

# JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATATYPES\*

BY ERIC F. LOCK, KATHERINE A. HOADLEY, J.S.  
MARRON AND ANDREW B. NOBEL

Research in a number of fields now requires the analysis of “multi-block” data, in which multiple high-dimensional, and fundamentally disparate, datatypes are available for a common set of objects. In this paper we introduce Joint and Individual Variation Explained (JIVE), a general method for the integrated analysis of multi-block data. The method decomposes a multi-block dataset into a sum of three terms: a low-rank approximation capturing joint variation across datatypes, low-rank approximations for structured variation individual to each datatype, and residual noise. This decomposition can be used to quantify the amount of joint variation between datatypes, visually explore joint and individual structure, and reduce the dimensionality of the data. The proposed method represents an extension of Principal Component Analysis (PCA) and has clear advantages over popular two-block methods such as Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS). We apply JIVE to data from The Cancer Genome Atlas (TCGA), where multiple genomic technologies are available for a common set of Glioblastoma Multiforme tumor samples.

Software is available at <https://genome.unc.edu/jive/>.

**1. Introduction.** Many fields of scientific research now involve the analysis of high-dimensional data, in which a large number of variables are measured for a given set of experimental objects. In particular, there is an increasing prevalence of *multi-block* data, in which multiple fundamentally disparate high-dimensional datasets are available for a common set of objects. In this paper we introduce JIVE, a general method for the integrated analysis of multi-block data.

While the JIVE method is broadly applicable, we focus on the analysis of biological data. In biomedical studies, a number of existing technologies may be used to collect diverse information on an organism or tissue sample. The amount of available biological data from multiple platforms and technologies is expanding rapidly. The 2011 Online Database collection of

---

\*This work was partially supported by NIH grant R01 MH090936-01, NSF grant DMS-0907177, NSF grant DMS-0854908 and NIH grant U24-CA143848.

*Keywords and phrases:* Vertical integration, Multi-block data, Principal Component Analysis

*Nucleic Acids Research* lists 1330 publicly available databases that measure different aspects of molecular and cell biology [Galberin and Cochrane, 2011]. Large online databases such as ArrayExpress [Parkinson et al., 2009] and the UCSC Genome-browser [Rhead et al., 2010] often contain multiple disparate datatypes collected from a common set of samples. Large-scale projects like The Human Connectome Project [Sporns et al., 2005] and The Cancer Genome Atlas [TCGA Research Network, 2008] focus on the integrated analysis of multi-block data.

Well established multivariate methods can be used to separately analyze different datatypes measured on the same set of objects. However, individual analysis of each datatype will not capture associations and potential causal relationships between datatypes. Furthermore, each datatype can impart unique and useful information. There is a need for new methods that explore associations between multiple datatypes and combine data from multiple disparate sources when making inference about the objects. This motivates an exciting new area of statistical research.

The JIVE method decomposes multi-block data into a sum of three components: a low rank approximation capturing joint structure between datatypes, low rank approximations capturing structure individual to each datatype, and residual noise. Analysis of individual structure provides a way to identify potentially useful information that exists in one datatype, but not others. Accounting for individual structure also allows for more accurate estimation of what is common between datatypes. JIVE can identify joint structure not found by existing methods, which are described in Section 1.2. Furthermore, JIVE is robust to the dimensionality of the data, applicable to more than two datatypes, and has a simple algebraic interpretation.

In Section 1.1 we formally introduce multi-block data, and in Section 1.2 we describe related existing methods for the integrated analysis of multiple datatypes. In Section 2 we give a detailed description of the JIVE method. In Section 3, we motivate and validate the JIVE method through a variety of simulated examples.

Section 4 describes an application of JIVE to multi-block data on Glioblastoma Multiforme (GBM) tumor samples from TCGA, an ongoing collaborative effort funded by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). A goal of TCGA is to molecularly characterize cancer through analysis and integration of multidimensional large scale genomic data [TCGA Research Network, 2008]. Information from disparate genomic datatypes can be integrated for a more comprehensive understanding of cancer genetics and cell biology. In addition, we would like to find distinguishing characteristics between tumor samples,

either across multiple datatypes or unique to a single datatype, that may be used to identify targeted therapies.

1.1. *Multi-block data.* Formally, a multi-block dataset consists of matrices  $X_1, X_2, \dots, X_k$  with  $k \geq 2$ . Each matrix has  $n$  columns, corresponding to a common set of  $n$  objects. The  $i$ th matrix  $X_i$  has  $p_i$  rows, each corresponding to a variable in a given measurement technology that varies from matrix to matrix. To give an example, for a given set of  $n$  biological samples, the rows of  $X_1$  might contain gene expression measurements (of dimension  $p_1$ ), the rows of  $X_2$  might contain genotype information, and the rows of  $X_3$  might contain the concentration of different metabolites. The data matrices in a multi-block data set may be unified vertically into a single data matrix

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} : p \times n,$$

where  $p = p_1 + p_2 + \dots + p_k$ .

Direct analysis of  $X$  can be problematic as the size and scale of the constituent datatypes are often significantly different. To remove baseline differences between datatypes, it is helpful to row-center the data by subtracting the mean within each row. Datatypes may also be of different dimension ( $p_i$ ) or differ in variability. To circumvent cases where “the largest dataset wins”, we scale each datatype by its total variation, or sum-of-squares. In particular, for each  $i$  define  $X_i^{\text{scaled}} = \frac{X_i}{\|X_i\|}$ , where  $\|\cdot\|$  defines the *Frobenious* norm  $\|A\|^2 = \sum_{i,j} a_{ij}^2$ . Then,  $\|X_i^{\text{scaled}}\| = 1$  for each  $i$ , and each datatype contributes equally to the total variation of the aggregated matrix

$$(1.1) \quad X^{\text{scaled}} = \begin{bmatrix} X_1^{\text{scaled}} \\ \vdots \\ X_k^{\text{scaled}} \end{bmatrix}.$$

1.2. *Existing methods.* One approach to the analysis of multi-block data is to mine the data for variable-by-variable associations. In computational biology, genome wide association and expression quantitative trait loci studies [Gilad et al., 2008] can identify millions of pairwise variable associations between genomic datatypes. Furthermore, network models can link associated variables across and within datatypes (see Adourian et al. [2008]). However, analysis of variable-by-variable associations alone does not identify the global modes of variation that drive associations across and within datatypes, which is the focus of this paper.

Principal Component Analysis (PCA) of the block-scaled matrix  $X^{\text{scaled}}$  coincides with *Consensus PCA* [Westerhuis et al., 1998, Wold et al., 1996]. This direct approach using the aggregated data matrix is also utilized by the iCluster method [Shen et al., 2009]. Designed to cluster samples based on information from multiple genomic datatypes, iCluster performs clustering based on a factor analysis of the aggregated matrix  $X$ . While these methods synthesize information from multiple datatypes, they do not distinguish between joint or individual effects.

Canonical Correlation Analysis (CCA) [Hotelling, 1936] is a popular method to globally examine the relationship between two sets of variables. If  $X_1$  and  $X_2$  are two datasets on a common set of samples, the first pair of canonical loadings (variable weights)  $u_1$  and  $u_2$  are unit vectors maximizing  $\text{Corr}(u_1^T X_1, u_2^T X_2)$ .

Geometrically,  $u_1$  and  $u_2$  can be interpreted as the pair of directions that maximize the correlation between  $X_1$  and  $X_2$ . Sample projections on the canonical loadings,  $u_1^T X_1$  and  $u_2^T X_2$ , give the canonical *scores* for  $X_1$  and  $X_2$ . Subsequent CCA directions can be found by enforcing orthogonality with previous directions. For datasets with  $p_1 > n$  or  $p_2 > n$  the CCA directions are not well defined, and over-fitting is often a problem even when  $p_1, p_2 < n$ . Hence, standard CCA is typically not applicable to high-dimensional data.

*Partial Least Squares* (PLS) [Wold, 1985] directions are defined similarly to CCA, but maximize covariance rather than correlation. PLS is appropriate for high-dimensional data. However, Trygg and Wold [2003] examine how structured variation in  $X_1$  not associated with  $X_2$  (and vice-versa) can drastically alter PLS scores, making the interpretation of such scores problematic. Their solution, called O2-PLS, seeks to remove structured variation in  $X_1$  not linearly correlated with  $X_2$  (and vice versa) from the PLS components. As such, O2-PLS components are often more representative of the true joint structure between two datatypes. However, the restriction of O2-PLS (and PLS) to pairwise comparisons limits their utility in finding common structure among more than two datatypes.

