



Published in final edited form as:

Ann Appl Stat. 2013 March 1; 7(1): 523–542. doi:10.1214/12-AOAS597.

JOINT AND INDIVIDUAL VARIATION EXPLAINED (JIVE) FOR INTEGRATED ANALYSIS OF MULTIPLE DATA TYPES*

Eric F. Lock,

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

Katherine A. Hoadley,

Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, 450 West Dr. Chapel Hill, NC 27599

J.S. Marron, and

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

Andrew B. Nobel

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

Eric F. Lock: lock@email.unc.edu; Katherine A. Hoadley: hoadley@med.unc.edu; J.S. Marron: marron@email.unc.edu; Andrew B. Nobel: nobel@email.unc.edu

Abstract

Research in several fields now requires the analysis of datasets in which multiple high-dimensional types of data are available for a common set of objects. In particular, The Cancer Genome Atlas (TCGA) includes data from several diverse genomic technologies on the same cancerous tumor samples. In this paper we introduce Joint and Individual Variation Explained (JIVE), a general decomposition of variation for the integrated analysis of such datasets. The decomposition consists of three terms: a low-rank approximation capturing joint variation across data types, low-rank approximations for structured variation individual to each data type, and residual noise. JIVE quantifies the amount of joint variation between data types, reduces the dimensionality of the data, and provides new directions for the visual exploration of joint and individual structure. The proposed method represents an extension of Principal Component Analysis and has clear advantages over popular two-block methods such as Canonical Correlation Analysis and Partial Least Squares. A JIVE analysis of gene expression and miRNA data on Glioblastoma Multiforme tumor samples reveals gene-miRNA associations and provides better characterization of tumor types.

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

Data and software are available at <https://genome.unc.edu/jive/>.

Supplement A: Additional Material

(doi: ??? ; .pdf). The supplementary article Lock et al. (2012) provides additional details and further validation of the JIVE method.

This includes:

A proof concerning the existence and uniqueness of the decomposition.

A description of the permutation approach to rank selection.

Pseudocode for the algorithm.

A discussion of computing time and efficiency.

A discussion of invariance properties.

Results from the application of JIVE to many diverse simulated datasets.

Keywords and phrases

Data integration; Multi-block data; Principal Component Analysis; Data fusion

1. Introduction

Many fields of scientific research now analyze high-dimensional data, in which a large number of variables are measured for a given set of experimental objects. Increasingly, those data include multiple high-dimensional datasets for a common set of objects. Table 1 gives very diverse examples of such data objects. In this context we refer to each dataset as a *data type* to indicate that it comes from a distinct mode of measurement or domain.

The motivation for this article is a particular application to biological data. In biomedical studies, a number of technologies now commonly 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 *Nucleic Acids Research* lists 1330 publicly available databases that measure various 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 data types collected from a common set of samples. Large-scale projects like The Human Connectome Project (Sporns, Tononi and Kotter, 2005) and The Cancer Genome Atlas (TCGA Research Network, 2008) focus on the integrated analysis of multiple data types.

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

1.1. Data

We describe an application to data 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 characterize cancer on a molecular level through the analysis and integration of multidimensional large scale genomic data. The integration of information from disparate genomic sources has the promise to provide a more comprehensive understanding of cancer genetics and cell biology.

We focus, in particular, on a set of 234 Glioblastoma Multiforme (GBM) tumor samples. GBM is a common and very fatal form of malignant brain tumor. However, GBM cases are not homogeneous, and an understanding of systematic distinctions between the tumor samples may lead to more targeted therapies. Verhaak et al. (2010) classified the TCGA GBM samples into four 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.

Copy number aberrations and somatic mutations, and their relation to 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 microRNA (miRNA) data in GBM

biology has not been well studied. In this article we focus on the integrated analysis of miRNA and gene expression data. These data are two distinct data types, as they are measured on different platforms and represent different biological components.

Current biological ideas suggest that miRNAs function primarily as post-transcriptional regulators of gene expression. Typically, they are considered negative regulators, decreasing gene expression levels. Many of the algorithms that predict miRNA targets (TargetScan, miRanda, PicTar, and rna22) have vastly different predicted gene lists (Peter, 2010). Therefore, miRNA and gene expression relations are not well understood. However, recent research suggests that miRNAs may be partly responsible for the expression of well-known tumor activating genes (oncogenes) and tumor suppressing genes.

Investigating each data type individually would lose some important relations, considering the inherent interactions between miRNAs and gene expression. An integrative, multivariate approach is desired. The data decomposition proposed in Section 1.2 gives new insight into the joint and individual variation between the miRNA and gene expression data. Although this decomposition is unsupervised with respect to the above GBM subtypes, we further investigate how it leads to a better characterization of these subtypes.

For each tumor sample, there are measures of intensity for 534 miRNAs and 23,293 genes (messenger RNA). These data are publicly available from TCGA. The preprocessed data used for this analysis are available at <https://genome.unc.edu/jive>.

1.2. Proposed Method

Given the biological relation between gene expression and miRNA, it is reasonable to expect shared patterns in the two sets of measurements. We refer to such shared patterns as *joint structure*. We also expect the gene expression data to have systematic variation that is unrelated to the miRNAs, and vice versa. This *individual structure* can be the result of technical artifacts, but may also be of biological interest. For example, miRNA regulation is just one of many factors that can influence gene expression. This structured individual variation can interfere with finding important joint signal, just as joint structure can obscure important signal that is individual to a data type.

