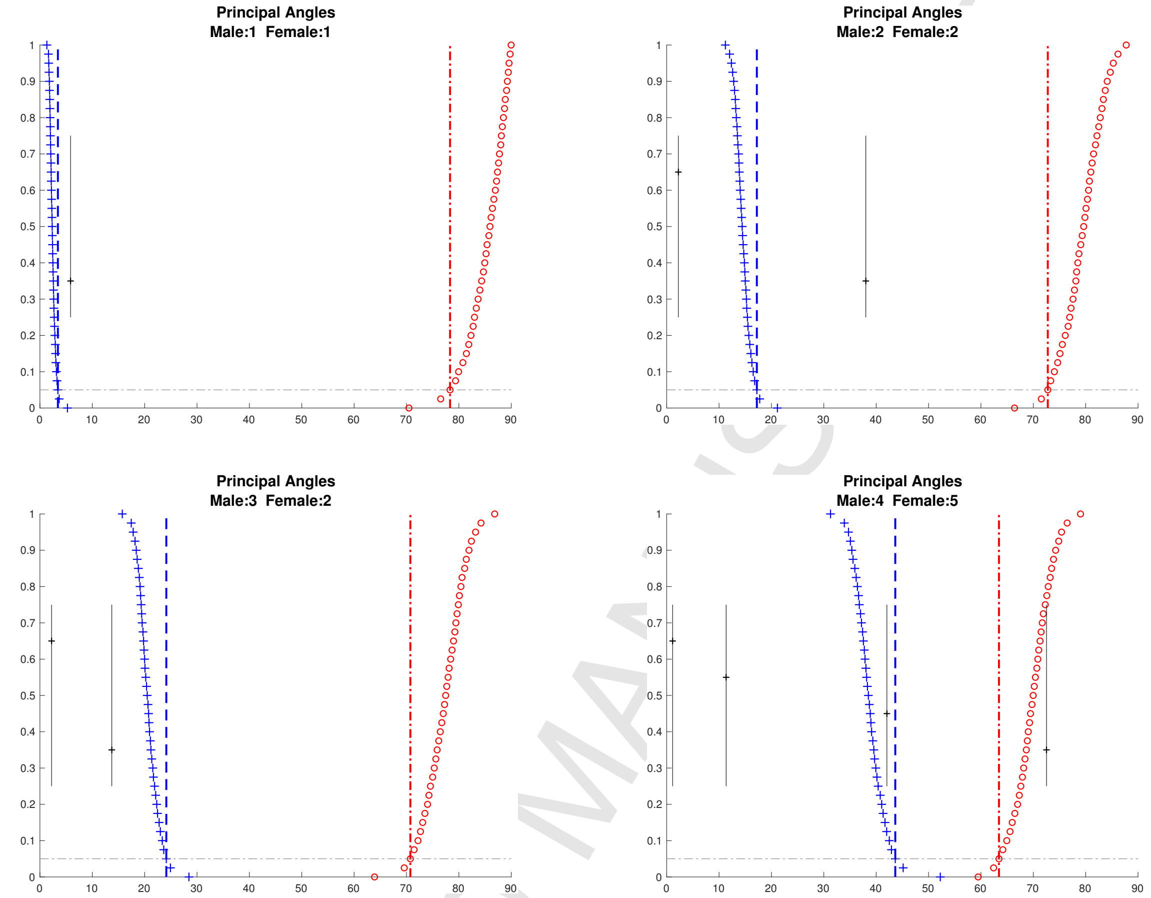

Figure 10: Principal angle diagnostic graphics for Spanish mortality data set over various rank choices. Provides the rationale of the rank choice, $\tilde{r}_1 = 3, \tilde{r}_2 = 2$.

single eigenvector. The corresponding coefficients are shown on the right. In conventional PCA/SVD terminology, the upper block specific coefficients are called *loadings*, and are in fact the entries of the left eigenvectors (colored using the heat color scale on the bottom). Similarly, the lower coefficients are called *scores* and are the entries of the right eigenvectors, colored using the rainbow bar shown in the top.

The scores plots together with the rows as curves plots in Figure 11 indicate a dramatic improvement in mortality over time for both males and females. The scores plots are bimodal indicating rapid overall improvement in mortality around the 1950s. This is also visible as the steepest part in the rows as curves plot. Thus the first mode of joint variation is driven by overall improvement in mortality. In addition to the overall improvement, the rows as curves and scores plots also show two major mortality events, the global flu pandemic of 1918 and the Spanish Civil war in the late 1930s. The loading plots together with the columns as curves plots present the different impacts of this common variation on different age groups for males and females. The loadings plot for males suggests the improvement in mortality is gradually increasing from older towards younger age groups. In contrast, the female block has a bimodal kernel density estimate of the loadings. This shows that females of child bearing age have received large benefits from improving health care. This effect is similarly visible from comparing the female versus male columns as curves.

The second block specific components of joint variation within each gender are similarly visualized in Figure 12. This common variation reflects the contrast between the years around 1950 and the years around 1980 which can be told from the curves in the left top and the colors in the right bottom subplots in both male and female panels. In the scores plots, the green circles, seen on the left end, represent the years around 1950 when automobile penetration