Witten and Tibshirani [2009] recently introduced *Multiple Canonical Correlation Analysis* (mCCA) to explore associations and common structure on two or more datasets. For  $X_1, X_2, \dots, X_k$  as in Section 1.1, each row centered and row standardized, the standard mCCA loading vectors  $u_1, u_2, \dots, u_k$  satisfy

$$\operatorname{argmax}_{\|u_1\|=\dots=\|u_k\|=1} \sum_{i<j} u_i^T X_i X_j^T u_j = \sum_{i<j} \text{Cov}(u_i^T X_i, u_j^T X_j).$$

As such, mCCA can be viewed as a natural extension of PLS to more than

two datatypes.

Di et al. [2009] develop multi-level functional PCA (MF-PCA) for the analysis of variation between and within grouped samples of functional data. Similar in spirit to JIVE, MF-PCA yields a sum of two PCA decompositions: one for variability between groups and one for variability within groups. We stress that JIVE is designed for analysis across disparate datatypes, while MF-PCA analyzes grouped observations on the same functional datatype. Furthermore, the global component in JIVE models similarities across datatypes, while the global component in MF-PCA models differences between sample groups.

**2. JIVE.** The JIVE method decomposes multi-block data into three terms: joint structure between datatypes, structure individual to each datatype, and residual noise. As in O2-PLS, accounting for individual structure can lead to better estimation of what is joint between multiple datatypes, and vice-versa. In addition, JIVE is robust to the dimensionality of the data (including  $n > p$  and  $p > n$ ), has a simple algebraic interpretation, and can be applied to more than two datatypes.

**2.1. Model.** Let  $X_1, X_2, \dots, X_k$  be datatypes as in Section 1.1. Variation that is consistent across datatypes in the aggregated matrix  $X$  is represented by a single  $p \times n$  matrix of rank  $r < \text{rank}(X)$ . This matrix represents the *joint structure* of  $X$ . For each  $X_i$ , structured variation in  $X_i$  unrelated to the other datatypes is represented by a  $p_i \times n$  matrix of rank  $r_i < \text{rank}(X_i)$ . These matrices represent the *individual structure* of each  $X_i$ . The sum of joint and individual structure gives a low-rank decomposition approximating the joint data matrix  $X$ . The general model for two datatypes  $X_1$  and  $X_2$  is shown in Figure 1.

More formally, let  $A_i$  be the matrix representing the individual structure of  $X_i$ , and let  $J_i$  be the submatrix of the joint structure matrix that is associated with  $X_i$ . Then, the unified JIVE model is

$$(2.1) \quad \begin{aligned} X_1 &= J_1 + A_1 + \epsilon_1 \\ &\vdots \\ X_k &= J_k + A_k + \epsilon_k, \end{aligned}$$

where  $\epsilon_i$  are  $p_i \times n$  error matrices of independent entries with  $\mathbb{E}(\epsilon_i) = 0_{p_i \times n}$ . Let

$$J = \begin{bmatrix} J_1 \\ \vdots \\ J_k \end{bmatrix},$$

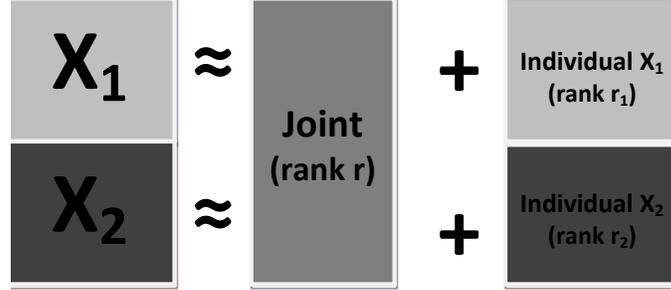


FIG 1. *Depiction of the JIVE decomposition for two datasets. The data are approximated by a low rank matrix of joint variation and low rank matrices giving structured variation unique to  $X_1$  and  $X_2$ .*

denote the joint structure matrix. The model imposes the rank constraints  $\text{rank}(J) = r$  and  $\text{rank}(A_i) = r_i$  for  $i = 1, \dots, k$ . Furthermore, we assume that the rows of joint and individual structure are orthogonal:  $JA_i^T = 0_{p \times p_i}$  for  $i = 1, \dots, k$ . Intuitively, this means that sample patterns related to joint variation between datatypes are unrelated to sample patterns responsible for structure in only one datatype. This assumption does not constrain the model, in that any set of datatypes in the form (2.1) can be written equivalently with orthogonality between joint and individual structure. Furthermore, the orthogonality constraint assures that the joint and individual components in the structure of the aggregated matrix  $X$  are uniquely determined. See Appendix A.1 for more details.

**2.2. Estimation.** Here we discuss estimation of joint and individual structure for fixed ranks  $r, r_1, \dots, r_k$ . Section 2.4 discusses the choice of ranks. Joint and individual structure are estimated by minimizing the sum of squared error. Let  $R$  be the  $p \times n$  matrix of residual noise after accounting for joint and individual structure:

$$R = \begin{bmatrix} R_1 \\ \vdots \\ R_k \end{bmatrix} = \begin{bmatrix} X_1 - J_1 - A_1 \\ \vdots \\ X_k - J_k - A_k \end{bmatrix}.$$

We estimate the matrices  $J$  and  $A_1, \dots, A_k$  by minimizing the sum of squared residuals  $\|R\|^2$  under the given ranks. This is accomplished by iteratively estimating joint and individual structure:

- For fixed  $J$ , find  $A_1, \dots, A_k$  to minimize  $\|R\|$ .
- For fixed  $A_1, \dots, A_k$ , find  $J$  to minimize  $\|R\|$ .

- Repeat until convergence.

The joint structure  $J$  minimizing  $\|R\|$  for fixed individual structure is equal to the first  $r$  terms in the singular value decomposition (SVD) of  $X$  with individual structure removed. For fixed joint structure, the estimated individual structure for  $X_i$  is equal to the first  $r_i$  terms of the SVD of  $X_i$  with the joint structure removed. The estimate of individual structure for  $X_i$  will not change those for  $X_j$ ,  $j \neq i$ , and hence the  $k$  individual approximations minimize  $\|R\|$  for fixed joint structure. Pseudocode for the iterative algorithm is given in Appendix A.2. We note that the iterative method is monotone in the sense that  $\|R\|$  decreases at each step. Thus  $\|R\|$  converges to a coordinate-wise minimum, that can't be improved by changing the estimated joint or individual, structure. Further convergence properties of the algorithm are currently under study.

2.3. *Relationship to PCA.* For a row-centered  $p \times n$  matrix  $X$ , the first  $r$  principal components yield the rank  $r$  approximation

$$X \approx US,$$

where  $S(r \times n)$  contains the sample scores and  $U(p_i \times r)$  contains the variable loadings for the first  $r$  components.

As in PCA, the rank  $r$  joint structure matrix  $J$  in the JIVE model can be written as  $US$ , where  $U$  is a  $p \times r$  loading matrix and  $S$  is an  $r \times n$  score matrix. Let

$$U = \begin{bmatrix} U_1 \\ \vdots \\ U_k \end{bmatrix}$$

where  $U_i$  gives the loadings of the joint structure corresponding to the rows of  $X_i$ . The rank  $r_i$  individual structure matrix  $A_i$  for  $X_i$  can be written as  $W_i S_i$ , where  $W_i$  is a  $p_i \times r_i$  loading matrix and  $S_i$  is an  $r_i \times n$  score matrix. Then, the low rank decomposition of  $X_i$  into joint and individual structure is given by  $X_i \approx U_i S + W_i S_i$ . This gives the unified model

$$(2.2) \quad \begin{aligned} X_1 &= U_1 S + W_1 S_1 + R_1 \\ &\vdots \\ X_k &= U_k S + W_k S_k + R_k. \end{aligned}$$

Joint structure is represented by the common score matrix  $S$ . These scores elicit patterns in the samples that explain variability across multiple datatypes. The loading matrices  $U_i$  indicate how these joint scores are expressed in the rows (variables) of datatype  $i$ . The score matrices  $S_i$  elicit sample patterns individual to datatype  $i$ , with variable loadings  $W_i$ .

2.4. *Rank Selection.* Section 2.2 describes the estimation of joint and individual structure in the JIVE model for a given set of ranks  $r, r_1, \dots, r_k$ . The choice of these ranks is important to accurately quantify the amount of joint variation and individual structure among datatypes. Furthermore, over or underestimation of what is joint can negatively effect estimation of what is individual, and vice-versa. Rank selection is integral to the estimation of joint and individual structure in JIVE. Indeed, rank constraints or another form of penalization is essential to simultaneously identify joint and individual variation.

Here, we describe a two-stage permutation testing approach to rank selection. First, the total underlying rank of structured (joint and individual) variation is estimated for each  $X_i$ . We refer to this as the *effective rank*, and denote it by  $\text{rank-eff}(X_i)$ . The joint structure rank  $r$ , and individual structure ranks  $r_i$  are then estimated under the restriction  $r + r_i = \text{rank-eff}(X_i)$  for  $i = 1, \dots, k$ .