To separate joint and individual effects, we introduce a method called Joint and Individual Variation Explained (JIVE). This exploratory method decomposes a dataset into a sum of three terms: a low-rank approximation capturing joint structure between data types, low-rank approximations capturing structure individual to each data type, and residual noise. Analysis of individual structure provides a way to identify potentially useful information that exists in one data type, but not others. Accounting for individual structure also allows for more accurate estimation of what is common between data types. As illustrated in Section 4.2, JIVE can identify joint structure not found by existing methods. It may be used regardless of whether the dimension of a dataset exceeds the sample size. Furthermore, JIVE is applicable to datasets with more than two data types, and has a simple algebraic interpretation.

A heatmap of joint structure and individual structure in the GBM data, identified by JIVE, is shown in Figure 1. Columns (corresponding to the 234 samples) are shown in the same order for both gene and miRNA data in the display of joint structure. This common ordering shows complex structure in both data types, and shared patterns are present in subsets of genes and miRNAs. JIVE also identifies a large amount of structured variation individual to the gene expression data, and a lesser degree of individual structure in the miRNA data. While the individual structure accounts for more variability in the data than the joint structure, our analysis suggests that the joint structure is more relevant to the cancer biology.

2. Model and Estimation

2.1. Formal Framework

Formally, we focus on data for multiple 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. For example, in our application to GBM data the rows of X_1 contain gene expression measurements (of dimension $p_1 = 23,293$) and the rows of X_2 contain miRNA measurements (of dimension $p_2 = 534$) for the same set of 234 tissue samples ($n = 234$). The k matrices may be combined 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 data types often differ. To remove baseline differences between data types, it is helpful to row-center the data by subtracting the mean within each row. Data types may also be of different dimension (p_i) or differ in variability. To circumvent cases where “the largest dataset wins”, it is helpful to scale each data type 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 Frobenius norm $\|A\|^2 = \sum_{i,j} a_{ij}^2$. Then, $\|X_i^{\text{scaled}}\| = 1$ for each i , and each data type contributes equally to the total variation of the concatenated matrix

$$X^{\text{scaled}} = \begin{bmatrix} X_1^{\text{scaled}} \\ \vdots \\ X_k^{\text{scaled}} \end{bmatrix}. \tag{2.1}$$

2.2. Model

Let X_1, X_2, \dots, X_k be matrices as in Section 2.1, scaled appropriately. Joint structure is represented by a single $p \times n$ matrix of rank $r < \text{rank}(X)$. Individual structure for each X_j is represented by a $p_j \times n$ matrix of rank $r_j < \text{rank}(X_j)$.

More formally, let A_j be the matrix representing the individual structure of X_j , and let J_j be the submatrix of the joint structure matrix that is associated with X_j . Then, the unified JIVE model is

$$\begin{aligned} X_1 &= J_1 + A_1 + \varepsilon_1 \\ &\vdots \\ X_k &= J_k + A_k + \varepsilon_k, \end{aligned} \tag{2.2}$$

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

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

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 require 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 responsible for joint structure between data types are unrelated to sample patterns responsible for individual structure. This requirement does not constrain the model, in that any matrix in the form (2.2) can be written equivalently with orthogonality between joint and individual structure. Furthermore, the orthogonality constraint assures that the joint and individual components in (2.2) are uniquely determined (see Supplement A for more details). It is remarkable that no further orthogonality constraints (say, between the column space of J_i and the column space of A_j , or between the row spaces of each A_j) are required to make the decomposition identifiable.

2.3. Estimation

Here we discuss estimation of the model for fixed ranks r, r_1, \dots, r_k . The choice of these ranks is important to accurately quantify the amount of joint and individual structure. Supplement A describes a permutation testing approach to rank selection.

Joint and individual structure are estimated by minimizing the sum of squared error. Let R be the $p \times n$ matrix of residuals 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 $\|R\|^2$ under the given ranks. This is accomplished by iteratively estimating joint and individual structure:

- Given J , find A_1, \dots, A_k to minimize $\|R\|$.
- Given A_1, \dots, A_k , find J to minimize $\|R\|$.
- Repeat until convergence.

The joint structure J minimizing $\|R\|$ is equal to the first r terms in the singular value decomposition (SVD) of X with individual structure removed. The estimated individual structure for X_j is equal to the first r_j terms of the SVD of X_j with the joint structure removed. The estimate of individual structure for X_j will not change those for X_i , $j \neq i$, and hence the k individual approximations minimize $\|R\|$ for fixed joint structure. Pseudocode for this iterative algorithm is given in Supplement A. Computing time can be improved by reducing the dimensionality of X_1, \dots, X_k at the outset via their SVD (see Supplement A).

The iterative method is monotone in the sense that $\|R\|$ decreases at each step. Thus $\|R\|$ converges to a coordinate-wise minimum, which can't be improved by changing the estimated joint or individual structure. Further convergence properties of the algorithm are currently under study.

2.4. 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 are depicted in Figure 2. Both X and Y are of dimension 50×100 , 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 randomly determined and are therefore independent 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 visually obscured.

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.

Figure 3 shows the JIVE estimate for joint structure, JIVE estimates for individual structure, and the fit given by the sum of joint and individual structure. Estimates closely resemble the true signal in Figure 2.

2.5. GBM Data

As the gene expression and miRNA data for the GBM samples differ in dimension and variability, they were scaled as in Section 2.1. Permutation testing (see Supplement A) 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 33 structure individual to gene expression.
- rank 13 structure individual to miRNA.