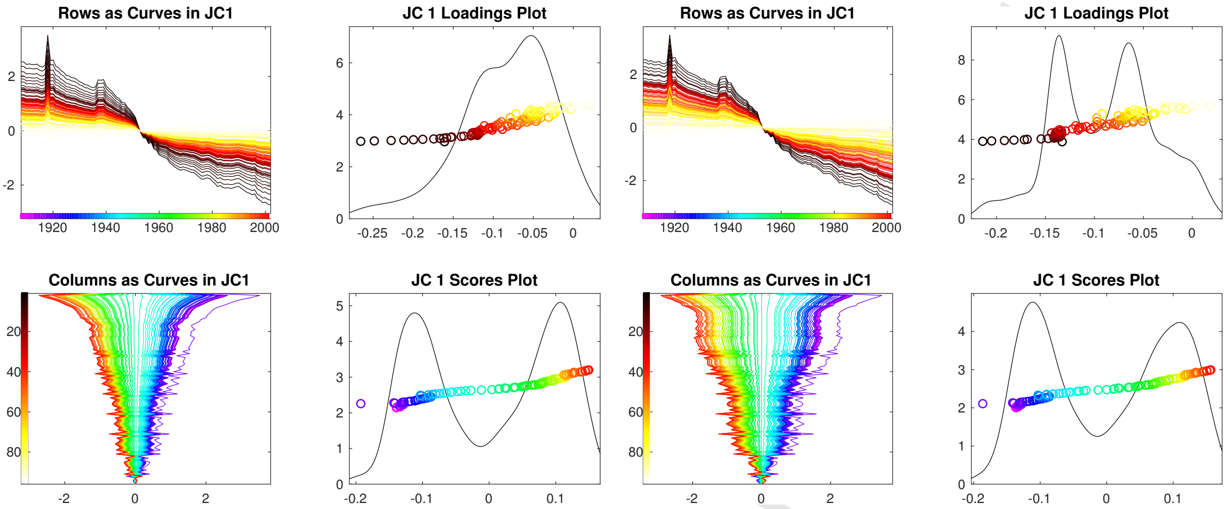

Figure 11: The first block specific joint components of male (left panel) and female (right panel) contain the common modes of variation caused by the overall improvement across different age groups, as can be seen from the scores plots in the right bottom of each panel. The dramatic decrease happened around the 1950s shown in the columns plots. The degree of decrease varies over age groups.

started. And the orange to red circles on the right end correspond to recent years, and much improved car and road safety. The loadings plot for males shows that these automobile events had a stronger influence on the 20–45 years old males in terms of both larger values and a second peak in the kernel density estimate. Although this contrast can also be seen in the loadings plot of females, it is not as strong as for the male block. Both loadings plots show an interesting outlier, the babies of age zero. We speculate this shows an improvement in post-natal care that coincidently happened around the same time.

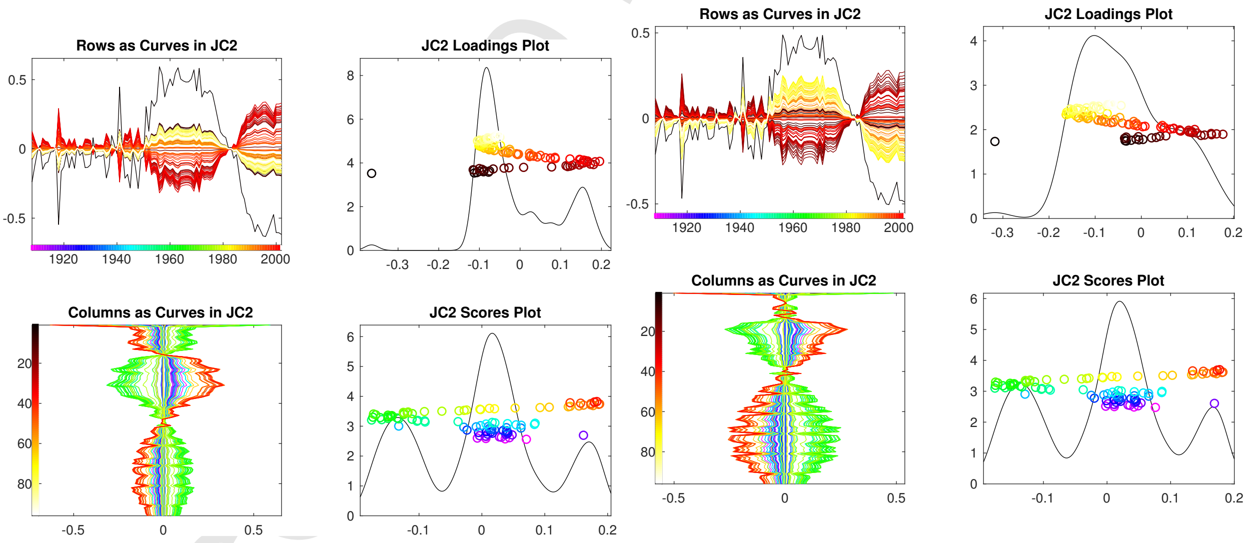

Figure 12: The second joint components of male (left) and female (right) contain the common modes of variation driven by the increase in fatalities caused by automobile penetration and later improvement due to safety improvements. This can be seen from the scores plots in the right bottom. The loadings plots show that this automobile event exerted a significantly stronger impact on the 20–45 males.

Another interesting result comes from the studying the first individual components for males, shown in Figure 13. In the scores plot of males, the blue circles stand out from the rest, corresponding to the years of the Spanish civil war