To estimate  $\text{rank-eff}(X_i)$  we use a permutation-based scheme that relies on the singular values of  $X_i$ , which is described in [Peres-Neto et al. \[2005\]](#). In this procedure, the first singular value of  $X_i$  is compared to the first singular values from several random permutations of  $X_i$ . In each permutation, columns are permuted independently within each row to maintain the distribution of each variable, while effectively removing between-variable associations. If only a small proportion (e.g.  $\alpha = 0.01$ ) of the largest singular values under permutation are greater than the observed value, the observed value is deemed significant, and the associated rank one matrix is subtracted from  $X_i$ . This process is repeated until significance is no longer achieved;  $\text{rank-eff}(X_i)$  is defined to be the number of significant singular values obtained in this way.

Given  $\text{rank-eff}(X_i)$  for each  $i$ , a similar permutation based test is used to determine the ranks  $r$  and  $r_1, \dots, r_k$ . For estimated joint structure  $J$  of rank  $r$  and individual structures  $A_i$  of rank  $r_i$  (where  $r + r_i = \text{rank-eff}(X_i)$ ), we test for remaining joint structure in the residuals  $X - J$ . This test is based on permuting columns within each datatype (across all rows), which maintains the multivariate distribution of each datatype while effectively removing between-datatype associations.  $\text{Rank}(J) = r$  is increased until the estimated individual structure and residual noise  $X - J$  has no significant joint structure. Detailed pseudocode for this permutation-based approach is provided in [Appendix A.3](#). Alternative approaches to rank selection are currently under study.

2.5. *Variable Sparsity.* In many practical applications, important structure between samples or objects is only present on a small subset of the measured variables. This motivates use of sparse methods, in which only a subset of variables contribute to a fitted model. Sparse versions of exploratory methods such as PCA [Shen and Huang, 2008], PLS [Le Cao et al., 2008] and CCA [Parkhomenko et al., 2009] already exist.

Here, we describe the use of a penalty term to induce variable sparsity in the JIVE decomposition. Sparsity is accomplished if some of the variable loadings for joint and individual structure ( $U$  and  $W_i$  in Section 2.3) are exactly 0. For weights  $\lambda$  and  $\lambda_i$ , we minimize the penalized sum of squares

$$\|R\|^2 + \lambda \text{Pen}(U) + \sum_i \lambda_i \text{Pen}(W_i),$$

where  $\text{Pen}$  is a penalty designed to induce sparsity in the loading vectors. In our implementation,  $\text{Pen}(\cdot)$  is an  $L1$  penalty analogous to Lasso regression [Tibshirani, 1996], namely

$$\text{Pen}(A) = \sum_{i,j} |a_{ij}|.$$

Under this penalty, loadings of variables with a small or insignificant contribution tend to shrink to 0. Other sparsity-inducing penalties (e.g. hard thresholding) may be substituted for  $L1$  penalization.

Estimation under sparsity penalization is accomplished by an iterative procedure analogous to that used for the non-sparse case:

- For fixed  $J$ , find  $A_i$  to minimize  $\|R_i\|^2 + \lambda_i \text{Pen}(W_i)$  for each  $i = 1, \dots, k$ .
- For fixed  $A_1, \dots, A_k$ , find  $J$  to minimize  $\|R\|^2 + \lambda \text{Pen}(U)$
- Repeat until convergence.

At each iteration, the sparsity penalty is incorporated through the use of a sparse singular value decomposition (SSVD), adapted from Lee et al. [2010]. The weights  $\lambda$ ,  $\lambda_i$  may be pre-specified or estimated via the *Bayesian Information Criterion* (BIC) [Schwarz, 1978] at each iteration.

Inducing sparsity in the joint structure effectively identifies subsets of variables within each datatype that are associated. Examination of the joint sample scores, in turn, reveals sample patterns that drive these associations. Section 4.3 illustrates the use of sparsity in the interpretation of associations across disparate genomic datatypes.

2.6. *Dimension Reducing Shortcut.* For high-dimensional data, where  $p_i > n$ , computing time can be improved by reducing the dimensionality of  $X_1, \dots, X_k$  at the outset. Rather than working with all variables, we consider a dimension-reducing transformation of the original data:  $X_i \rightarrow X_i^\perp$  where  $X_i^\perp$  is an  $n \times n$  matrix derived from the SVD of  $X_i$ . In particular, if

$$SVD(X_i) = U_i \Lambda_i V_i',$$

then  $X_i^\perp = \Lambda_i V_i'$ . Covariance and Euclidian distance between columns (samples) of  $X_i$  are preserved in  $X_i^\perp$ . Applying the JIVE algorithm to the transformed datasets  $X_1^\perp, \dots, X_k^\perp$  can be substantially faster for high-dimensional data. Estimated joint ( $J_i^\perp$ ) and individual ( $A_i^\perp$ ) structure for  $X_i^\perp$  can then be transformed back to the original variable space through the left singular vectors  $U_i$ :  $J_i = U_i J_i^\perp$  and  $A_i = U_i A_i^\perp$ . Applying the iterative estimation method to  $X$  directly or estimating joint and individual structure for  $X_i^\perp$  and mapping back to the original variable space will yield identical results. Hence, high-dimensional data is always transformed via SVD before estimation of joint and individual structure.

**3. Illustrative Examples and Simulated Data.** Here we motivate and demonstrate the use of JIVE through three examples. In Section 3.1, a simple example illustrates how JIVE can identify structure not found by existing methods such as Consensus PCA, PLS or CCA. In Section 3.2, we add an artificial signal to real data from three genomic platforms in order to illustrate the use of JIVE to identify joint signal among multiple datatypes with complex structure. In Section 3.3, the iterative estimation method is applied to hundreds of diverse datasets from the model (2.1), in order to demonstrate its robustness.

3.1. *Illustrative Example.* As a basic illustration we generate two matrices,  $X$  and  $Y$ , with simple patterns corresponding to joint and individual structure. The simulated data is depicted in Figure 2. Both  $X$  and  $Y$  are of dimension  $50 \times 100$ , i.e., each has 50 variables measured for the same 100 objects. A common pattern  $V$  of 100 independent standard normal variables is added to half of the rows in  $X$  and half of the rows in  $Y$ . This represents the joint structure between the two datasets. Structure individual to  $X$  is generated by partitioning the objects into five groups, each of size twenty. Those columns corresponding to group 1, 2, 3, 4, or 5 have -2, -1, 0, 1, 2 added to each row of  $X$ , respectively. Structure individual to  $Y$  is generated similarly, but the groups are defined independently of the groups in  $X$ . Finally, independent  $N(0,1)$  noise is added to both  $X$  and  $Y$ . Note that the important joint structure is not visually apparent.

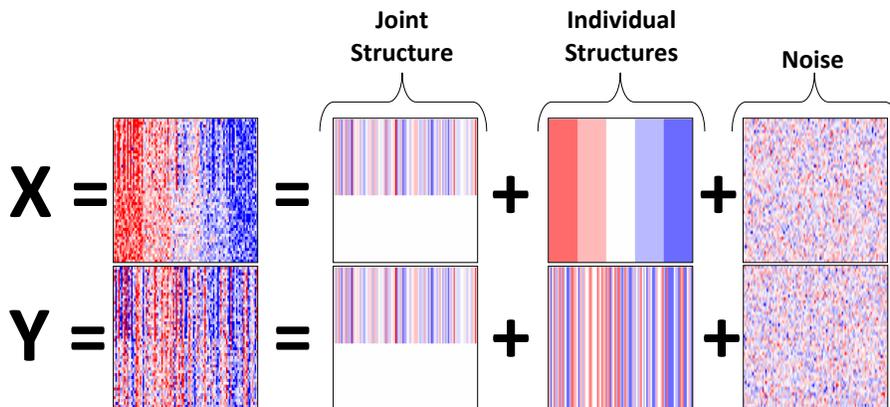


FIG 2.  $X$  and  $Y$  are generated by adding together joint structure, individual structure, and noise. Blue corresponds to negative values, red positive values.

The common pattern  $V$  represents an underlying phenomenon that contributes to several variables in both  $X$  and  $Y$ . Practically, the individual structure in  $X$  (or  $Y$ ) may correspond to an experimental grouping of the measured variables in  $X$  ( $Y$ ) not present in  $Y$  ( $X$ ), e.g., *batch* effects in microarray data. Our goal is to identify both the common underlying phenomenon and individual group effects.

Consensus PCA of the aggregated matrix  $[X' Y']'$  does a poor job of finding the joint structure. The scatterplot in Figure 3 shows a weak association between the first principal component scores and the joint response  $V$ . This is because PCA of the aggregated data is driven by all variation in the data, joint or individual.