The percentage of variation (sum of squares) explained in each dataset by joint structure, individual structure, and residual noise is shown in Figure 4. This illustrates how the JIVE decomposition can be used to quantify and compare the amount of shared and individual variation between data types. As shown in Figure 4, joint structure is responsible for more variation in miRNA than in gene expression (23% and 14%, respectively), and the gene expression data has a considerable amount of structured variation (58%) that is unrelated to miRNA. This is consistent with current biological understanding, as miRNAs are just one of several factors that can influence gene expression.

Heatmaps of the low-rank estimates for joint structure, individual structure, and residual noise are shown in Figure 5. Columns have the same order in all heatmaps. This reveals the shared patterns present in the joint structure from Figure 1, but little of the structure that is present in the individual estimates.

3. Model Factorization

3.1. Relation to PCA

The JIVE model can be factorized as in *Principal Component Analysis* (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 $U(p \times r)$ 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 factorized model

$$\begin{aligned} X_1 &= U_1 S + W_1 S_1 + R_1 \\ &\vdots \\ X_k &= U_k S + W_k S_k + R_k. \end{aligned} \quad (3.1)$$

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

3.2. GBM Data

Sample scores for joint structure, matrix S in Equation 3.1, reveal sample patterns that are present across the miRNA and gene expression data. Sample scores for individual structure, matrices S_1 and S_2 in Equation 3.1, reveal sample patterns that are individual to each data type. Figure 6 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. Subtype distinctions are clearly present in the scatter-plot 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 6 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 4) is small.

To numerically compare the extent to which subtype distinctions are present, we consider the standardized within-subtype sum of squares

$$\text{SWISS}(A) = \frac{\sum_i \sum_j (A_{ij} - \bar{A}_{i,s(j)})^2}{\sum_i \sum_j (A_{ij} - \bar{A}_{i.})^2},$$

where $s(j) = \{k: \text{samples } j \text{ and } k \text{ belong to the same subtype}\}$. This represents the variability within subtypes (across all rows) as a proportion of total variability. Table 2 gives SWISS scores for the gene expression and miRNA data, and SWISS scores for the JIVE estimates of joint and individual structure. A permutation test described in Cabanski et al. (2010) concludes that the four subtypes are significantly more distinguished on the estimated joint structure than on the gene expression and miRNA data ($p < 0.001$; 10, 000 permutations). SWISS scores for individual structure in gene expression and miRNA are close to one, as subtype distinctions are almost entirely represented in the joint structure between the two data types. 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 data types can result in a better distinction between subtypes or other biological classes. One could conduct 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 data types.

4. Comparison with Existing Methods

4.1. Existing Methods

One approach to the analysis of multiple datasets is to mine the data for variable-by-variable associations. In computational biology, large-scale correlation studies can identify millions of pairwise variable associations between genomic data types (see, for example, Gilad, Rifkin and Pritchard (2008)). Furthermore, network models can link associated variables across and within data types (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 data types, which is the focus of this paper.

PCA of the block-scaled matrix X^{scaled} in (2.1) coincides with *Consensus PCA* (Wold, Kettaneh and Tjessem, 1996 ; Westerhuis, Kourti and MacGregor, 1998). This direct approach is also used by the iCluster method (Shen, Olshen and Ladanyi, 2009), which performs clustering based on a factor analysis of the concatenated matrix X . While these methods synthesize information from multiple data types, they do not distinguish between joint or individual effects.

Canonical Correlation Analysis (CCA) (Hotelling, 1936) is a popular method to globally examine the relation 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 related to 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 data types. However, the restriction of O2-PLS (and PLS) to pairwise comparisons limits their utility in finding common structure among more than two data types.

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 2.1, standardized so that each row has mean 0 and standard deviation 1, the standard mCCA loading vectors u_1, u_2, \dots, u_k satisfy

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

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

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 data types, while MF-PCA analyzes grouped observations on the same functional data type. More substantively, MF-PCA does not model shared structure and individual structure. Rather, MF-PCA models structure among main effects (e.g., at the group level) and structured variation about those main effects (e.g., at the sample level), in the context of functional data.

4.2. Illustrative Example

We return to the example introduced in Section 2.4. Consensus PCA of the concatenated matrix $[X' Y']'$ does a poor job of finding the joint structure. The scatterplot in Figure 7 shows a weak association between the first principal component scores and the joint response V . This is because PCA of the concatenated data is driven by all variation in the data, joint or individual.

Figure 8 shows an analysis of PLS and CCA for X and Y . The scores for the first PLS direction for X and Y show a weak association (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 groups, indicating how for PLS individual structure can interfere with the identification of joint structure. The 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 . Scores and loadings for the joint component and both individual components are shown in Figure 9. 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 the groups 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 groups were defined on all 50 variables for both X and Y , which is reflected in the individual loadings (panels E and G). Note that joint and individual loadings are not constrained to be orthogonal, which gives the analysis more flexibility.

5. Variable Sparsity

In many practical applications, important structure between samples or objects is present on only a small subset of the measured variables. This motivates the 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, Tritchler and Beyene, 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_j in Equation 3.1) are exactly 0. For weights λ and λ_j , we minimize the penalized sum of squares

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

where Pen is a penalty designed to induce sparsity in the loading vectors. In our implementation, Pen 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 with sparsity is accomplished by an iterative procedure analogous to that used for the non-sparse case:

- Given J , find A_j to minimize $\|R_j\|^2 + \lambda_j \text{Pen}(W_j)$ for each $i = 1, \dots, k$.
- Given 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 λ , λ_j 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 data type that are associated. Examination of the joint sample scores, in turn, reveals sample patterns that drive these associations.