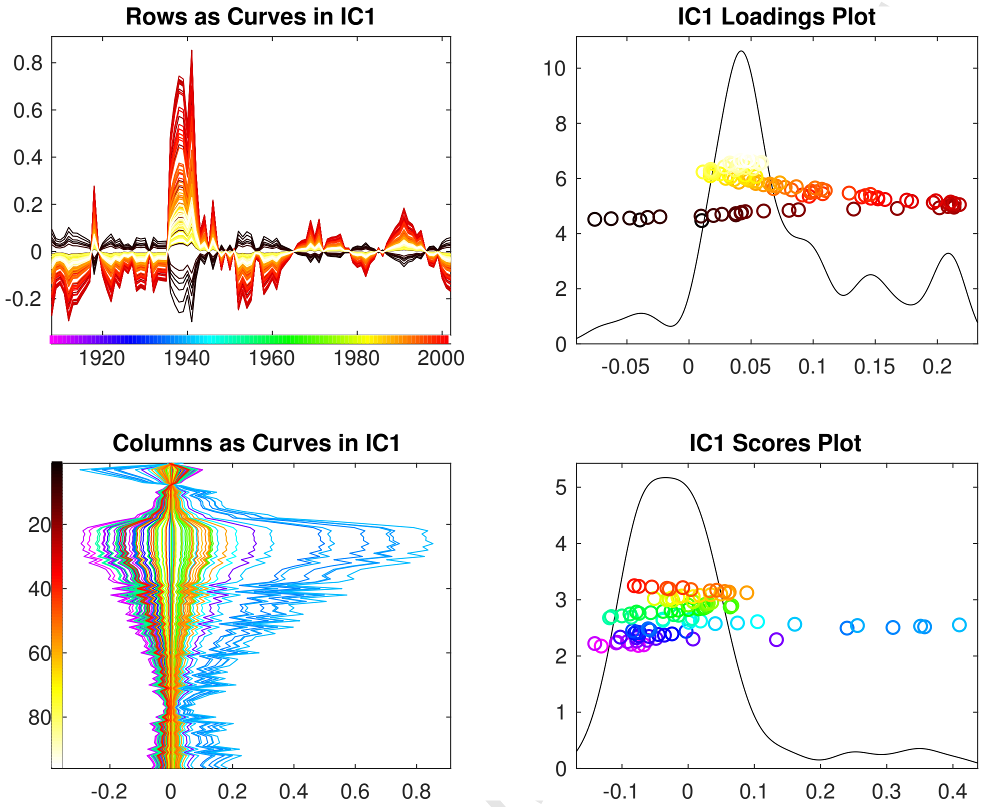

Figure 13: The individual component of male contains the variation driven by the Spanish civil war which can be seen from the blue circles on the right end of the right bottom plot. The Spanish civil war mainly affected the young to middle age males.

when a significant spike can be seen in the rows as curves plot. Young to middle age groups (typical military age) are affected more than the others as seen in the loadings plot and columns as curves plot.

4. Optimization perspective

In this section we will investigate how AJIVE compares to PLS, CCA and COBE using the optimization problems that each method is based on. Recall that $X_1, \ldots, X_K$ are $(d_k \times n)$ data matrices, with SVD decompositions $X_k = U_{X_k}\Sigma_{X_k}V_{X_k}^\top$, where Σ_{X_k} contains no zeros on its diagonal. To be compatible with AJIVE, we will consider these three algorithms in a non-standard configuration using row spaces. In Sections 4.1 and 4.2, we assume that the matrices X_k are row centered. We will also use the following notation: for $\vec{a}_1 \in \mathbb{R}^{d_1}, \vec{a}_2 \in \mathbb{R}^{d_2}$,

$$\langle \vec{a}_1 X_1, \vec{a}_2 X_2 \rangle = \operatorname{cov}(\vec{a}_1 X_1, \vec{a}_2 X_2) = \sqrt{\operatorname{var}(\vec{a}_1 X_1)\operatorname{var}(\vec{a}_2 X_2)}\operatorname{corr}(\vec{a}_1 X_1, \vec{a}_2 X_2).$$

4.1. Partial least squares

The PLS finds linear combinations of rows of X_1 and X_2 maximizing their sample covariance. More precisely, the PLS identifies a set of pairs of principal vectors, indexed by i, obtained sequentially from the following maximization problems:

$$\begin{gathered}\{\vec{a}_1^{(i)}, \vec{a}_2^{(i)}\} = \underset{\vec{a}_1 \in \mathbb{R}^{d_1}, \vec{a}_2 \in \mathbb{R}^{d_2}}{\operatorname{argmax}} \langle \vec{a}_1 X_1, \vec{a}_2 X_2 \rangle \\ \text{subject to the constraints: } \|\vec{a}_1\| = 1, \|\vec{a}_2\| = 1, \\ \langle \vec{a}_1 X_1, \vec{a}_1^{(j)} X_1 \rangle = 0, \langle \vec{a}_2 X_2, \vec{a}_2^{(j)} X_2 \rangle = 0, \text{ for all } j \in \{1, \ldots, i-1\}.\end{gathered} \tag{9}$$

Unlike AJIVE, the directions from PLS are influenced by both variance within data blocks and correlation between the data blocks. In particular, if the signal strength of the individual structure is sufficiently large it might be mistakenly classified as a joint structure by being found ahead of the real joint structure. This phenomenon can be seen in the analysis of the toy example of Section 1.1 in Appendix B.

4.2. Canonical correlation analysis/ Principal angle analysis

Similar to PLS, the CCA finds linear combinations of rows of X_1 and X_2 maximizing their sample correlation. In particular, CCA identifies a set of pairs of canonical vectors obtained sequentially from the optimization problem:

$$\begin{aligned} \{\vec{a}_1^{(i)}, \vec{a}_2^{(i)}\} = & \underset{\vec{a}_1 \in \mathbb{R}^{d_1}, \vec{a}_2 \in \mathbb{R}^{d_2}}{\operatorname{argmax}} \langle \vec{a}_1 X_1, \vec{a}_2 X_2 \rangle \\ \text{subject to the constraints: } & \|\vec{a}_1 X_1\| = 1, \|\vec{a}_2 X_2\| = 1 \\ & \langle \vec{a}_1 X_1, \vec{a}_1^{(j)} X_1 \rangle = 0, \langle \vec{a}_2 X_2, \vec{a}_2^{(j)} X_2 \rangle = 0, \text{ for all } j \in \{1, \ldots, i-1\}. \end{aligned} \tag{10}$$

This form makes the relationship between (9) and (10) clear and is equivalent to the usual formulation of optimizing the correlation.

There is an important relationship between CCA and PAA [2], i.e., if $\rho_i = \langle \vec{a}_1^{(i)} X_1, \vec{a}_2^{(i)} X_2 \rangle$ is the ith canonical correlation, $\rho_i = \cos(\theta_i)$, where θ_i is the ith principal angle between row spaces of X_1 and X_2. The principal vector pairs $\{\vec{x}_{1,i}, \vec{x}_{2,i}\} = \{\vec{a}_1^{(i)} X_1, \vec{a}_2^{(i)} X_2\}$ are often obtained through SVD of $V_{X_1}^\top V_{X_2}$. In particular, let $\vec{u}_{X_1,i}, \vec{u}_{X_2,i}$ be the ith left and right singular vectors of $V_{X_1}^\top V_{X_2}$. Then, the ith pair of principal vectors are

$$\vec{x}_{1,i} = \vec{u}_{X_1,i}^\top V_{X_1}^\top, \quad \vec{x}_{2,i} = \vec{u}_{X_2,i}^\top V_{X_2}^\top.$$

