Exploratory Analysis of Multi-Source Genomic Data

Eric F. Lock
University of Minnesota
Division of Biostatistics

With
AB Nobel, JS Marron, K Hoadley, and I Rusyn, UNC Chapel Hill
and
DB Dunson, AA Koch, Duke University

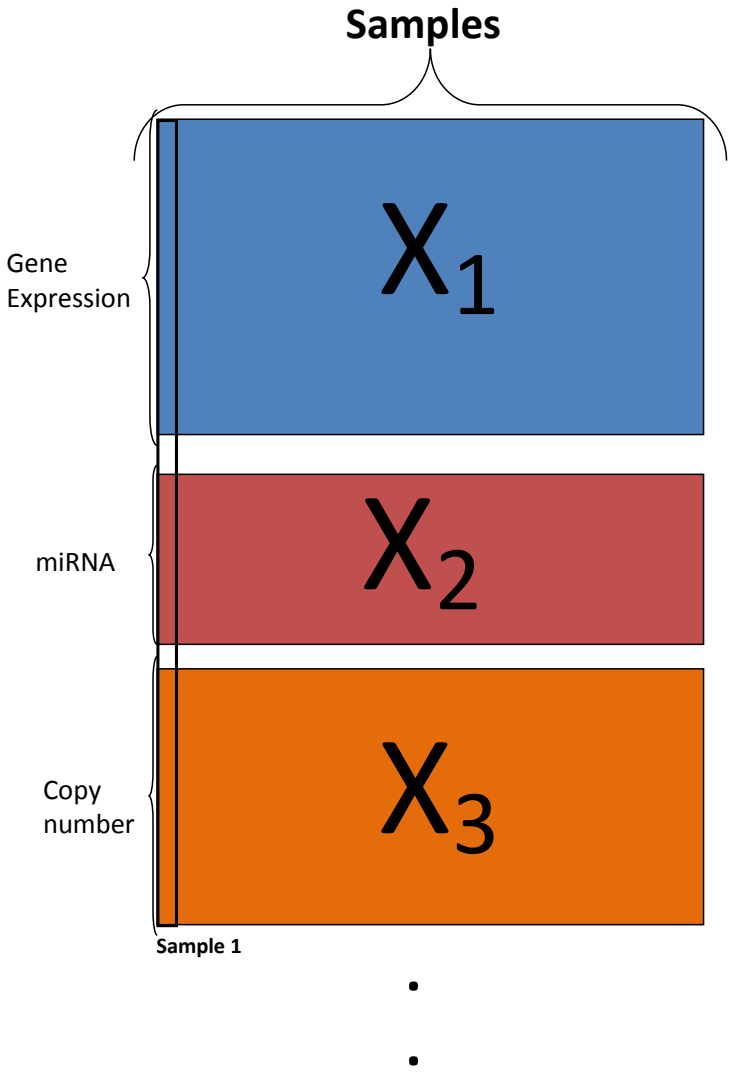
Johns Hopkins University, 09/28/2015

Motivating example

- Publicly available data from The Cancer Genome Atlas (TCGA)
- Multiple kinds of data for the same set of 348 breast cancer tumors:
  - **GE**: Gene expression data (17814 genes)
  - **miRNA**: miRNA data (655 miRNAs)
  - **CN**: Copy number data (200,000 probes / 19,780 genes)
  - **ME**: Methylation data (21,986 CG regions)
  - **MUT**: Mutation data (12,481 genes)
  - **RPPA**: Reverse phase protein array data (171 proteins)

Multi-source data example

- Multiple high-dimensional *data sources* for the same objects.



Exploratory analysis of multi-source datasets

- "Joint" analyses ignore features specific to each data source

Exploratory analysis of multi-source datasets

- "Joint" analyses ignore features specific to each data source
- "Separate" analyses sacrifice power, miss inter-source dependencies

Exploratory analysis of multi-source datasets

- "Joint" analyses ignore features specific to each data source
- "Separate" analyses sacrifice power, miss inter-source dependencies
- Goal: simultaneously model dependence and heterogeneity of data sources

Exploratory analysis of multi-source datasets

- "Joint" analyses ignore features specific to each data source
- "Separate" analyses sacrifice power, miss inter-source dependencies
- Goal: simultaneously model dependence and heterogeneity of data sources
- Extend exploratory methods to the multi-source context.
  - Clustering
  - PCA

Exploratory analysis of multi-source datasets

- "Joint" analyses ignore features specific to each data source
- "Separate" analyses sacrifice power, miss inter-source dependencies
- Goal: simultaneously model dependence and heterogeneity of data sources
- Extend exploratory methods to the multi-source context.
  - Clustering
  - PCA

Multi-source clustering

- **Joint clustering**
  - A single clustering of the objects, based on all sources
    - Shen, Olshen, & Ladanyi, *Bioinformatics*, 2009
    - Rey & Roth, *ICML*, 2012
    - Kormaksson et. al., *Annals of Applied Statistics*, 2012

Multi-source clustering

- **Joint clustering**
  - A single clustering of the objects, based on all sources
    - Shen, Olshen, & Ladanyi, *Bioinformatics*, 2009
    - Rey & Roth, *ICML*, 2012
    - Kormaksson et. al., *Annals of Applied Statistics*, 2012
- **Separate clustering**
  - Separate clustering for each source
  - Post-hoc integration
    - Assess cluster agreement (Hubert & Arabie, 1985)
    - Consensus clustering (TCGA research network, *Nature*, 2012)

Multi-source clustering

- **Joint clustering**
  - A single clustering of the objects, based on all sources
    - Shen, Olshen, & Ladanyi, *Bioinformatics*, 2009
    - Rey & Roth, *ICML*, 2012
    - Kormaksson et. al., *Annals of Applied Statistics*, 2012
- **Separate clustering**
  - Separate clustering for each source
  - Post-hoc integration
    - Assess cluster agreement (Hubert & Arabie, 1985)
    - Consensus clustering (TCGA research network, *Nature*, 2012)
- **Dependent clustering**
  - Pairwise-dependence model
    - Kirk et. al., *Bioinformatics*, 2012
  - Bayesian consensus clustering (BCC)

Bayesian consensus clustering

- Separate clustering for each data source
  - Adhere loosely to an overall clustering
- Level of adherence is estimated from the data
- Overall and source clusterings are estimated simultaneously

Bayesian consensus clustering

- Separate clustering for each data source
  - Adhere loosely to an overall clustering
- Level of adherence is estimated from the data
- Overall and source clusterings are estimated simultaneously
- Advantages over traditional consensus clustering:
  1. Models uncertainty in both the source and overall clusterings.
  2. Permits borrowing of information across sources.
  3. Level of adherence is learned for each source.

Bayesian consensus clustering

- Sources $X_1, \ldots, X_M$, for a common set of N samples
- Overall cluster index $C_n \in \{1, \ldots, K\}$ for samples $n = 1, \ldots, N$.
- Source cluster index $L_{mn} \in \{1, \ldots, K\}$ for sources $m = 1, \ldots, M$, samples $n = 1, \ldots, N$.
- Source clusters depend partially on overall clusters:

$$P(L_{mn} = k | C_n) = \begin{cases} \alpha_m \text{ if } C_n = k \\ \frac{1-\alpha_m}{K-1} \text{ otherwise} \end{cases}$$

where $\alpha_m \in [\frac{1}{K}, 1]$ controls level of adherence.

Bayesian consensus clustering

- Probability model f_m, with cluster-specific parameters θ_{mk}:

$$P(L_{mn} = k|X_{mn}, C_n, \Theta_m) \propto P(L_{mn} = k|C_n) f_m(X_{mn}|\theta_{mk})$$

- Overall cluster mixture probabilities $\Pi = (\pi_1, \ldots, \pi_k)$:

$$P(C_n = k|\Pi, \{L_{mn}, \alpha_m\}_{m=1}^M) \propto \pi_k \prod_{m=1}^M P(L_{mn} = k|C_n)$$

- Give prior for $\Pi, \alpha_m's$, and $\Theta_m's$.
  - Uniform Dirichlet for Π
  - Uniform $[\frac{1}{K}, 1]$ for α_m
  - Normal-Gamma conjugate prior distribution for Θ_m, f_m

- Estimate full posterior via MCMC
  - Iteratively sample from conditional posteriors of $\Pi, \{\alpha_m\}_{m=1}^M, \{\Theta_m\}_{m=1}^M, \mathbb{C}$ and $\{\mathbb{L}_m\}_{m=1}^M$.

Simulation example

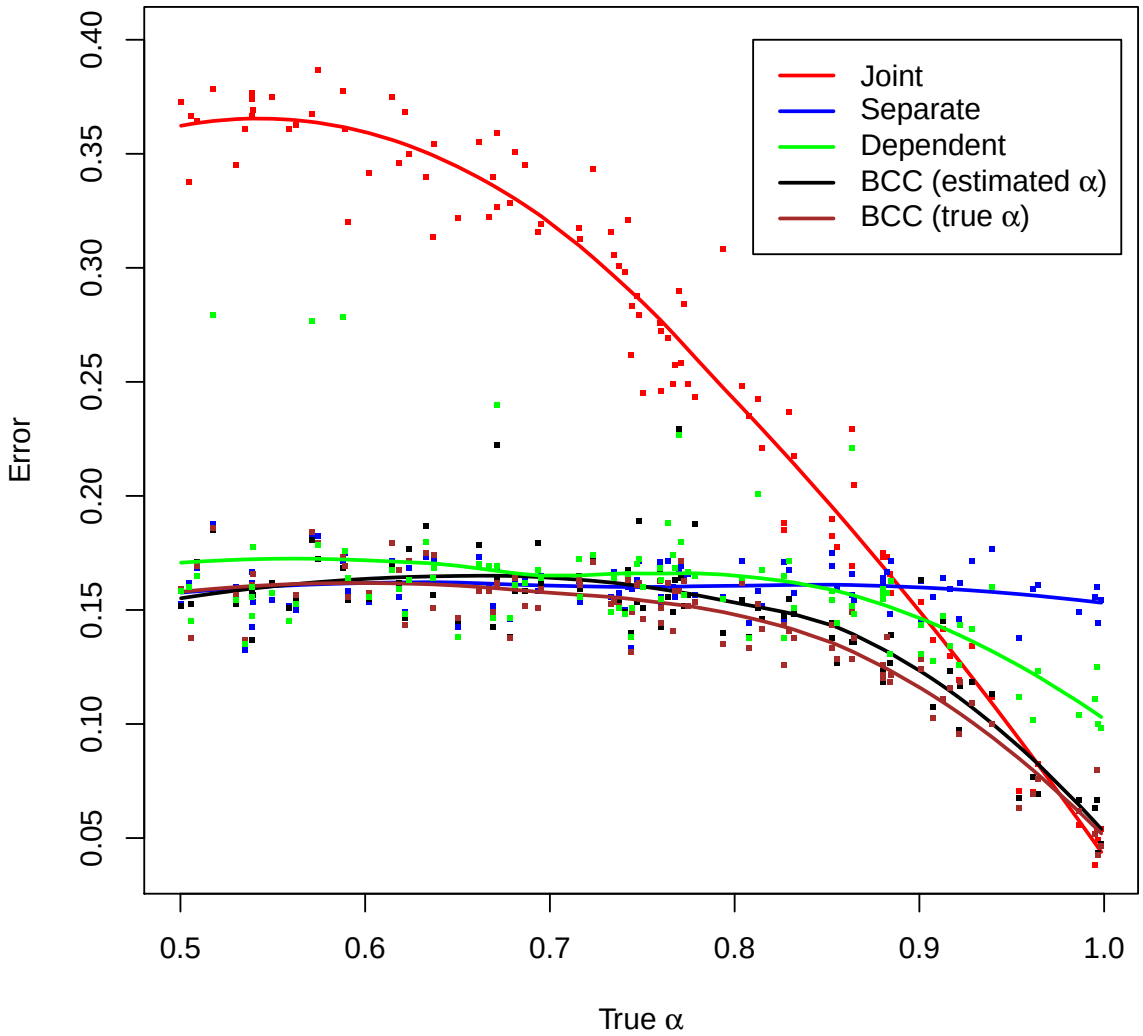
- Simulate data for $M = 3$ univariate sources.
  - $N = 200$ samples
  - $K = 2$ Gaussian clusters for each source
    - Cluster means: $\mu_1 = -1$ and $\mu_2 = 1$
    - Standard deviation: $\sigma = 1$
  - Draw α uniformly from 0.5 to 1 ($\alpha_1 = \alpha_2 = \alpha_3$).