5.1. GBM Data

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. Panel A of Figure 10 shows a heatmap of the significant gene-miRNA correlations.

A sparse implementation of JIVE provides an alternative approach to 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.

Of primary interest are the biological relations between the genes and miRNAs. Figure 11 displays a network of possible gene-miRNA interactions for each of the first two joint components, constructed from genes and miRNAs with large absolute loadings. A gene-miRNA pair is linked if the miRNA is predicted to regulate the expression of the gene, based on its DNA sequence. In particular, we use the mirWalk target prediction module (Dweep et al., 2011) and include those pairs that are given in two or more of the miRNA target databases *miRanda*, *Pictar*, *RNA22* and *TargetScan*. We caution that current methods to predict miRNA targets are inexact, and linked gene-miRNA pairs only indicate potential causal relations. Nevertheless, each joint component includes several predicted gene-miRNA interactions.

We further examine the individual genes and miRNAs that contribute the most to the first two joint components. The *POSTN* gene, which has the largest loading among genes in the first joint component, encodes the protein Periostine. Over-expression of Periostine is frequently reported in cancerous tumor cells, and is suspected to facilitate cell motility (the ability of a cancer cell to migrate quickly and spontaneously) (Gillan et al., 2002). In GBM, the downregulation of *POSTN* expression by mir-219 has recently been linked to differences in survival and time to disease progression (Zinn et al., 2011). Interestingly, the miRNA mir-124a has the largest loading in both components, and a recent study suggests that this miRNA may also play an important role in GBM cell motility (Fowler et al., 2011).

6. Summary and Discussion

TCGA and similar projects are providing researchers with access to an increasing number of datasets that consist of multiple high-dimensional data types. However, there are relatively few general statistical methods for the analysis of such integrated datasets, and the unique features of JIVE provide a powerful new approach. JIVE finds both the coordinated activities of multiple data types, as well as those features unique to a particular data type. We demonstrate how accounting for joint structure can lead to better estimation of individual structure, and vice-versa. Our application of JIVE to gene and miRNA data on GBM tumor samples has provided better characterization of tumor types, and better understanding of the biological interactions between the given data types.

JIVE does not require that the data are ordered, as in *functional data analysis*. Regularization techniques such as smoothing may improve the method for functional data. As pointed out by an Associate Editor, JIVE estimates for joint and individual structure are not robust to outliers. Exploratory methods such as PCA suffer similarly, and robust versions of PCA have recently been developed (see, for example, Candes et al. (2009)). A robust version of JIVE is an interesting potential extension. Further, any missing values must be imputed prior to computing JIVE estimates. An approach that explicitly accounts for missing values in the estimation of joint and individual structure is another potential extension.

The statistical properties of the algorithm also deserve further attention. In particular, measures of confidence (e.g., for variation explained by joint and individual structure) would be useful. Standard resampling techniques such as bootstrapping may help in this regard. However, factors such as the discrete nature of the ranks must be considered carefully.

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 data type. In finance, JIVE has the potential to improve on current models that explain variation across and within disparate markets (see Bekaert, Hodrick and Zhang (2009)). These applications are currently under study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to thank the Editors and referees for their constructive comments. We especially appreciate the diligent work of the Associate Editor on this article.

References

- Adourian A, Jennings E, Balasubramanian R, Hines W, Damian D, Plasterer T, Clish C, Stroobant P, McBurney R, Verheij E, Bobeldijk I, Greef J, Lindberg J, Kenne K, Andersson U, Hellmold H, Nilsson K, Salterd H, Schuppe-Koistinenc I. Correlation network analysis for data integration and biomarker selection. *Molecular BioSystems*. 2008; 4:249–259. [PubMed: 18437268]
- Bekaert G, Hodrick R, Zhang X. International stock return comovements. *The Journal of Finance*. 2009; 64:2591–2626.
- Bredel M, Scholtens D, Harsh G, Bredel C, Chandler J, Renfrow J, Yadav A, Vogel H, Scheck A, Tibshirani R, Sikic B. A network model of a cooperative genetic landscape in brain tumors. *Journal of the American Medical Association*. 2009; 302:261–275. [PubMed: 19602686]
- Cabanski C, Qi Y, Yin X, Bair E, Hayward M, Fan C, Li J, Wilkerson M, Marron J, Perou C, Hayes D. SWISS MADE: Standardized within class sum of squares to evaluate methodologies and dataset elements. *PLoS ONE*. 2010; 5:e9905. [PubMed: 20360852]
- Candes E, Li X, Ma Y, Wright J. Robust principal component analysis? arXiv. 2009:0912.3599.
- Di C, Crainiceanu C, Caffo B, Punjabi N. Multilevel functional principal component analysis. *The Annals of Applied Statistics*. 2009; 3:458–488. [PubMed: 20221415]
- Dweep H, Sticht C, Pandey P, Gretz N. miRWalk - database: prediction of possible miRNA binding sites by “walking” the genes of 3 genomes. *Journal of Biomedical Informatics*. 2011; 44:839–847. [PubMed: 21605702]
- Fowler A, Thompson D, Giles K, Maleki S, Mreich E, Wheeler H, Leedman P, Biggs M, Cook R, Little N, Robinson B, McDonald K. miR-124a is frequently down-regulated in glioblastoma and is

