Joint and Individual Variation Explained (JIVE) for the Integrated Analysis of Multiple Data Types

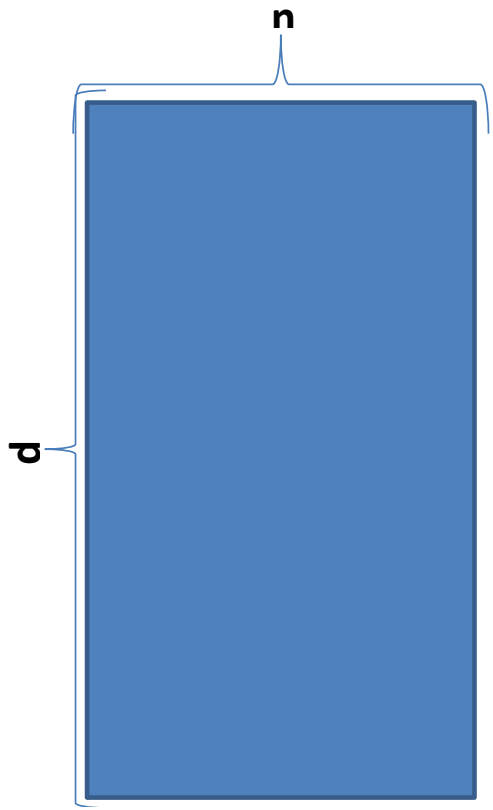
Eric F. Lock

Joint work with J.S. Marron, Andrew Nobel, and Katherine Hoadley

11/15/2012

High-Dimensional Data

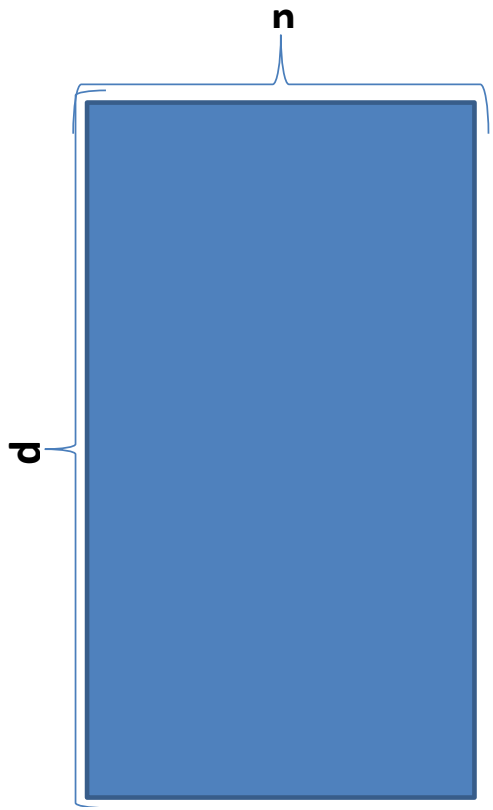
- Dimension d is very large (often $d > n$):



- Exploratory analysis

High-Dimensional Data

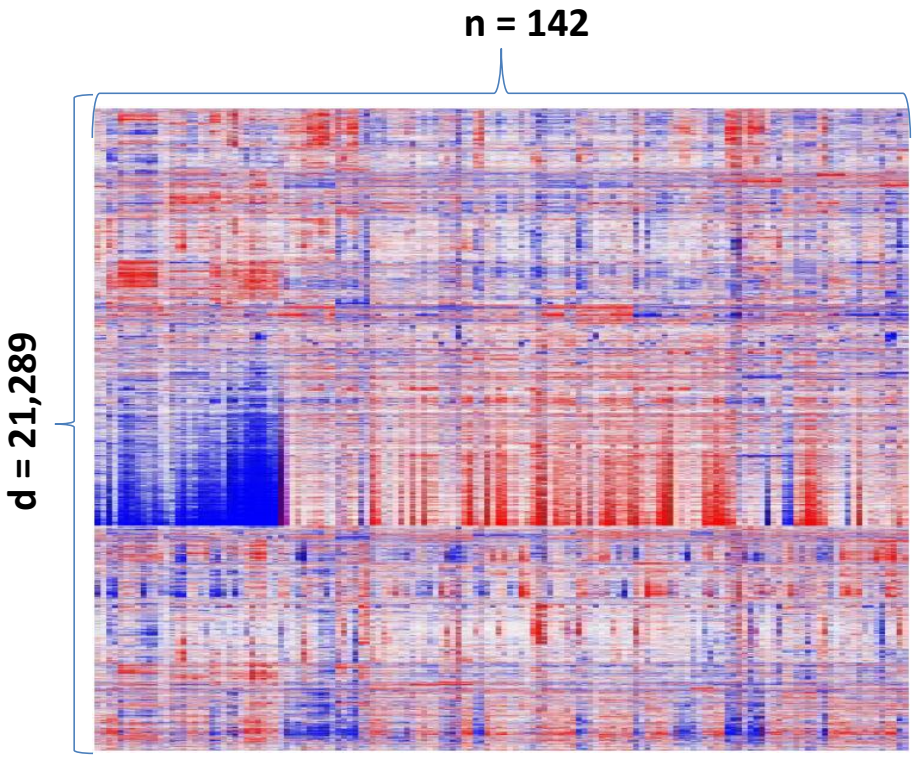
- Dimension d is very large (often $d > n$):



- Exploratory analysis
  - Heatmaps, Principal components analysis (PCA), projection pursuit, clustering, etc...

Example: Mice

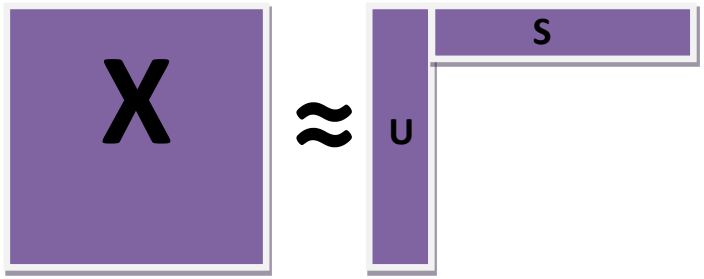
- Expression data available for 21,289 genes ($d = 21,289$) on 142 mice ($n = 142$).
  - Mice from 21 genetic strains.
  - 79 mice given dose of alcohol.



- Heatmap (red = high values; blue=low values)

Principal Components Analysis (PCA)

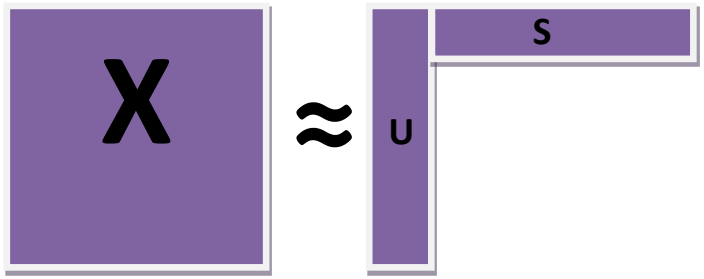
- Data matrix $X : d \times n$
- Approximate X in factorized form:



  - $U : d \times r$ are the variable "loadings"
  - $S : r \times n$ are the sample "scores"

Principal Components Analysis (PCA)

- Data matrix $X : d \times n$
- Approximate X in factorized form:



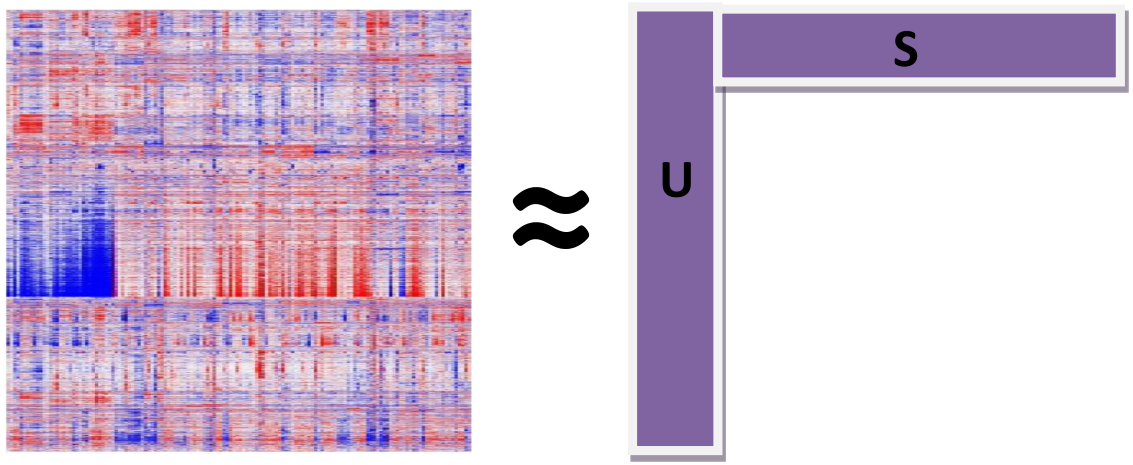
  - $U : d \times r$ are the variable "loadings"
  - $S : r \times n$ are the sample "scores"
- $\tilde{X} = US$ is the rank r matrix that minimizes

$$||X - \tilde{X}||_F^2 = \sum_{i,j} (x_{ij} - \tilde{x}_{ij})^2.$$

- Computation
  - Eigen-analysis of $X'X$.
  - Singular Value Decomposition (SVD) of X.

PCA: Mice (r=3)

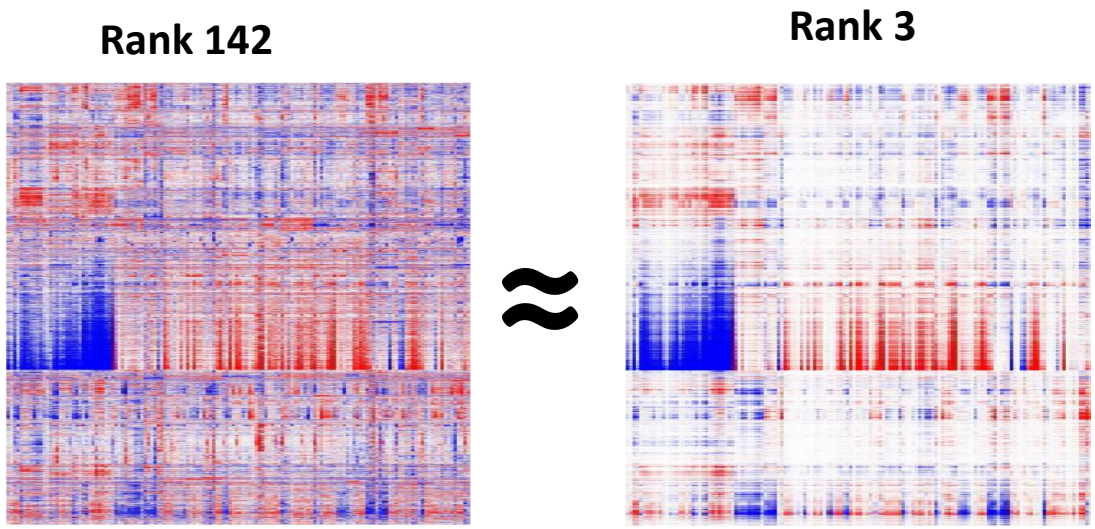
- Data matrix $X : 21,289 \times 142$
- Approximate X in factorized form:



- $U : 21,289 \times 3$ are the variable "loadings"
- $S : 3 \times 142$ are the sample "scores"

PCA: Mice (r=3)

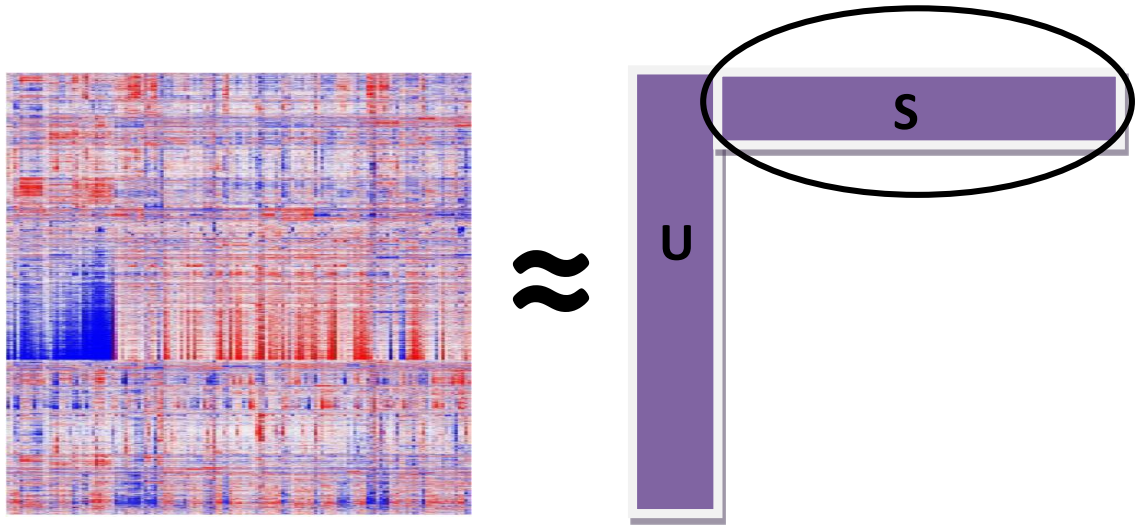
- Data matrix $X : 21,289 \times 142$
- Approximate X in factorized form:



- $U : 21,289 \times 3$ are the variable "loadings"
- $S : 3 \times 142$ are the sample "scores"

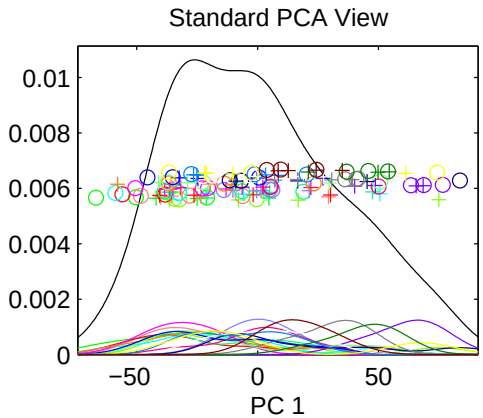
PCA: Mice (r=3)

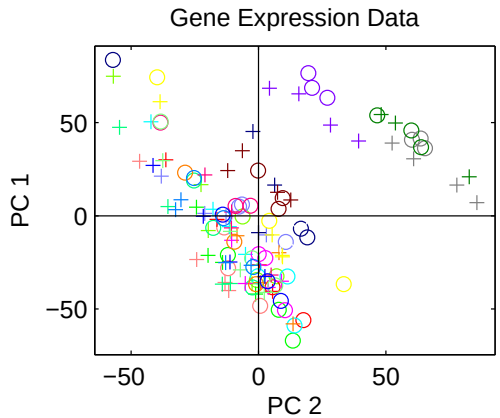
- Data matrix $X : 21,289 \times 142$
- Approximate X in factorized form:

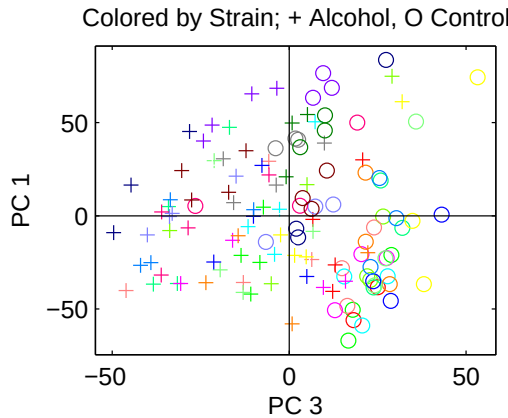


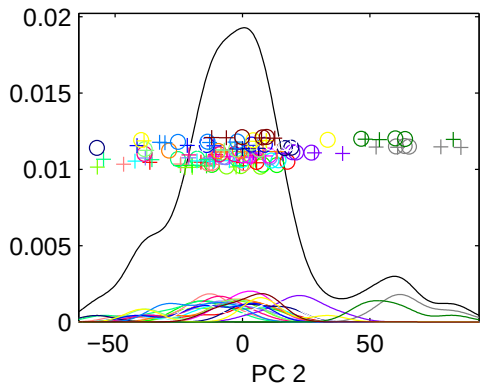
- $U : 21,289 \times 3$ are the variable "loadings"
- $S : 3 \times 142$ are the sample "scores"

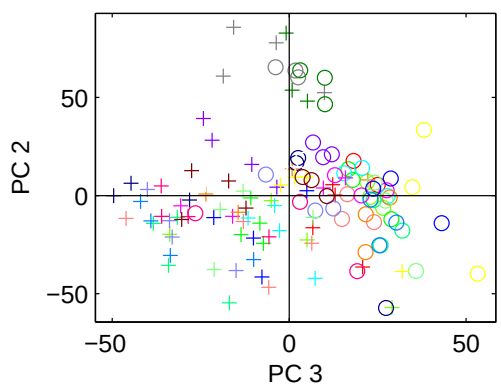
Mice PCA scores

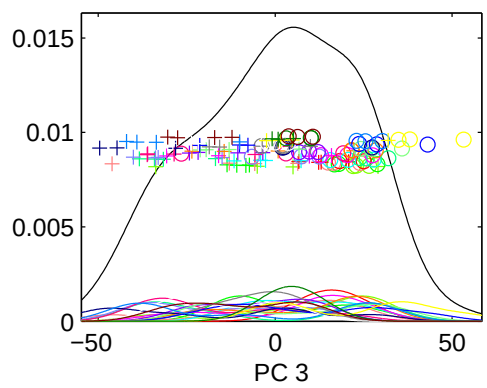










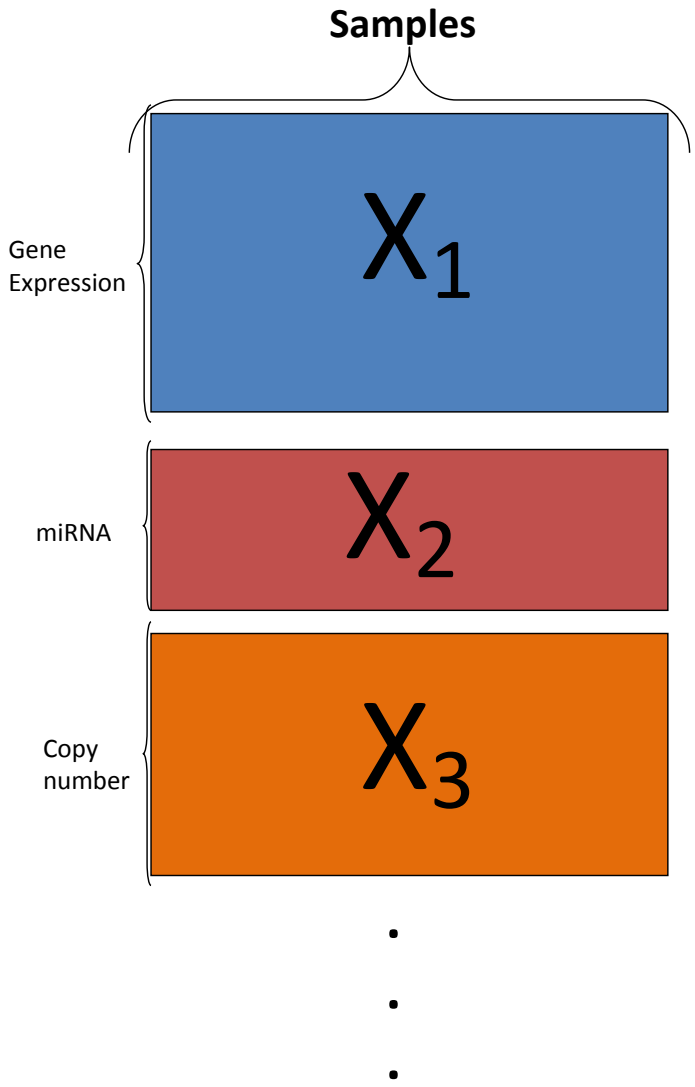


Challenge

- Multiple high-dimensional *data types* from the same objects.