Figure 4 shows an analysis of PLS and CCA for  $X$  and  $Y$ . The scores for the first pair of PLS directions show a weak association between  $X$  and  $Y$  (panel C). Furthermore, the PLS scores are not strongly related to the joint response  $V$  (panels A and B). Scores for  $X$  and  $Y$  show a stronger association with  $V$  within classes, indicating how for PLS individual structure can interfere with the estimation of joint structure. The first pair of CCA scores are very highly correlated (panel F), but show nearly no association with the joint or individual structure (panels D and E). This illustrates the tendency of CCA to overfit on high-dimensional data.

Next we consider the JIVE analysis of  $X$  and  $Y$ . The permutation testing approach described in Section 2.4 suggests a rank one approximation for joint structure and rank one approximations of individual structure for both  $X$  and  $Y$ . Scores and loadings for the joint component and both individual

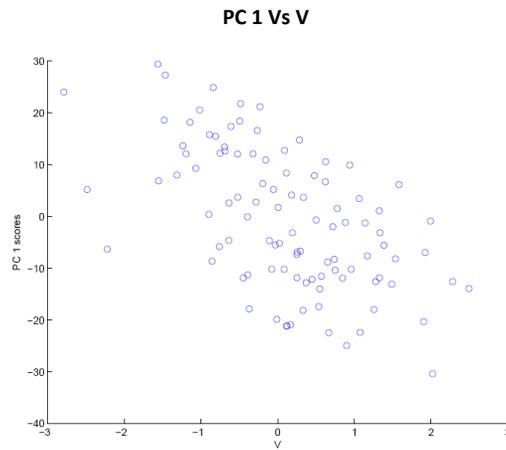


FIG 3. Scatterplot of the first consensus principal component scores Vs the joint signal  $V$ . The scores are weakly associated with the joint signal.

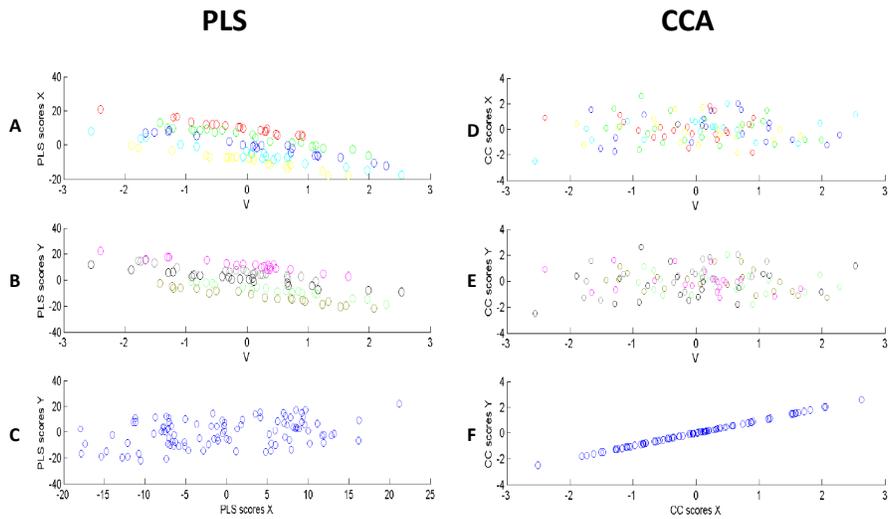


FIG 4. The first pair of PLS, and CCA, directions for  $X$  and  $Y$ . Panels (A) and (B) show a weak association between the PLS scores and the joint signal  $V$ . Points are colored by simulated class in both  $X$  and  $Y$ , and are more highly associated with  $V$  within each class. The first pair of CCA directions correlate with each other (F), but not the common signal  $V$  (D and E). This illustrates the tendency of CCA to overfit.

components are shown in Figure 5. JIVE is able to find the true joint signal between the two datasets, as joint scores are closely associated with the common response  $V$  (panel A). Furthermore, individual scores do a good job of distinguishing classes specific to  $X$  and  $Y$  (panels D and F). The joint signal was added to only the first 25 variables in  $X$  and  $Y$ , which is reflected in the joint loadings (panels B and C). The individual classes were defined on all 50 variables for both  $X$  and  $Y$ , which is reflected in the individual loadings (panels E and F). Note that joint and individual loadings are not constrained to be orthogonal, which gives the analysis more flexibility.

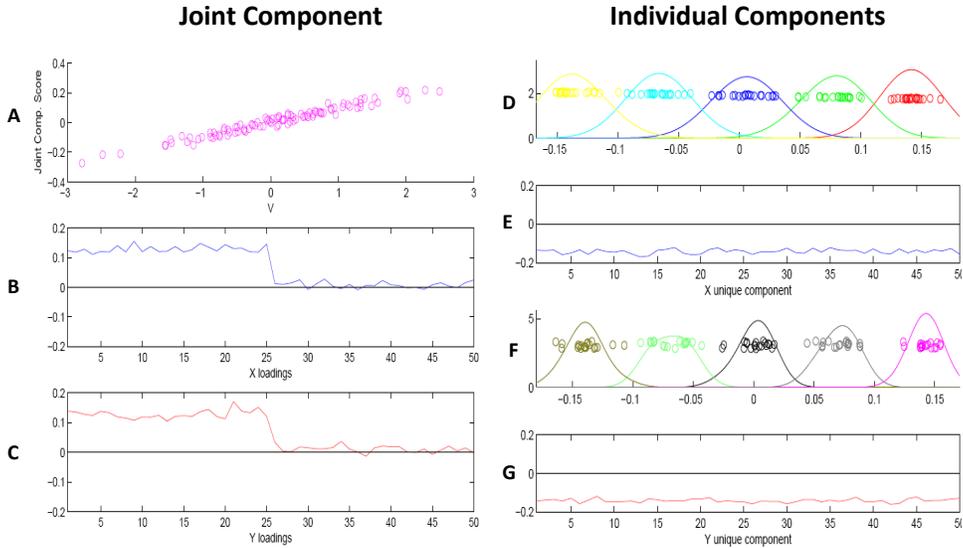


FIG 5. Scores and loadings for joint and individual components in the JIVE decomposition. Joint scores are highly associated with the common signal  $V$  (panel A). Individual scores distinguish classes specific to  $X$  and  $Y$  (D and F). Joint loadings (B and C) show a strong effect (difference from zero) on half of the variables in  $X$  and  $Y$ . Individual loadings (E and G) show a similar effect on all variables in  $X$  and  $Y$ .

Figure 6 shows the resulting low rank approximations for joint structure, individual structure and residual noise obtained by JIVE. Estimates closely resemble the true signal in Figure 2.

3.2. *Example Based on Genomic Data.* We now use real genomic data to illustrate the extraction of a known joint signal from datatypes with realistic individual variation. We begin with gene expression (GE), copy number

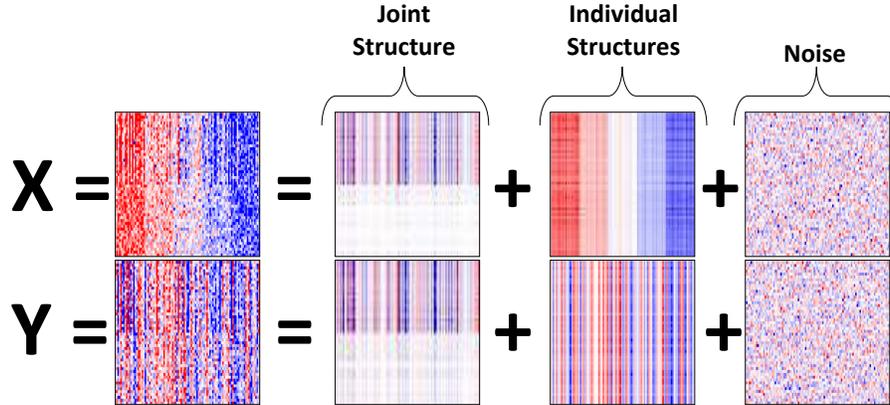


FIG 6. *JIVE* estimates for joint structure, individual structure, and noise. *Blue* corresponds to negative values, *red* positive values.

variation (CN), and microRNA (miRNA) data available for the same set of 234 GBM tumor tissue samples from TCGA [TCGA Research Network, 2008]. The GE and CN data are both of dimension 14556 (expression intensity and average copy number variation for the same 14556 genes), the miRNA data is of dimension 535 (535 miRNA intensities measured). The three datasets are normalized as in Section 1.1.