An issue with CCA of high-dimensional data is related to the fact that CCA is interested in the canonical vectors $\vec{a}_i$ rather than the principal vectors $\vec{x}_i$. In particular, when $d_1 > n, d_2 > n$, the values of $\vec{a}_i$ in (10) are not identifiable due to the singularity of $X_1 X_1^\top$ and $X_2 X_2^\top$. Several approaches have been taken to solve this problem. One approach is to use the Moore–Penrose pseudo inverse to replace the inverse of $X_1 X_1^\top$ and $X_2 X_2^\top$. A second approach is to add a ridge penalty on $X_1 X_1^\top$ and $X_2 X_2^\top$ [38]. A third approach called penalized CCA is to add penalty functions on $\{\vec{a}_1^{(i)}, \vec{a}_2^{(i)}\}$, such as an ℓ_1 penalty [16, 29], an elastic net [39] or a fused lasso [43]. Another approach called diagonal penalized CCA is to replace $X_1 X_1^\top$ and $X_2 X_2^\top$ by $\operatorname{diag}(X_1 X_1^\top)$ and $\operatorname{diag}(X_2 X_2^\top)$ [30, 43].

Another important issue with CCA, which is directly related to AJIVE, is that when $d_1 > n, d_2 > n$, CCA is generally driven by noise. Lee [17, 18], Samarov [32] study the asymptotic behavior of CCA in the high-dimension low sample size context and point out the inconsistency phenomenon in this case. One can solve this issue, like AJIVE and COBE, by replacing X_k by its low-rank approximation $\tilde{A}_1, \tilde{A}_2$. Recall notation from Eq. (1). The ith principal vectors are $\vec{p}_i = \tilde{V}_1 \vec{u}_{1,i}$, $\vec{q}_i = \tilde{V}_2 \vec{u}_{2,i}$, where $\vec{u}_{j,i}$ is the ith singular vector of $\tilde{U}_i$ of the SVD of $\tilde{V}_1^\top \tilde{V}_2$ respectively.

As discussed in Section 2, AJIVE uses an equivalent principal angle calculation based on SVD of $M = [\tilde{V}_1, \tilde{V}_2]^\top = U_M \Sigma_M V_M^\top$ [23]. AJIVE uses the transpose of the ith right singular vector, $V_{M,i}^\top$, as the estimated ith basis vector of the joint space, provided that the ith principal angle is smaller than the threshold derived in Section 2.3.1. Moreover, if the ith principal angle has a value distinct from other principal angles, then the ith left singular vector of M can be written as $U_{M,i} = [\vec{u}_{1,i}^\top, \vec{u}_{2,i}^\top]^\top / \sqrt{2}$. Consequently

$$V_{M,i}^\top = \frac{1}{\sigma_{M,i}} U_{M,i}^\top M = \frac{1}{\sqrt{2}\sigma_{M,i}} (\vec{u}_{1,i}^\top \tilde{V}_1^\top + \vec{u}_{2,i}^\top \tilde{V}_2^\top) = \frac{1}{\sqrt{2}\sigma_{M,i}} (\vec{p}_i^\top + \vec{q}_i^\top).$$

This shows that the AJIVE direction $V_{M,i}^\top$ is the scaled sum of the ith pair of principal vectors.

CCA applied to the low-rank approximations $\tilde{A}_k$ and AJIVE are therefore closely related. However, AJIVE provides one joint vector per two distinct principal vectors that by the virtue of being an average should be a better estimate of the joint space than either of the principal vectors. More importantly, AJIVE uses a theoretically sound threshold of the principal angles that allows us to segment individual and joint variation.

The AJIVE formulation allows for a natural extension to multi-block situations. Several approaches of Multiset Canonical Correlation Analysis (mCCA) have been developed as extensions of CCA [8, 12, 26]. There is no general consensus on which of these extensions is preferable. We point out that AJIVE is closely related to one of the mCCA discussed in Nielsen [26].

This version of mCCA is defined using the optimization problem for the ith set of canonical vectors $\{\vec{a}_1^{(i)}, \ldots, \vec{a}_K^{(i)}\}$ and corresponding principal vectors (also called canonical variables) $\{\vec{a}_1^{(i)} X_1, \ldots, \vec{a}_K^{(i)} X_K\}$:

$$\begin{aligned}
\{\vec{a}_1^{(i)}, \ldots, \vec{a}_K^{(i)}\} &= \underset{\vec{a}_1,\ldots,\vec{a}_K}{\operatorname{argmax}} \sum_{1 \le k,l \le K} \langle \vec{a}_k X_k, \vec{a}_l X_l \rangle \\
\text{subject to the constraints: } &\sum_{k=1}^{K} \|\vec{a}_k X_k\|_2^2 = 1, \qquad (11) \\
\langle \vec{a}_k X_k, \vec{a}_k^{(j)} X_k \rangle = 0, &\quad \text{for all } k \in \{1, \ldots, K\}, \quad j \in \{1, \ldots, i-1\}.
\end{aligned}$$

Notice that the constraint in (11) is different than the perhaps more natural $\|\vec{a}_k X_k\|_2^2 = 1$ for all $k \in \{1, \ldots, K\}$.

If the ith singular value corresponding to the AJIVE direction $V_{M,i}^{\top}$ has a value distinct from other singular values in the AJIVE SVD, then calculations similar to the two block case show that the ith basis vector of the joint space from AJIVE

$$V_{M,i}^{\top} = \frac{1}{\sigma_{M,i}} \sum_{k=1}^{K} \vec{a}_k^{(i)} X_k$$

is the scaled sum of the corresponding canonical variables. In fact, $V_{M,i}^{\top}$ is the ith flag mean of the row spaces of $X_1, \ldots, X_K$, as defined by Draper et al. [5], which thus is a building block of AJIVE.