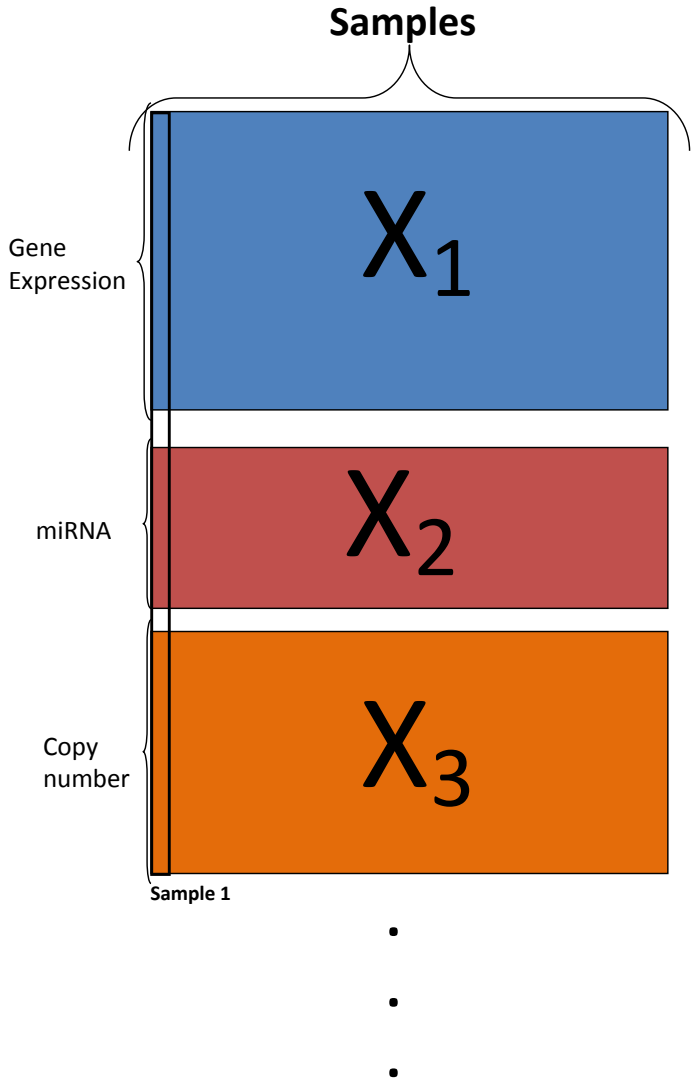
Challenge

- Multiple high-dimensional *data types* from the same objects.
- Example:



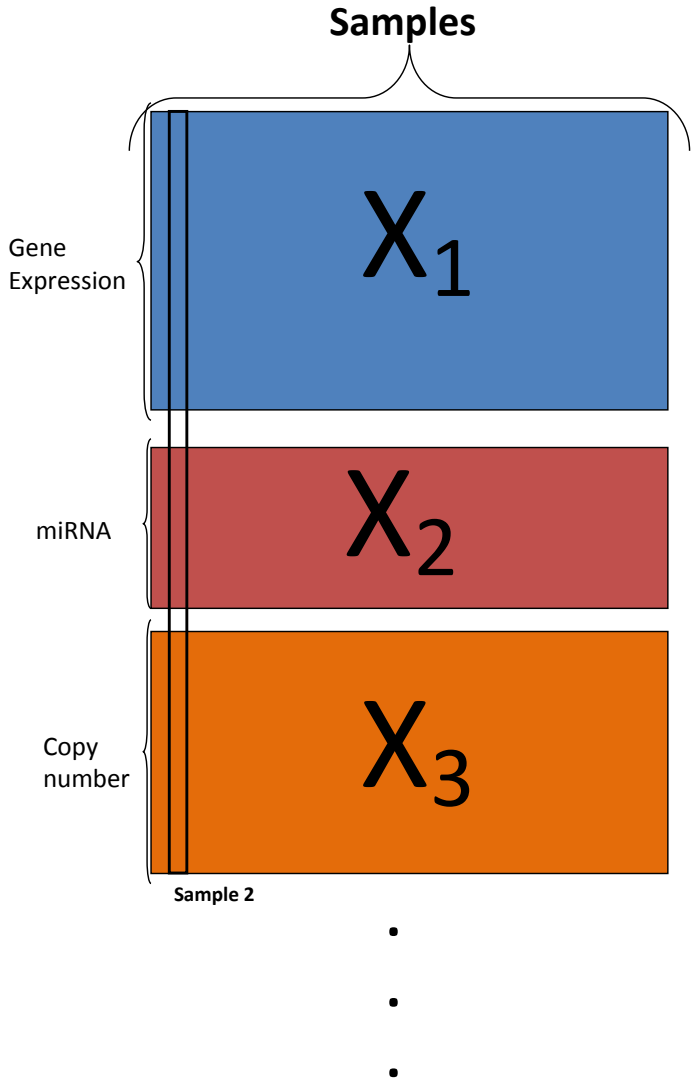
Challenge

- Multiple high-dimensional *data types* from the same objects.
- Example:



Challenge

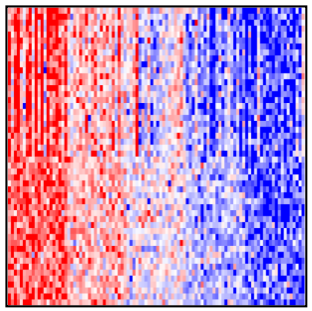
- Multiple high-dimensional *data types* from the same objects.
- Example:

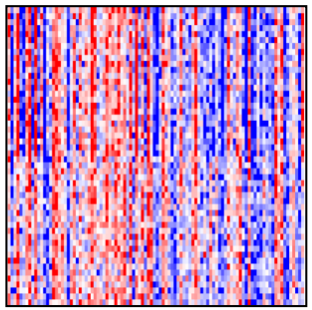


Integrated Analysis

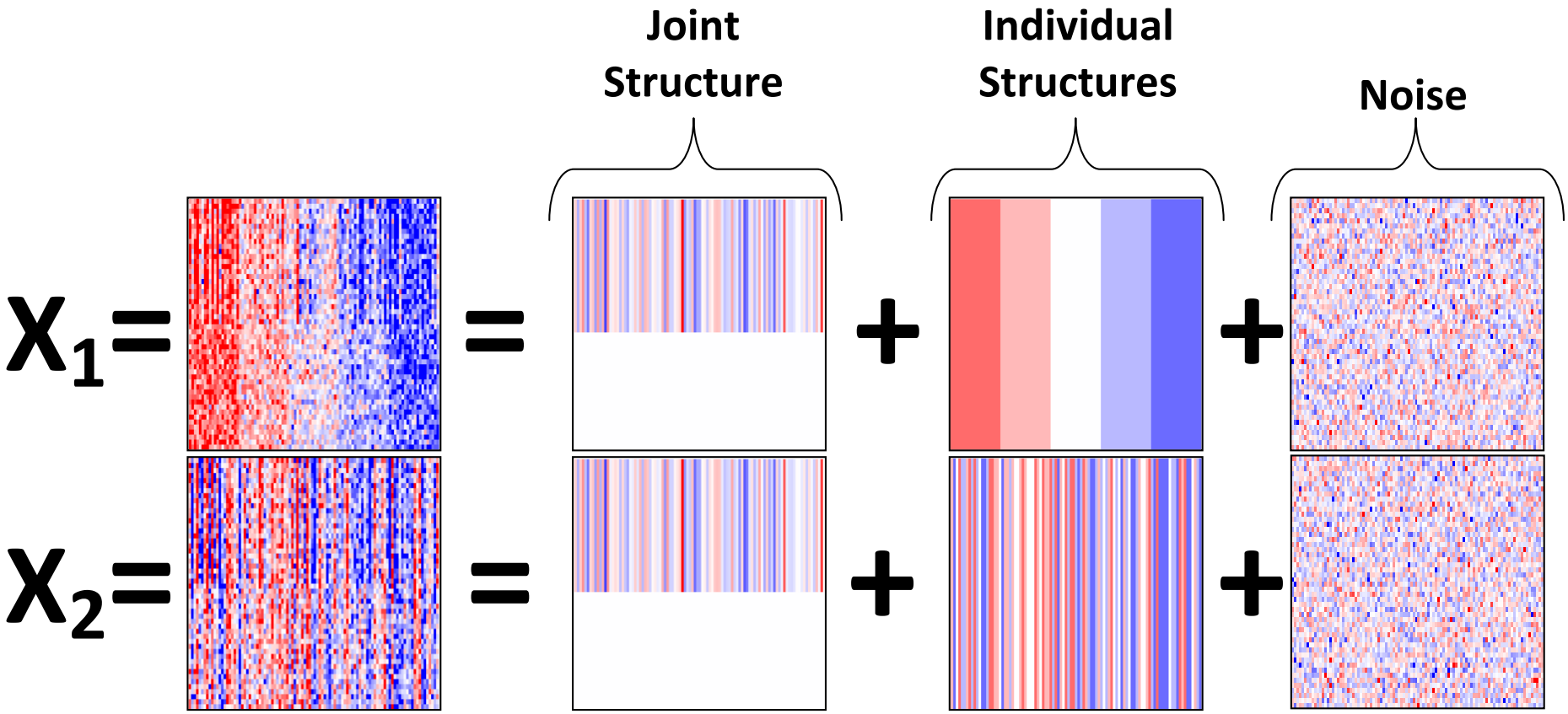
- Goals
  - Examine global associations across datatypes.
  - Identify sample patterns consistent across multiple datatypes.
  - Identify patterns unique to a particular datatype.

Toy Example: Two Datatypes

$X_1 =$ 

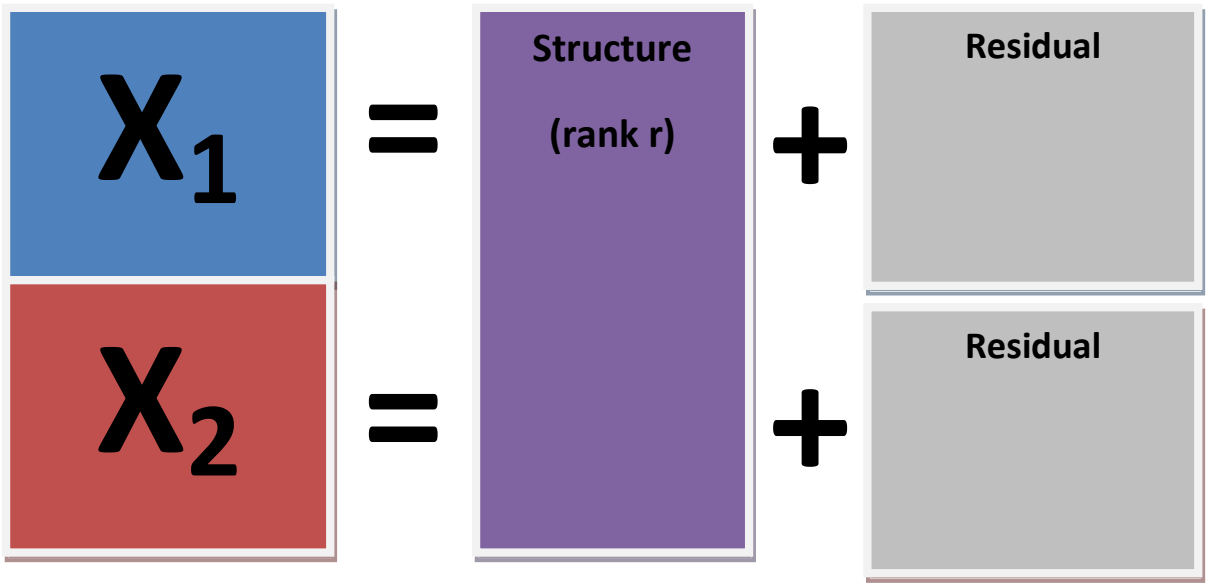
$X_2 =$ 

Toy Example: Two Datatypes

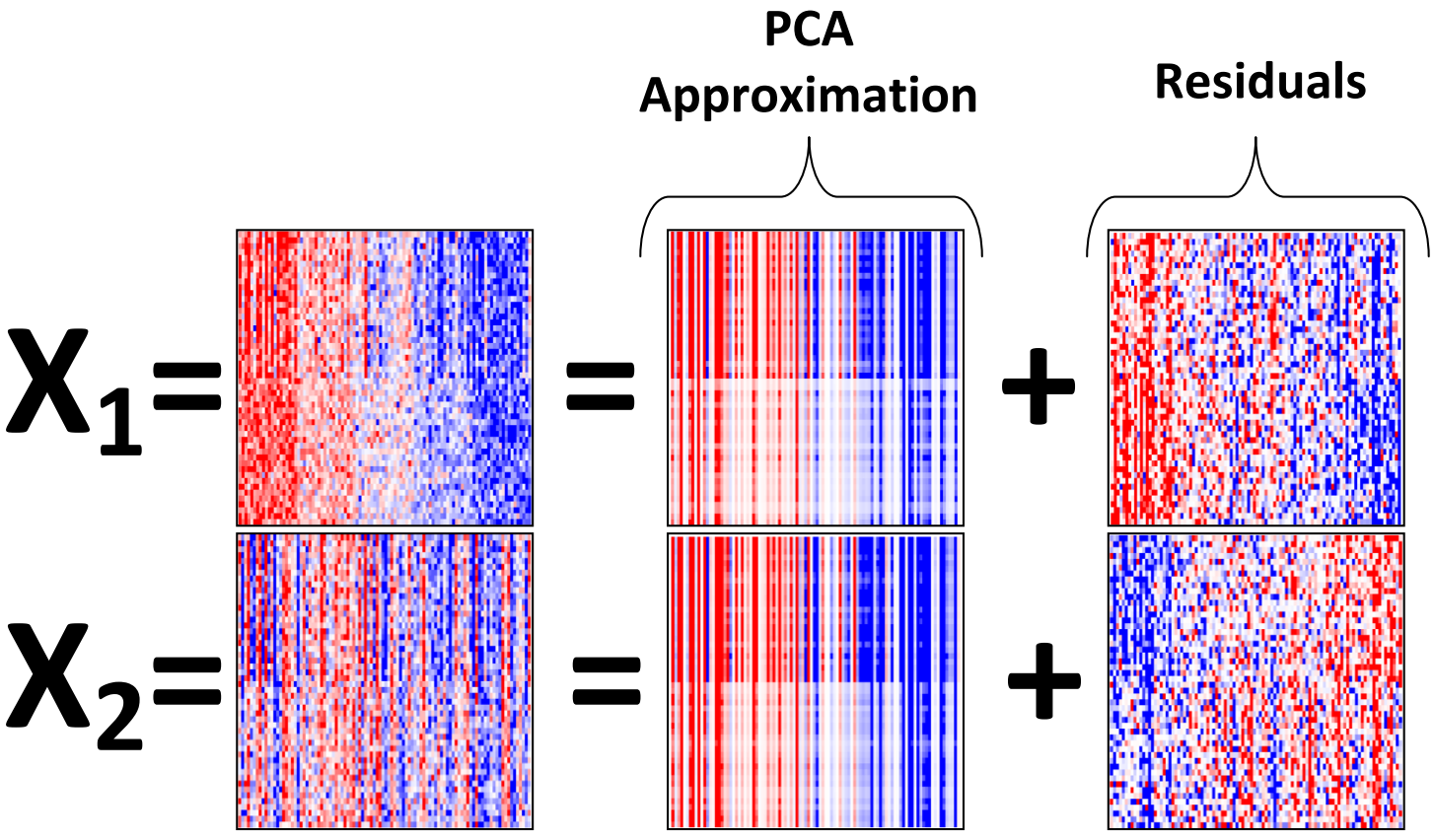


PCA Approximation

- PCA as a low rank approximation:

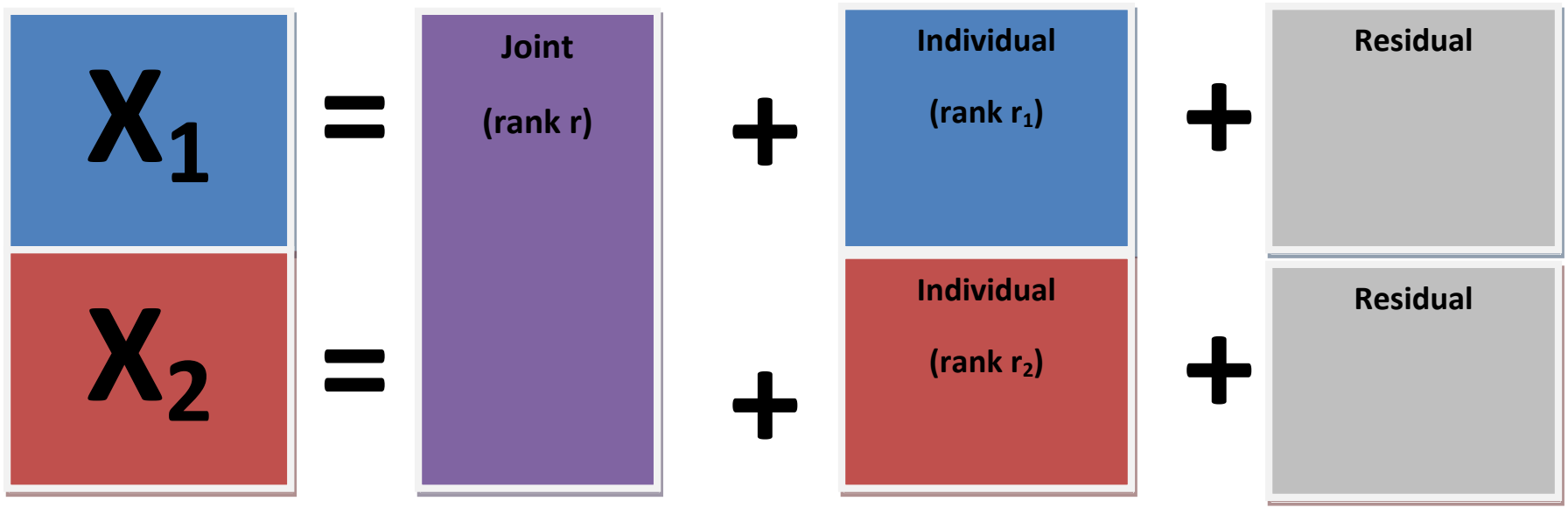


PCA Approximation ($r = 1$)