- Estimate source clusterings via
  - **Separate clustering**: no dependence model between sources.
  - **Joint clustering**: assume all sources have same clustering.
  - **Dependent clustering**: model pairwise clustering dependence between sources.
  - **Bayesian consensus clustering**

- Repeat simulation and estimation 100 times for varying α

Simulation example

- Simulation study (clustering error by adherence level):



TCGA example

- Applied BCC to GE, ME, miRNA & RPPA data for 348 TCGA breast samples

TCGA example

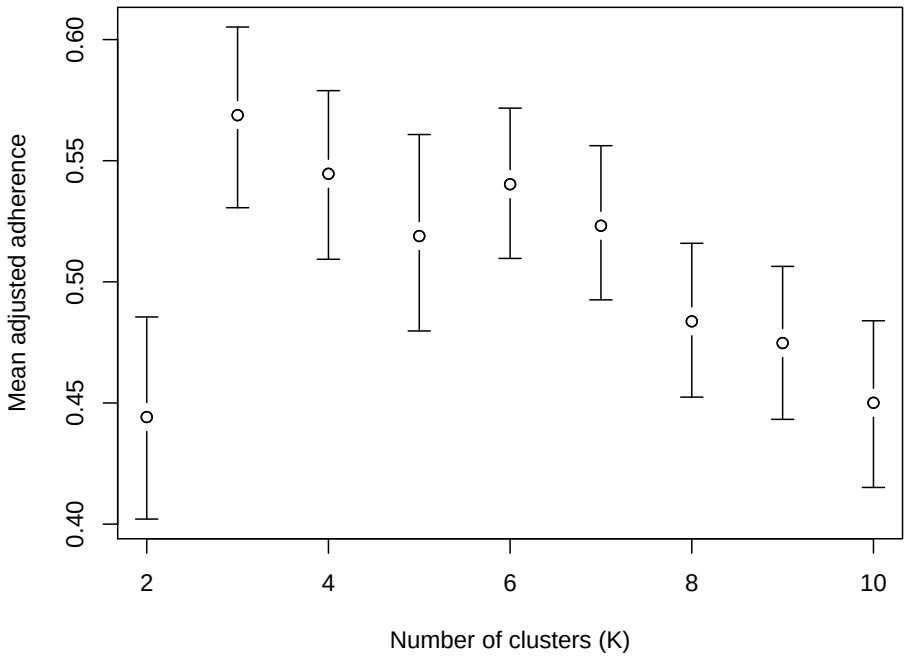
- Applied BCC to GE, ME, miRNA & RPPA data for 348 TCGA breast samples
- Choose K to maximize mean adjusted adherence

$$\frac{1}{M}\sum_{m=1}^{M}\frac{K\alpha_m - 1}{K-1}.$$

TCGA example

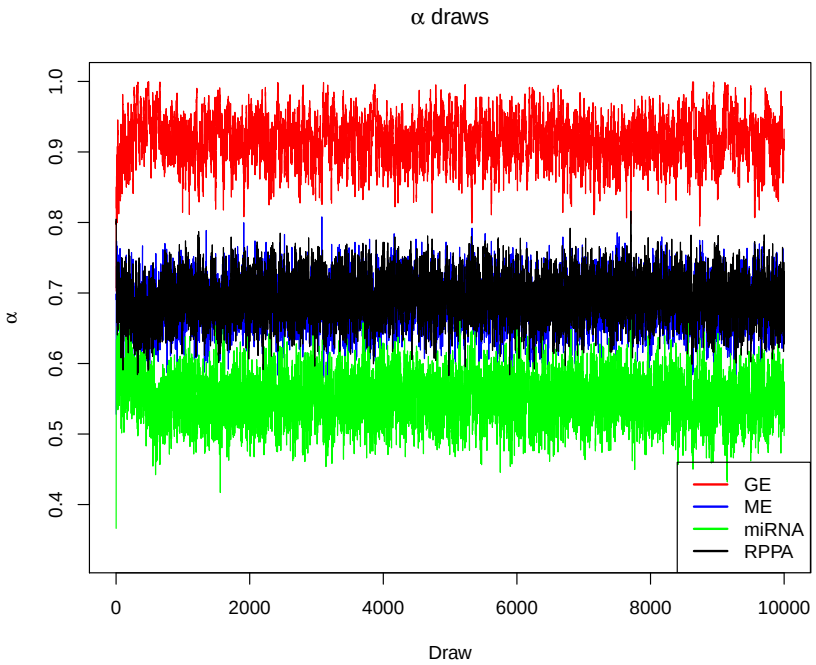
- Applied BCC to GE, ME, miRNA & RPPA data for 348 TCGA breast samples
- Choose K to maximize mean adjusted adherence

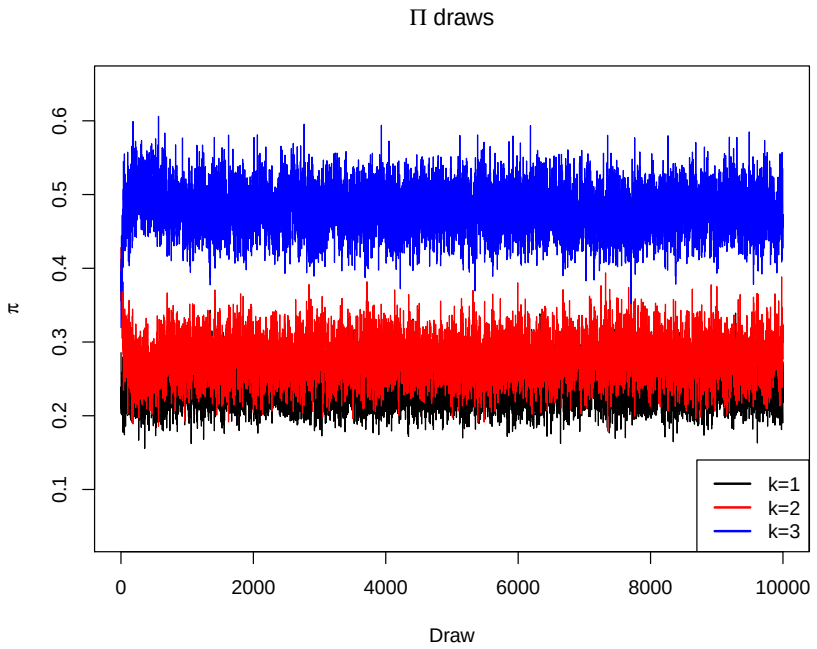
$$\frac{1}{M}\sum_{m=1}^{M}\frac{K\alpha_m - 1}{K-1}.$$



TCGA example

- MCMC mixing (K=3)





TCGA example

- Applied BCC to GE, ME,miRNA & RPPA data

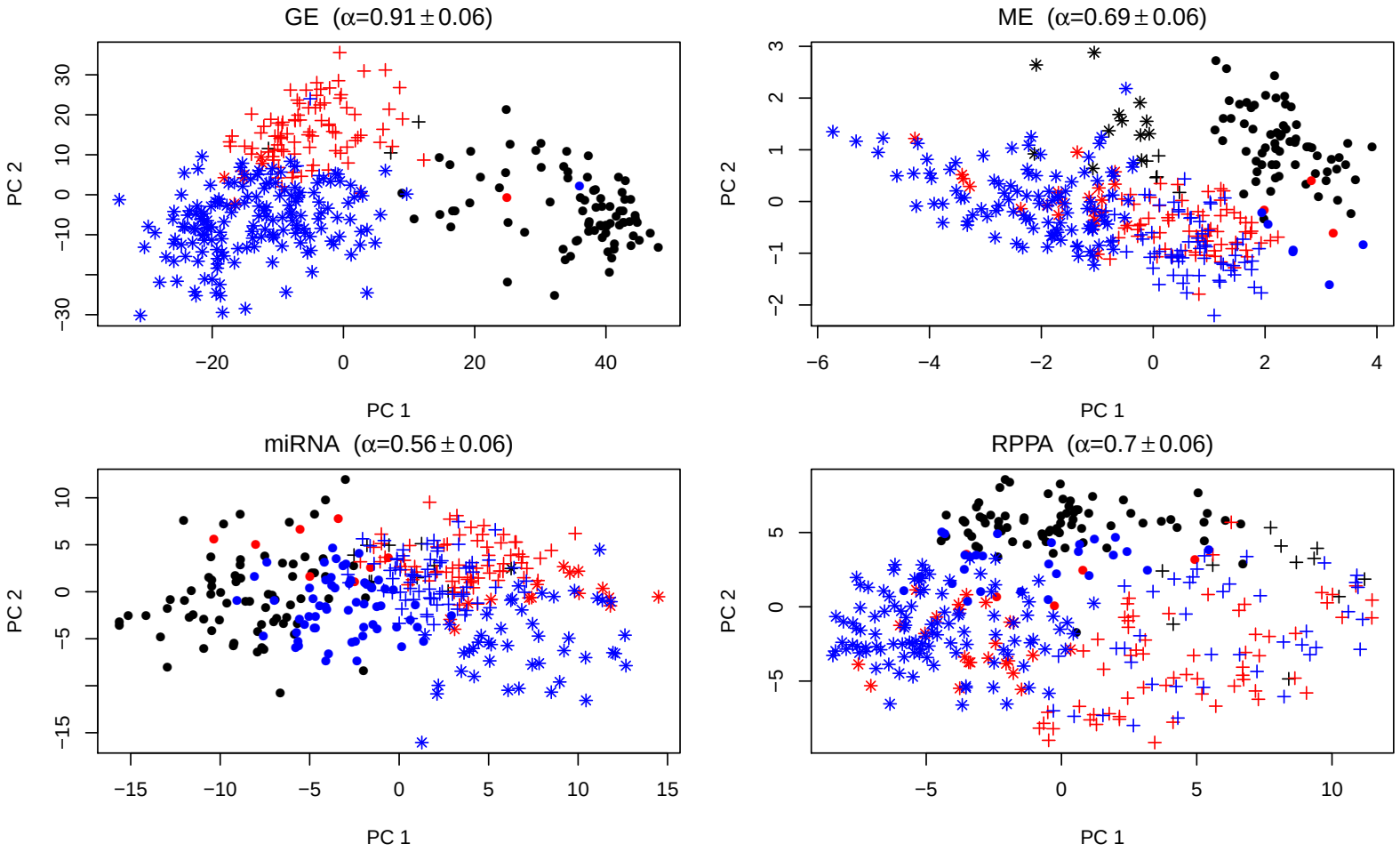
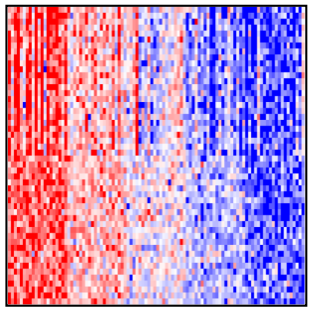


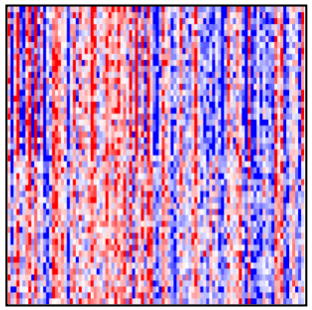
Figure: PCA plots. Samples are colored by overall cluster; cluster 1 is **black**, cluster 2 is red, cluster 3 is blue. Symbols indicate source-specific cluster; cluster 1 is '•', cluster 2 is '+', cluster 3 is '∗' .

Exploratory analysis of multi-source datasets

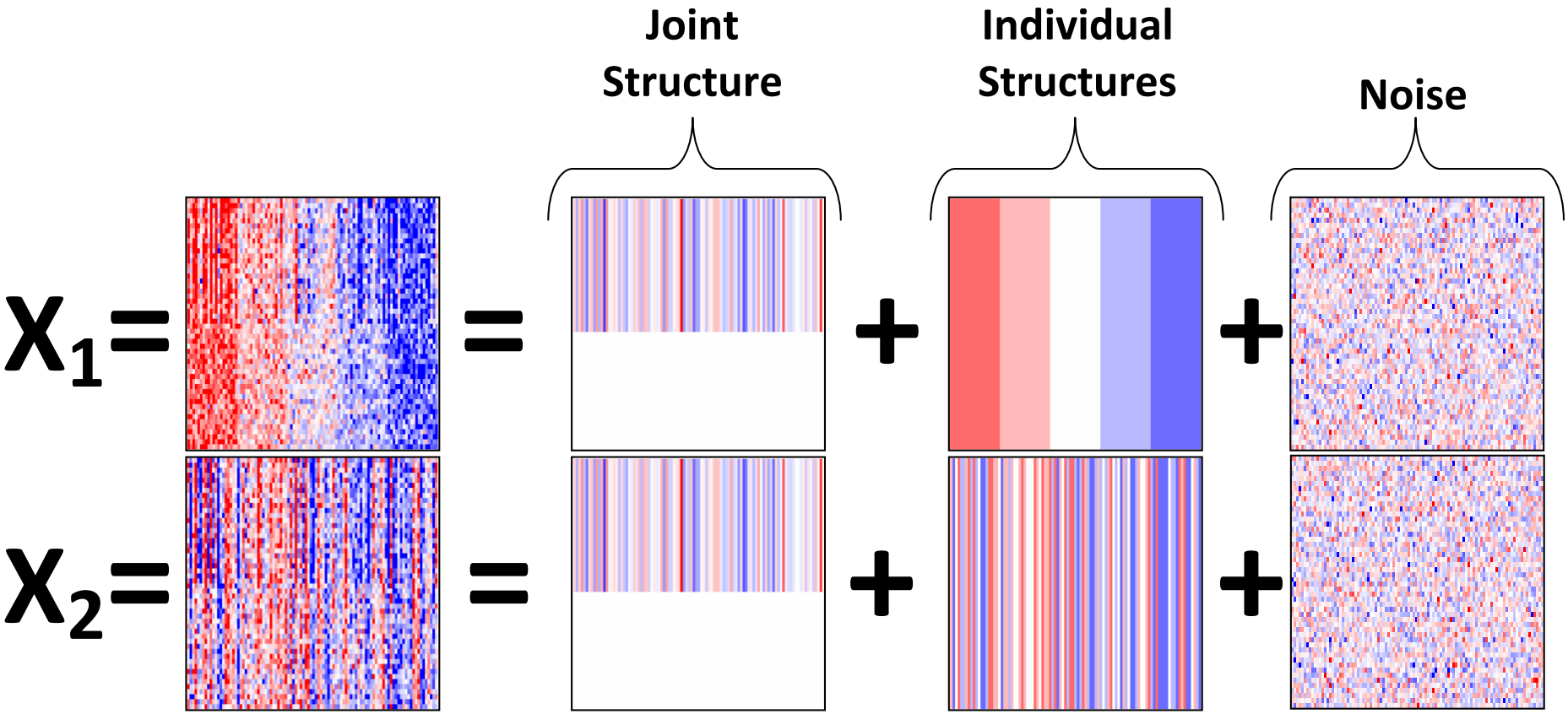
- Extend exploratory methods to the multi-source context.
  - Clustering
  - Principal components analysis (PCA)