4.3. Common orthogonal basis extraction

Zhou et al. [50] proposed a compelling optimization problem for finding the common orthogonal basis (COBE). It is based on iteratively solving

$$\begin{aligned}
\bar{a}_i = \underset{\bar{a},\ z_{i,k}, k=1,\ldots,K}{\operatorname{argmin}} \sum_{k=1}^{K} \|\tilde{V}_k z_{i,k} - \bar{a}\|^2 \qquad (12) \\
\text{subject to the constraints: } \|\bar{a}\|_2 = 1, \langle \bar{a}, \bar{a}_j \rangle = 0, \text{ for all } j \in \{1, \ldots, i-1\}.
\end{aligned}$$

To compare COBE to AJIVE we first simplify the objective function of (12) to

$$\sum_{k=1}^{K} \|\tilde{V}_k z_{i,k} - \bar{a}_i\|_2^2 = \sum_{k=1}^{K} \|\tilde{V}_k z_{i,k}\|_2^2 + K\|\bar{a}_i\|_2^2 - 2\sum_{k=1}^{K} \langle \tilde{V}_k z_{i,k}, \bar{a}_i \rangle = \|z_i\|_2^2 + K\|\bar{a}_i\|_2^2 - 2z_i^{\top} M \bar{a}_i.$$

where $z_i = [z_{i,1}, \ldots, z_{i,K}]$. If we fix the value of $\|z_i\|$ we see that the solution to the optimization problem (12) is the same as SVD of M with $\bar{a}_i = V_{M,i}$. Moreover this solution is invariant in $\|z_i\|$.

Thus the optimization problem (12) gives the same result as AJIVE. However, because AJIVE uses well optimized SVD rather than a heuristic iteration algorithm, AJIVE is much faster than the COBE algorithm. Moreover, COBE lacks any principally based standard on how to choose the threshold for selecting the joint space.

To understand why this is a serious issue consider the results of applying COBE to the TCGA data discussed in detail in the next section. To make comparisons fair we provided COBE the same selected first stage ranks for each data block as AJIVE. COBE's default threshold for separating joint and individual structure of 0.01 is too low to find any joint component. Therefore we tried raising the default threshold 0.01 to 1, in which case COBE fails to finish on our computer due to its inefficient handling of high dimensional data.

Acknowledgments

This research was supported in part by the National Science Foundation under Grant Nos 1016441, 1512945 and 1633074.

Appendix A. Proofs

Proof of Lemma 1. Define the row subspaces respectively for each matrix A_k as $\text{row}(A_k) \subseteq \mathbb{R}^n$. For non-trivial cases, define a subspace $\text{row}(J) \neq \{\vec{0}\}$ as the intersection of the row spaces $\{\text{row}(A_1), \ldots, \text{row}(A_K)\}$, i.e.,

$$\text{row}(J) \triangleq \bigcap_{k=1}^{K} \text{row}(A_k).$$

For each matrix A_k, two matrices J_k, I_k can be obtained by projection of A_k on $\text{row}(J)$ and its orthogonal complement in the row space $\text{row}(A_k)$. Thus the two matrices satisfy $J_k + I_k = A_k$ and their row subspaces are orthogonal with each other, i.e., $\text{row}(J) \perp \text{row}(I_k)$, for all $k \in \{1, \ldots, K\}$. Then the intersection of the row subspaces $\{\text{row}(I_1), \ldots, \text{row}(I_K)\}$, $\bigcap\limits_{k=1}^{K} \text{row}(I_k)$, has a zero projection matrix. Therefore, we have $\bigcap\limits_{k=1}^{K} \text{row}(I_k) = \{\vec{0}\}$ and have obtained a set of matrices simultaneously satisfying the stated constraints.

On the other hand, it follows from the assumptions that the row space $\text{row}(A_k)$ is spanned by the union of basis vectors of $\text{row}(J_k)$ and $\text{row}(I_k)$, which indicates

$$\text{row}(J) = \bigcap_{k=1}^{K} \text{row}(A_k).$$

Accordingly, the matrices $J_1, \ldots, J_K$ and $I_1, \ldots, I_K$ are also uniquely defined.

□

Proof of Lemma 2. Let P_1 and P_2 be the projection matrices onto the individually perturbed joint row spaces. And let P be the projection matrix onto the common joint row space J. Thus, we have

$$\begin{aligned}\sin\theta = \|(I - P_1)P_2\| &\leq \|(I - P_1)(I - P)P_2\| + \|(I - P_1)PP_2\| \\ &\leq \|(I - P_1)(I - P)\| \, \|(I - P)P_2\| + \|(I - P_1)P\| \, \|PP_2\|,\end{aligned}$$

in which $\|(I - P_1)P\| = \sin\theta_{11}$, $\|(I - P_1)(I - P)\| = \cos\theta_1$, $\|(I - P_2)P\| = \sin\theta_{2,1}$ and $\|(I - P_2)(I - P)\| = \cos\theta_{2,1}$. Therefore,

$$\sin\phi \leq \cos\theta_{1,1}\sin\theta_{2,1} + \sin\theta_{1,1}\cos\theta_{2,1} = \sin(\theta_{1,1} + \theta_{2,1}).$$

□

Proof of Lemma 3. Notation from (4) and (6) is used here. For each singular value $\sigma_{M,i}$, it can be formulated as a sequential optimization problem i.e

$$\sigma_{M,i}^2 = \max_Q \|MQ\|_F^2 = \max_Q \sum_{k=1}^{K} \|\tilde{V}_1^\top Q\|_F^2,$$

where Q is a rank 1 projection matrix that is orthogonal to the previous $i - 1$ optima, i.e., $Q_1, \ldots, Q_{i-1}$. The Q that maximizes the Frobenius norm of MQ is denoted as Q_i.