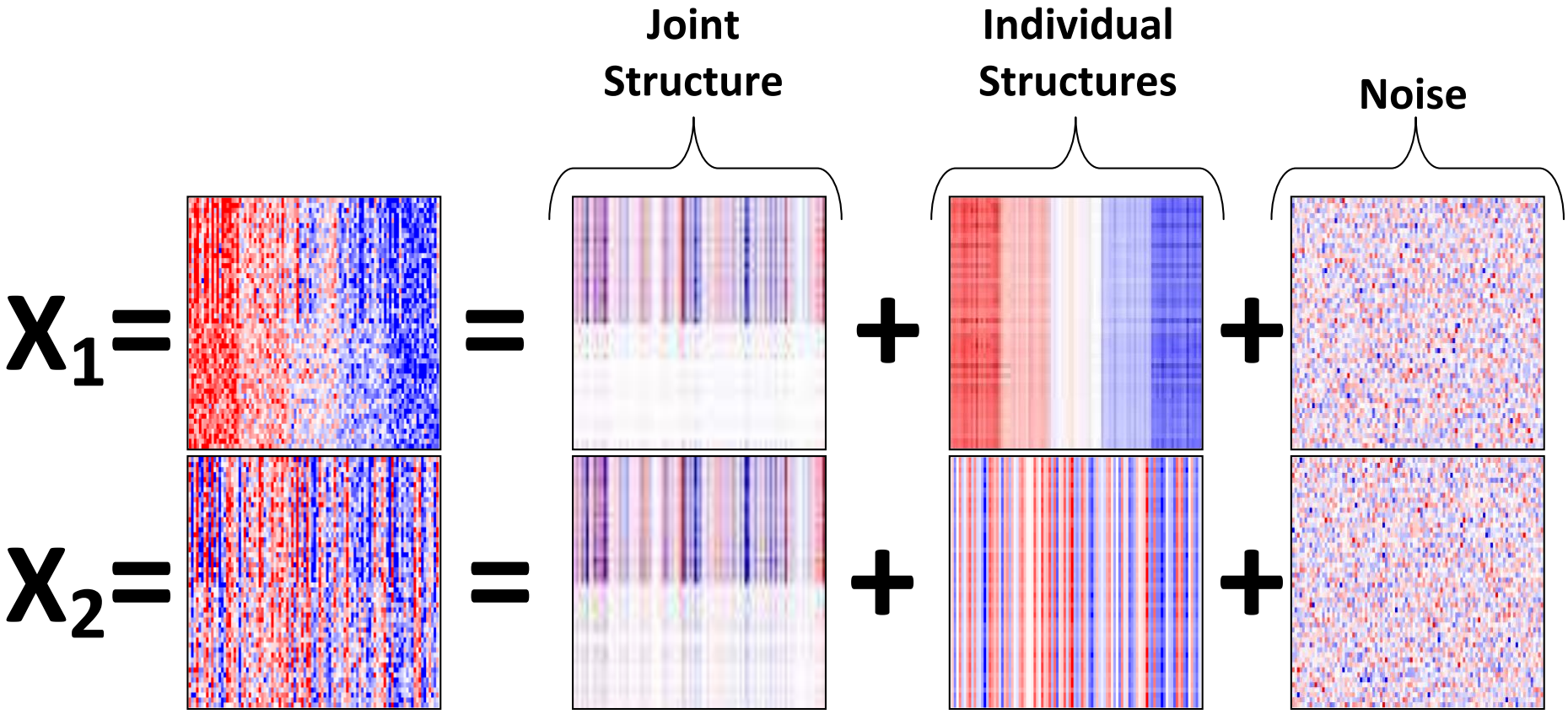


JIVE decomposition

- Joint and Individual Variation Explained (JIVE):

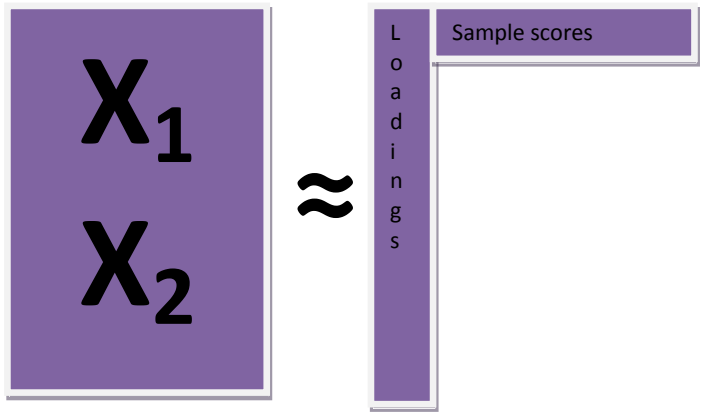


JIVE decomposition ($r = r_1 = r_2 = 1$)

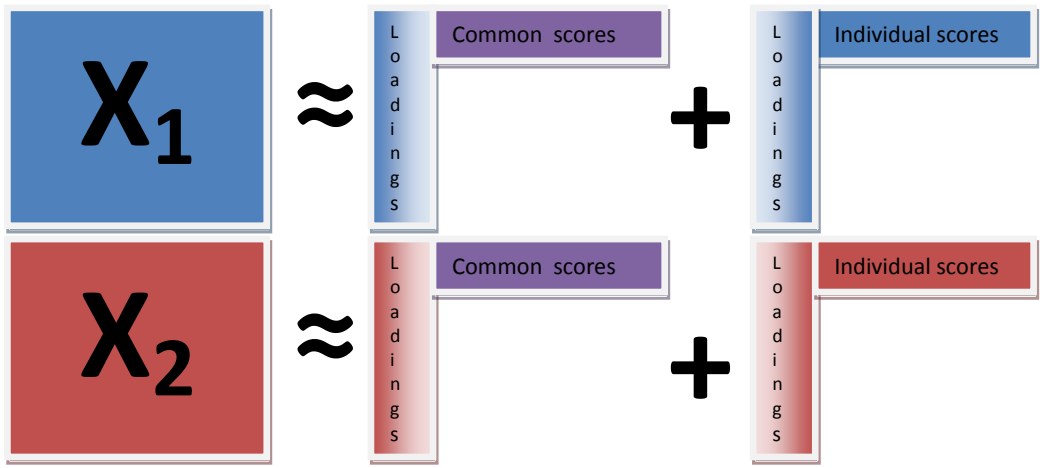


PCA vs JIVE

- PCA:



- JIVE:



JIVE decomposition

- Multiple datatypes $X_1, ..., X_k$ of dimension $p_1, ..., p_k$ on the same set of n samples.
- Decomposition:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} = \overbrace{\begin{bmatrix} J_1 \\ J_2 \\ \vdots \\ J_k \end{bmatrix}}^{J} + \overbrace{\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_k \end{bmatrix}}^{A} + \overbrace{\begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_k \end{bmatrix}}^{R}$$

- $J : p \times n$ is rank r.
- $A_i : p_i \times n$ are rank r_i.
- $R_i : p_i \times n$ are residual matrices.

JIVE decomposition (factorized form)

- Relationship to PCA:

$$\begin{aligned} X_1 &= \overbrace{U_1 S}^{J_1} + \overbrace{W_1 S_1}^{A_1} + R_1 \\ &\vdots \\ X_k &= U_k S + W_k S_k + R_k. \end{aligned}$$

  - S is an $r \times n$ score matrix explaining joint variation across datatypes.
  - U_i are $p_i \times r$ loading matrices.
  - S_i are $r_i \times n$ score matrices explaining unique variation.
  - W_i are $p_i \times r_i$ loading matrices.

Estimation

- Fixed ranks $r, r_1, \ldots, r_k$.
- Minimize sum of squared residuals $||R||_F^2$, where

$$R = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_k \end{bmatrix} = \begin{bmatrix} X_1 - J_1 - A_1 \\ X_2 - J_2 - A_2 \\ \vdots \\ X_k - J_k - A_k \end{bmatrix}.$$

Estimation

- Fixed ranks $r, r_1, \ldots, r_k$.
- Minimize sum of squared residuals $||R||_F^2$, where

$$R = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_k \end{bmatrix} = \begin{bmatrix} X_1 - J_1 - A_1 \\ X_2 - J_2 - A_2 \\ \vdots \\ X_k - J_k - A_k \end{bmatrix}.$$

- Iterative approach:
  - Fix J. Find $A_1, A_2, \ldots, A_k$ to minimize $||R||_F^2$
  - Fix $A_1, A_2, \ldots, A_k$. Find J to minimize $||R||_F^2$.

Estimation

- Fixed ranks $r, r_1, \ldots, r_k$.
- Minimize sum of squared residuals $||R||_F^2$, where

$$R = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_k \end{bmatrix} = \begin{bmatrix} X_1 - J_1 - A_1 \\ X_2 - J_2 - A_2 \\ \vdots \\ X_k - J_k - A_k \end{bmatrix}.$$

- Iterative approach:
  - Fix J. Find $A_1, A_2, \ldots, A_k$ to minimize $||R||_F^2$
  - Fix $A_1, A_2, \ldots, A_k$. Find J to minimize $||R||_F^2$.
- WLOG may enforce orthogonality of J and $A_1, \ldots, A_k$:

$$JA' = 0_{p \times p}.$$

Dimension Reducing Shortcut

- Given singular value decompositions

$$\begin{aligned} SVD(X_1) &= U_1 \Lambda_1 V_1^T \\ &\vdots \\ SVD(X_k) &= U_k \Lambda_k V_k^T. \end{aligned}$$

  define $X_i^{\perp} = \Lambda_i V_i^T$ for each $i = 1, ..., k$.

- Then, $X_i^{\perp}$ are $n \times n$ (assuming $p_i > n$) and preserve covariance and Euclidian distance between columns (samples).

- Performing iterative process on $X_i^{\perp}$ instead of X_i can be substantially faster and gives identical results.

Key Issue: Scaling of Individual Datasets

- $X_1, X_2, \ldots, X_k$ of different scale and dimension.

Key Issue: Scaling of Individual Datasets

- $X_1, X_2, \ldots, X_k$ of different scale and dimension.
- Suggest centering and scaling by total variation.
  - Subtract mean from each row: $X_i \to X_i^{\text{centered}}$
  - Divide by $||X_i^{\text{centered}}||_F$:

$$X_i^{\text{scaled}} = \frac{X_i^{\text{centered}}}{||X_i^{\text{centered}}||_F}$$

Key Issue: Scaling of Individual Datasets

- $X_1, X_2, \ldots, X_k$ of different scale and dimension.
- Suggest centering and scaling by total variation.
  - Subtract mean from each row: $X_i \to X_i^{\text{centered}}$
  - Divide by $||X_i^{\text{centered}}||_F$:

$$X_i^{\text{scaled}} = \frac{X_i^{\text{centered}}}{||X_i^{\text{centered}}||_F}$$

- Gives each dataset same total signal power.

Rank Selection: Permutation Testing Approach

- Extends Peres-Neto et al. (2005)...
- To estimate rank of joint structure
  - Compare
    - Singular values of concatenated matrix
    - Singular values after permuting samples within each datatype.
- To estimate rank of individual structure
  - Compare:
    - Singular values of individual matrix
    - Singular values after permuting samples within each row.

The Cancer Genome Atlas (TCGA) Data

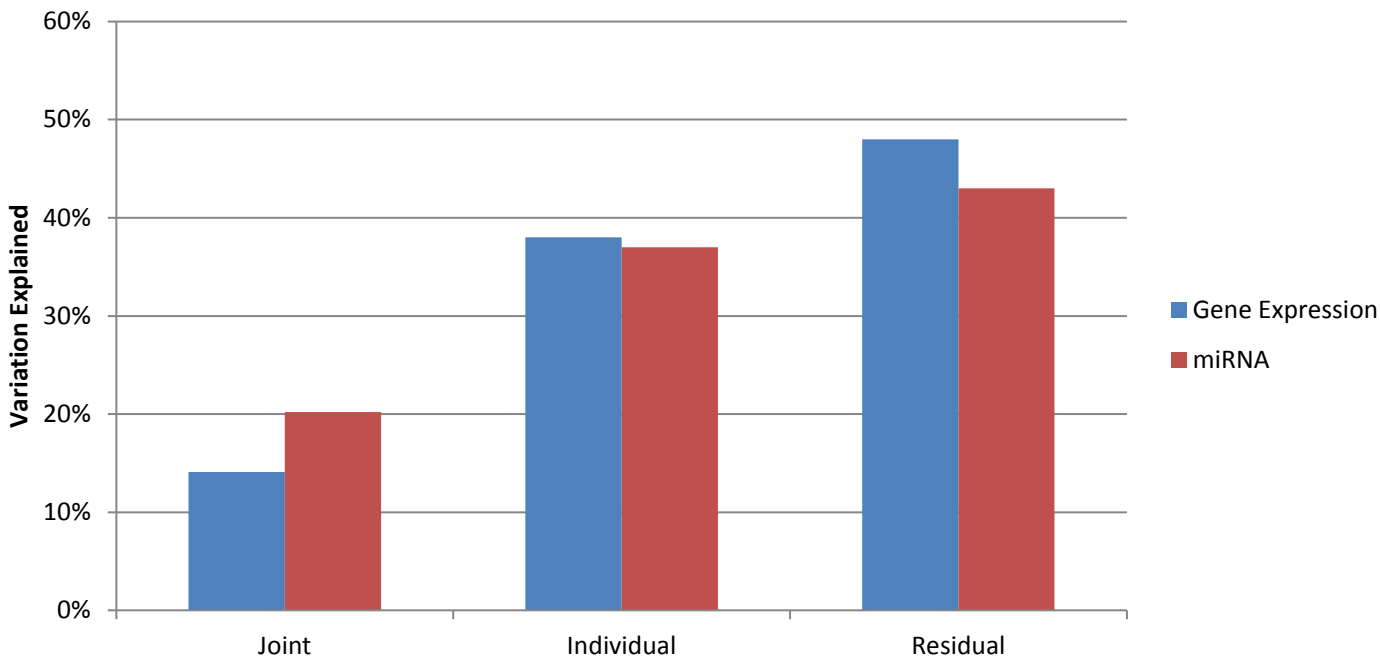
- Multiple kinds of data for the same set of 348 breast cancer tumors, from TCGA.
  - Gene expression data (17814 genes)
  - miRNA data (655 miRNAs)
  - Copy number data (200,000 probes / 19,780 genes)
  - Methylation data (21,986 CG regions)
  - Mutation data (12,481 genes)
  - Protein data
- Tumors classified into 5 subtypes based on the expression data:
  - Basal (66 samples)
  - Her2 (42 samples)
  - Luminal A (154 samples)
  - Luminal B (81 samples)
  - Normal (5 samples)

The Cancer Genome Atlas (TCGA) Data

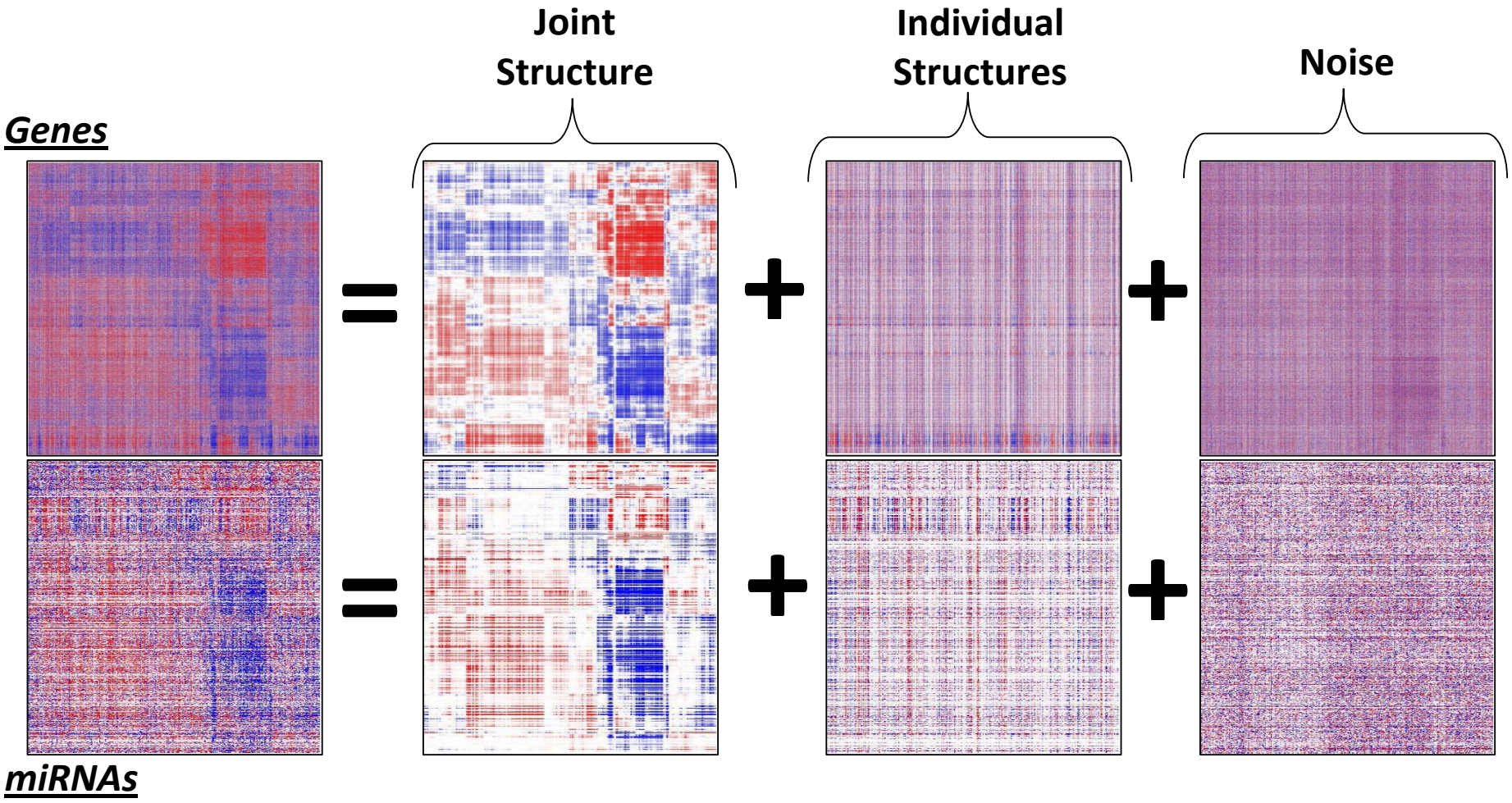
- Multiple kinds of data for the same set of 348 breast cancer tumors, from TCGA.
  - **Gene expression data (17814 genes)**
  - **miRNA data (655 miRNAs)**
  - Copy number data (200,000 probes / 19,780 genes)
  - Methylation data (21,986 CG regions)
  - Mutation data (12,481 genes)
  - Protein data
- Tumors classified into 5 subtypes based on the expression data:
  - Basal (66 samples)
  - Her2 (42 samples)
  - Luminal A (154 samples)
  - Luminal B (81 samples)
  - Normal (5 samples)