Toy Example: Two Sources

$X_1 =$ 

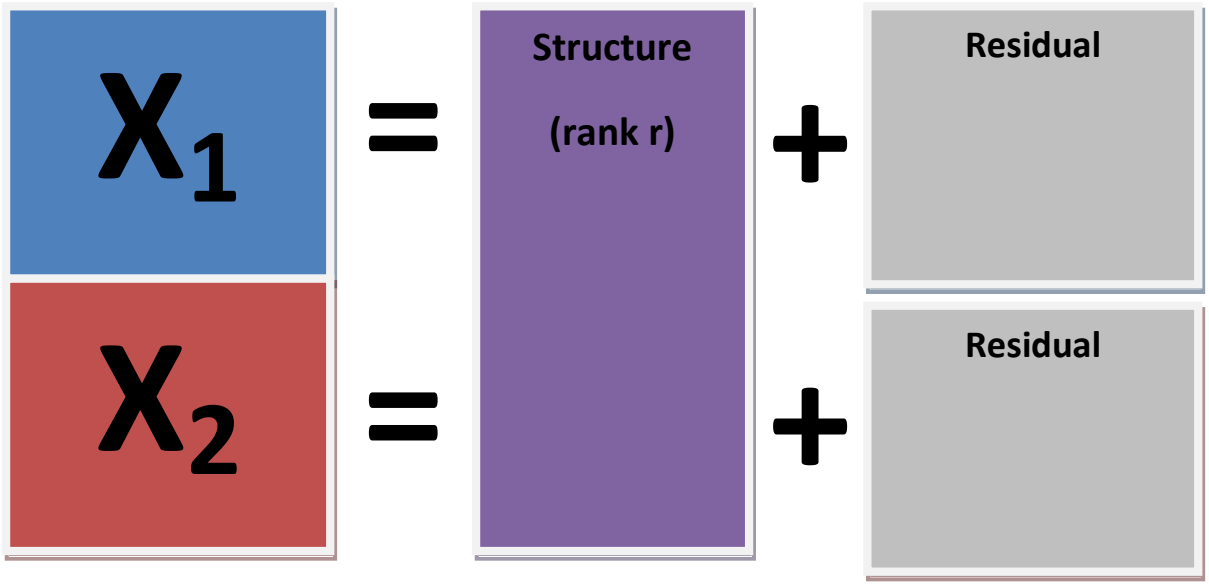
$X_2 =$ 

Toy Example: Two Sources

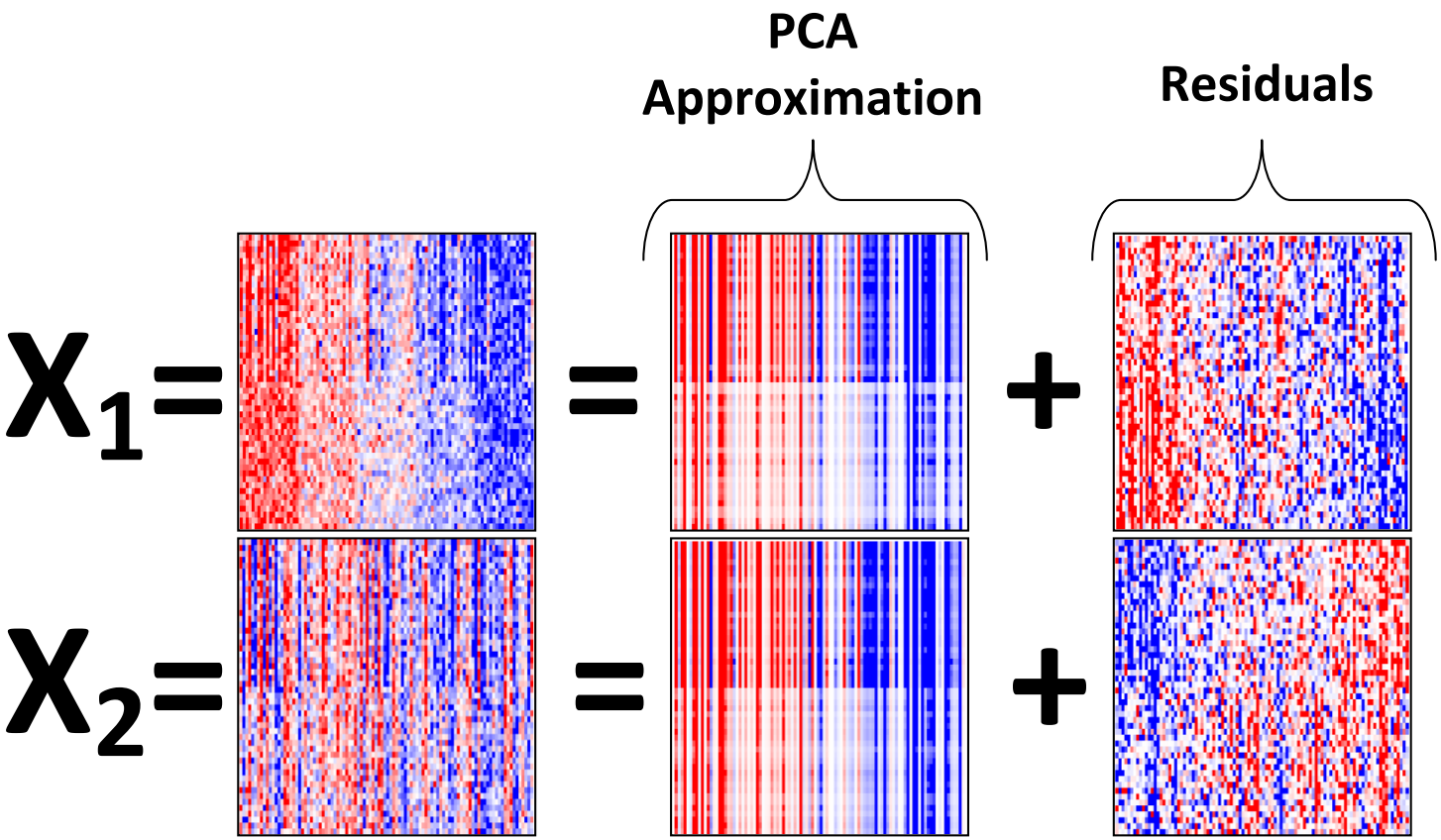


PCA Approximation

- PCA as a low rank approximation:

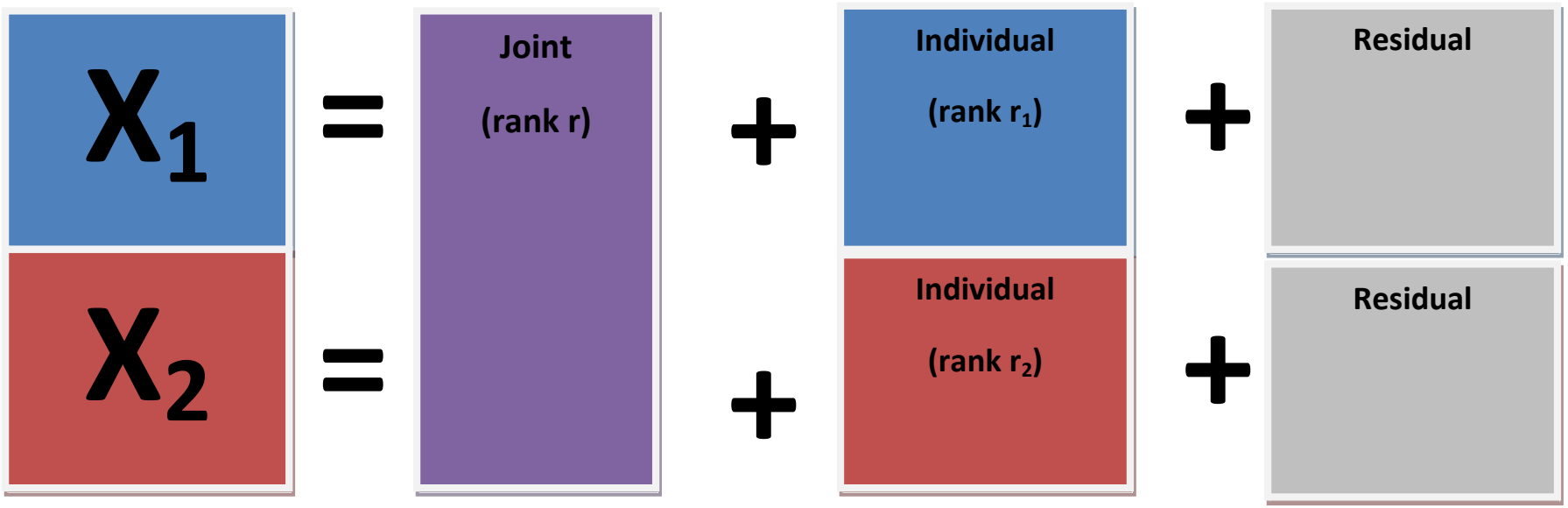


PCA Approximation ($r = 1$)

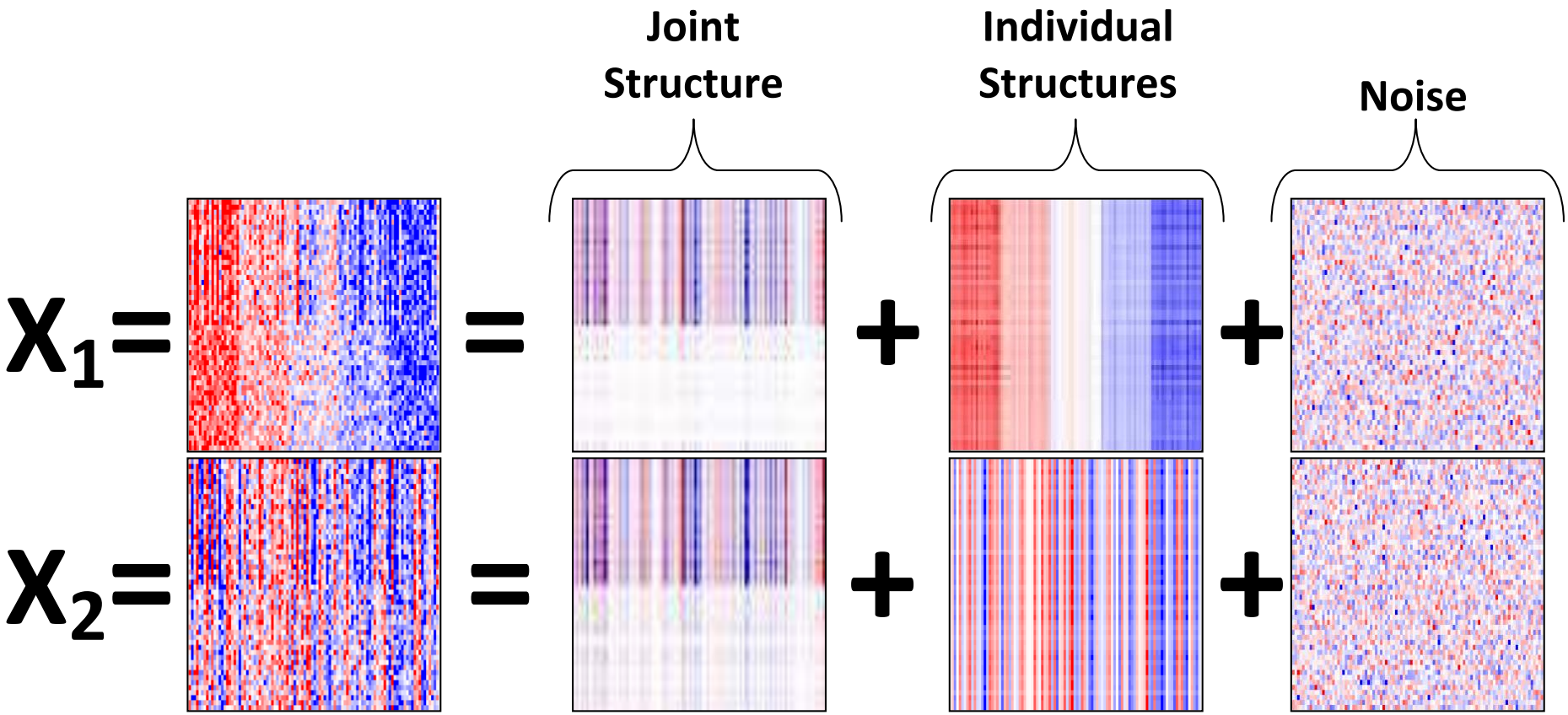


JIVE decomposition

- Joint and Individual Variation Explained (JIVE):

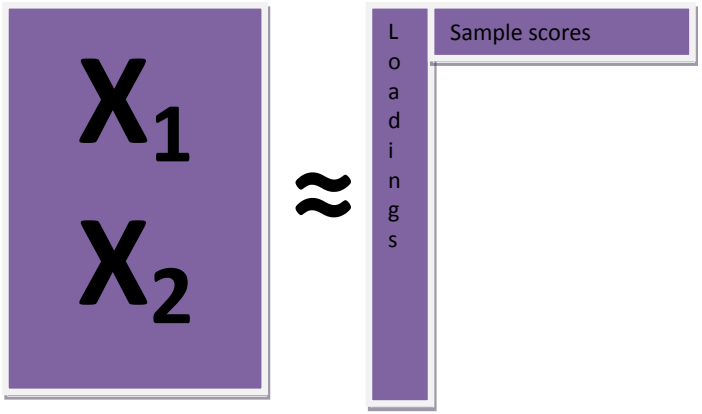


JIVE decomposition ($r = r_1 = r_2 = 1$)