For an arbitrary component in the theoretical joint score subspace $\text{row}(J)$, write its projection matrix as $P_J^{(1)}$. The Frobenius norm of M projected onto $P_J^{(1)}$ is

$$\|MP_J^{(1)}\|_F^2 = \begin{bmatrix} \tilde{V}_1^\top P_J^{(1)} \\ \vdots \\ \tilde{V}_K^\top P_J^{(1)} \end{bmatrix}_F^2 \geq \begin{bmatrix} \cos\theta_1 \\ \vdots \\ \cos\theta_K \end{bmatrix}_F^2 = \sum_{k=1}^{K} \cos^2\theta_k$$

Considering the mechanism of SVD, $\sigma_{M,1}^2$ is the maximal norm obtained from the optimal projection matrix $Q_1 \subseteq \bigcup_{k=1}^{K} \text{row}(\tilde{A}_k) \subseteq \mathbb{R}^n$. If all $\tilde{A}_k$ contain all components obtained by noise perturbation of the common row space $\text{row}(J)$, then we have

$$\sigma_{M,1}^2 \geq \|MP_J^{(1)}\|_F^2 \geq \sum_{k=1}^{K} \cos^2\theta_k.$$

to be considered as a component of the joint score subspace.

This argument can be applied sequentially. For the $Q_2 \in Q_1^{\perp} \cap \{\bigcup_{k=1}^{K} \mathrm{row}(\tilde{A}_k)\}$, there exist a non-empty joint subspace ($\subseteq \mathrm{row}(J)$) such that all $Q_1^{\perp} \cap \mathrm{row}(\tilde{A}_k)$ contain perturbed directions of a joint component other than the one above. Therefore this joint component with projection matrix $P_J^{(2)}$ should have

$$\sigma_{M,2}^2 \geq \|MP_J^{(2)}\|_F^2 \geq \sum_{k=1}^{K} \cos^2 \theta_k.$$

Thus the singular values corresponding to the joint components satisfies (7) and this procedure can continue through at least r_J steps. □

Appendix B. Details of the toy example

Section 1.1 introduces a toy example of two data blocks, X (100 × 100) and Y (10,000 × 100), with patterns corresponding to joint and individual structures. For details see Figure 2.

A naive attempt at integrative analysis can be done by concatenating X and Y on columns and performing a singular value decomposition on this concatenated matrix. Figure B.14 shows the results for three choices of rank. The rank-2 approximation essentially captures the joint variation component and the individual variation component of X, but the Y components are hard to interpret. The bottom 2000 rows show the joint variation but the top half of Y reveals signal from the individual component of X. One might hope that the Y individual components would show up in the rank-3 and rank-4 approximations. However, because the noise in the X matrix is so large, a random noise component from X dominates the Y signal, so the important latter component disappears from this low-rank representation unlike the AJIVE result in Figure 3. In this example, this naive approach completely fails to give a meaningful joint analysis.

Figure B.15 presents the PLS approximations with different numbers of components selected. PLS completely fails to separate joint and individual components. Instead it provides mixtures of the joint, and some of the individual components. Increasing the rank of the PLS approximation only includes more noise.

The Lock et al. [20] method, called JIVE here, is applied to this toy data set. We implemented the JIVE algorithm using the R package `r.jive` [27] without the orthogonality constraint. The `jive` function provides two options for rank selection: using a permutation test and the Bayesian Information Criterion, respectively. However, neither of them segmented joint signal properly. When using the Bayesian Information Criterion approach, no joint signal was identified and the true joint signals were labeled as noise. The permutation test approach gave a reasonable approximation of the total signal variation within each data block as in the left panel of Figure B.16. However, the Lock et al. [20] method gave rank-2 approximations to the joint matrices shown in the middle panel. The approximation consists of the real joint component together with the individual component of X. Consequently, the approximation of the X individual matrix is a zero matrix and a wrong approximation of the Y individual matrix is shown in the top half of the right panel. We speculate that failure to correctly apportion the joint and individual variation is caused by the fact that the individual spaces are correlated, because the permutation test does not handle correlated individual signals very well. We also manually specified the correct joint and individual ranks for X and Y, which results in the correct results.

We finally remark that the Zhou et al. [50] method COBE correctly segments the toy example. However it takes significantly (39 times) longer time than AJIVE to do so.

References

[1] H. Abdi, L.J. Williams, D. Valentin, Multiple factor analysis: Principal component analysis for multitable and multiblock data sets, Wiley Interdisciplinary Reviews: Computational Statistics 5 (2013) 149–179.

[2] Å. Björck, G.H. Golub, Numerical methods for computing angles between linear subspaces, Mathematics of Computation 27 (1973) 579–594.

[3] T.T. Cai, A. Zhang, Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics, arXiv preprint arXiv:1605.00353 .

[4] G. Ciriello, M.L. Gatza, A.H. Beck, M.D. Wilkerson, S.K. Rhie, A. Pastore, H. Zhang, M. McLellan, C. Yau, C. Kandoth, R. Bowlby, H. Shen, S. Hayat, R. Fieldhouse, S.C. Lester, G.M. Tse, R.E. Factor, L.C. Collins, K.H. Allison, Y.Y. Chen, K. Jensen, N.B. Johnson, S. Oesterreich, G.B. Mills, A.D. Cherniack, G. Robertson, C. Benz, C. Sander, P.W. Laird, K.A. Hoadley, T.A. King, TCGA Research Network, C.M. Perou, Comprehensive molecular portraits of invasive lobular breast cancer, Cell 163 (2015) 506–519.

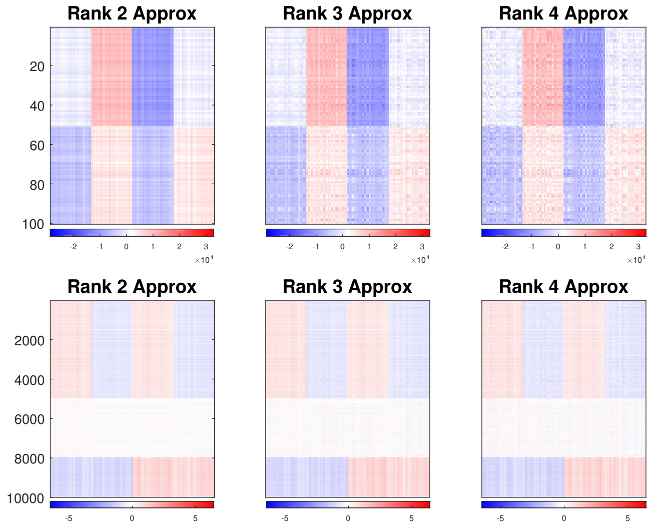

Figure B.14: Shows the concatenation SVD approximation of each block for rank 2 (left), 3 (center) and 4 (right). Although block X has a relatively accurate approximation when the rank is chosen as 2, the individual pattern in block Y has never been captured due to the heterogeneity between X and Y.