JIVE application: Gene expression and miRNA

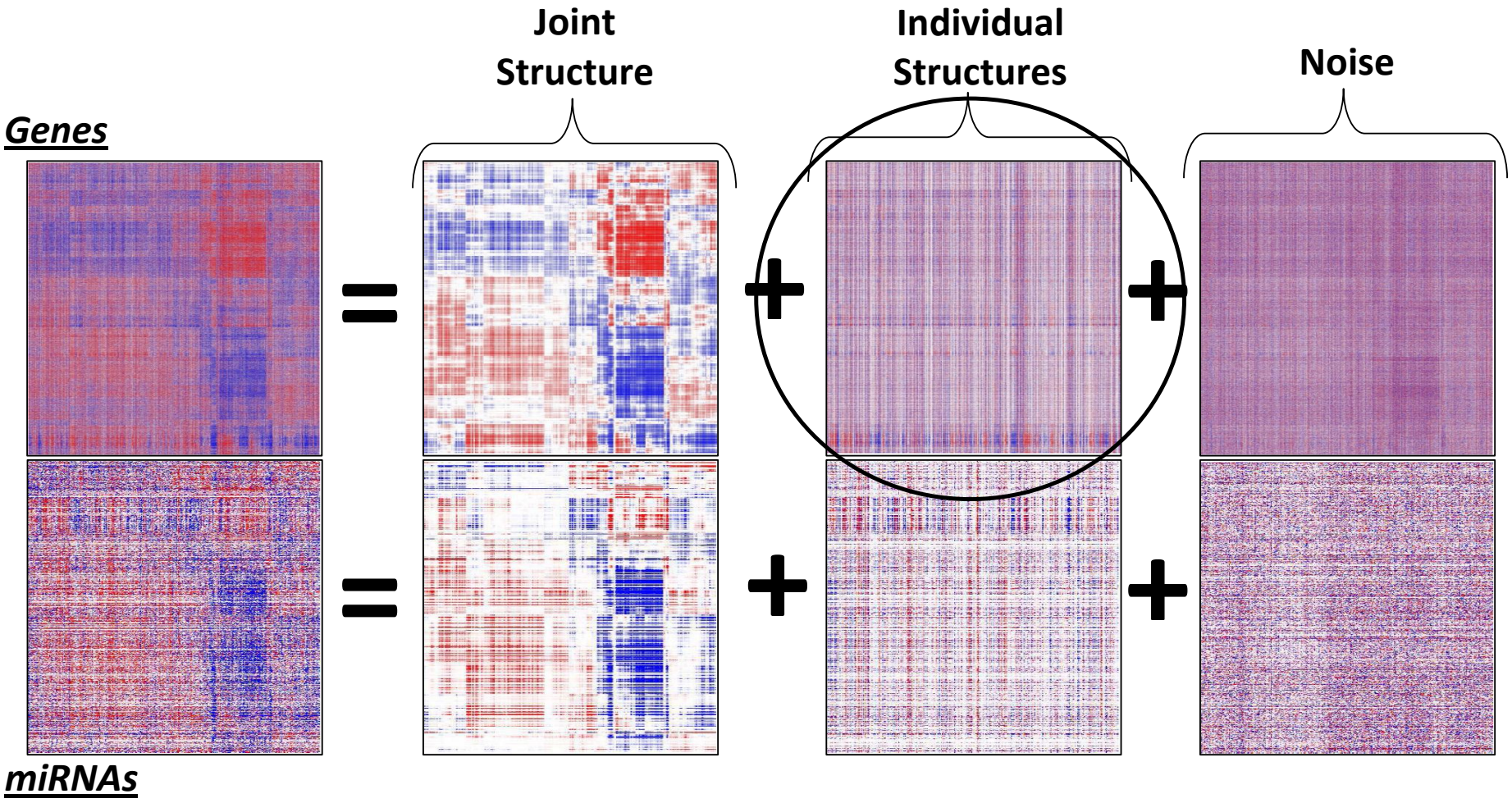
- Applied JIVE decomposition to Gene expression and miRNA.
- Permutation testing identifies
  - Rank 4 joint structure
  - Rank 22 structure individual to gene expression
  - Rank 9 structure individual to miRNA
- Variation decomposition:



JIVE Estimates

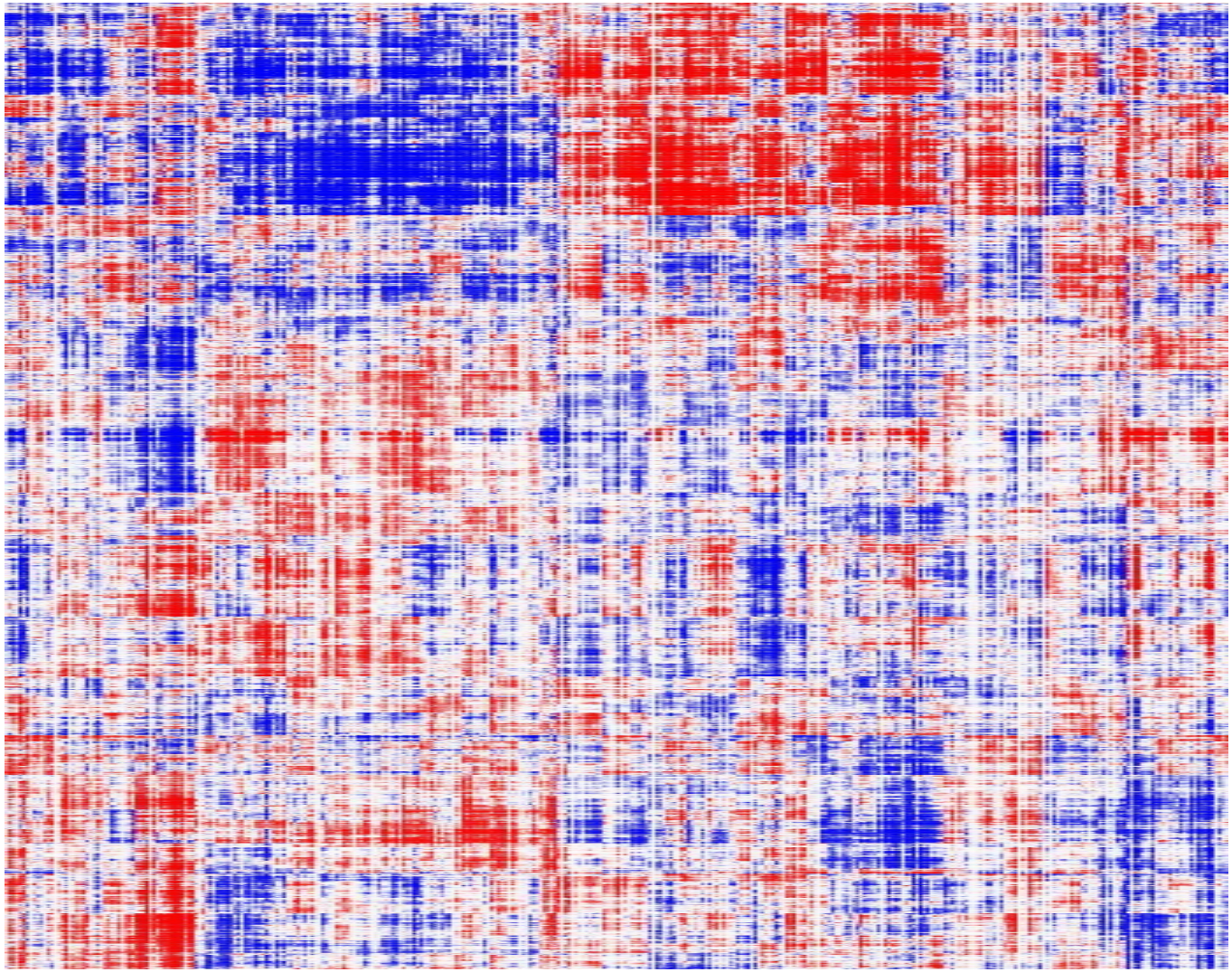


JIVE Estimates

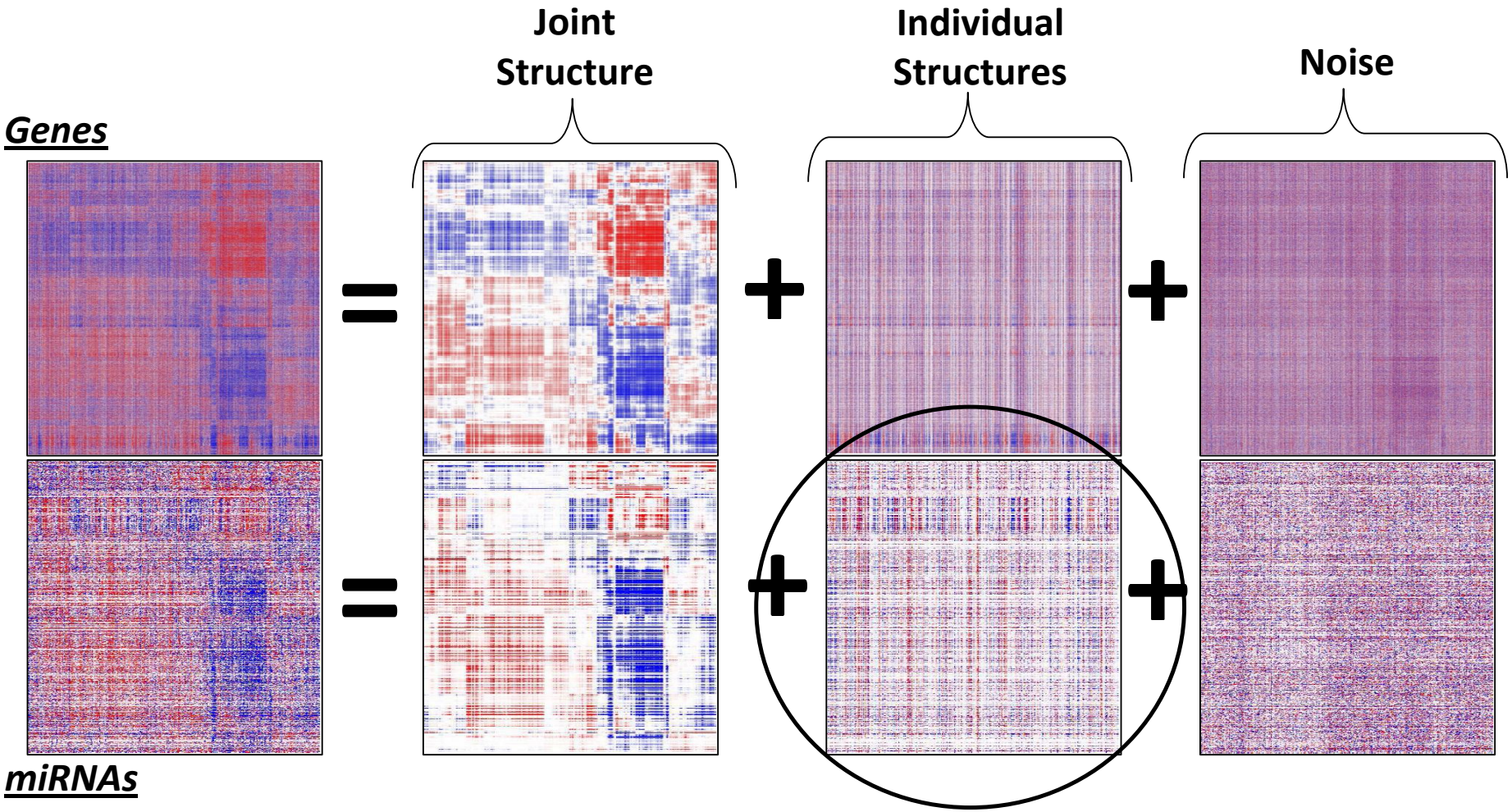


JIVE Estimates

- Gene individual (reorder rows and columns)