These three genomic datasets are used to simulate data with realistic background variation (see Figure 7). First, samples (columns) are randomly permuted separately within each datatype. This effectively removes joint structure between datatypes, but maintains the structure within each datatype. An artificial joint signal is then added to 5% of the rows in each of the three datatypes. This joint signal is generated by adding a constant equal to the row standard deviation to the first half of the samples, and subtracting the same constant from the other half. This yields two sample clusters that are common to each of the three datatypes. We would like to identify this common signal among the variation individual to each datatype.

The first Consensus PC of the three datasets is shown in Figure 8. It shows a slight association between the two joint clusters, but does not clearly separate them. In this case, the joint signal does not dominate the variation in the data, and is hard to extract without accounting for individual structure. Application of PLS and CCA is complicated by the fact that these methods are designed for two datatypes. Moreover, the pairwise applications of these methods is also negatively affected by individual structure and over-fitting.

A *JIVE* analysis of the data successfully distinguishes joint and individ-

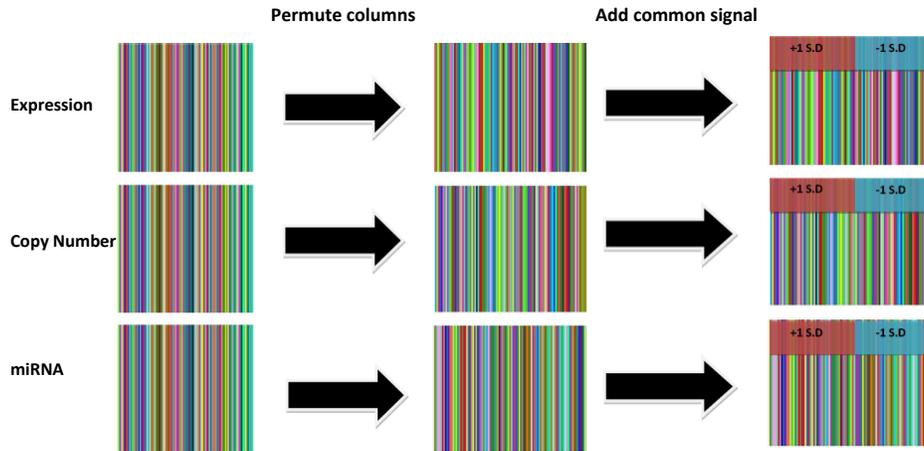


FIG 7. Simulation based on miRNA, GE and CN data. Columns are permuted within each datatype so that samples are not associated. A common signal is then added to 5% of the rows in each datatypes.

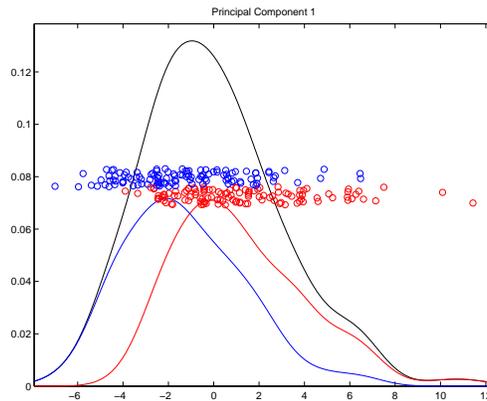


FIG 8. Scores for the first Consensus PC of the aggregated data, colored by cluster. The two clusters (representing common signal) are not well distinguished.

ual structure. The permutation approach described in Section 2.4 suggests estimating rank 1 joint structure, and rank 39, 35, and 13 individual structure for GE, CN and miRNA, respectively. Sample scores and GE, CN and miRNA loadings for the joint component are shown in Figure 9. Joint scores now clearly separate the two artificial clusters. Furthermore, loadings clearly

indicate which rows contain the joint signal in each of the three datatypes.

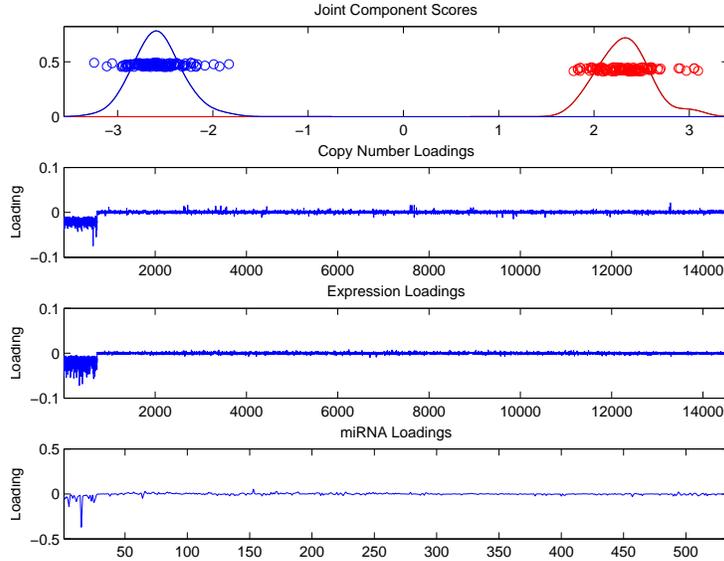


FIG 9. Joint component scores and loadings. Scores are colored by artificial cluster and show good separability. Loadings for CN, GE, and miRNA are ordered so that the first 5% have joint signal added, and these show the strongest contribution (difference from 0).

3.3. *Extensive simulations from model.* To test the robustness of the iterative estimation method, we apply it to a diverse set of 200 randomly generated models. In our simulations, two data matrices  $X_1$  and  $X_2$  are generated with varying sample size and dimensions. Joint and individual structure are generated from different probability distributions, and have varying ranks. These randomly generated models are tested both with and without noise.

We first test models with joint and individual structure, and no additional noise. The sample size  $n$  and dimensions  $p_1$  and  $p_2$  for  $X_1$  and  $X_2$  are drawn at random uniformly from  $\{10, 11, \dots, 100\}$ . The rank of joint structure,  $r$ , and individual structure,  $r_1$  and  $r_2$ , are each drawn from  $\{0, 1, \dots, 4\}$ . Low-rank structure is then generated in factorized form, as in (2.2):

$$\begin{aligned} X_1 &= U_1 S + W_1 S_1 \\ X_2 &= U_2 S + W_2 S_2. \end{aligned}$$

For each realization of  $U_1, U_2, S, W_1, W_2, S_1$ , and  $S_2$ , the entries are generated from a random choice among the distributions  $N(0, 1)$ ,  $\text{Uniform}(0, 1)$  and  $\text{Bernoulli}(\frac{1}{2})$ .

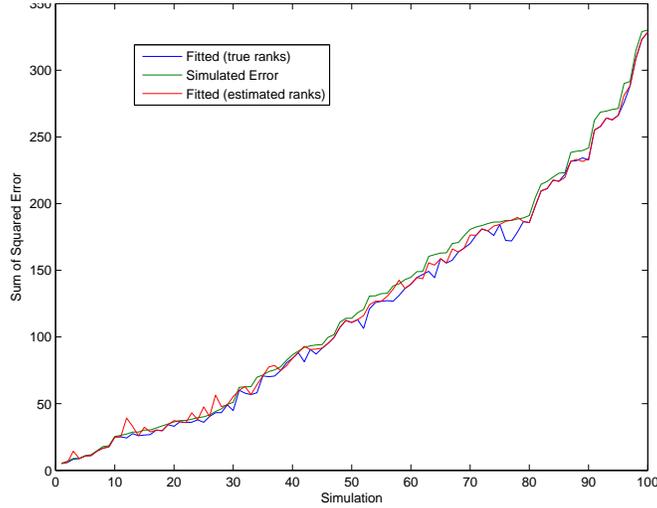


FIG 10. Sum of squared residuals in the simulated model (green), the fitted model with true ranks (blue) and the fitted model under permutation testing (red) in 100 randomly generated simulations.

We applied the iterative estimation method to  $X_1$  and  $X_2$  for 100 simulated examples, generated as described above. In all cases, the fitted model with the true ranks  $r$ ,  $r_1$  and  $r_2$  converged to the simulated data  $X_1$  and  $X_2$  ( $\|X - \hat{X}\|^2 < 10^{-12}$ ). This is evidence that, in the absence of noise, low rank joint and individual structure can be recovered exactly.

We then included error in our simulations, using the model

$$\begin{aligned} X_1 &= U_1 S + W_1 S_1 + E_1 \\ X_2 &= U_2 S + W_2 S_2 + E_2, \end{aligned}$$

where  $U_1, U_2, S, W_1, W_2, S_1, S_2$  are generated as above, and  $E_1, E_2$  are error matrices with independent entries from  $N(0, \sigma^2)$ . The standard deviation,  $\sigma$ , of the noise is randomly determined from a Uniform(0, 2) distribution. In 100 randomly generated examples, joint and individual structure are estimated given the simulated ranks  $r, r_1$  and  $r_2$ , and also with ranks estimated by the approach described in Section 2.4. We expect the sum of squared error  $\|E_1\|^2 + \|E_2\|^2$  to be close to the sum of squared residuals after estimating joint and individual structure,  $\|R_1\|^2 + \|R_2\|^2$ .