[5] B. Draper, M. Kirby, J. Marks, T. Marrinan, C. Peterson, A flag representation for finite collections of subspaces of mixed dimensions, Linear Algebra Appl. 451 (2014) 15–32.

[6] M. Hanafi, A. Kohler, E.-M. Qannari, Connections between multiple co-inertia analysis and consensus principal component analysis, Chemom. Intell. Lab. Syst. 106 (2011) 37–40.

[7] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[8] P. Horst, Relations among m sets of measures, Psychometrika 26 (1961) 129–149.

[9] H. Hotelling, Relations between two sets of variates, Biometrika 28 (1936) 321–377.

[10] S. Jere, J. Dauwels, M.T. Asif, N.M. Vie, A. Cichocki, P. Jaillet, Extracting commuting patterns in railway networks through matrix decompositions, In: Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on, IEEE, pp. 541–546, 2014.

[11] C. Jordan, Essai sur la géométrie à n dimensions, Bull. Soc. Math. France 3 (1875) 103–174.

[12] J.R. Kettenring, Canonical analysis of several sets of variables, Biometrika (1971) 433–451.

[13] S. Kotz, S. Nadarajah, Multivariate t-distributions and Their Applications, Cambridge University Press, 2004.

[14] O. Kühnle, Integration of multiple high-throughput data-types in cancer research, Doctoral dissertation, Ludwig Maximilian Universität München, Germany, 2011.

[15] J. Kuligowski, D. Pérez-Guaita, Á. Sánchez-Illana, Z. León-González, M. de la Guardia, M. Vento, E.F. Lock, G. Quintás, Analysis of

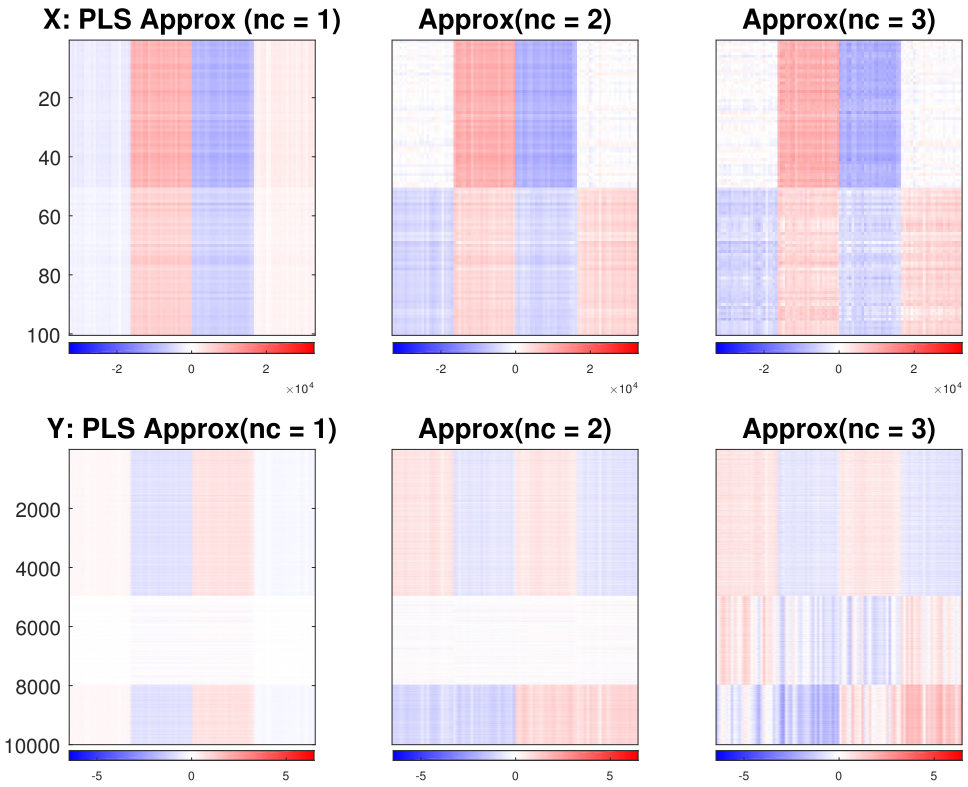

Figure B.15: PLS approximations of each block for numbers of components as 1 (left), 2 (center) and 3 (right). PLS fails to distinguish the joint and individual variation structure.

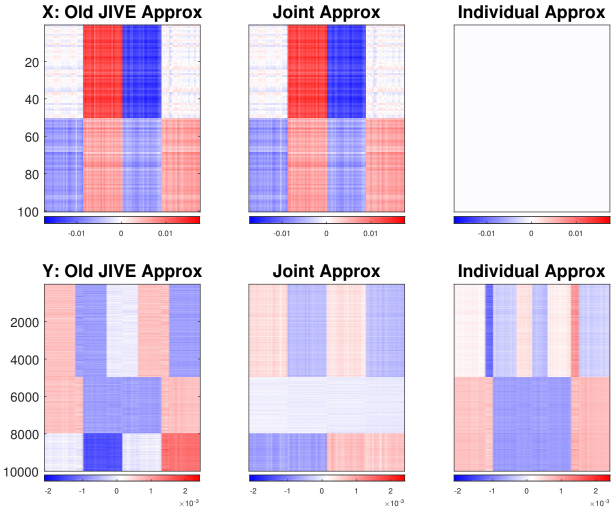

Figure B.16: The Lock et al. [20] JIVE method approximation of the data blocks X and Y in the toy example are shown in the first panel of figures. The joint matrix approximations (middle panel) incorrectly contain the individual component of X because of the failure of the permutation test to correctly select ranks in the presence of correlated individual components.

multi-source metabolomic data using joint and individual variation explained (JIVE), Analyst.