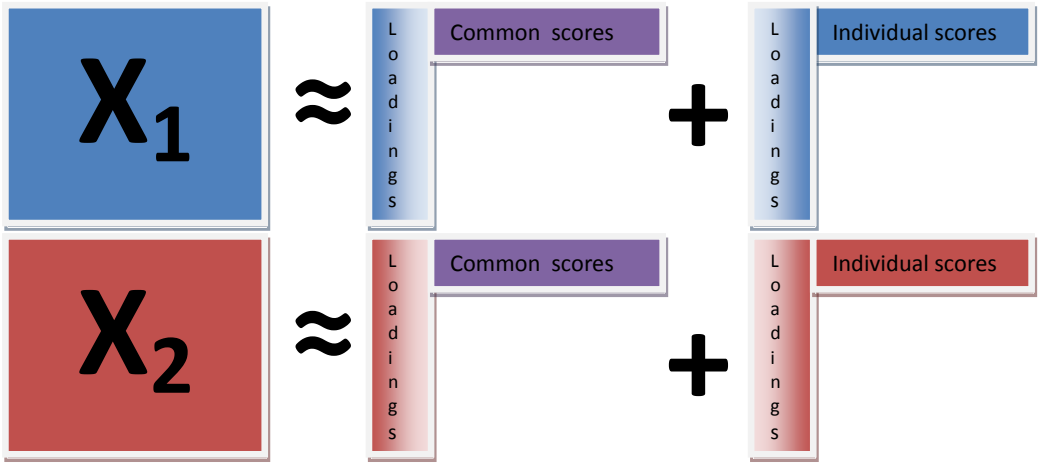


PCA vs JIVE

- PCA:



- JIVE:



JIVE decomposition

- Sources $X_1, ..., X_M$ of dimension $d_1, ..., d_M$ for n samples.
- Decomposition:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix} = \overbrace{\begin{bmatrix} J_1 \\ J_2 \\ \vdots \\ J_M \end{bmatrix}}^{J} + \overbrace{\begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_M \end{bmatrix}}^{A} + \overbrace{\begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_M \end{bmatrix}}^{R}$$

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

JIVE decomposition (factorized form)

- Relation to PCA:

$$\begin{aligned} X_1 &= \overbrace{U_1 S}^{J_1} + \overbrace{W_1 S_1}^{A_1} + R_1 \\ &\vdots \\ X_M &= U_M S + W_M S_M + R_M. \end{aligned}$$

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

Estimation

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

$$R = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_M \end{bmatrix} = \begin{bmatrix} X_1 - J_1 - A_1 \\ X_2 - J_2 - A_2 \\ \vdots \\ X_M - J_M - A_M \end{bmatrix}.$$

Estimation

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

$$R = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_M \end{bmatrix} = \begin{bmatrix} X_1 - J_1 - A_1 \\ X_2 - J_2 - A_2 \\ \vdots \\ X_M - J_M - A_M \end{bmatrix}.$$

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

Estimation

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

$$R = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_M \end{bmatrix} = \begin{bmatrix} X_1 - J_1 - A_1 \\ X_2 - J_2 - A_2 \\ \vdots \\ X_M - J_M - A_M \end{bmatrix}.$$

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

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

Key Issue: Scaling of Individual Datasets

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

Key Issue: Scaling of Individual Datasets

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

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

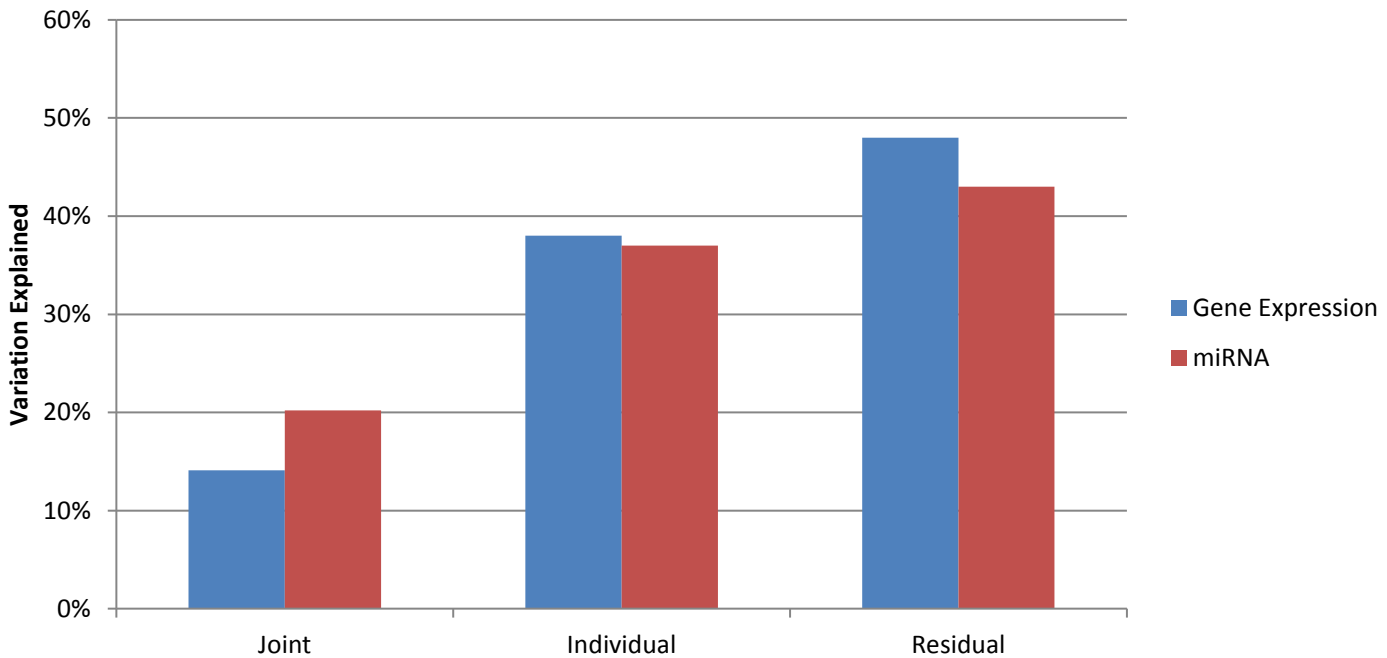
- Multiple kinds of data for the same set of 348 breast cancer tumors, from TCGA.
  - Gene expression data (17814 genes)
  - miRNA data (655 miRNAs)
  - Copy number data (200,000 probes / 19,780 genes)
  - Methylation data (21,986 CG regions)
  - Mutation data (12,481 genes)
  - Protein data
- Tumors 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

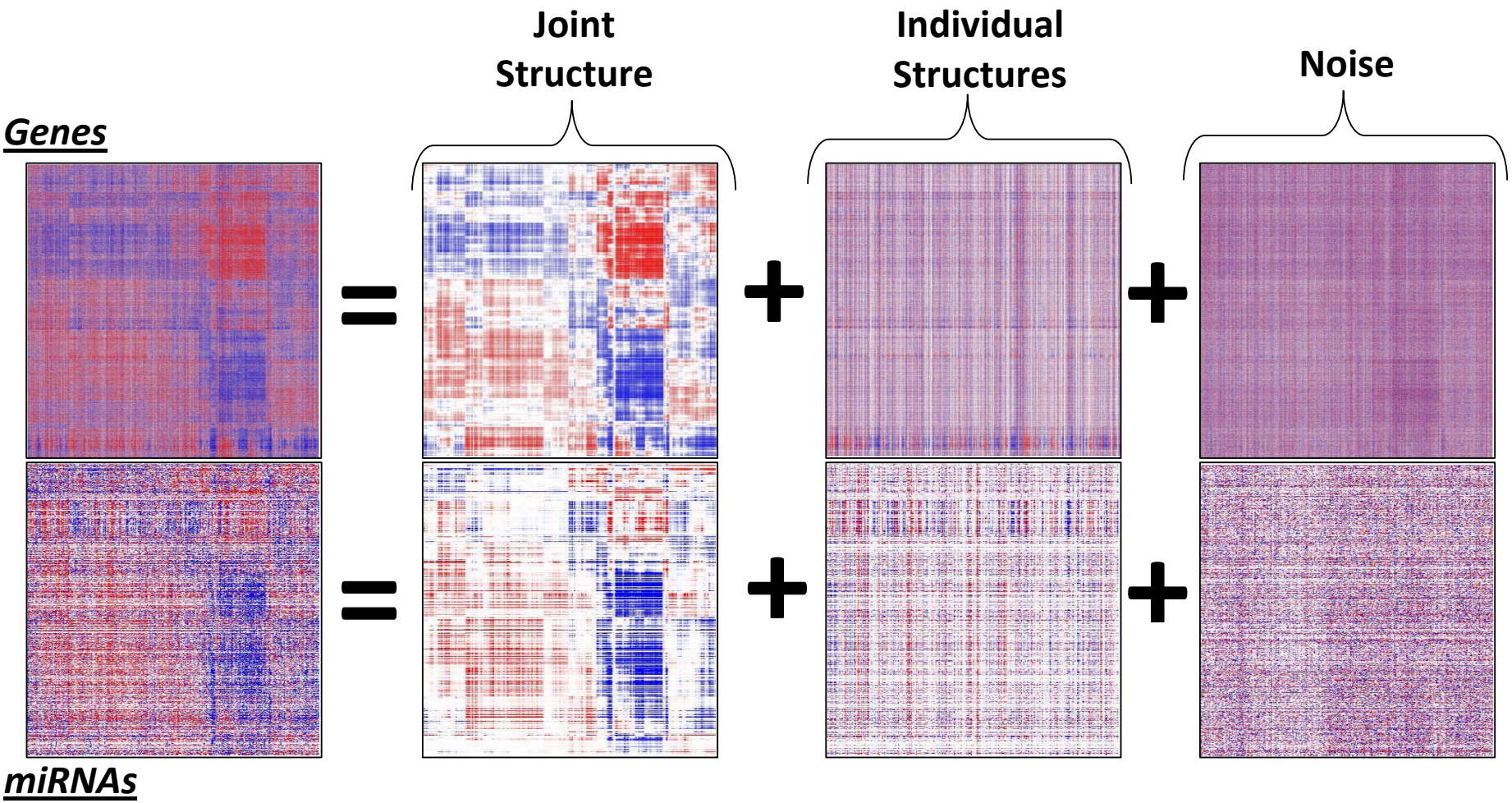
- Multiple kinds of data for the same set of 348 breast cancer tumors, from TCGA.
  - **Gene expression data (17814 genes)**
  - **miRNA data (655 miRNAs)**
  - Copy number data (200,000 probes / 19,780 genes)
  - Methylation data (21,986 CG regions)
  - Mutation data (12,481 genes)
  - Protein data
- 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

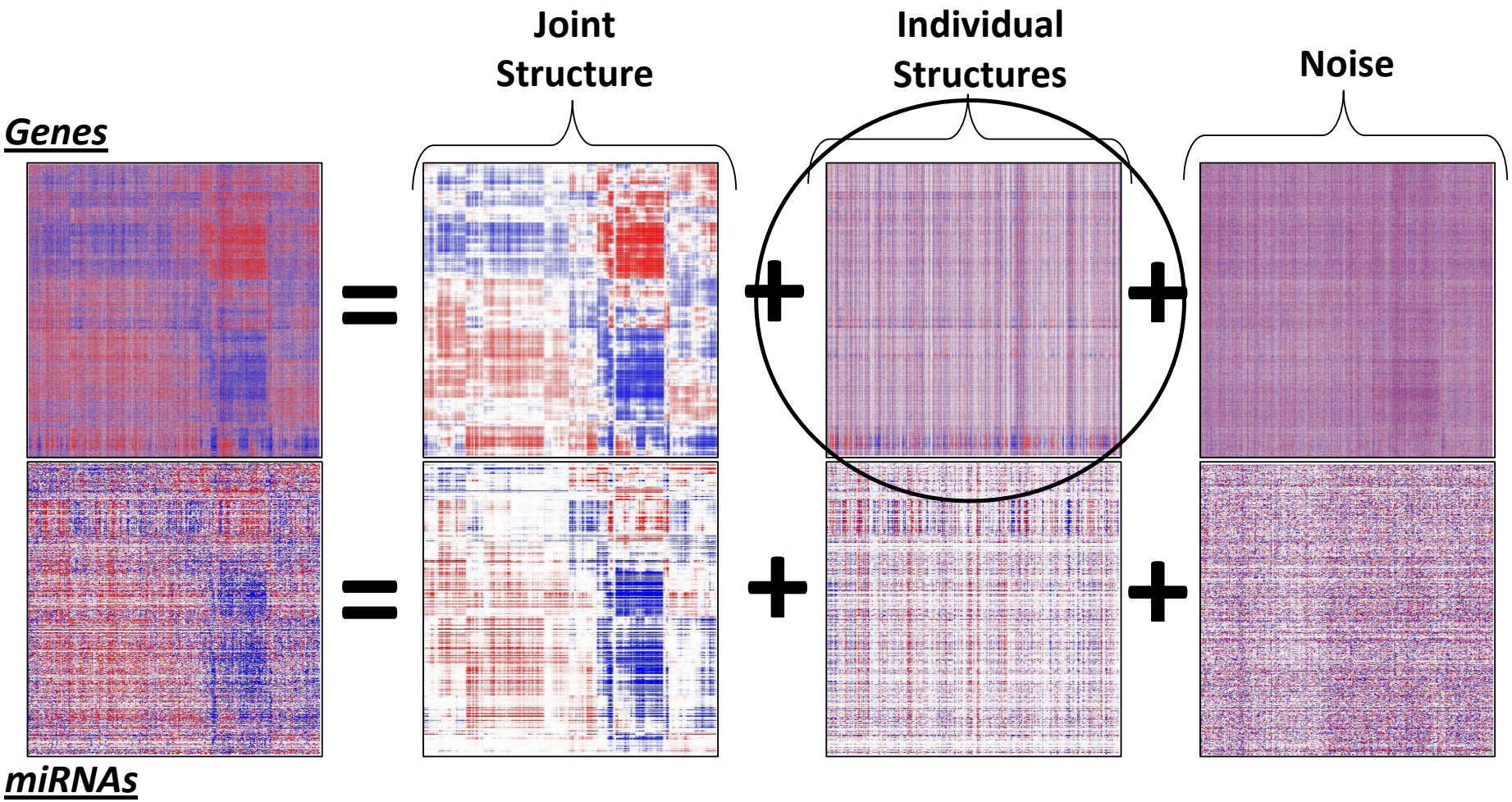
- Applied JIVE decomposition to Gene expression and miRNA.
- Permutation testing identifies
  - Rank 4 joint structure
  - Rank 22 structure individual to gene expression
  - Rank 9 structure individual to miRNA
- Variation decomposition:



JIVE Estimates

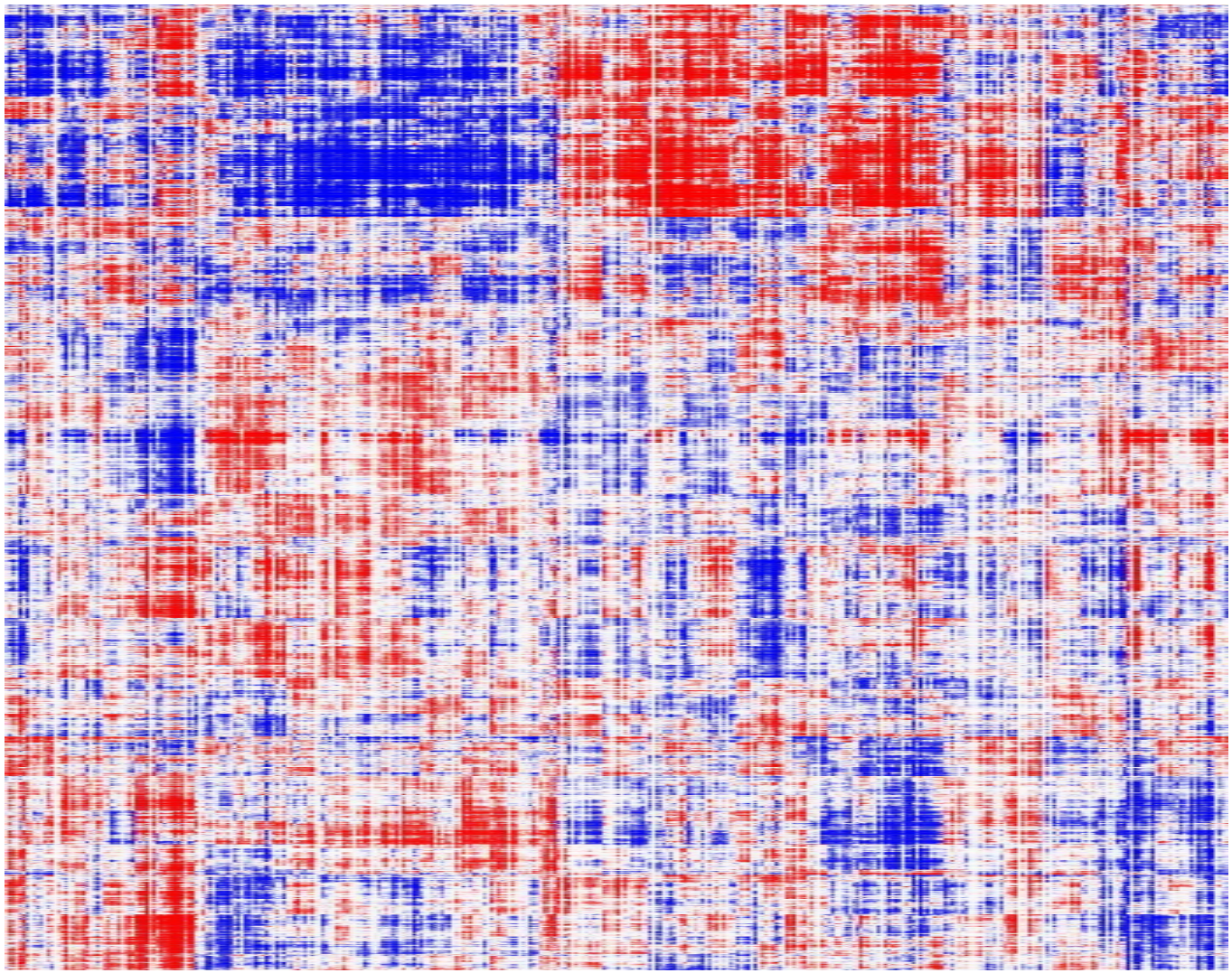


JIVE Estimates

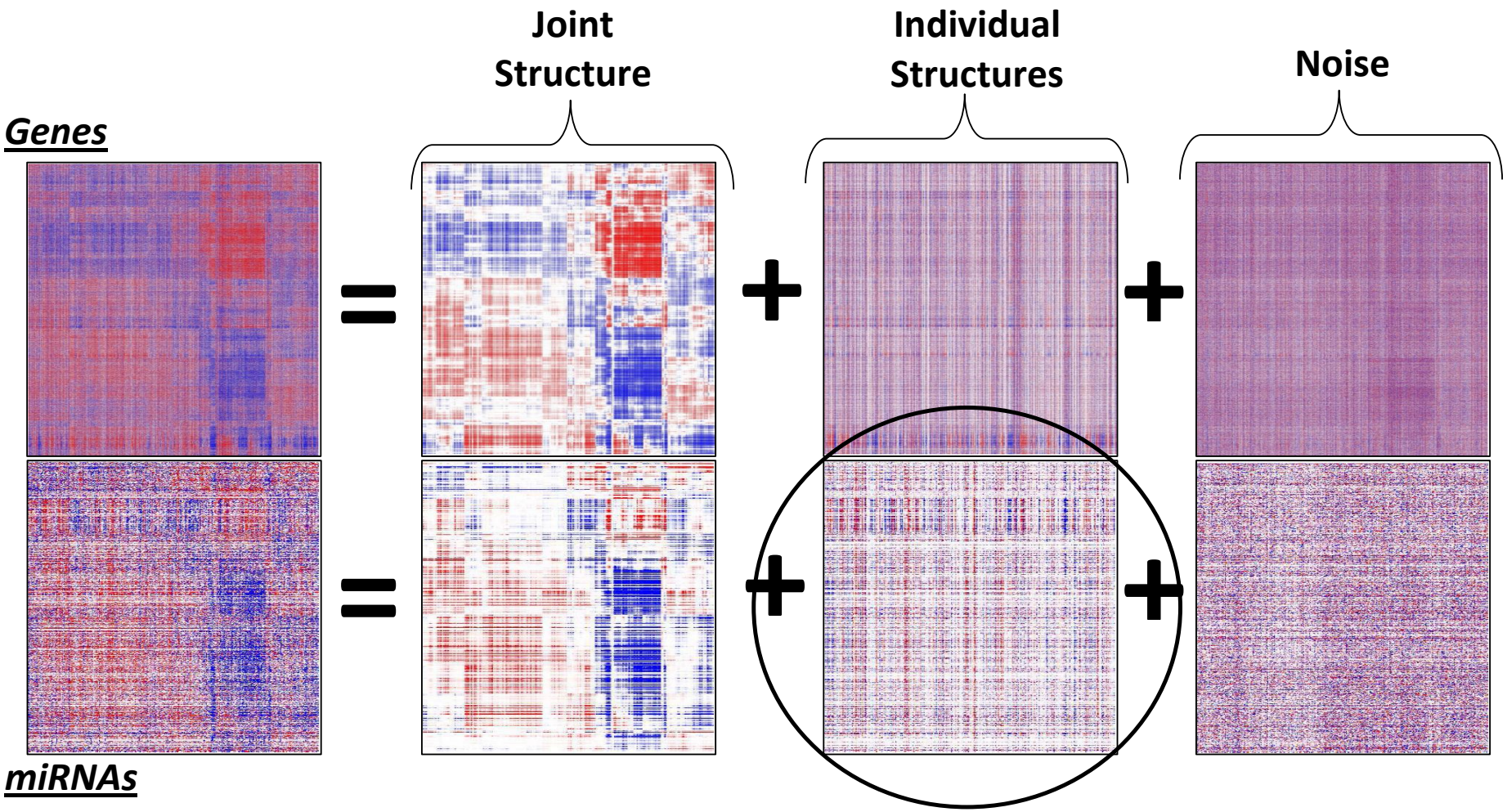


JIVE Estimates

- Gene individual (reorder rows and columns)