Figure 10 plots the sum of squared errors in the simulated model and the sum of squared residuals from the fitted model, both with and without estimated ranks, for the 100 examples. When the simulated ranks are used,

the sum of squared residuals for the fitted model is always smaller than the sum of squared error in the simulation. This is evidence that the iterative approach is successful in minimizing the sum of squared residuals. Estimated ranks for joint and individual structure agree completely with the simulated ranks in 60% of simulations. The permutation testing approach tends to underestimate the simulated ranks, if noise overwhelms the low rank signal.

**4. Application to Genomic Data.** Here we exhibit how JIVE can be used to explore global associations across multiple genomic datatypes. We examine a set of 234 GBM tumor samples. For each tumor sample, 535 miRNA intensities and 24,350 gene expression intensities are available. These data are publicly available from The Cancer Genome Atlas (TCGA) [TCGA Research Network, 2008]. The pre-processed data used for this analysis is available at <https://genome.unc.edu/jive>. Verhaak et al. [2010] classified these GBM samples into four gene expression subtypes: Neural, Mesenchymal, Proneural and Classical. These subtypes have distinct expression characteristics, copy number alterations, and gene mutations. In addition, there were clinical differences across subtypes in response to aggressive therapy. The classification of GBMs into these subtypes may be exploited for more targeted therapy.

Copy number aberrations and somatic mutations, and their relationship with gene expression, have been recognized as important aspects of GBM biology (see, e.g., Bredel et al. [2009] and TCGA Research Network [2008]). However, the role of miRNA in GBM biology has not been well studied. We applied JIVE to the miRNA and gene expression data in order to identify joint and individual variation among the two datatypes, and we further investigated how this variation is related to the GBM subtypes.

4.1. *Quantifying joint and individual variation.* Permutation testing (Section 2.4) was used to determine the ranks of estimated joint and individual structure. The test (using  $\alpha = 0.01$ , and 1000 permutations) identified

- rank 5 joint structure
- rank 7 structure individual to miRNA
- rank 27 structure individual to gene expression.

The percentage of variation (sum of squares) explained in each dataset by joint structure, individual structure, and residual noise is shown in Figure 11. This illustrates how the JIVE decomposition can be used to quantify and compare the amount of shared and individual variation between datatypes. As shown in Figure 11, joint structure is responsible for more variation in miRNA than in gene expression (31% and 16%, respectively), and the gene

expression data has a considerable amount of structured variation (53%) that is unrelated to miRNA. This is consistent with current biological understanding, as the primary function of miRNA is thought to be modification of gene expression, whereas gene expression is involved in many more biological functions.

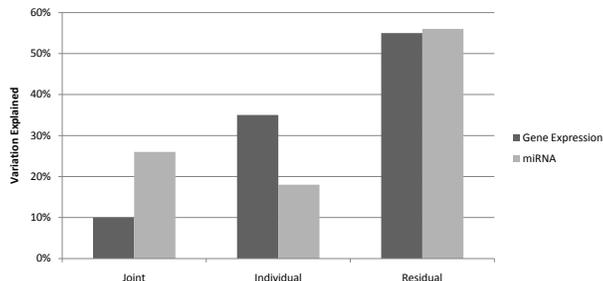


FIG 11. Percentage of variation (sum of squares) explained by estimated joint structure, individual structure and residual noise for miRNA and gene expression data.

4.2. *Sample distinctions on joint and individual structure.* Sample scores for joint and individual structure, matrices  $S$  and  $S_i$  in equation 2.2, reveal sample patterns that are present across datatypes, and patterns that are individual to each datatype. Figure 12 shows separate scatterplots of the sample scores for the first two principal components of estimated joint structure, the first two components individual to miRNA, and the first two components individual to gene expression. All four subtypes are clearly distinguished in the scatterplot of joint scores, but a subtype effect is not visually apparent in either of the individual scatterplots.

Since the subtypes are defined by gene expression clustering, their appearance in Figure 12 is not surprising. However, the clustering apparent in the joint plot shows involvement of miRNA in the differentiation of these subtypes. It is interesting that a subtype effect is not apparent in either scatterplot for individual structure, suggesting that this variation is driven by other biological components. This is remarkable, as the fraction of gene expression variation explained by joint structure (see Figure 11) is small.

To numerically compare the extent to which subtype distinctions are present, we considered the variability within subtypes (across all rows) as a proportion of total variability. Table 1 gives SWISS scores (Standardized WithIn subtype Sum of Squares) for the gene expression and miRNA data, SWISS scores for the JIVE estimates of joint and individual structure, and SWISS scores for the JIVE estimates of joint and individual struc-

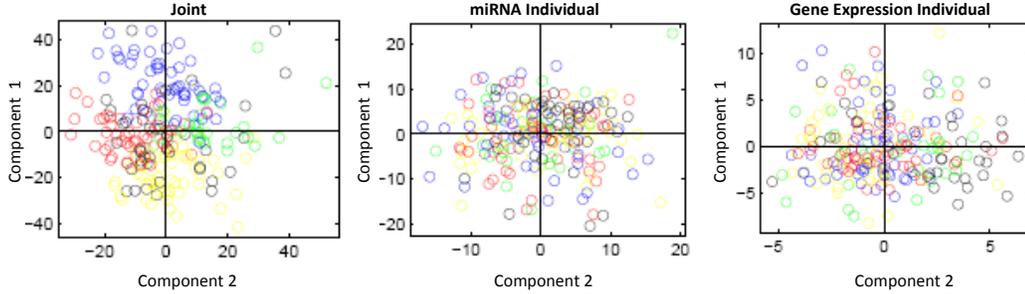


FIG 12. Scatterplots of sample scores for the first two joint components, first two individual miRNA components, and first two individual gene expression components. Samples are colored by subtype: Mesenchymal (yellow), Proneural (blue), Neural (green) and Classical (red).

ture with sparsity. Sparsity is enforced as described in Section 2.5, and the weight parameter  $\lambda$  is determined via the Bayesian Information Criterion. A permutation test described in Cabanski et al. [2010] concludes that the four subtypes are significantly more distinguished on the estimated joint structure, both with and without sparsity, than on the gene expression and miRNA data ( $p < 0.001$ ). SWISS scores for individual structure in gene expression and miRNA are close to one, as differences between subtypes are almost entirely represented in the joint structure between the two datatypes. This suggests that miRNA may play a greater role in GBM biology than previously thought.

In general, these analyses illustrate how an unsupervised, integrated analysis across multiple datatypes can result in a better distinction between subtypes or other biological classes. Note that one could carry out a similar analysis to investigate how the JIVE components relate to survival or other clinical factors, rather than subtype. Furthermore, a direct cluster analysis on the JIVE components could be used to identify sample groups that are distinguished across multiple datatypes.

4.3. *Exploring gene-miRNA associations through sparsity.* A natural way to explore associations between individual genes and miRNAs is to compute the matrix of all gene-miRNA correlations, and then examine the set of significant correlations. A heat map showing significant gene-miRNA correlations is shown in Panel A of Figure 13.

A sparse implementation of JIVE provides an alternative approach to

TABLE 1

*SWISS scores for TCGA subtypes. Lower scores indicate more subtype distinction.*

Data	Gene expression miRNA	0.8431 0.8763
JIVE	Joint	0.7952
	Gene expression individual miRNA individual	0.9613 0.9698
JIVE with sparsity	Joint	0.6797
	Gene expression individual miRNA individual	0.9800 0.9865

identifying gene-miRNA associations, and can reveal additional structure. Panel B shows the sample scores in the first joint component resulting from a sparse JIVE analysis of the data. Panel C shows all the gene-miRNA pairs with the property that both that gene and miRNA have non-zero loadings in the first joint component. Thus, the non-zero entries of the heat map have the form of a Cartesian product. We note that the non-zero entries in Panel C closely match those in the correlation map of Panel A, and that the signs of these entries also show good agreement. Scores for the first joint component (Panel B) distinguish the Mesenchymal and Proneural subtypes, suggesting that differences between these sample groups are driving the first joint component, and appear to influence the correlation structure of the data as well.

Panels D and E display sample scores and non-zero loadings for the second joint component. Panel D shows that the second joint component distinguishes the Neural and Classical subtypes. We note that Panel E is markedly different from Panel A, indicating that the second joint component is capturing associations between the expression of genes and miRNA that are not immediately apparent from the consideration of correlations alone. Indeed, these associations appear to be masked by variation captured in the first joint component.