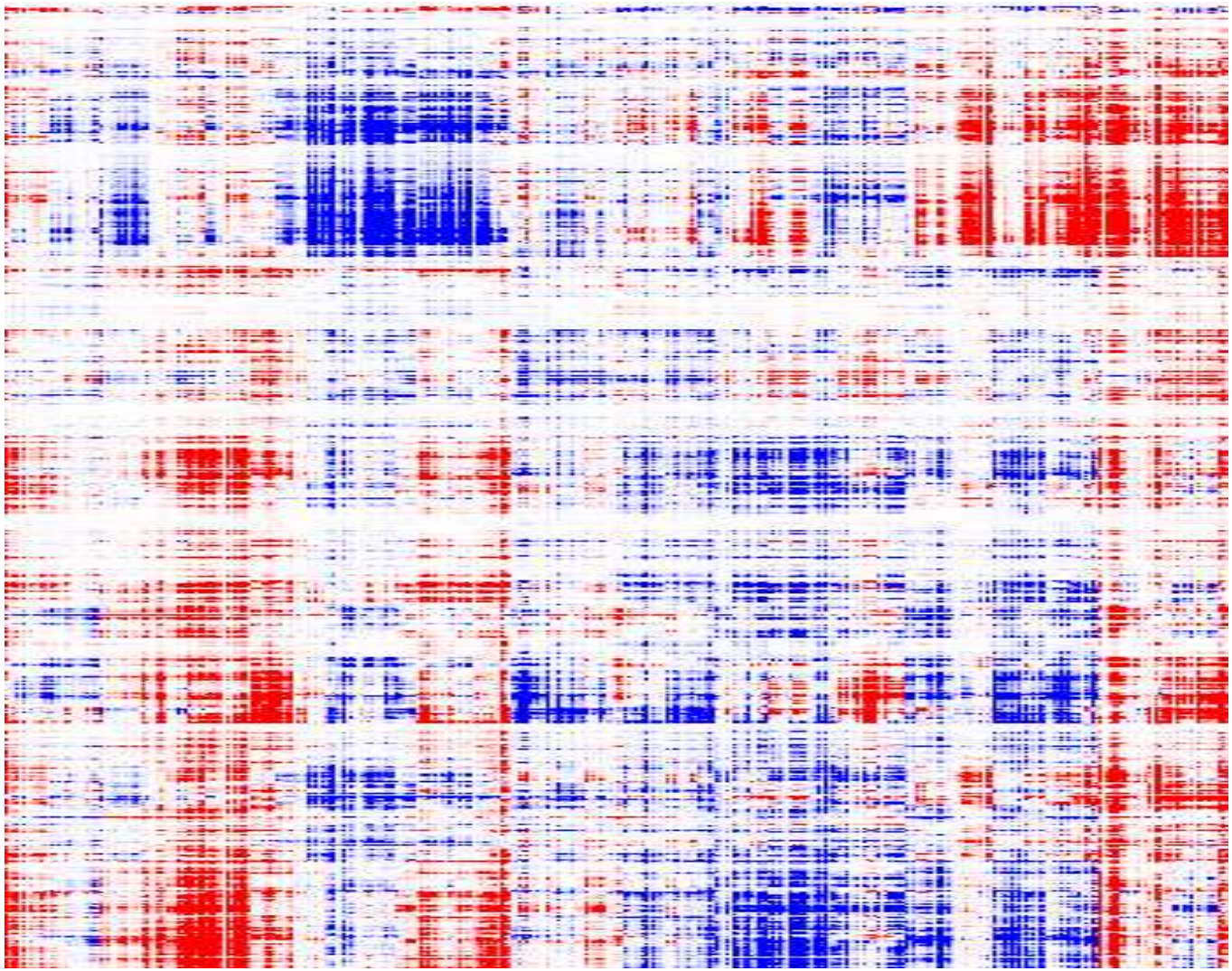


JIVE Estimates

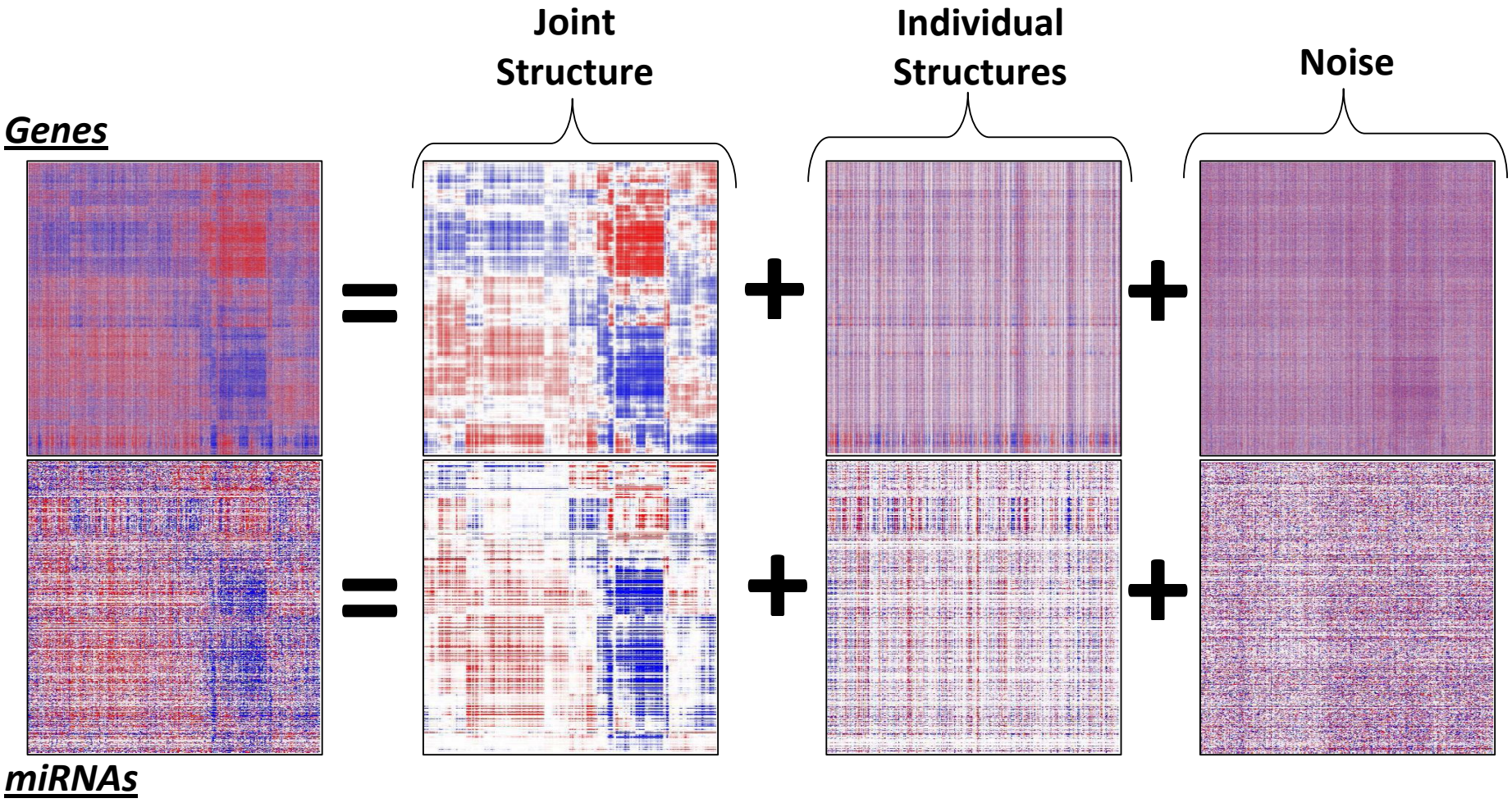


JIVE Estimates

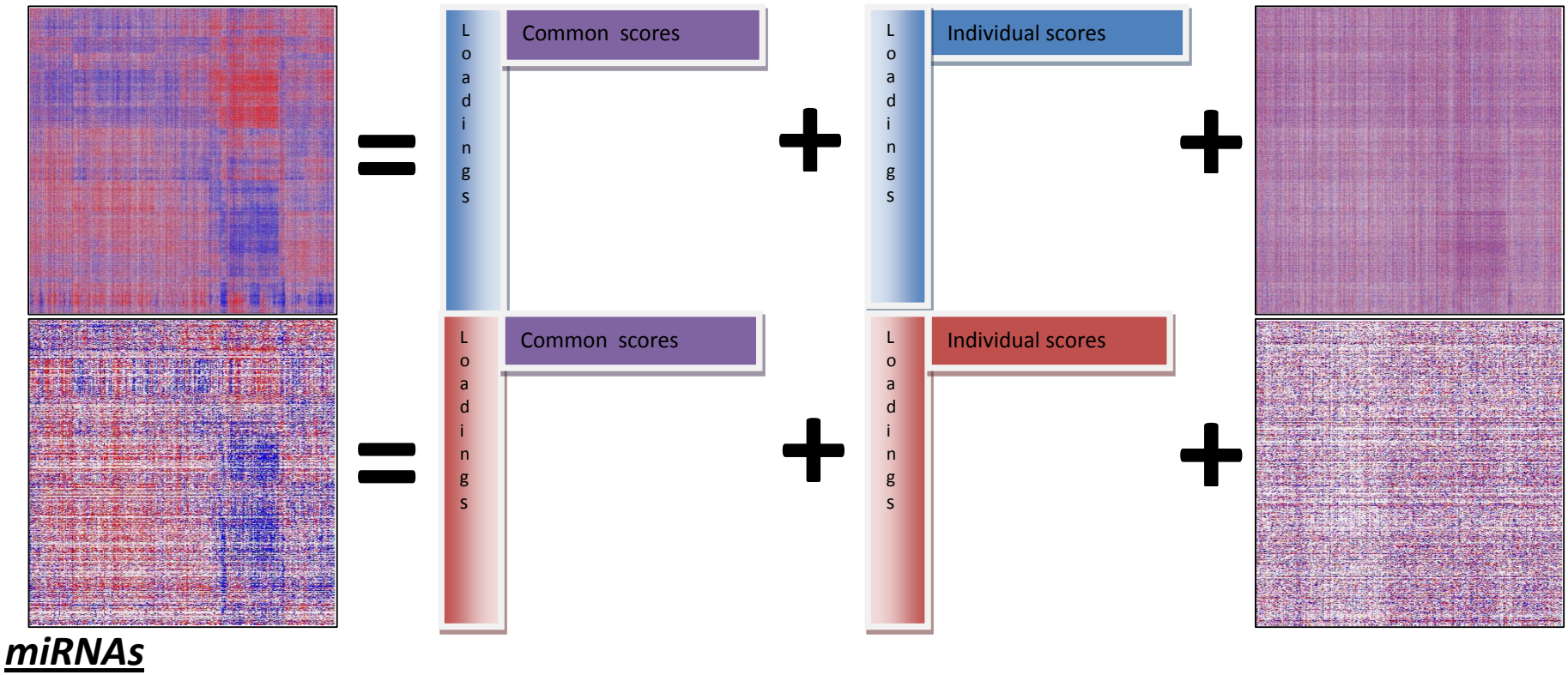
- miRNA individual (reorder rows and columns)



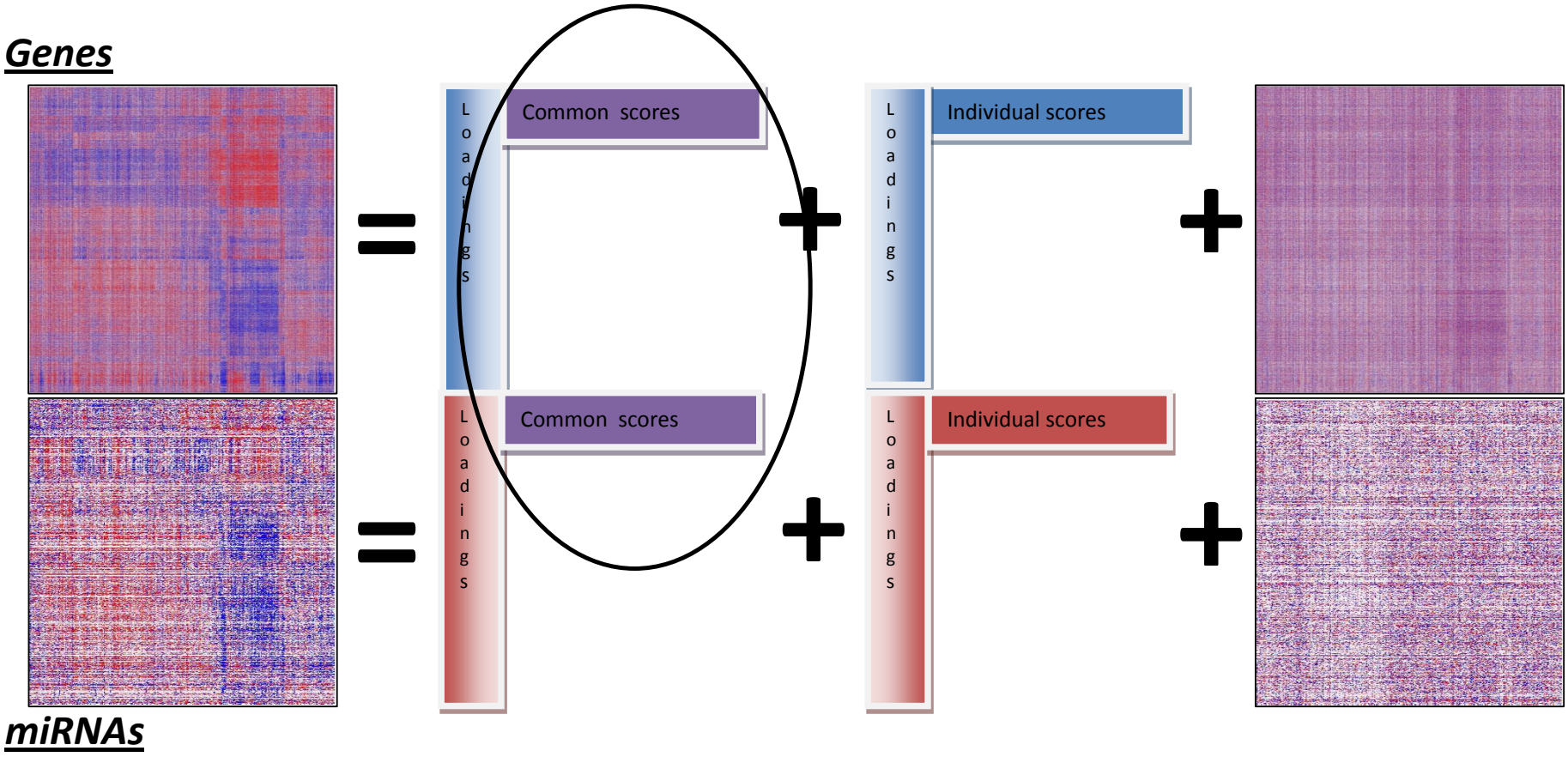
JIVE Estimates



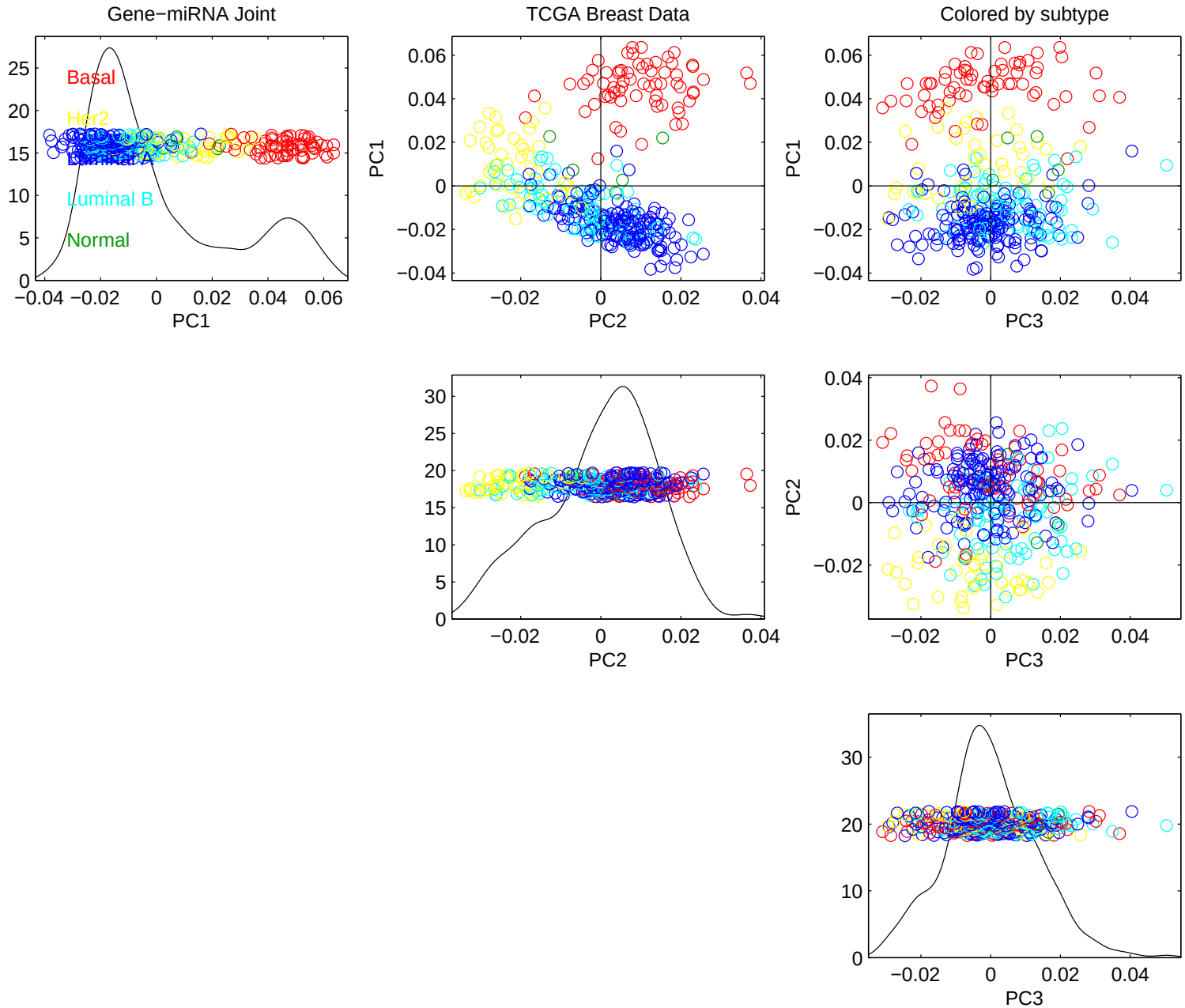
JIVE Estimates (factorized)



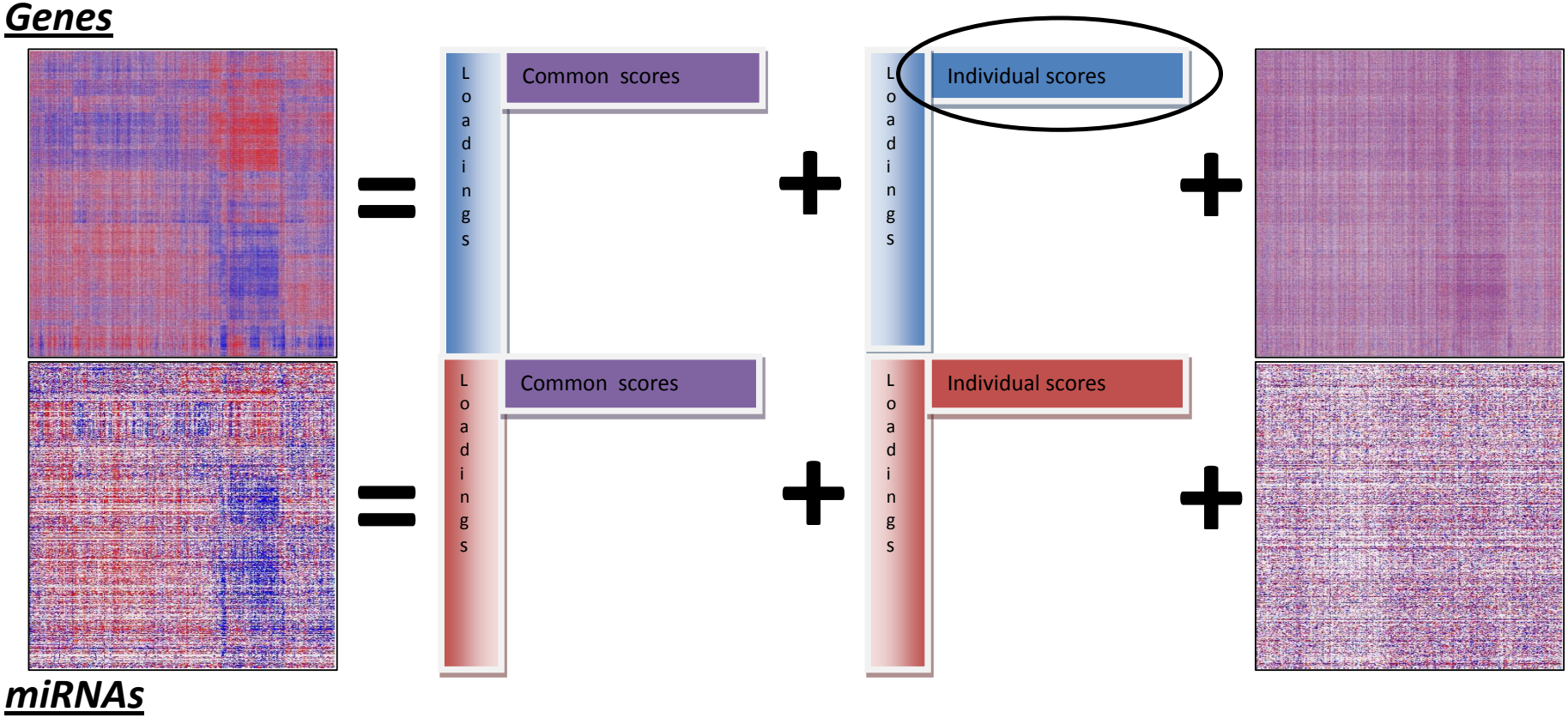
JIVE Estimates (factorized)



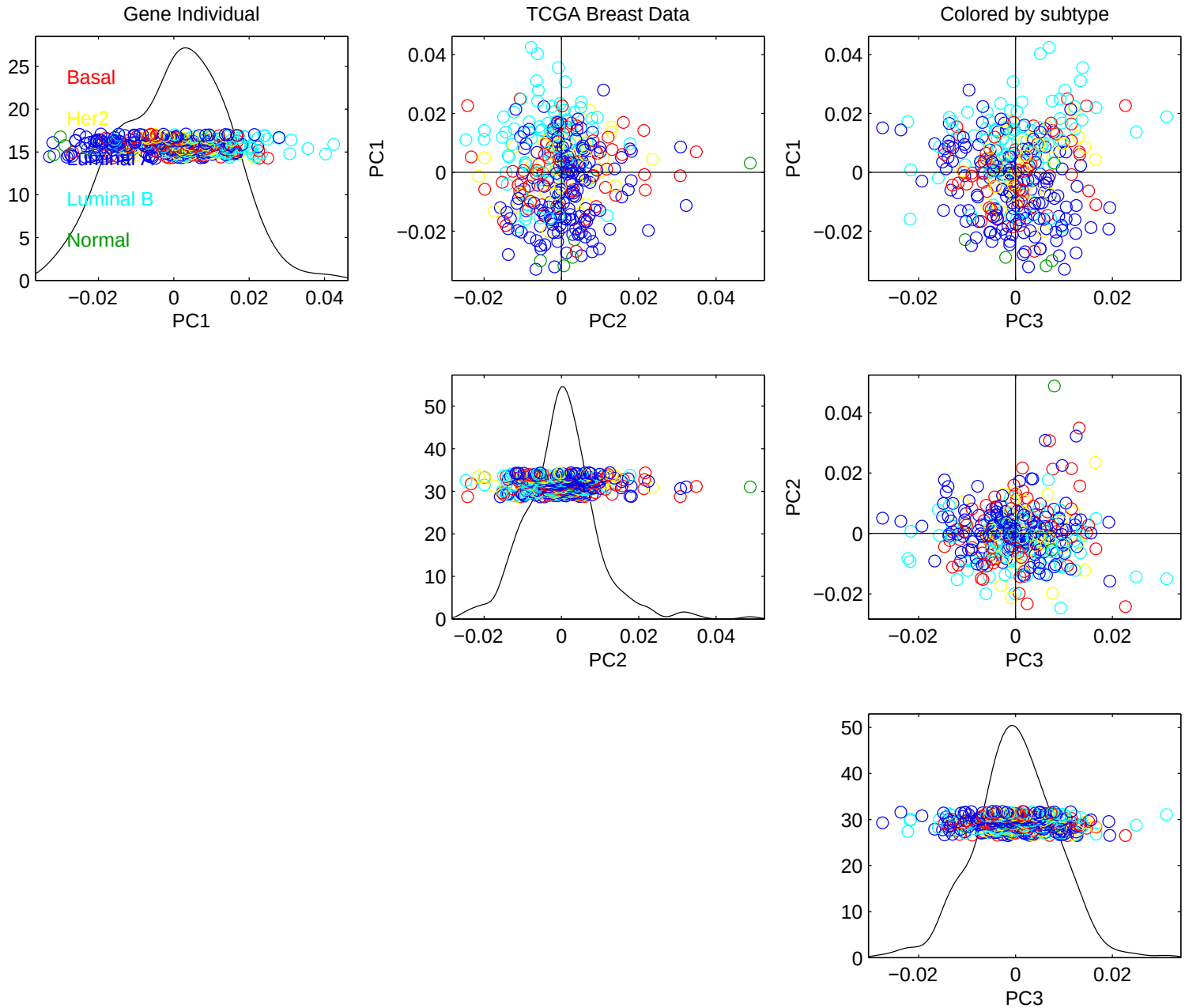
Joint PCs


Gene−miRNA Joint
TCGA Breast Data
Colored by subtype
Basal
Her2
Luminal A
Luminal B
Normal
PC1
PC2
PC3