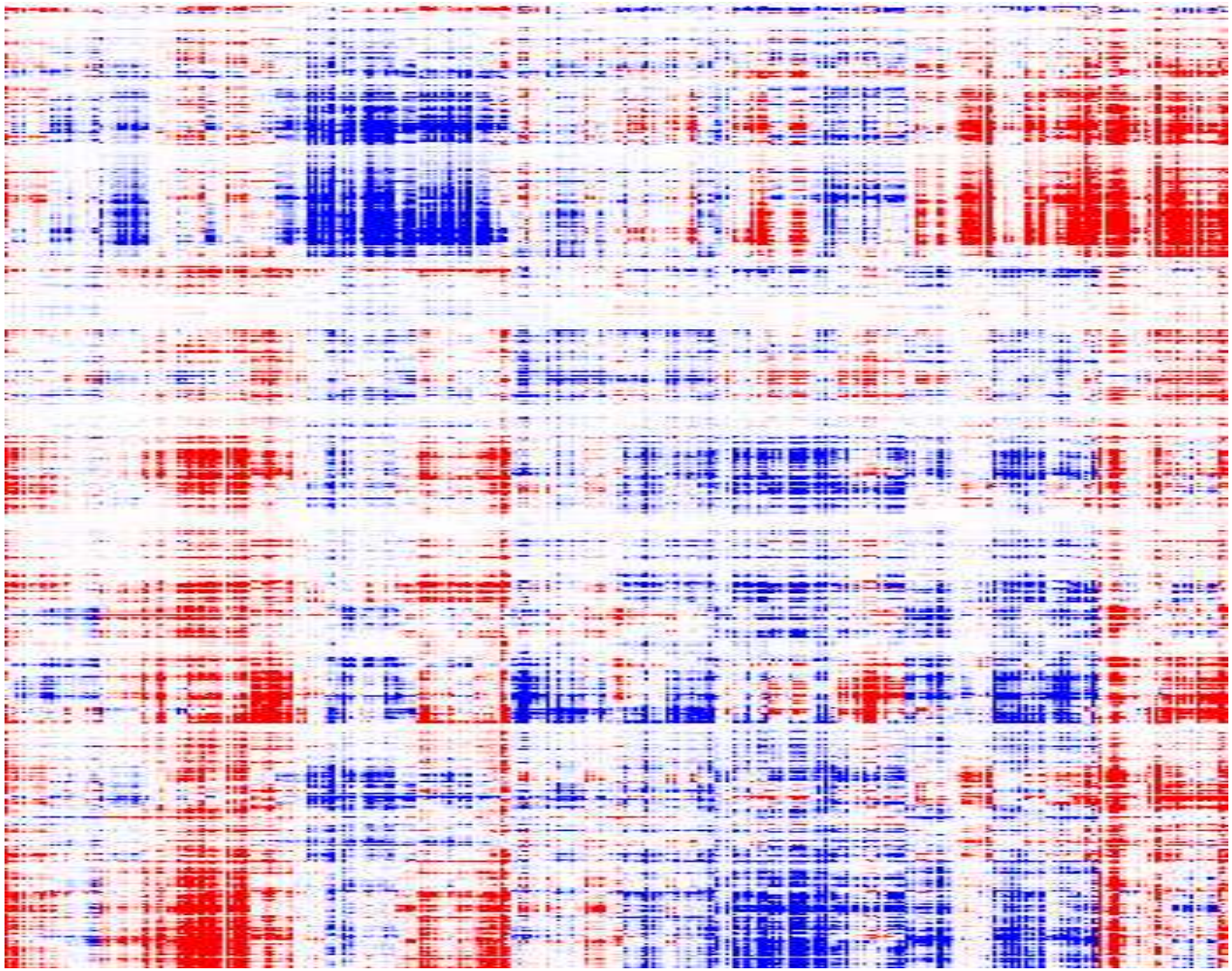


JIVE Estimates

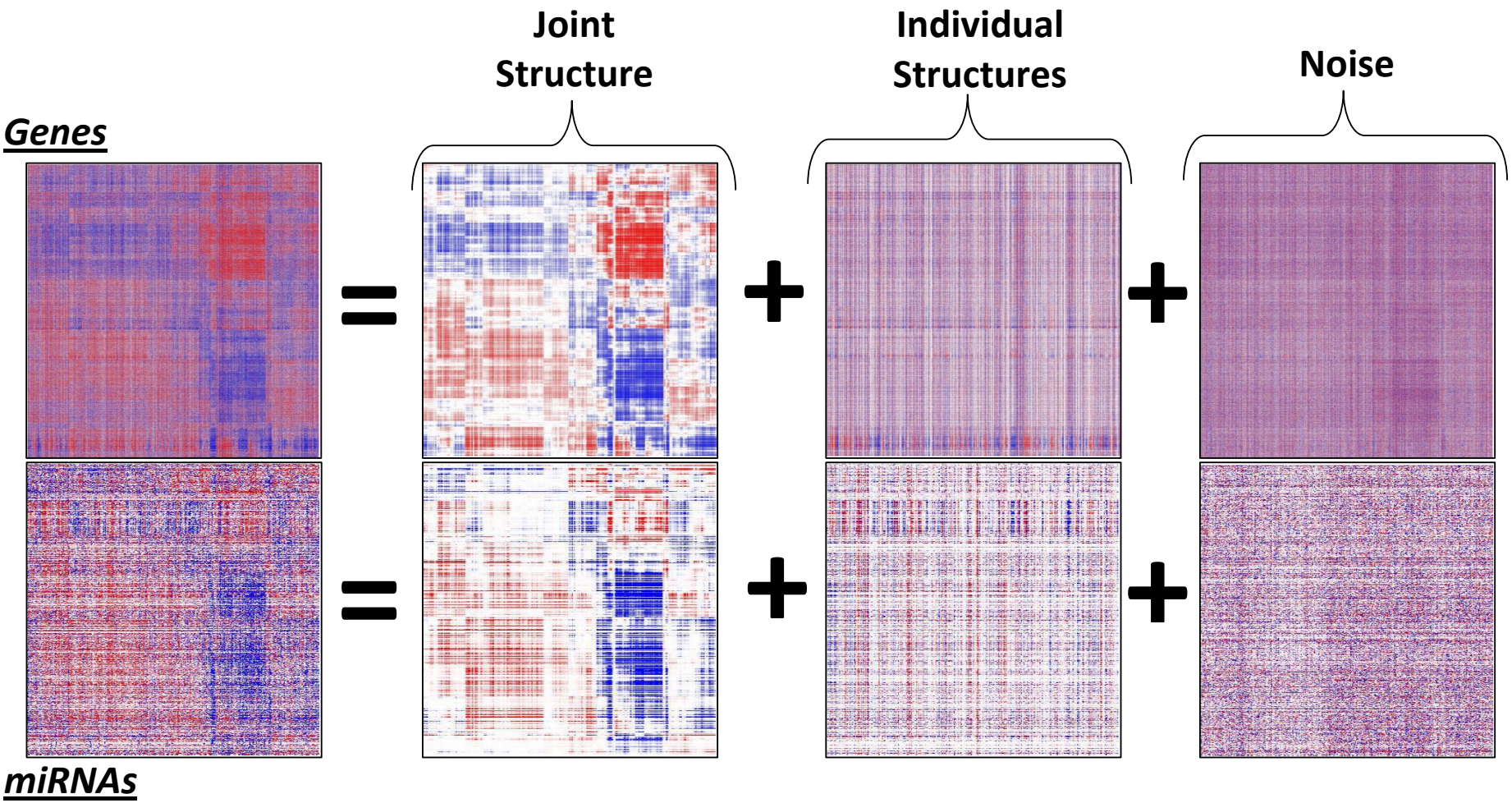


JIVE Estimates

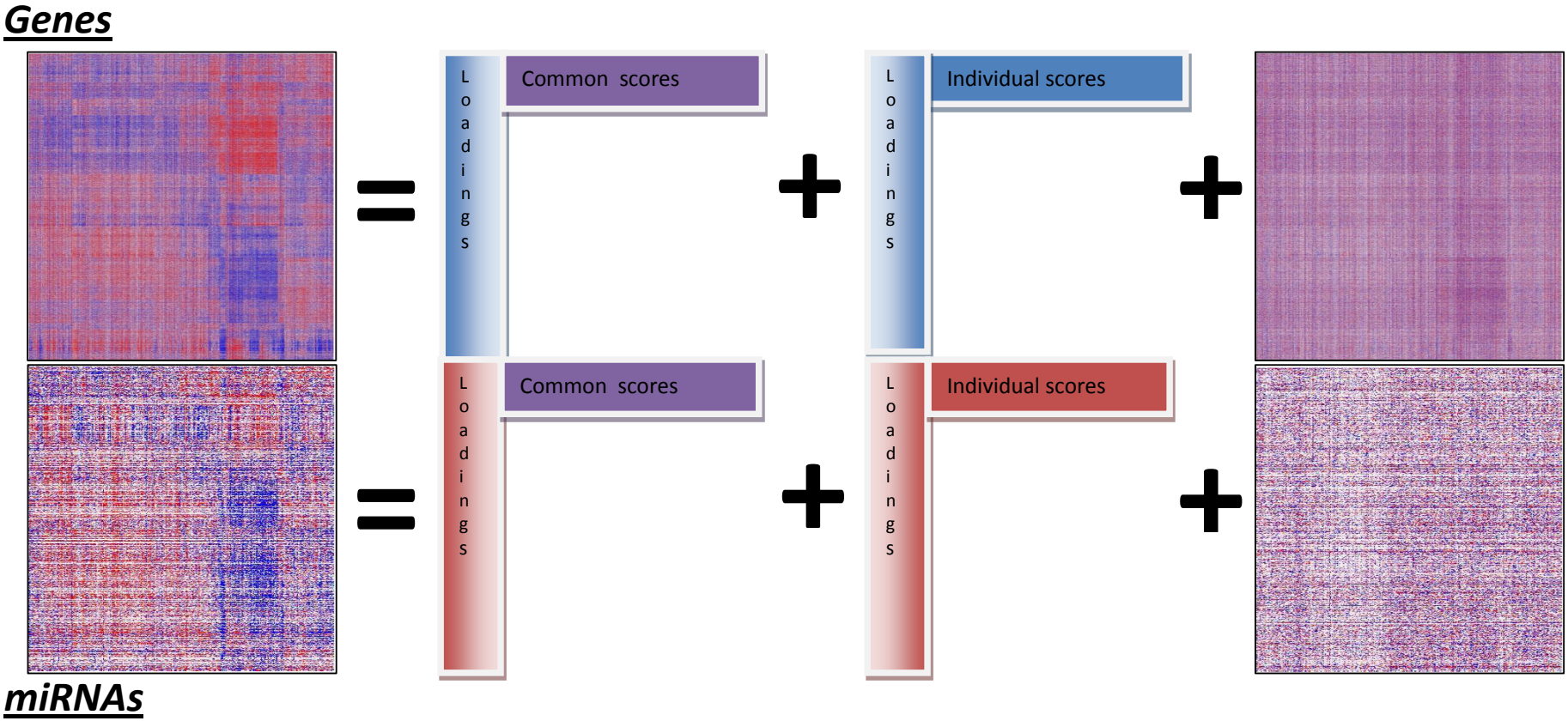
- miRNA individual (reorder rows and columns)



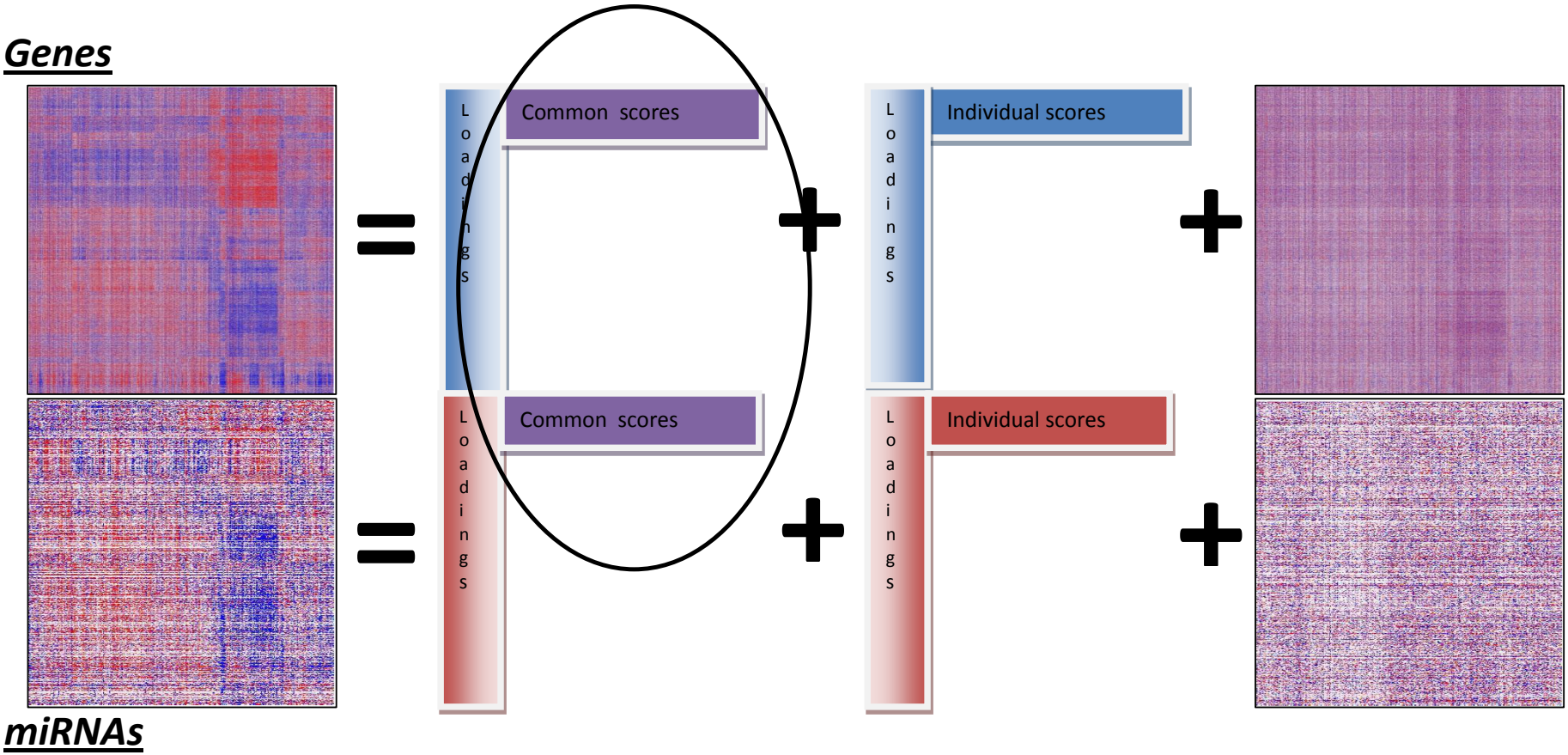
JIVE Estimates



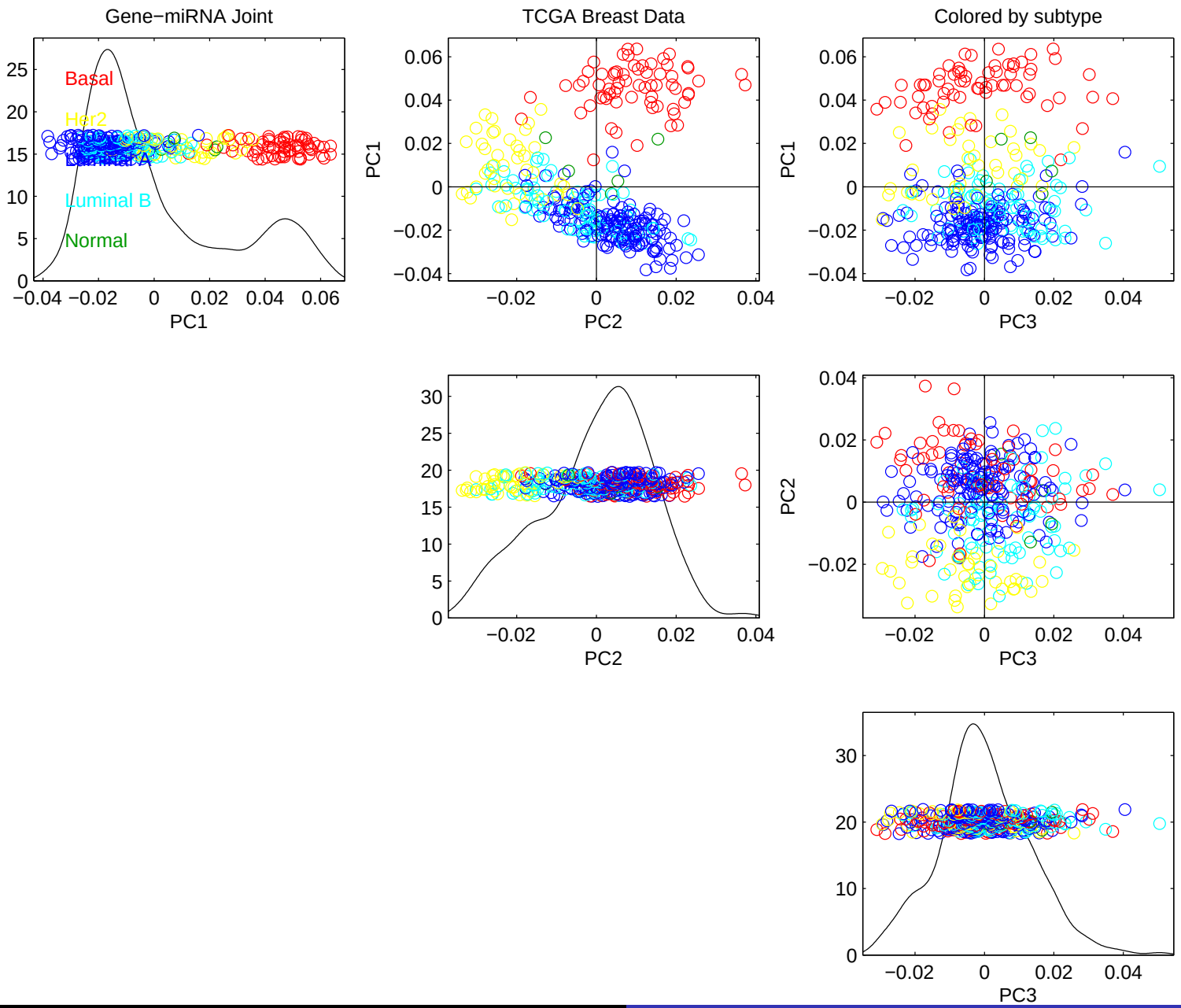
JIVE Estimates (factorized)