[16] K.-A. Lê Cao, P. G. Martin, C. Robert-Granié, P. Besse, Sparse canonical methods for biological data integration: Application to a cross-platform study, BMC Bioinformatics 10 (2009) 34.

[17] M.H. Lee, Continuum direction vectors in high dimensional low sample size data, Doctoral dissertation, University of North Carolina at Chapel Hill, Chapel Hill, NC, 2007.

[18] S. Lee, High-dimension, low sample size asymptotics of canonical correlation analysis, arXiv preprint arXiv:1609.02992 .

[19] E.F. Lock, D.B. Dunson, Bayesian consensus clustering, Bioinformatics 29 (2013) 2610–2616.

[20] E.F. Lock, K.A. Hoadley, J.S. Marron, A.B. Nobel, Joint and individual variation explained (JIVE) for integrated analysis of multiple data types, Ann. Appl. Statist. 7 (2013) 523–542.

[21] T. Löfstedt, D. Hoffman, J. Trygg, Global, local and unique decompositions in OnPLS for multiblock data analysis, Analytica Chimica Acta 791 (2013) 13–24.

[22] J.S. Marron, A.M. Alonso, Overview of object oriented data analysis, Biometrical J. 56 (2014) 732–753.

[23] J. Miao, A. Ben-Israel, On principal angles between subspaces in $\mathbb{R}^n$, Linear Algebra Appl. 171 (1992) 81–98.

[24] Q. Mo, S. Wang, V.E. Seshan, A.B. Olshen, N. Schultz, C. Sander, R.S. Powers, M. Ladanyi, R. Shen, Pattern discovery and cancer gene identification in integrated cancer genomic data, Proc. Natl. Acad. Sci. USA 110 (2013) 4245–4250.

[25] C.G.A. Network, et al., Comprehensive molecular portraits of human breast tumours, Nature 490 (2012) 61–70.

[26] A.A. Nielsen, Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data, IEEE Trans. Image Process. 11 (2002) 293–305.

[27] M.J. O'Connell, E.F. Lock, R. JIVE for exploration of multi-source molecular data, Bioinformatics 32 (2016) 2877–2879.

[28] S. O'Rourke, V. Vu, K. Wang, Random perturbation of low rank matrices: Improving classical bounds, arXiv preprint arXiv:1311.2657.

[29] E. Parkhomenko, D. Tritchler, J. Beyene, Genome-wide sparse canonical correlation of gene expression with genotypes, in: BMC Proceedings, vol. 1, BioMed Central, S119, 2007.

[30] E. Parkhomenko, D. Tritchler, J. Beyene, et al., Sparse canonical correlation analysis with application to genomic data integration, Stat. Appl. Genet. Mol. Biol. 8 (2009) 1–34.

[31] P. Ray, L. Zheng, J. Lucas, L. Carin, Bayesian joint analysis of heterogeneous genomics data, Bioinformatics 30 (10) (2014) 1370–1376.

[32] D.V. Samarov, The analysis and advanced extensions of canonical correlation analysis, Doctoral dissertation, University of North Carolina at Chapel Hill, Chapel Hill, NC 2009.

[33] M. Schouteden, K. Van Deun, S. Pattyn, I. Van Mechelen, SCA with rotation to distinguish common and distinctive information in linked data, Behav. Res. Methods 45 (2013) 822–833.

[34] M. Schouteden, K. Van Deun, T.F. Wilderjans, I. Van Mechelen, Performing DISCO-SCA to search for distinctive and common information in linked data, Behavior Research Methods 46 (2014) 576–587.

[35] A.K. Smilde, J.A. Westerhuis, S. de Jong, A framework for sequential multiblock component methods, J. Chemometrics 17 (2003) 323–337.

[36] G. Stewart, J.-G. Sun, Matrix Perturbation Theory, Computer Science and Scientific Computing, Academic Press, ISBN 9780126702309, 1990.

[37] J. Trygg, S. Wold, O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter, J. Chemometrics 17 (2003) 53–64.

[38] H.D. Vinod, Canonical ridge and econometrics of joint production, J. Econometrics 4 (1976) 147–166.

[39] S. Waaijenborg, P.V. de Witt Hamer, A.H. Zwinderman, et al., Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis, Stat. Appl. Genet. Mol. Biol. 7 (2008) Article 3.

[40] P.-A. Wedin, Perturbation bounds in connection with singular value decomposition, BIT Numerical Mathematics 12 (1972) 99–111.

[41] S. Wei, C. Lee, L. Wichers, J.S. Marron, Direction-projection-permutation for high-dimensional hypothesis tests, J. Comput. Graphical Stat. 25 (2016) 549–569.

[42] J.A. Westerhuis, T. Kourti, J.F. MacGregor, Analysis of multiblock and hierarchical PCA and PLS models, J. Chemometrics 12 (1998) 301–321.

[43] D.M. Witten, R.J. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (2009) 515–534.

[44] H. Wold, Partial least squares, In: S. Kotz, N.L. Johnson, Eds., Encyclopedia of Statistical Sciences, Vol. 6, Wiley, New York, pp. 581–591.

[45] S. Wold, P. Geladi, K. Esbensen, J. Öhman, Multi-way principal components-and PLS-analysis, J. Chemometrics 1 (1987) 41–56.

[46] S. Wold, N. Kettaneh, K. Tjessem, Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection, J. Chemometrics 10 (1996) 463–482.

[47] Z. Yang, G. Michailidis, A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data, Bioinformatics 32 (2015) 1–8.

[48] Q. Yu, B.B. Risk, K. Zhang, J.S. Marron, JIVE integration of imaging and behavioral data, NeuroImage 152 (2017) 38–49.

[49] Y. Zhang, G. Zhou, J. Jin, X. Wang, A. Cichocki, SSVEP recognition using common feature analysis in brain-computer interface, J. Neuroscience Methods 244 (2015) 8–15.

[50] G. Zhou, A. Cichocki, Y. Zhang, D. Mandic, Group component analysis for multiblock data: Common and individual feature extraction, IEEE Trans. Neural Netw. Learn. Syst. 17 (2016) 2426–2439.