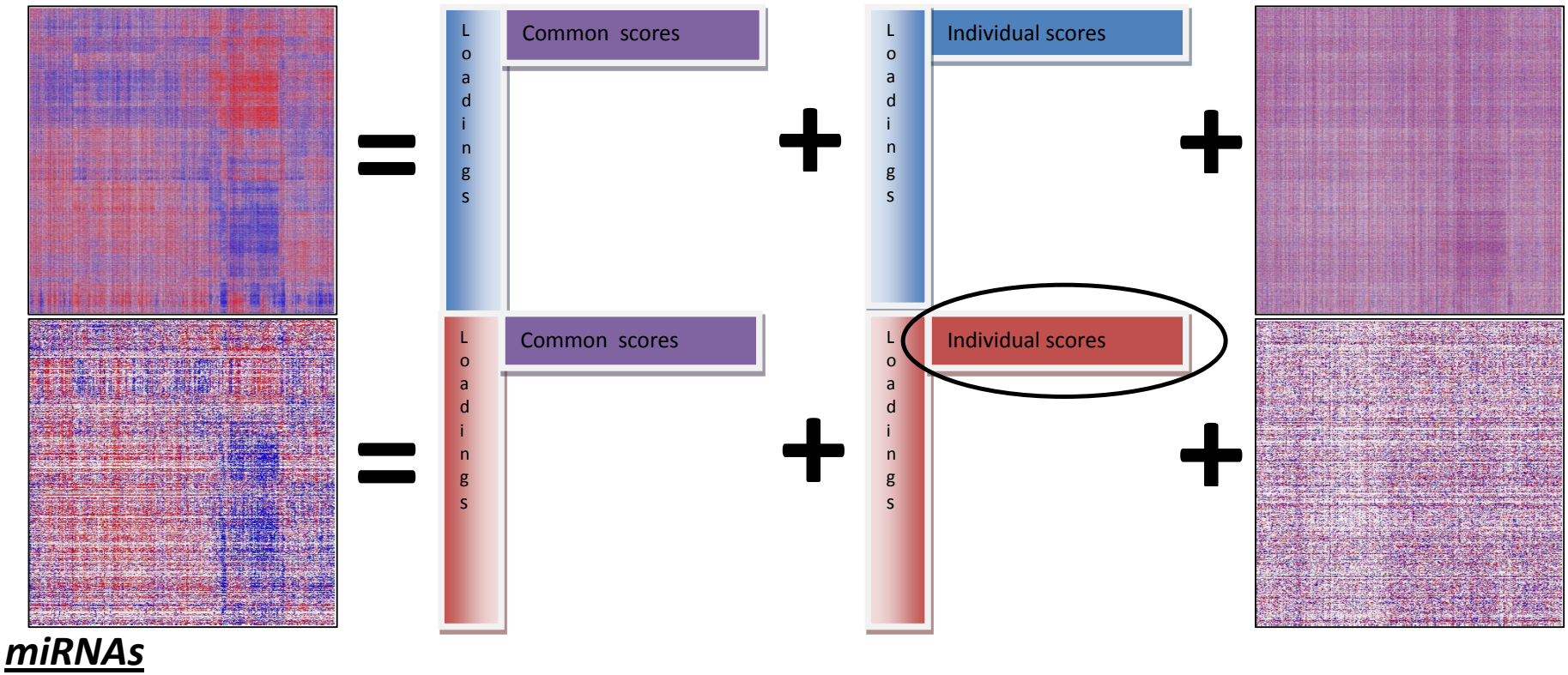
JIVE Estimates (factorized)



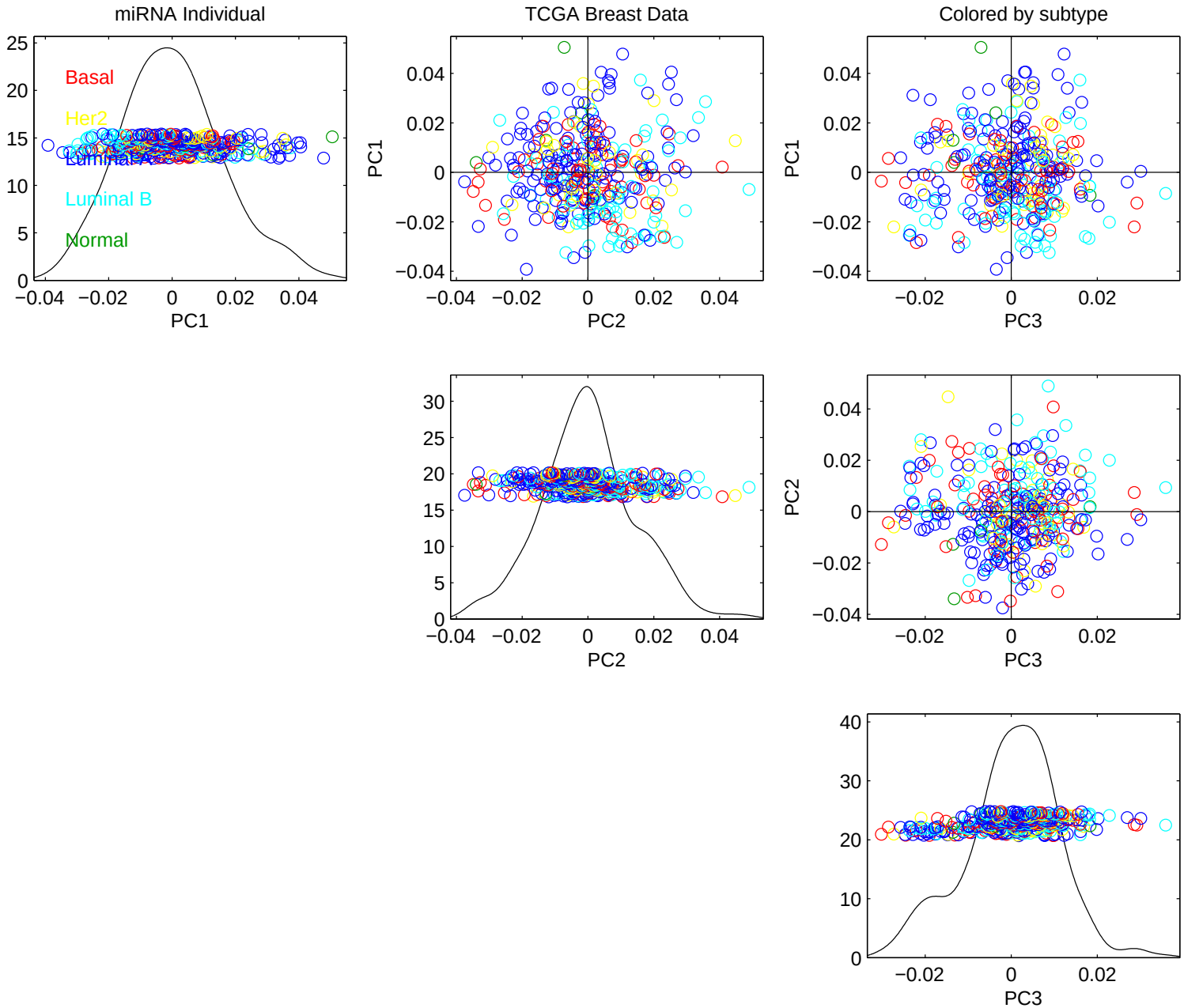
Individual PCs: Expression



JIVE Estimates (factorized)



Individual PCs: miRNA



Variable sparsity

- Important signal only on a subset of variables
- Motivates use of a *sparse* model
- Can aid results and interpretation.

Variable Sparsity

- Penalized sum-of-squares criterion

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

where Pen is a penalty designed to induce sparsity in the loading vectors and λ, λ_i are weights.

Variable Sparsity

- Penalized sum-of-squares criterion

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

  where Pen is a penalty designed to induce sparsity in the loading vectors and λ, λ_i are weights.

- E.g, Pen may be an L_1 penalty, corresponding to the Lasso:

$$\text{Pen}(U) = \sum |u_{ij}|.$$

Variable Sparsity

- Penalized sum-of-squares criterion

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

  where Pen is a penalty designed to induce sparsity in the loading vectors and λ, λ_i are weights.

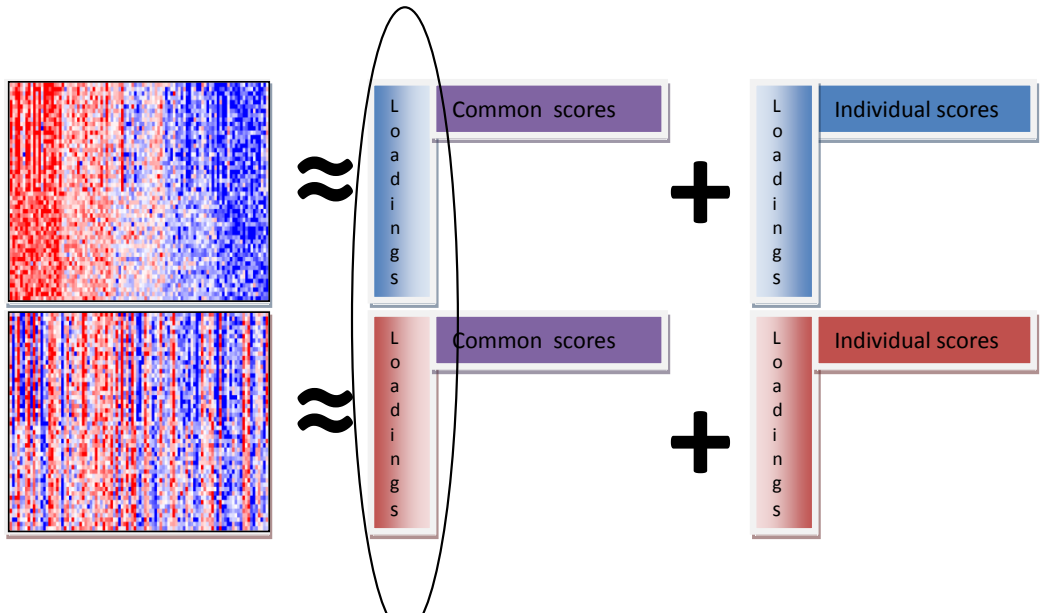
- E.g, Pen may be an L_1 penalty, corresponding to the Lasso:

$$\text{Pen}(U) = \sum |u_{ij}|.$$

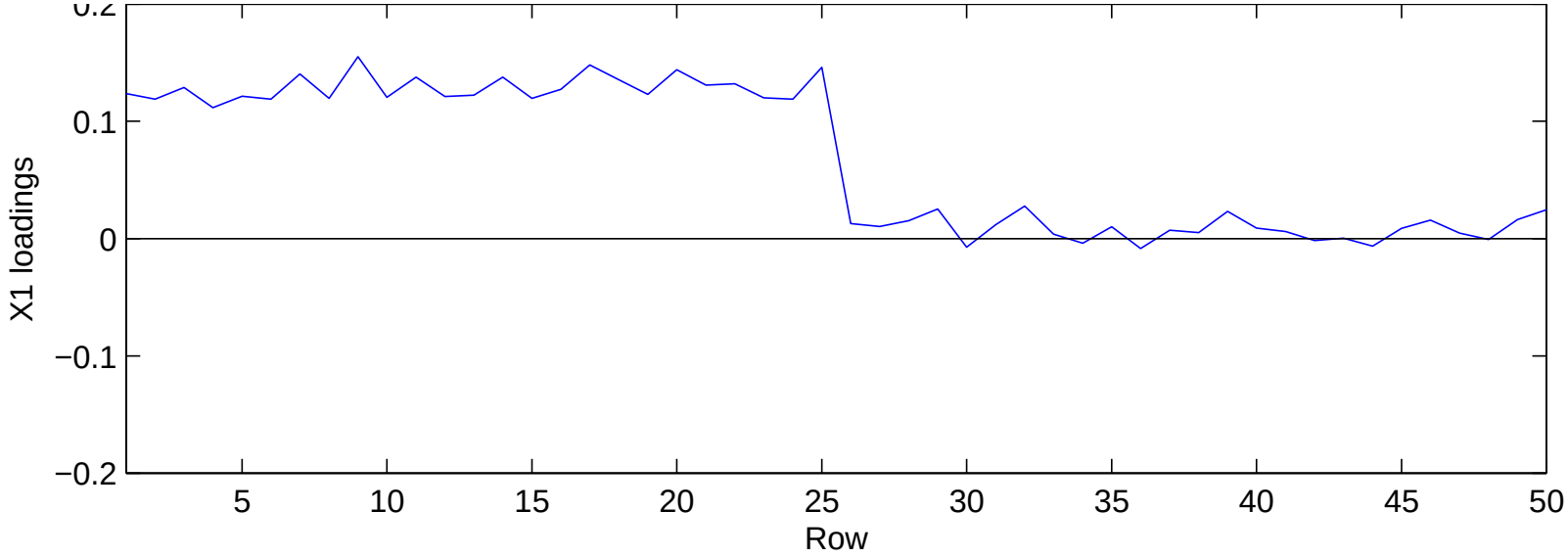
- Iterative approach:
  - Fix U, S: Find W_i, S_i to minimize $||R_i||_F^2 - \lambda_i \, \text{Pen}(W_i)$, for each $i = 1, ..., k$.
  - Fix $W_1, ..., W_k, S_1, ..., S_k$: Find U, S to minimize $||R||_F^2 - \lambda \text{Pen}(U)$.

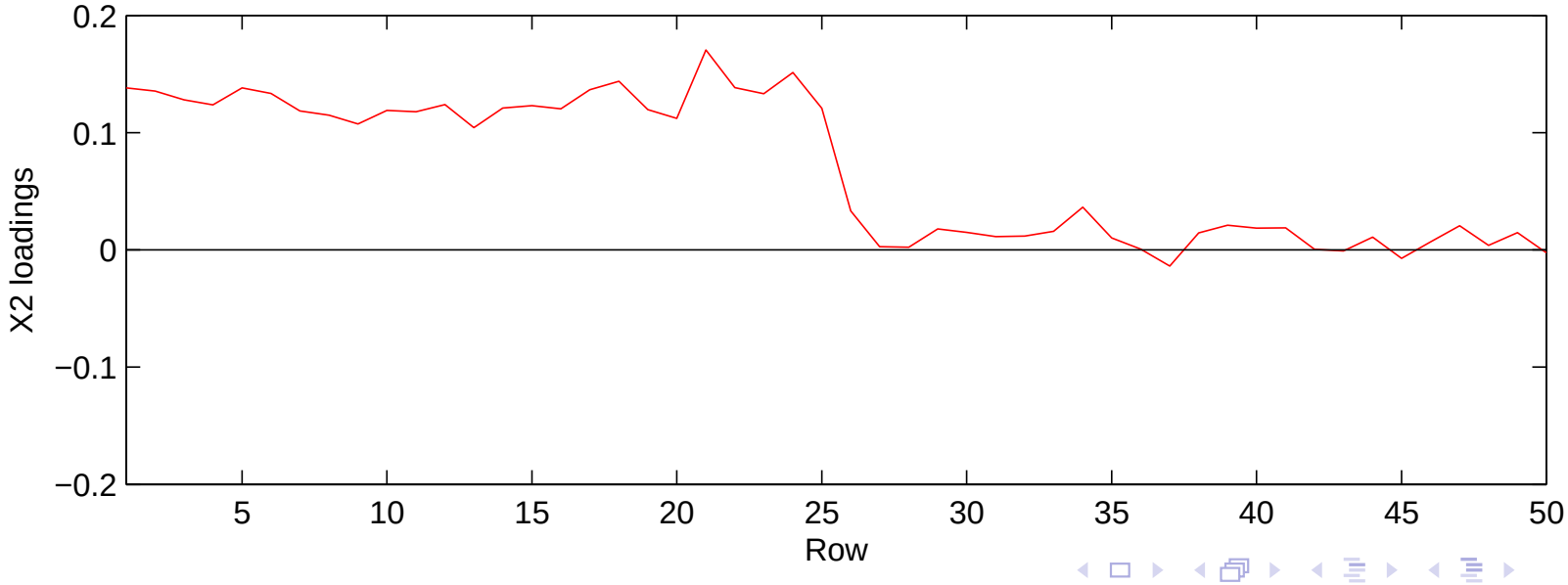
Sparsity Illustration

- JIVE:

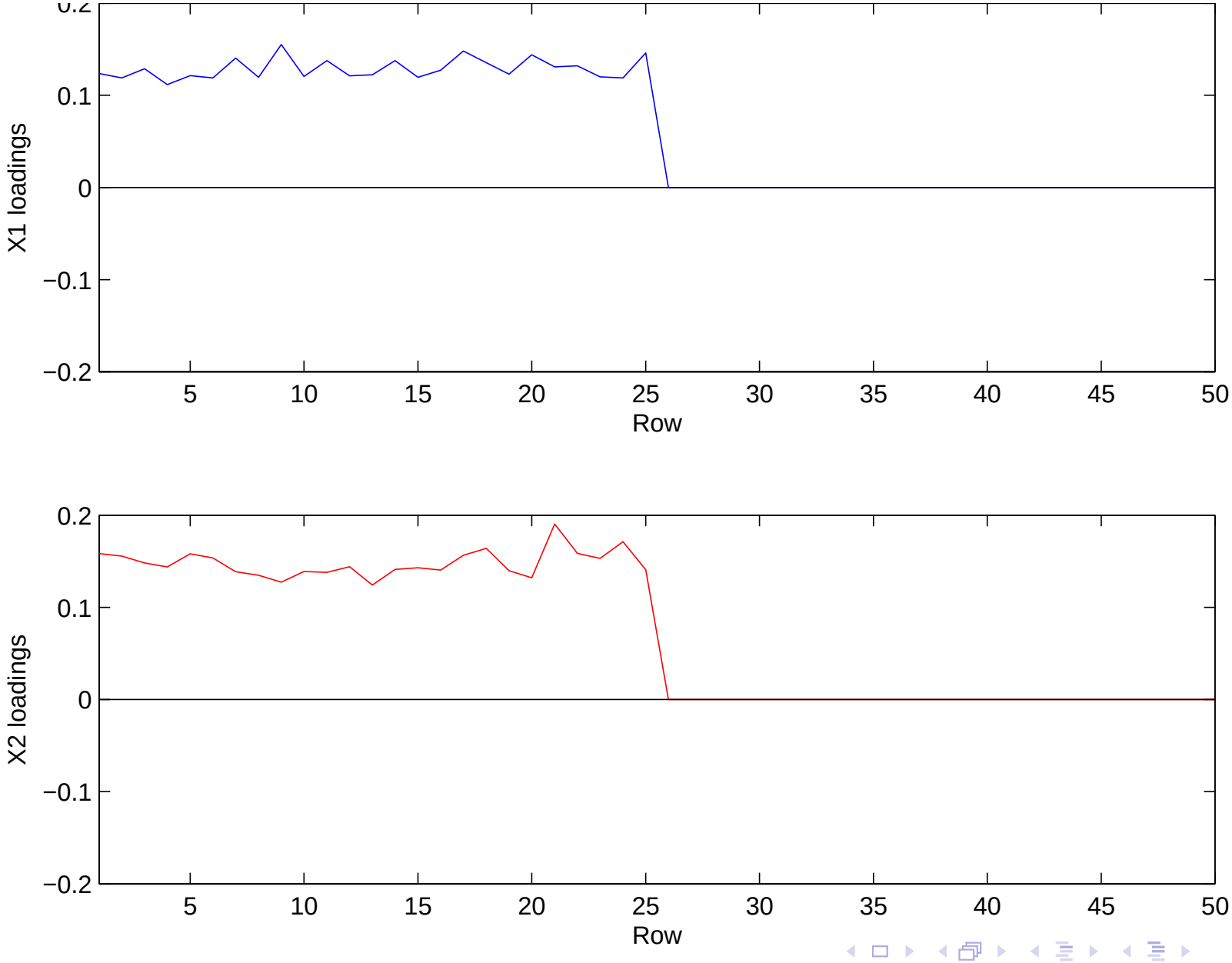


Joint component row loadings (without sparsity)