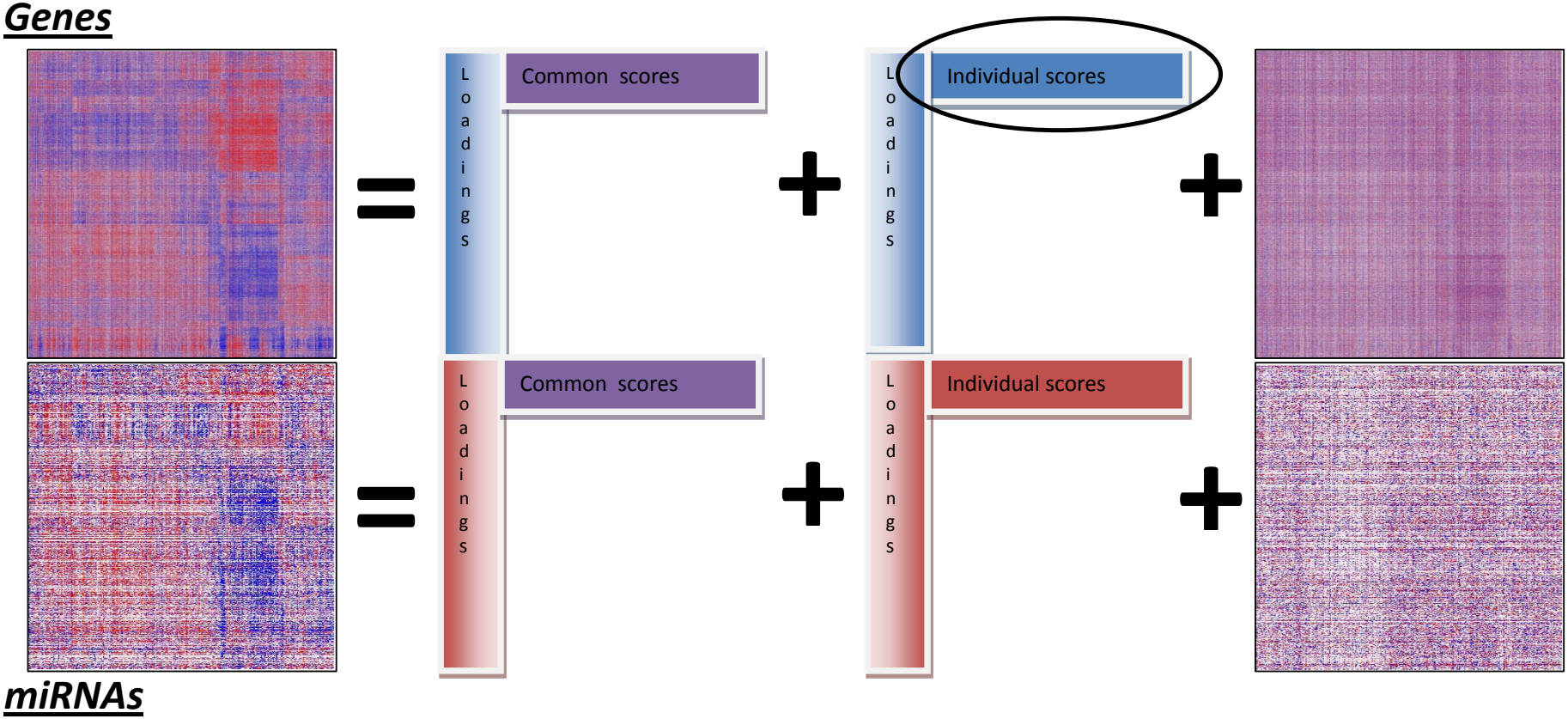
JIVE Estimates (factorized)



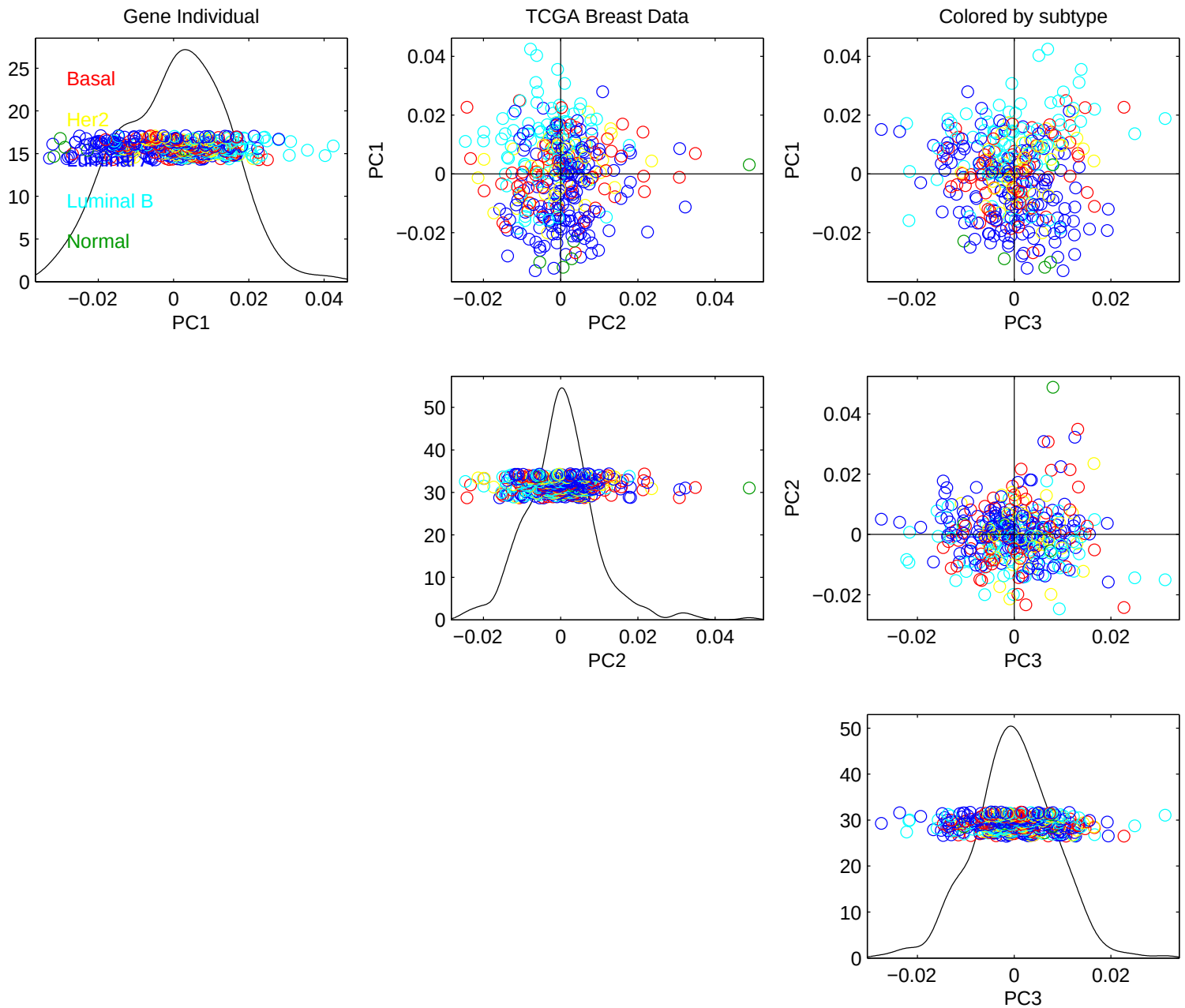
Joint PCs


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

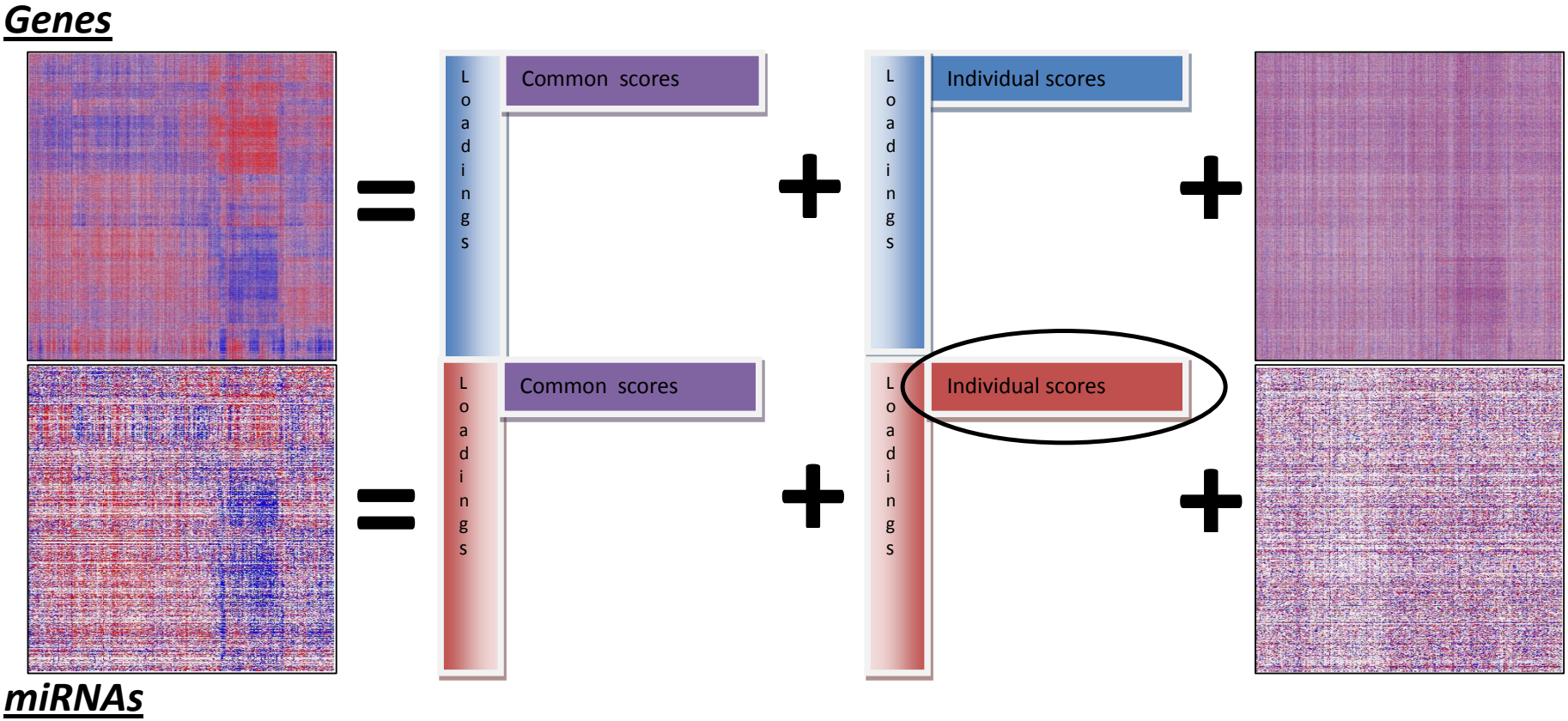
JIVE Estimates (factorized)



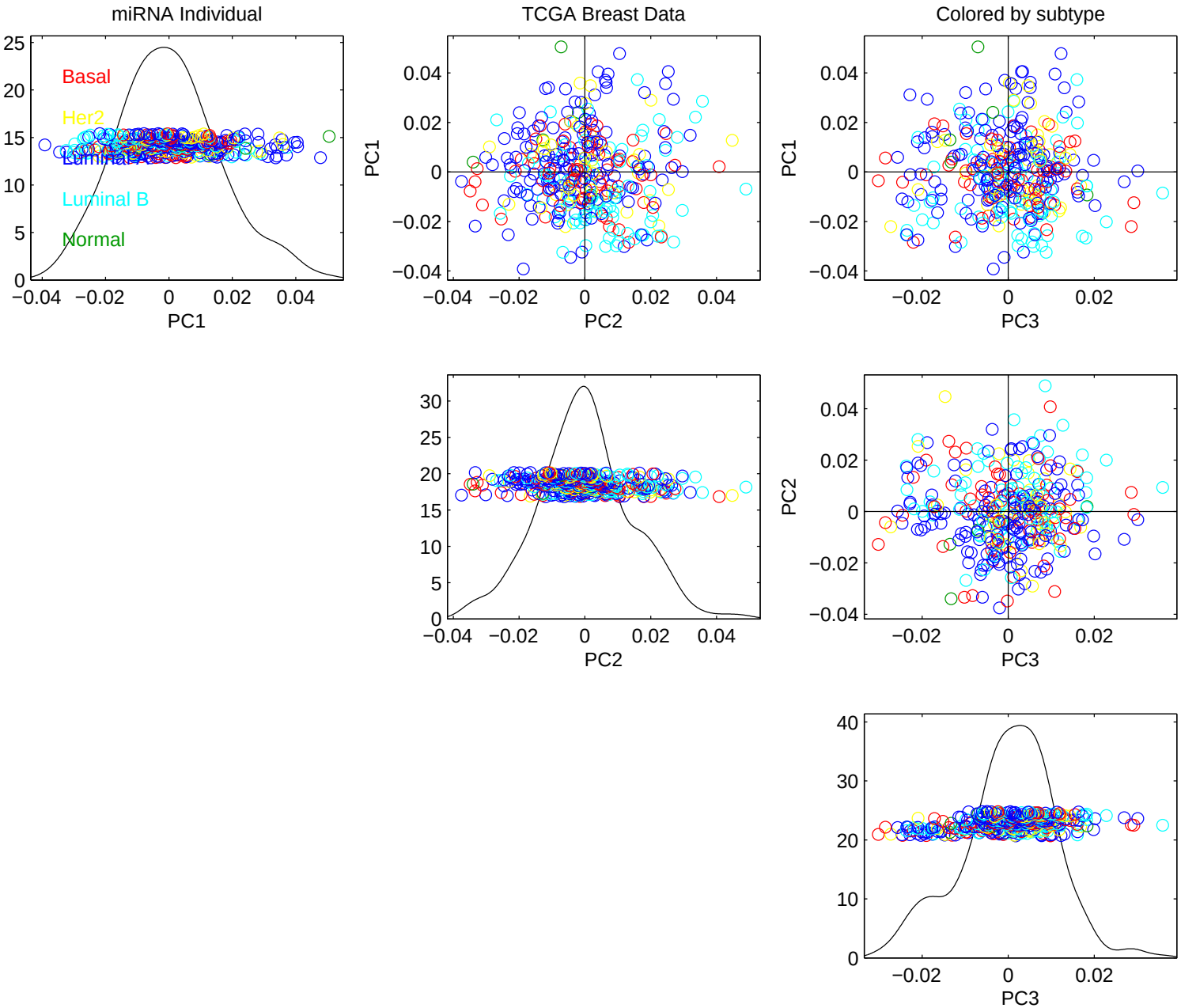
Individual PCs: Expression



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.