**5. Summary and Discussion.** TCGA and similar projects are providing researchers with access to an increasing number of multi-block datasets. However, there are relatively few general statistical methods for the analysis of such integrated datasets. The unique features of JIVE provide a powerful new approach to analyzing multi-block data. JIVE finds both the coordinated activities of multiple datatypes, as well as those features unique to a particular datatype. We demonstrate how accounting for joint structure can lead to better estimation of individual structure, and vice-versa. Our application of JIVE to multi-block data on GBM tumor samples has provided better characterization of tumor types, and better understanding of

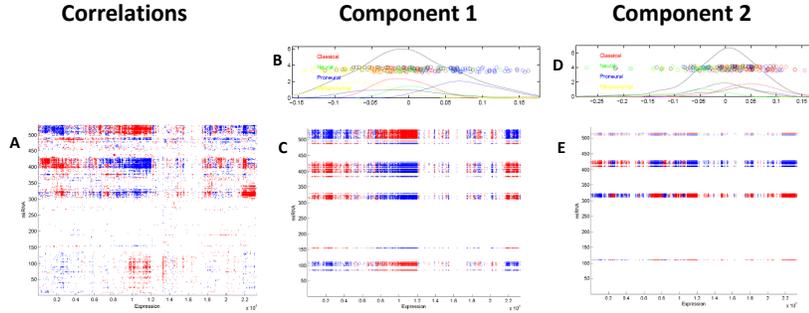


FIG 13. Plot of gene-miRNA correlations (A), and scores and loadings for the first two sparse joint components (B-E). In (A), gene-miRNA pairs are colored *red* if they have a significant positive correlation and *blue* if they have a significant negative correlation ( $P < 10^{-5}$ ). Panels (B) and (D) show sample scores for the first two joint components, colored by subtype. Panels (C) and (E) display gene-miRNA pairs where each have non-zero loadings. Pairs are colored *red* if both gene and miRNA loadings have the same sign, *blue* otherwise. In panels (A), (C) and (E) genes and miRNAs are ordered separately by average linkage correlation clustering.

the biological interactions between the given datatypes.

While this paper focuses on vertically integrated biomedical data, the JIVE model and algorithm are very general and may be useful in other contexts. A similar approach can be applied to horizontally integrated data, in which disparate sets of samples (e.g. sick and healthy patients) are available on the same datatype. In finance, JIVE has the potential to improve on current models that explain variation across and within disparate markets (see Bekaert et al. [2009]). These applications are currently under study.

## APPENDIX A

**A.1. Notes on Orthogonality.** Enforcing orthogonality between joint and individual structure in the JIVE decomposition does not constrain the solution, and is a sufficient statement of condition for uniqueness of each component under mild assumptions. Here we formally present and prove this result.

THEOREM A.1. *Let*

$$T = \begin{bmatrix} T_1 \\ \vdots \\ T_k \end{bmatrix} = \overbrace{\begin{bmatrix} J_1 \\ \vdots \\ J_k \end{bmatrix}}^J + \overbrace{\begin{bmatrix} A_1 \\ \vdots \\ A_k \end{bmatrix}}^A,$$

where  $\text{rank}(T_i) = \text{rank}(J) + \text{rank}(A_i) \forall i$ . Then,

1. There exists  $J^\perp, A^\perp$  such that

$$T = \overbrace{\begin{bmatrix} J_1^\perp \\ J_2^\perp \\ \vdots \\ J_k^\perp \end{bmatrix}}^{J^\perp} + \overbrace{\begin{bmatrix} A_1^\perp \\ A_2^\perp \\ \vdots \\ A_k^\perp \end{bmatrix}}^{A^\perp},$$

$\text{rank}(J^\perp) = \text{rank}(J)$ ,  $\text{rank}(A_i^\perp) = \text{rank}(A_i) \forall i$ , and  $J^\perp(A^\perp)' = 0_{p \times p}$ .

2. If  $\text{row}(A_1) \cap \dots \cap \text{row}(A_k) = \emptyset$ , then  $J^\perp$  and  $A^\perp$  in (1) are uniquely defined.

PROOF.

1. Let  $P_J$  define the projection matrix onto the row space of  $J$ ,  $\text{row}(J)$ . For  $i = 1, \dots, k$ , define

$$J_i^\perp = J_i + A_i P_J \quad \text{and} \quad A_i^\perp = A_i(I - P_J).$$

Then,  $J_i^\perp + A_i^\perp = J_i + A_i P_J + A_i - A_i P_J = J_i + A_i \forall i$ , and hence

$$T = \begin{bmatrix} J_1^\perp \\ \vdots \\ J_k^\perp \end{bmatrix} + \begin{bmatrix} A_1^\perp \\ \vdots \\ A_k^\perp \end{bmatrix}.$$

Furthermore,

$$J^\perp A^{\perp'} = J^\perp(I - P_J)A^{\perp'} = (J^\perp - J^\perp)A^{\perp'} = 0_{p \times p}.$$

Note  $\text{row}(J^\perp) \subseteq \text{row}(J)$ , so  $\text{rank}(J^\perp) \leq \text{rank}(J)$ .

Also  $\text{rank}(A_i^\perp) \leq \min\{\text{rank}(A_i), \text{rank}(I - P_J)\} \leq \text{rank}(A_i)$ .

Hence, as

$$\text{rank}(J) + \text{rank}(A_i) = \text{rank}(T_i) = \text{rank}(J^\perp) + \text{rank}(A_i^\perp),$$

$\text{rank}(J^\perp) = \text{rank}(J)$  and  $\text{rank}(A_i^\perp) = \text{rank}(A_i) \forall i$ .

2. Assume

$$T = \overbrace{\begin{bmatrix} \tilde{J}_1 \\ \tilde{J}_2 \\ \vdots \\ \tilde{J}_k \end{bmatrix}}^{\tilde{J}} + \overbrace{\begin{bmatrix} \tilde{A}_1 \\ \tilde{A}_2 \\ \vdots \\ \tilde{A}_k \end{bmatrix}}^{\tilde{A}},$$

where  $\text{rank}(\tilde{J}) = \text{rank}(J)$ ,  $\text{rank}(\tilde{A}_i) = \text{rank}(A_i) \forall i$ , and  $\tilde{J}(\tilde{A})' = 0_{p \times p}$ . Then,

$$\text{rank}(\tilde{J}_i) = \text{rank}(T_i) - \text{rank}(\tilde{A}_i) = \text{rank}(\tilde{J}).$$

So, as  $\text{row}(\tilde{J}_i) \subseteq \text{row}(\tilde{J}) \forall i$ ,

$$\text{row}(\tilde{J}_1) = \dots = \text{row}(\tilde{J}_k) = \text{row}(\tilde{J}).$$

Note that

$$\tilde{J}_i = \tilde{J}_i P_{J_i^\perp} + \tilde{J}_i P_{A_i^\perp} \quad \forall i.$$

We will show  $\tilde{J}_i P_{A_i^\perp} = 0$ .

First, take  $c' A_i^\perp \in \text{row}(\tilde{J}_i P_{A_i^\perp})$ . Then, for any  $j = 1, \dots, k$ ,

$$\text{row}(c' A_i^\perp) \subseteq \text{row}(\tilde{J}) \subseteq \text{row}(T_j) = \text{row}(J^\perp) \cup \text{row}(A_j^\perp),$$

hence there exists  $a_j, b_j$  such that

$$\begin{aligned} a'_j J^\perp + b'_j A_j^\perp &= c' A_i^\perp \\ \rightarrow a'_j J^\perp &= c' A_i^\perp - b'_j A_j^\perp \\ \rightarrow a'_j J^\perp &= 0, \end{aligned}$$

since  $\text{row}(J^\perp) \perp \{\text{row}(A_i^\perp) \cup \text{row}(A_j^\perp)\}$ . So,

$$b'_j A_j^\perp = c' A_i^\perp \quad \forall j.$$

Note  $\text{row}(A_i^\perp) \subseteq \text{row}(A_i) \forall i$ , so by assumption

$\text{row}(A_1^\perp) \cap \dots \cap \text{row}(A_k^\perp) = \emptyset$  and hence  $\text{row}(c' A_i^\perp) = 0$ .

So  $\tilde{J}_i P_{A_i^\perp} = 0$ , and hence  $\tilde{J}_i = \tilde{J}_i P_{J^\perp}$ ,  $\forall i$ . So we conclude

$\text{row}(\tilde{J}) \subseteq \text{row}(J^\perp)$ , and by an analogous argument  $\text{row}(J^\perp) \subseteq \text{row}(\tilde{J})$ .