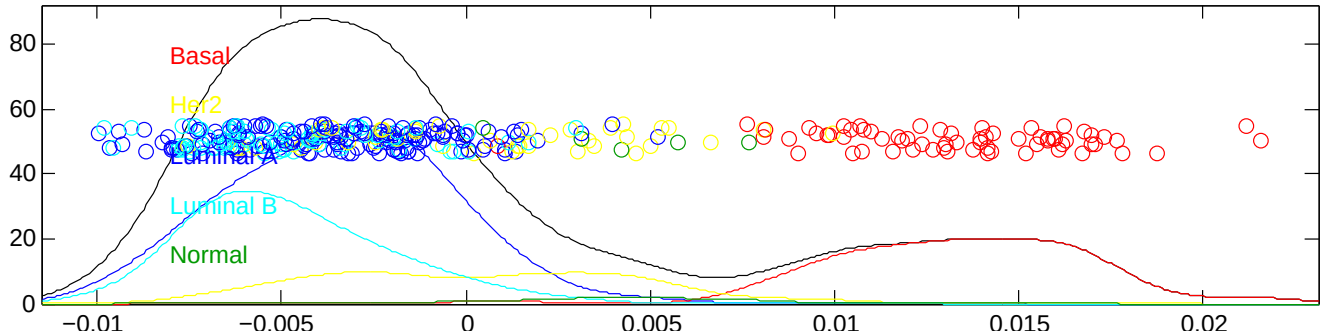


Joint component row loadings (with sparsity)



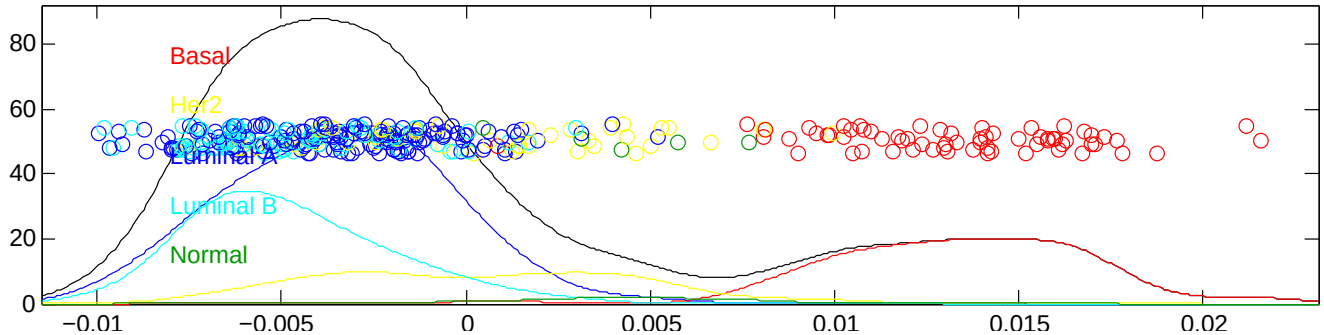
Gene-miRNA Sparse JIVE

- First "Sparse" joint component sample scores:

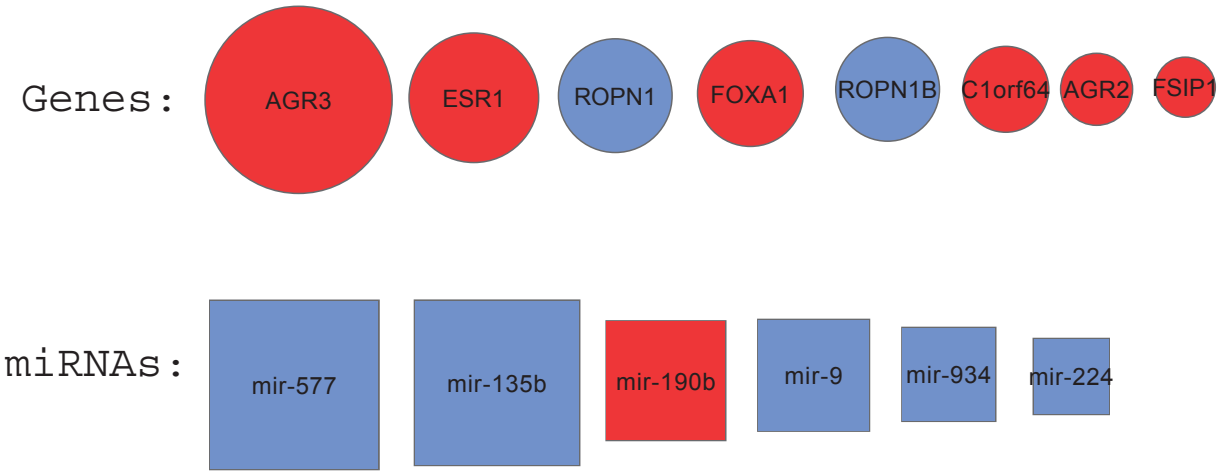


Gene-miRNA Sparse JIVE

- First "Sparse" joint component sample scores:



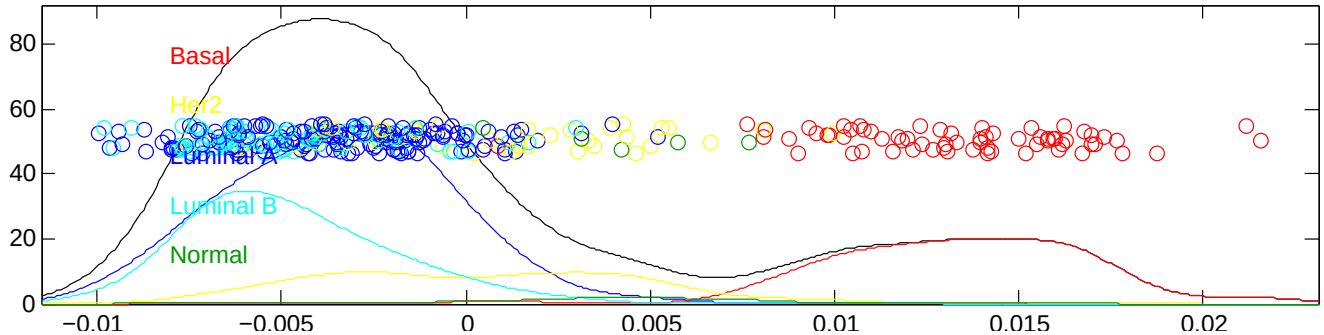
- Genes and miRNAs with non-zero loadings:



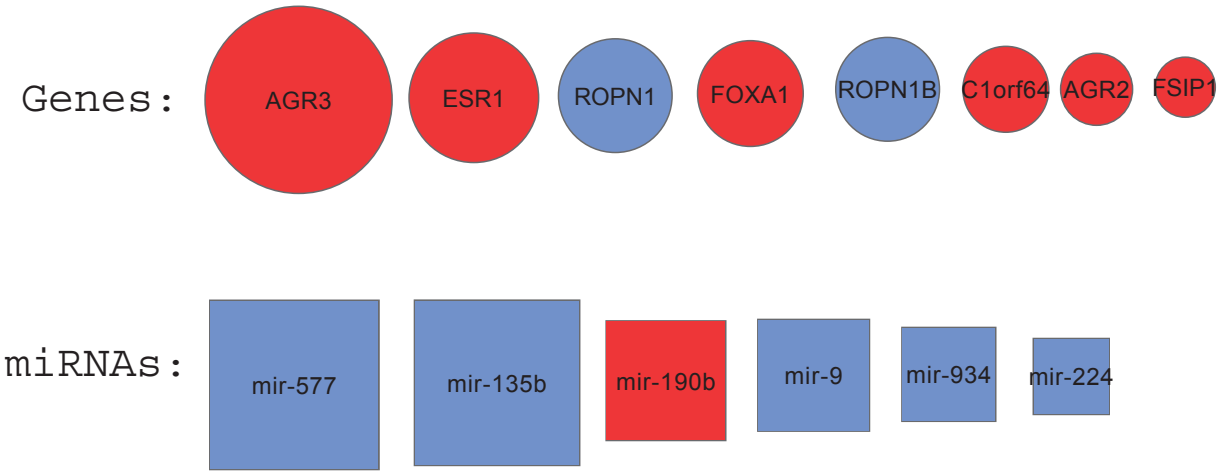
- red: positive loading; blue: negative loading

Gene-miRNA Sparse JIVE

- First "Sparse" joint component sample scores:



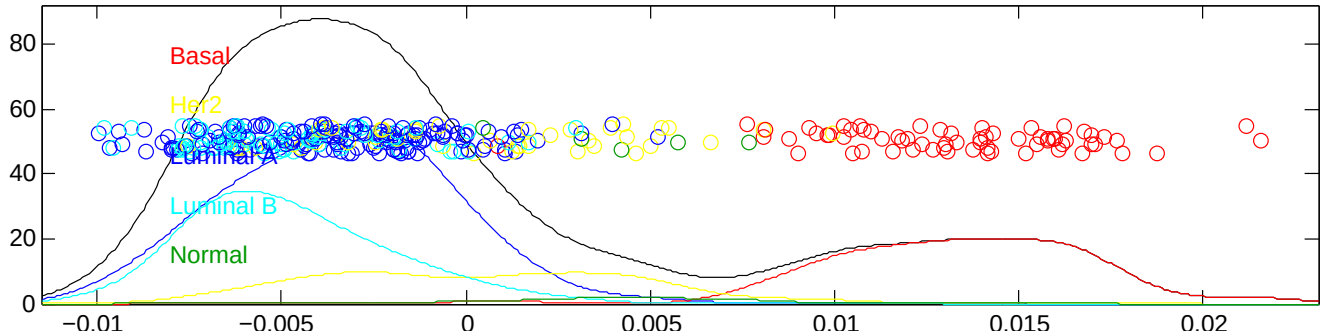
- Genes and miRNAs with non-zero loadings:



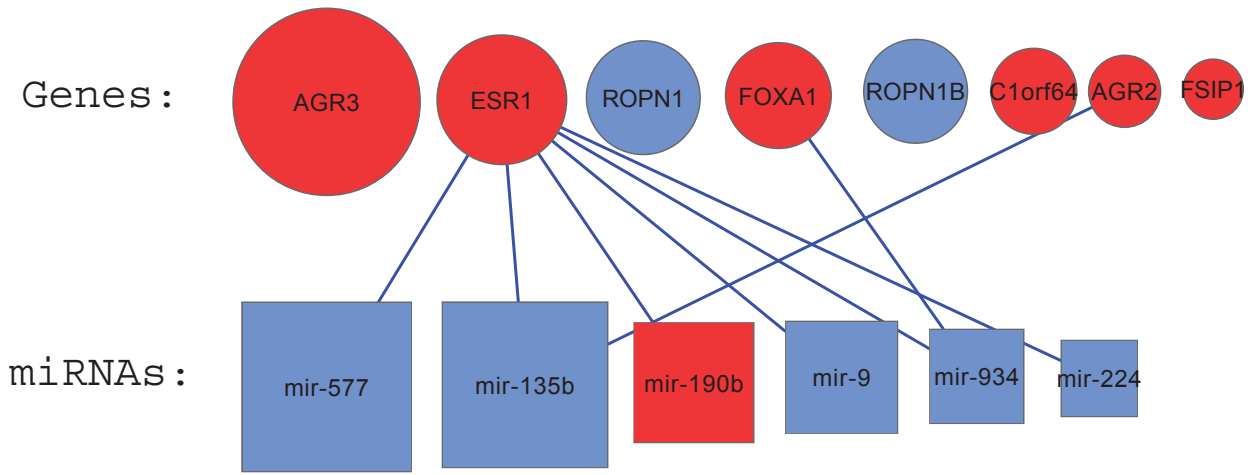
- red: positive loading; blue: negative loading
- miRNA linked if gene is a predicted target in at least two of *Pictar*, *miRanda*, *TargetScan* and *RNA22*

Gene-miRNA Sparse JIVE

- First "Sparse" joint component sample scores:



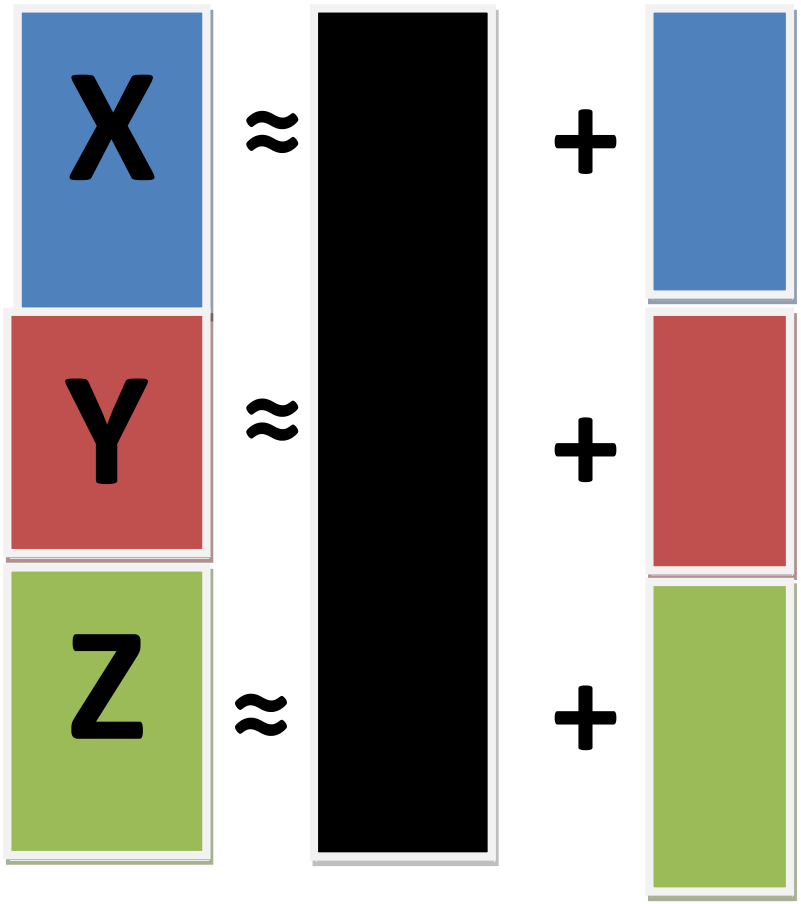
- Genes and miRNAs with non-zero loadings:



- red: positive loading; blue: negative loading
- miRNA linked if gene is a predicted target in at least two of *Pictar*, *miRanda*, *TargetScan* and *RNA22*

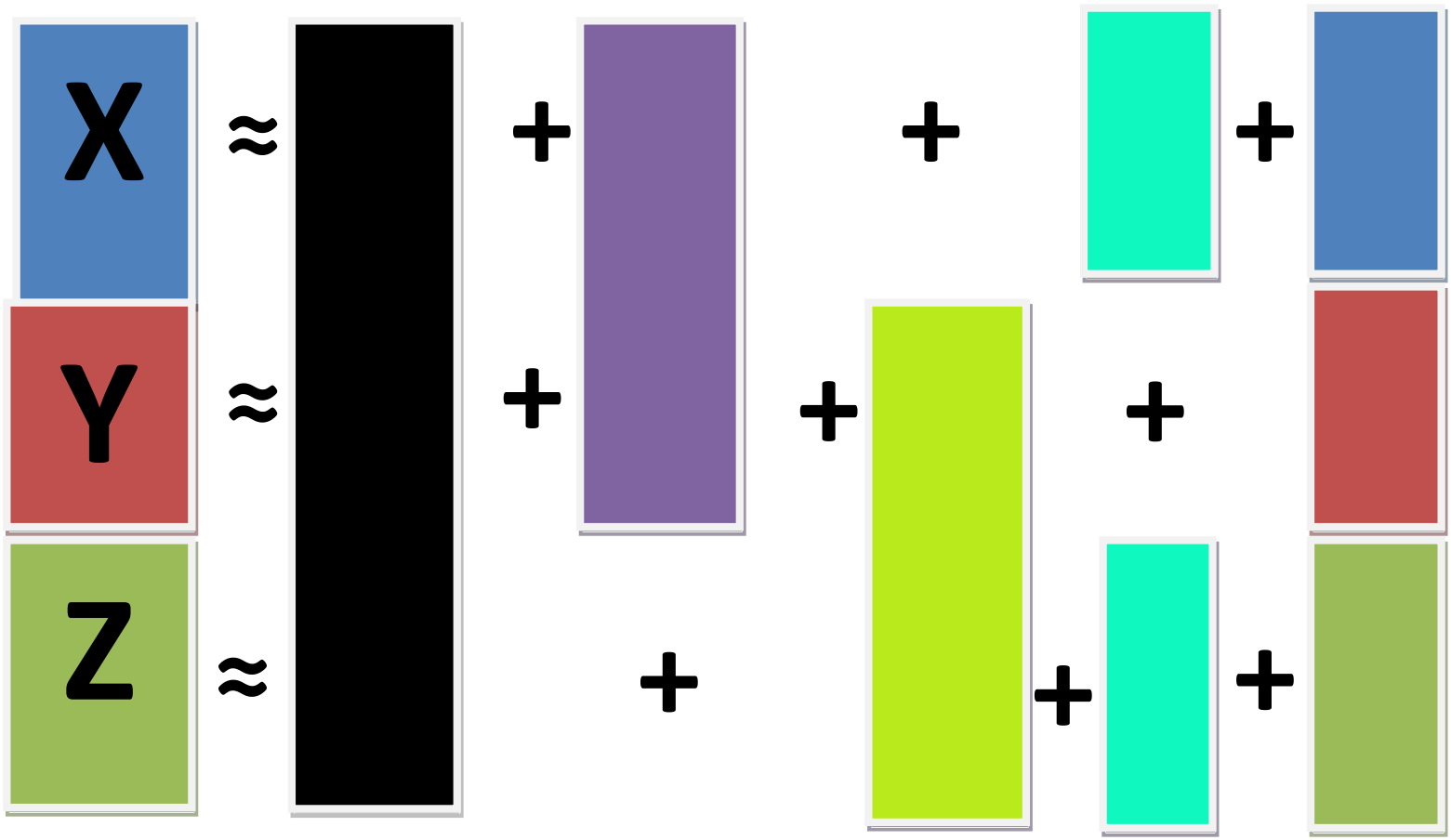
Future work: Factorial JIVE

- More than two datasets (standard JIVE):