- involved in migration and invasion. *European Journal of Cancer*. 2011; 47:953–963. [PubMed: 21196113]
- Galberin M, Cochrane G. The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*. 2011; 39:D1–D6. [PubMed: 21177655]
- Gilad Y, Rifkin S, Pritchard J. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*. 2008; 24:408–415. [PubMed: 18597885]
- Gillan L, Matei D, Fishman D, Gerbin C, Karlan B, Chang D. Periostin secreted by epithelial ovarian carcinoma is a ligand for alpha(V)beta(3) and alpha(V)beta(5) integrins and promotes cell motility. (2002). *Cancer Research*. 62:5358–5364. [PubMed: 12235007]
- Hotelling H. Relations between two sets of variates. *Biometrika*. 1936; 28:321–377.
- Le Cao K, Rossouw D, Robert-Granie C, Besse P. A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*. 2008; 7:Article 35. [PubMed: 19049491]
- Lee M, Shen H, Huang J, Marron J. Biclustering via sparse singular value decomposition. *Biometrics*. 2010; 66:1087–1095. [PubMed: 20163403]
- Lock E, Hoadley K, Marron J, Nobel A. Supplement to “Joint and Individual Variation Explained (JIVE) for integrated analysis of multiple data types”. 2012
- Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*. 2009; 8:1–34.
- Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner T, Rezwan F, Sharma A, Williams E, Bradley X, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi S, Rocca-Serra P, Sansone S, Sklyar N, Zhao M, Sarkans U, Brazma A. ArrayExpress update{from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*. 2009; 37:868–872.
- Peter M. Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene*. 2010; 29:2161–2164. [PubMed: 20190803]
- Rhead B, Karolchik D, Kuhn R, Hinrichs A, Zweig A, Fujita P, Diekhans M, Smith K, Rosenbloom K, Raney B, Pohl A, Pheasant M, Meyer L, Learned K, Hsu F, Hillman-Jackson J, Harte R, Giardine B, Dreszer T, Clawson H, Barber G, Haussler D, Kent W. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*. 2010; 38:613–619.
- Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6:461–464.
- Shen H, Huang J. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*. 2008; 99:1015–1034.
- Shen R, Olshen A, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009; 25:2906–2912. [PubMed: 19759197]
- Sporns O, Tononi G, Kotter R. The human connectome: A structural description of the human brain. *PLoS Computational Biology*. 2005; 1:e42. [PubMed: 16201007]
- TCGA Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008; 455:1061–1068. [PubMed: 18772890]
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
- Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics*. 2003; 17:53–64.
- Verhaak R, Hoadley K, Purdom E, Wang V, Qi Y, Wilkerson M, Miller C, Ding L, Golub T, Mesirov J, Alexe G, Lawrence M, O’Kelly M, Tamayo P, Weir B, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler H, Hodgson J, James C, Sarkaria J, Brennan C, Kahn A, Spellman P, Wilson R, Speed T, Gray J, Meyerson M, Getz G, Perou C, Hayes D. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010; 17:98–110. [PubMed: 20129251]
- Westerhuis J, Kourti T, MacGregor J. Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*. 1998; 12:301–321.

- Witten D, Tibshirani R. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*. 2009; 8:Article 28.
- Wold, H. Partial Least Squares. In: Kotz, S.; Johnson, N., editors. *Encyclopedia of Statistical Sciences*. Vol. 6. Wiley; New York: p. 581-591.
- Wold S, Kettaneh N, Tjessem K. Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*. 1996; 10:463–482.
- Zinn P, Majadan B, Sathyan P, Singh K, Majumder S, Jolesz F, Colen R. Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. *PLoS One*. 2011; 6:e25451. [PubMed: 21998659]

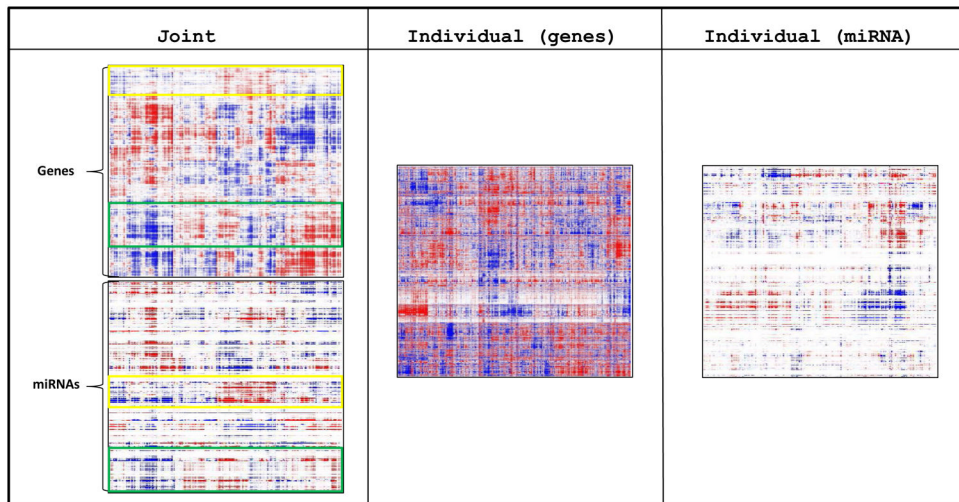


Fig 1. JIVE estimates for joint structure and individual structure in the GBM data. Blue corresponds to negative values, red positive values. Columns are given in the same order in the joint structure, and subsets of genes and miRNAs that share similar patterns are highlighted in green and yellow.