It follows that  $J_i^\perp = \tilde{J}_i \forall i$  since

$$T_i P_{J^\perp} = (J_i^\perp + A_i^\perp) P_{J^\perp} = J_i^\perp$$

and

$$T_i P_{J^\perp} = (\tilde{T}_i + \tilde{A}_i) P_{J^\perp} = \tilde{J}_i.$$

Furthermore,  $A_i^\perp = T_i - J_i^\perp = T_i - \tilde{J}_i = \tilde{A}_i \forall i$ . □

**A.2. Pseudocode: Estimation.** Pseudocode for the iterative estimation procedure described in Section 2.2:

- Initialize  $X^{\text{Joint}} = X = [X_1 \dots X_k]'$
- Loop:
  - Give  $J = U\Lambda V^T$  by rank  $r$  SVD of  $X^{\text{Joint}}$
  - For  $i = 1, \dots, k$ :
    - \* Set  $X_i^{\text{Unique}} = X_i - U_i S$
    - \* Give  $A_i = W_i \Lambda V_i^T$  by rank  $r_i$  SVD of  $X_i^{\text{Unique}}(I - VV^T)$ , set  $S_i = \Lambda V_i^T$
    - \* Set  $X_i^{\text{Joint}} = X_i - W_i S_i$
  - Set  $X^{\text{Joint}} = [X_1^{\text{Joint}} \dots X_k^{\text{Joint}}]'$

Note that the orthogonality constraint is imposed in the estimation of individual structure.

**A.3. Pseudocode: Rank Estimation.** Pseudocode for the two-stage permutation based approach to rank selection described in Section 2.4, for given number of permutations  $n_{\text{perm}}$  and significance threshold  $\alpha$ .

To estimate  $\text{rank-eff}(X_i)$ :

1. Initialize  $\text{rank-eff}(X_i) = 0$
2. Determine the first singular value of  $X_i$ ,  $d_1(X_i)$ .
3. Determine the first singular value, after permuting columns within each row of  $X_i$ . Repeat  $n_{\text{perm}}$  times.
4. Test if the proportion of singular values after permutation greater than  $d_1(X_i)$  is less than  $\alpha$ .
5. If significant, remove variation due to  $d_1(X_i)$ , set  $\text{rank-eff}(X_i) = \text{rank-eff}(X_i) + 1$ , and repeat from Step 2.

To estimate  $r, r_1, \dots, r_k$ :

1. Initialize  $r = 0$
2. Let  $J, A_1, \dots, A_k$  be the JIVE estimates to  $X$ , where  $\text{rank}(J) = r$  and  $\text{rank}(A_i) = r_i$  ( $r + r_i = \text{rank-eff}(X_i)$ ). Set  $X_i^u = X_i - J_i \forall i$ .
3. Test for remaining joint structure between  $X_1^u, \dots, X_k^u$ .
  - Determine the total variation explained by fitting rank 1 joint and rank  $r_i - 1$  individual structure to  $[X_1^u \dots X_k^u]'$ .
  - Determine the total variation explained by fitting rank 1 joint and rank  $r_i - 1$  individual structure, after permuting columns within each  $X_i^u$ . Repeat  $n_{\text{perm}}$  times.

- Test if the proportion of models on the permuted data that explain more variability than the original model is less than  $\alpha$ .
4. If significant joint structure remains, set  $r = r + 1$  and repeat from Step 2.

### ACKNOWLEDGEMENTS

This work was partially supported by NIH grant R01 MH090936-01, NSF grant DMS-0907177, NSF grant DMS-0854908 and NIH grant U24-CA143848.

### REFERENCES

- A Adourian, E Jennings, R Balasubramanian, WM Hines, D Damian, TN Plasterer, CB Clish, P Stroobant, R McBurney, ER Verheij, I Bobeldijk, J Greef, J Lindberg, K Kenne, U Andersson, H Hellmold, K Nilsson, H Salterd, and I Schuppe-Koistinenc. Correlation network analysis for data integration and biomarker selection. *Molecular BioSystems*, 4:249–259, 2008.
- G Bekaert, RJ Hodrick, and X Zhang. International stock return comovements. *The Journal of Finance*, 64(6):2591–2626, 2009.
- M Bredel, DM Scholtens, GR Harsh, C Bredel, JP Chandler, JJ Renfrow, AK Yadav, H Vogel, AC Scheck, R Tibshirani, and BI Sikic. A network model of a cooperative genetic landscape in brain tumors. *Journal of the American Medical Association*, 302(3):261–275, 2009.
- CR Cabanski, Y Qi, X Yin, E Bair, MC Hayward, C Fan, J Li, M Wilkerson, JS Marron, CM Perou, and DN Hayes. Swiss made: Standardized within class sum of squares to evaluate methodologies and dataset elements. *PLoS ONE*, 5(3):e9905, 2010.
- CZ Di, CM Crainiceanu, BS Caffo, and NM Punjabi. Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1):458–488, 2009.
- MY Galberin and GR Cochrane. The 2011 nucleic acids research database issue and the online molecular biology database collection. *Nucleic Acids Research*, 39:D1–D6, 2011.
- Y Gilad, SA Rifkin, and JK Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends Genet.*, 24(8):408–415, 2008.
- H Hotelling. Relations between two sets of variants. *Biometrika*, 28:321–377, 1936.
- KA Le Cao, D Rossouw, C Robert-Granie, and P Besse. A sparse pls for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7:Article 35, 2008.
- M Lee, H Shen, JZ Huang, and JS Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 2010.
- E Parkhomenko, D Tritchler, and J Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.
- H Parkinson, M Kapushesky, N Kolesnikov, G Rustici, M Shojatalab, N Abeygunawardena, H Berube, M Dylag, I Emam, A Farne, E Holloway, M Lukk, J Malone, R Mani, E Plicheva, TF Rayner, F Rezwan, A Sharma, E Williams, XZ Bradley, T Adamusiak, M Brandizi, T Burdett, R Coulson, M Krestyaninova, P Kurnosov, E Maguire, SG Neogi, P Rocca-Serra, SA Sansone, N Sklyar, M Zhao, U Sarkans, and A Brazma. Arrayexpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, 37:868–872, 2009.

- PR Peres-Neto, DA Jackson, and KM Somers. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis*, 49:974–997, 2005.
- B Rhead, D Karolchik, RM Kuhn, AS Hinrichs, AS Zweig, PA Fujita, M Diekhans, KE Smith, KR Rosenbloom, BJ Raney, A Pohl, M Pheasant, LR Meyer, K Learned, F Hsu, J Hillman-Jackson, RA Harte, B Giardine, TR Dreszer, H Clawson, GP Barber, D Haussler, and WJ Kent. The ucsc genome browser database: update 2010. *Nucleic Acids Res.*, 38:613–619, 2010.
- GE Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.
- H Shen and JZ Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99:1015–1034, 2008.
- R Shen, AB Olshen, and M Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.
- O Sporns, G Tononi, and R Kotter. The human connectome: A structural description of the human brain. *PLoS Computational Biology*, 1(4):e42, 2005.
- TCGA Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455:1061–1068, 2008.
- R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996.
- J Trygg and S Wold. O2-pls, a two-block (xyc) latent variable regression (lvr) method with an integral osc filter. *Journal of Chemometrics*, 17(1):53–64, 2003.
- RG Verhaak, KA Hoadley, E Purdom, V Wang, Y Qi, MD Wilkerson, CR Miller, L Ding, T Golub, JP Mesirov, G Alexe, M Lawrence, M O’Kelly, P Tamayo, BA Weir, S Gabriel, W Winckler, S Gupta, L Jakkula, HS Feiler, JG Hodgson, CD James, JN Sarkaria, C Brennan, A Kahn, PT Spellman, RK Wilson, TP Speed, JW Gray, M Meyerson, G Getz, CM Perou, and DN Hayes. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *pdgfra*, *idh1*, *egfr*, and *nf1*. *Cancer Cell*, 17(1):98–110, 2010.
- JA Westerhuis, T Kourti, and JF MacGregor. Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics*, 12(5):301–321, 1998.
- DM Witten and RJ Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.
- H Wold. Partial least squares. In S Kotz and NL Johnson, editors, *Encyclopedia of Statistical Sciences (Vol. 6)*, pages 581–591. Wiley: New York, 1985.
- S Wold, N Kettaneh, and K Tjessem. Hierarchical multiblock pls and pc models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, 10(5-6):463–482, 1996.

E.F. LOCK  
 J.S. MARRON  
 A.B. NOBEL  
 DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH  
 UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL  
 CHAPEL HILL, NC 27599  
 E-MAIL: [lock@email.unc.edu](mailto:lock@email.unc.edu)  
 E-MAIL: [marron@email.unc.edu](mailto:marron@email.unc.edu)  
 E-MAIL: [nobel@email.unc.edu](mailto:nobel@email.unc.edu)

K.A. HOADLEY  
 LINEBERGER COMPREHENSIVE CANCER CENTER  
 UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL  
 450 WEST DR.  
 CHAPEL HILL, NC 27599 E-MAIL: [hoadley@med.unc.edu](mailto:hoadley@med.unc.edu)