Future work: Factorial JIVE

- Factorial model:



Related work

- Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS)
  - H Hotelling, 1936; H. Wold, 1965.
  - Find pairs of direction vectors to maximize correlation (CCA) or covariance (PLS)
  - Limited to two datasets
  - Overfitting in high-dimensionsal cases (esp. CCA)
  - Interference from individual structure (esp. PLS)
- Integrative Network Models
  - A Adourian et al., 2008; C Xing and DB Dunson, 2011.
  - Focused on pairwise relationships, not global variation
- Hierarchical Latent Variable Models
  - V. Baladandayuthapani et al., 2008; C Di, 2009; L Zhou et al., 2010.
  - Analysis of different sample groups on the same kind of data
  - Models differences between groups, not shared structure across datatypes

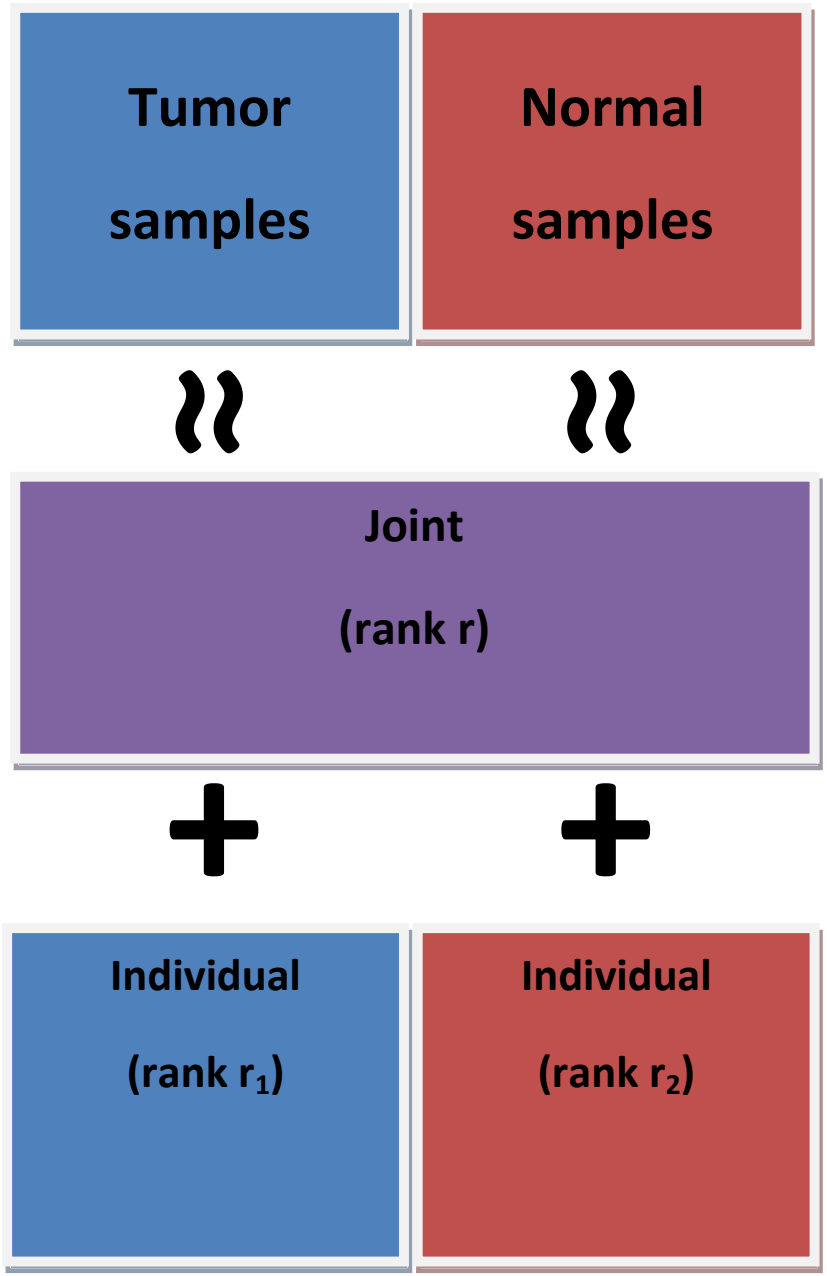
JIVE: additional applications

- For a single datatype, could look over different sample sets
  - Sick vs healthy
  - Treatment vs control
- Image analysis
  - Estimate "background" and unique characteristics from collection of images
- Financial data
  - Explore variation across and within financial markets

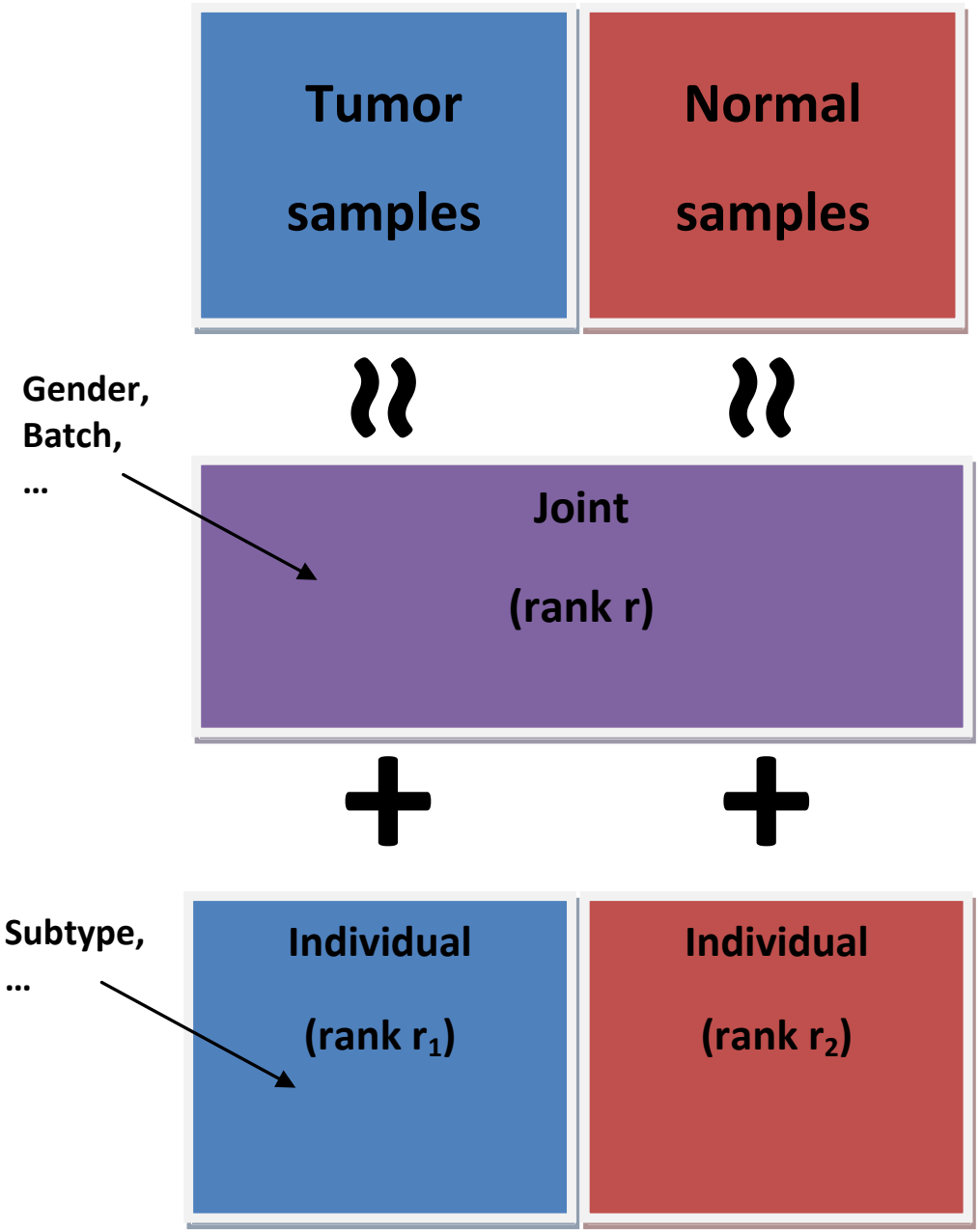
JIVE: additional applications

- For a single datatype, could look over different sample sets
  - Sick vs healthy
  - Treatment vs control
- Image analysis
  - Estimate "background" and unique characteristics from collection of images
- Financial data
  - Explore variation across and within financial markets

Horizontal JIVE



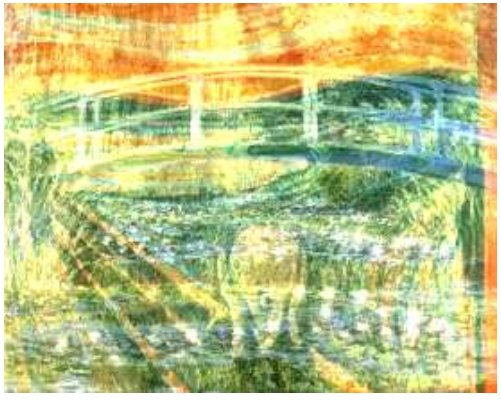
Horizontal JIVE



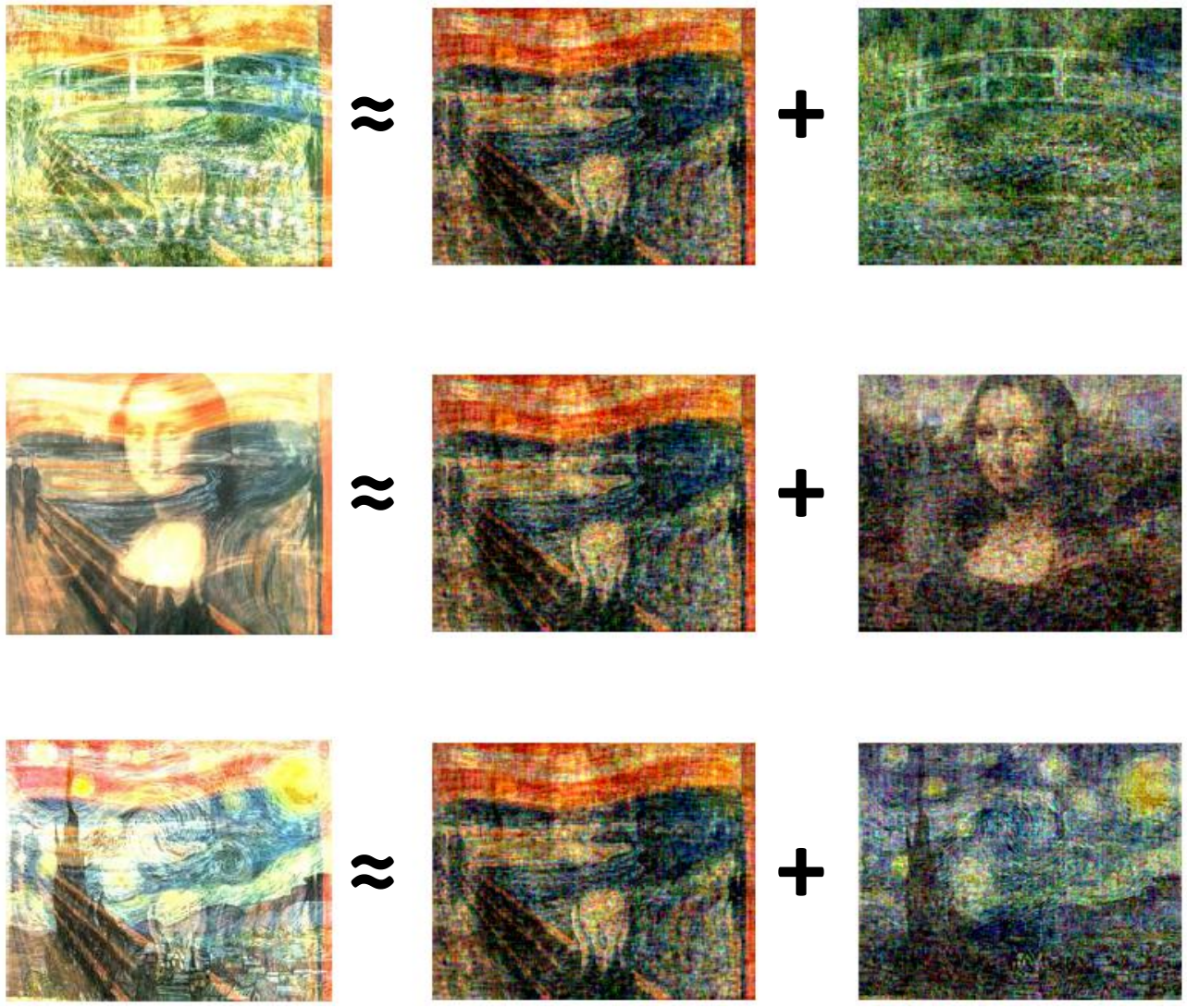
JIVE: additional applications

- For a single datatype, could look over different sample sets
  - Sick vs healthy
  - Treatment vs control
- Image analysis
  - Estimate "background" and unique characteristics from collection of images
- Financial data
  - Explore variation across and within financial markets

Mixed Art



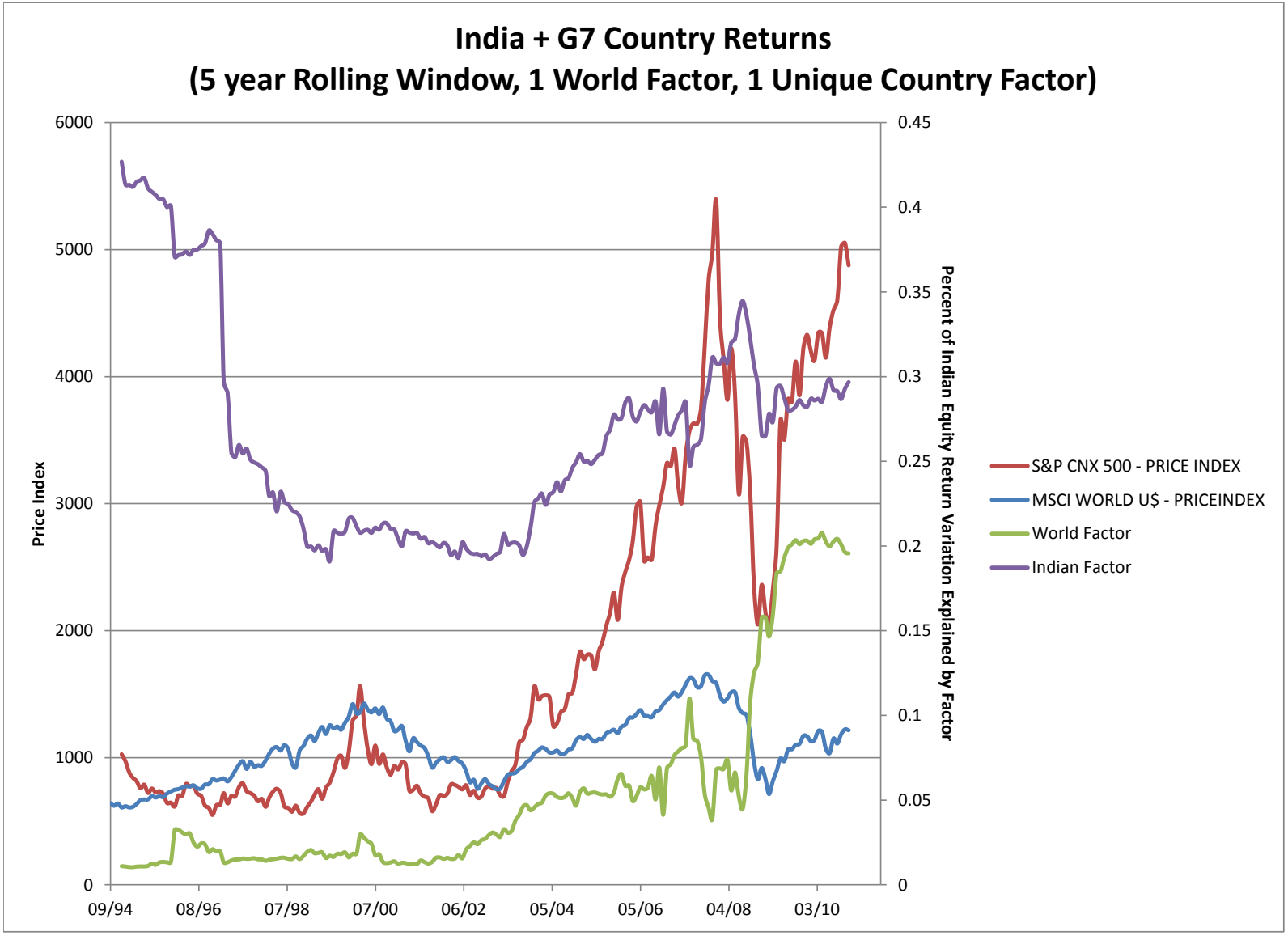
Mixed Art: Estimated decomposition



Mixed Art: Actual decomposition

JIVE: additional applications

- For a single datatype, could look over different sample sets
  - Sick vs healthy
  - Treatment vs control
- Image analysis
  - Estimate "background" and unique characteristics from collection of images
- Financial data
  - Explore variation across and within financial markets



Thanks to Philip Howard, UNC Kenan-Flagler Business School

THANK YOU!

References

- Mice-Alcohol Project
  - BU Bradford, EF Lock, O Kosyk, S Kim, T Uehara, D Harbourt, M DeSimone, DW Threadgill, V Tryndyak, IP Pogribny, L Bleyle, DR Koop, and I Rusyn. **Inter-strain differences in the liver effects of trichloroethylene in a multi-strain panel of inbred mice**. Toxicological Sciences, 120(1), 2010.
- JIVE
  - EF Lock, KA Hoadley, JS Marron, and AB Nobel. **Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Datatypes**. *AOAS*, to appear.
- miRNA target predications
  - H Dweep, C Sticht, P Pandey, and N Gretz. **miRWalk - database: prediction of possible miRNA binding sites by "walking" the genes of 3 genomes**. Journal of Biomedical Informatics, 44(5):839-847, 2011.
- Image data / multi-way tensor decompositions
  - EF Lock, AB Nobel, and JS Marron. **Comment: Population Value Decomposition, a Framework for the Analysis of Image Populations**. JASA, 106(495), 2011.