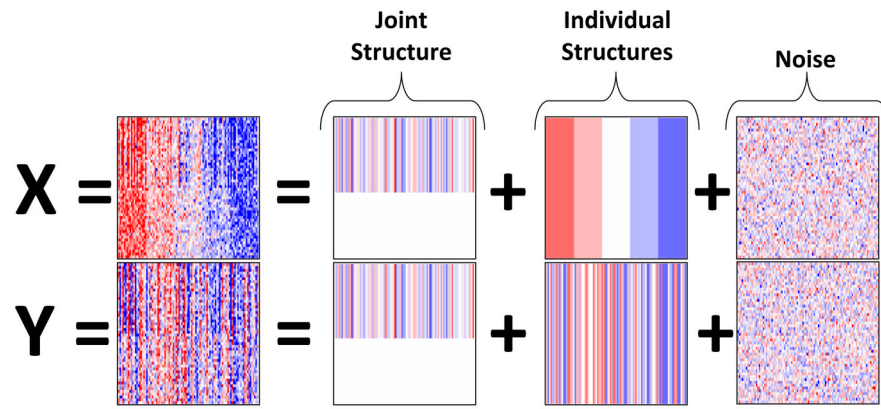


Fig 2.
 X and Y are generated by adding together joint structure, individual structure, and noise.
 Blue corresponds to negative values, red positive values.

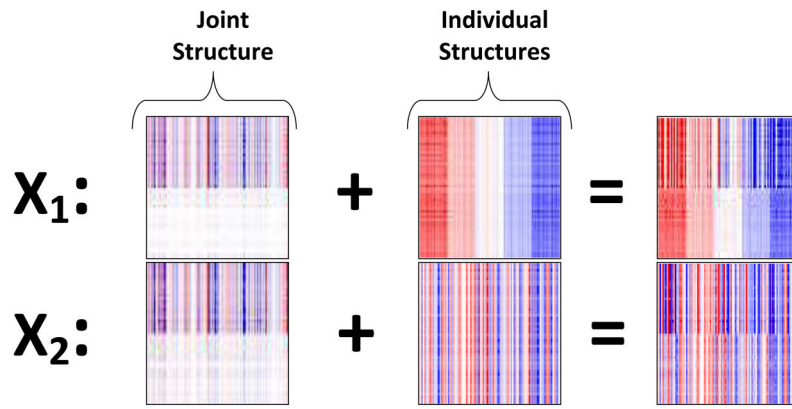


Fig 3. JIVE estimates for joint structure and individual structure. Blue corresponds to negative values, red positive values.

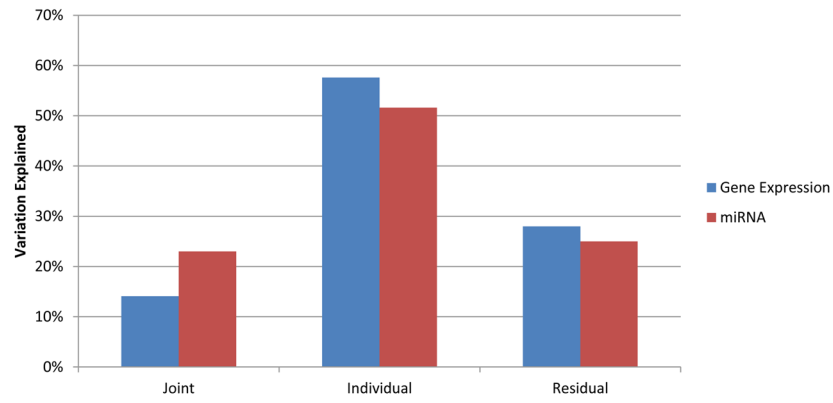


Fig 4. Percentage of variation (sum of squares) explained by estimated joint structure, individual structure and residual noise for miRNA and gene expression data.

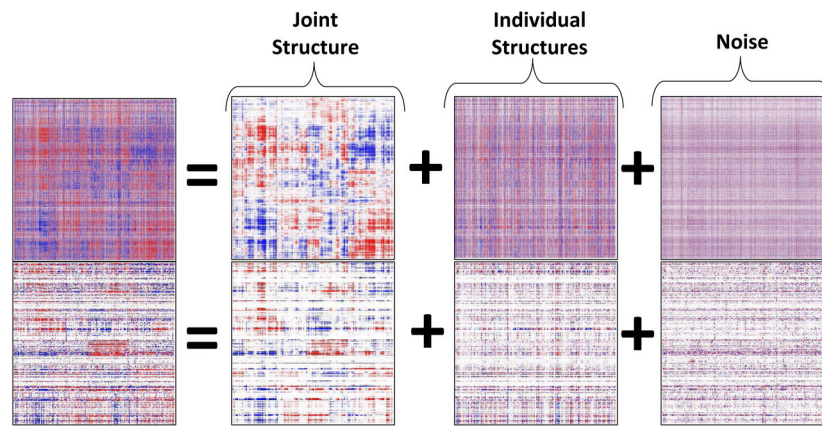
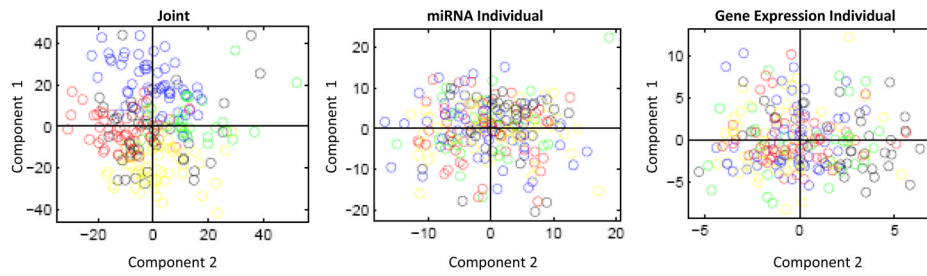


Fig 5. Heatmaps of low-rank estimates for joint structure, individual structure and residual noise in the gene expression (top) and miRNA (bottom) data. Blue corresponds to negative values, Red positive values. Columns have the same order in all heatmaps.

**Fig 6.**

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). Samples colored black were assayed after the initial subtype analysis, and are considered unclassified.

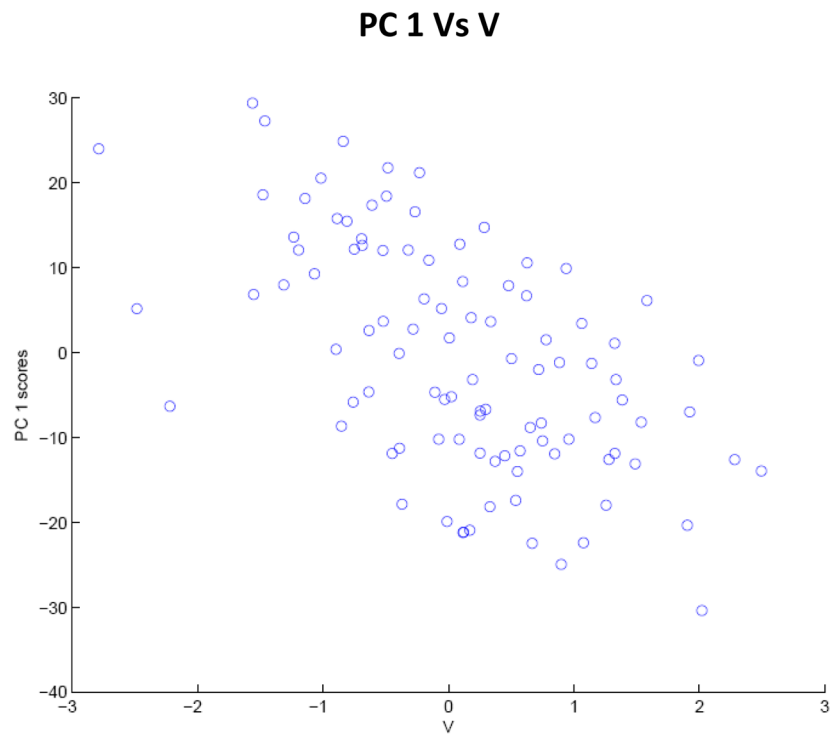
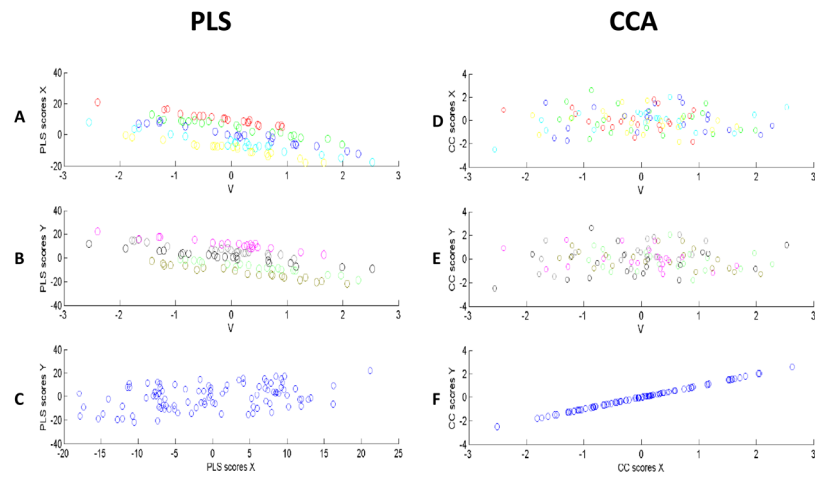


Fig 7. Scatterplot of the first consensus principal component scores vs the joint signal V . The scores are weakly associated with the joint signal.

**Fig 8.**

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

Joint Component

Individual Components

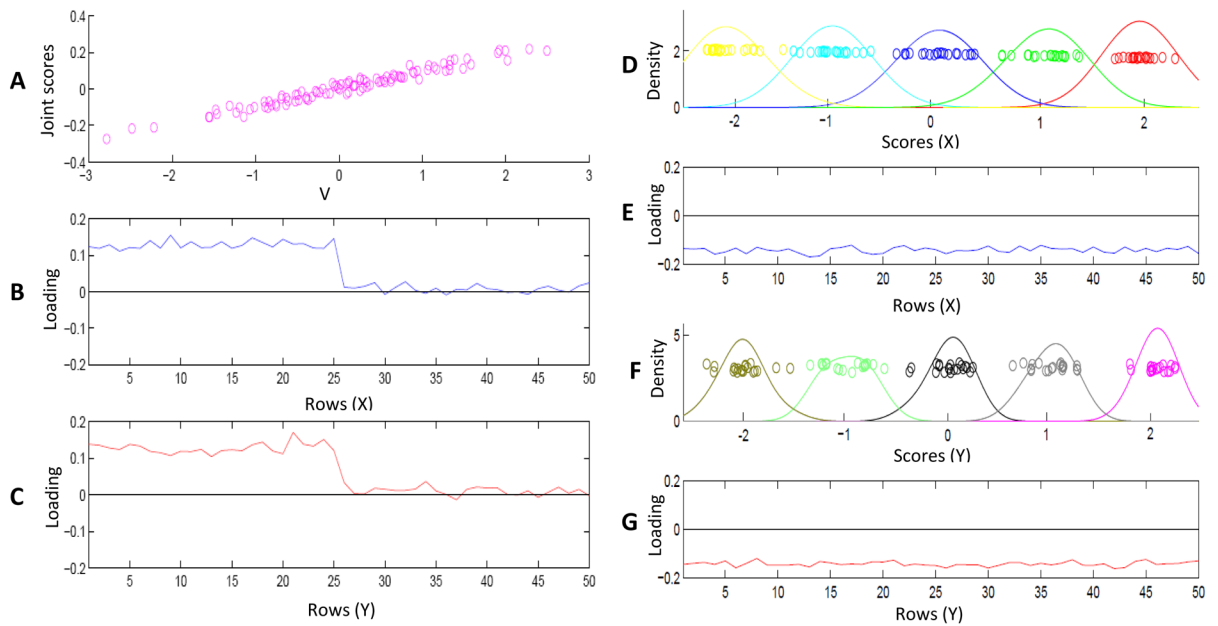


Fig 9. Scores and loadings for joint and individual components in the JIVE decomposition. Joint scores are highly associated with the common signal V (panel A). Individual scores distinguish groups specific to X and Y (D and F). Joint loadings (B and C) show a strong effect (difference from zero) on half of the variables in X and Y . Individual loadings (E and G) show a similar effect on all variables in X and Y .

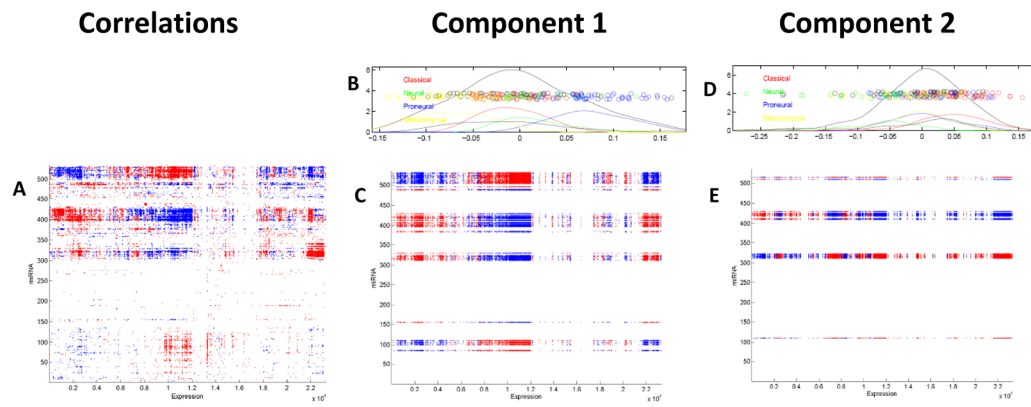


Fig 10. Plot of gene-miRNA correlations (A), and scores and loadings for the first two sparse joint components (B–E). In (A), gene-miRNA pairs are colored red if they have a significant positive correlation and blue if they have a significant negative correlation ($P < 10^{-5}$). Panels (B) and (D) show sample scores for the first two joint components, colored by subtype. Panels (C) and (E) display gene-miRNA pairs where both have nonzero 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.

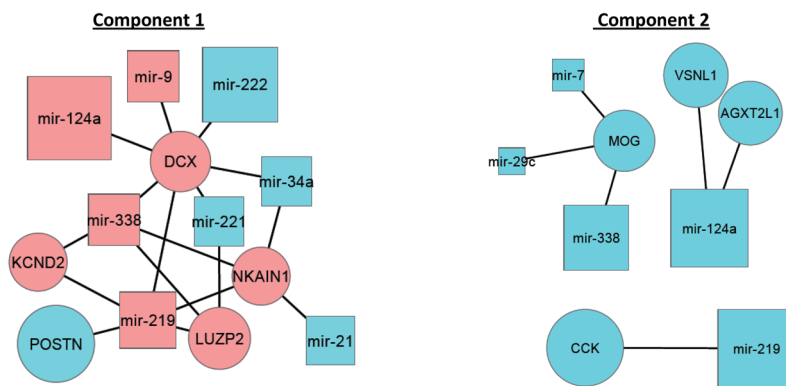


Fig 11. Network of predicted gene-miRNA interactions for the first two sparse joint components. Gene-miRNA pairs are linked if the miRNA is predicted to target the gene in two or more of the four databases miRanda, Pictar, RNA22 and TargetScan. Genes are shown as circles, and miRNAs are shown as squares. All predicted targets are shown between the 10 miRNAs with the largest absolute loading and the 10 genes with the largest absolute loading, for both components (genes and miRNAs with no targets are not shown). Genes and miRNAs with a positive loading are colored red, and with a negative loading are colored blue. The icon size is proportional to the absolute loading.

Table 1

Examples with multiple high-dimensional data types.

Field	Object	Data types
Computational biology	Tissue samples	Gene expression, microRNA, genotype, protein abundance/activity
Chemometrics	Chemicals	Mass spectra, NMR spectra, atomic composition
Atmospheric sciences	Locations	Temperature, humidity, particle concentrations over time
Internet traffic	Websites	Word frequencies, visitor demographics, linked pages

Table 2

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

Data	Gene expression	0.8431
	miRNA	0.8763
<hr/>		
	Joint	0.7678
JIVE	Gene expression individual	0.9019
	miRNA individual	0.9284