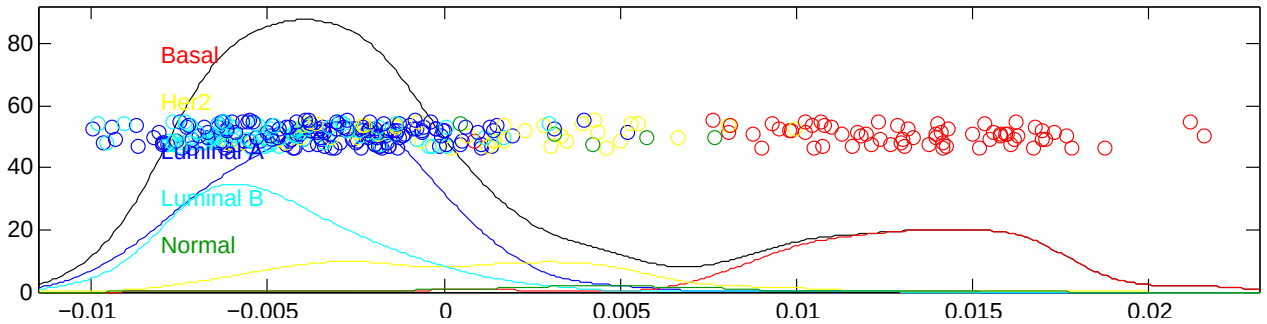
- E.g, Pen may be an L_1 penalty, corresponding to the Lasso:

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

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

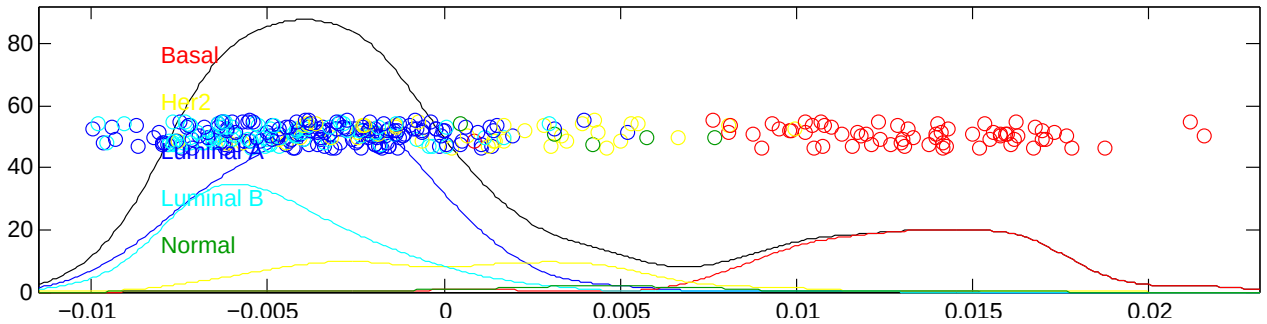
Gene-miRNA Sparse JIVE

- First "Sparse" joint component sample scores:

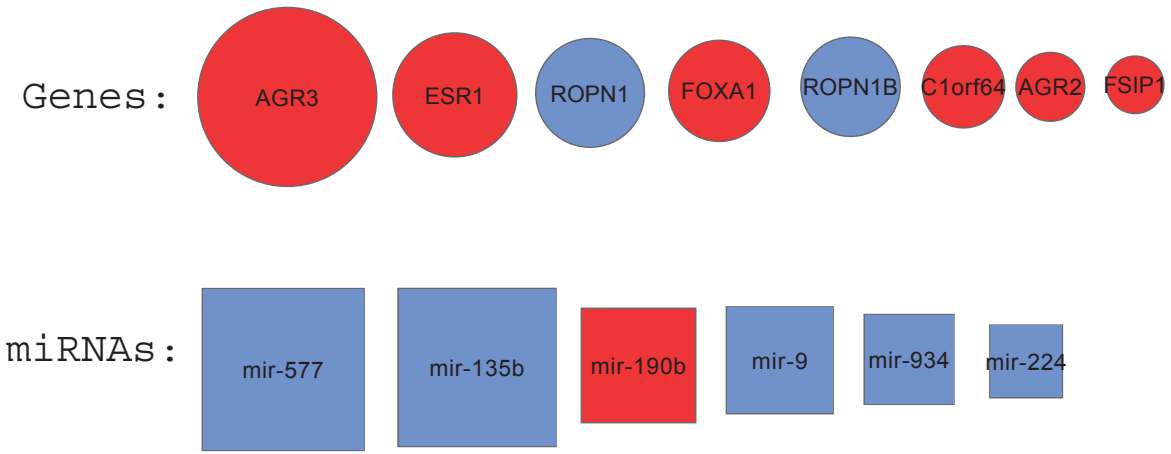


Gene-miRNA Sparse JIVE

- First "Sparse" joint component sample scores:



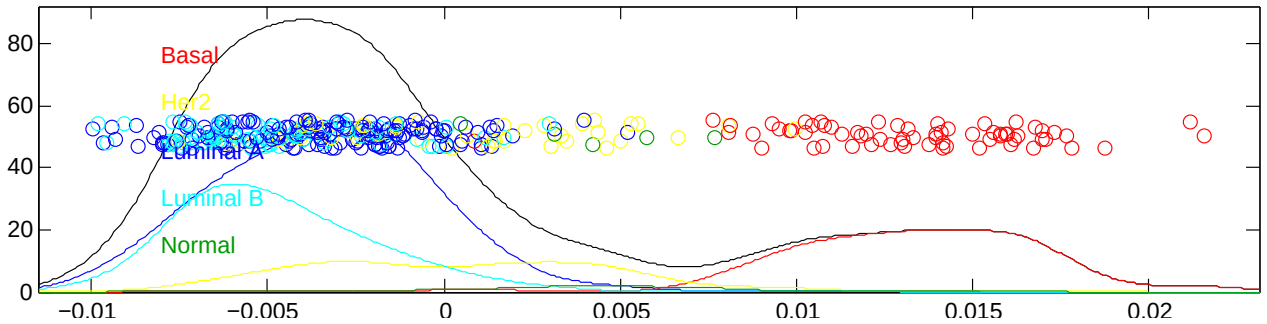
- Genes and miRNAs with non-zero loadings:



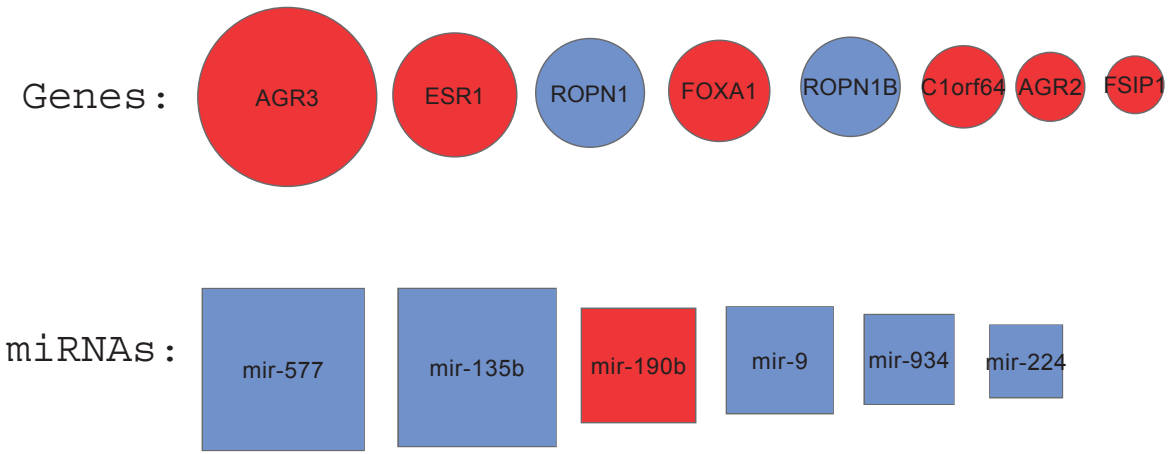
- red: positive loading; blue: negative loading

Gene-miRNA Sparse JIVE

- First "Sparse" joint component sample scores:



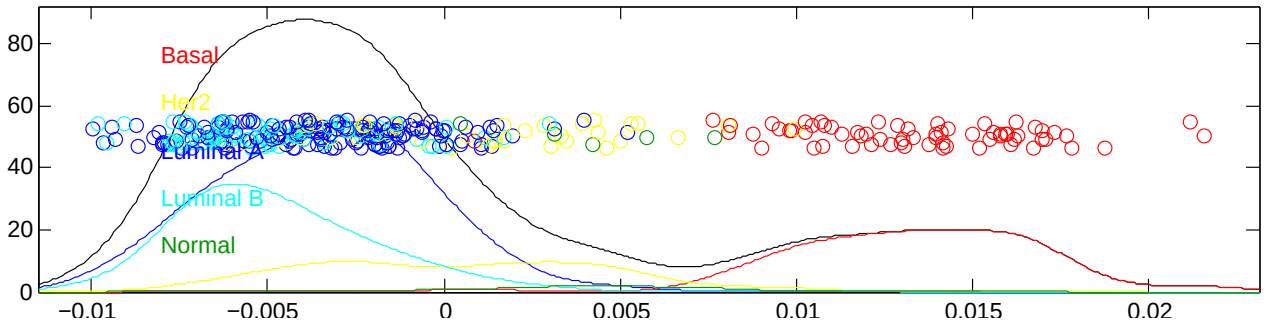
- Genes and miRNAs with non-zero loadings:



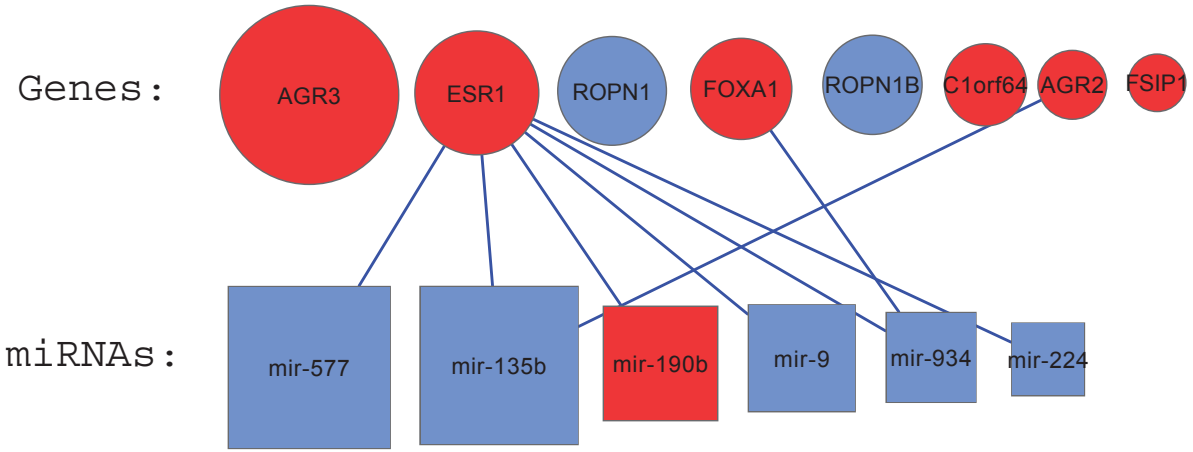
- red: positive loading; blue: negative loading
- miRNA linked if gene is a predicted target in at least two of *Pictar*, *miRanda*, *TargetScan* and *RNA22*

Gene-miRNA Sparse JIVE

- First "Sparse" joint component sample scores:



- Genes and miRNAs with non-zero loadings:



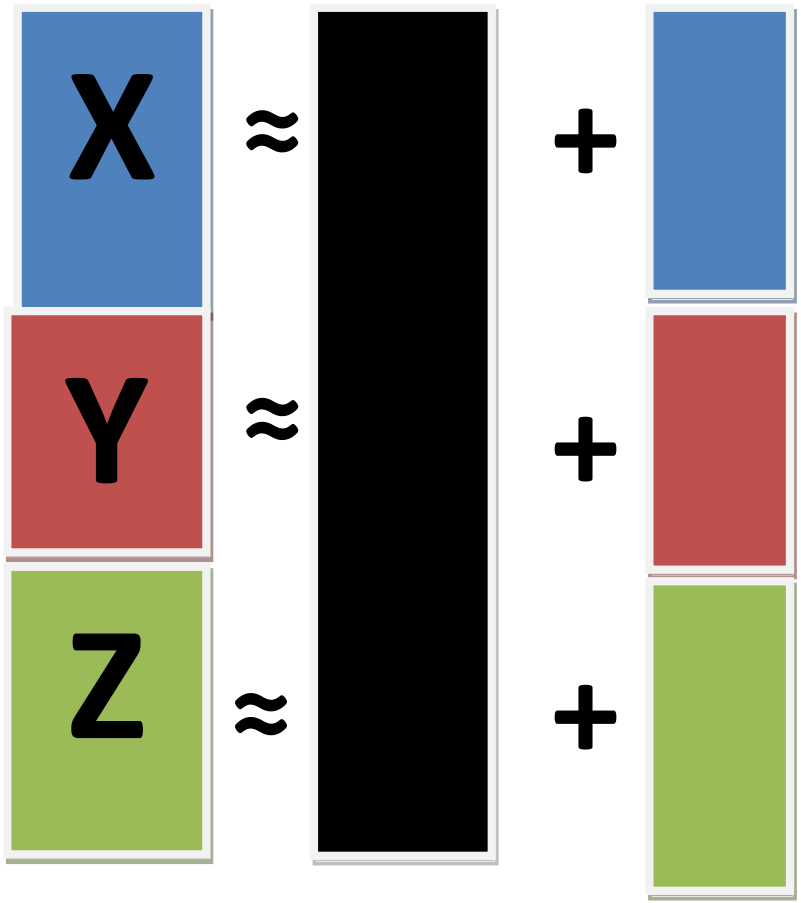
- red: positive loading; blue: negative loading
- miRNA linked if gene is a predicted target in at least two of *Pictar*, *miRanda*, *TargetScan* and *RNA22*

JIVE: 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-dimensional cases (esp. CCA)
  - Interference from individual structure (esp. PLS)
- Multi-level PCA models
  - C Di et al., 2009; L Zhou et al., 2010.
  - Analysis of hierarchical sampling structure, same data source
  - Global component models differences between sampling groups, not shared structure
- Related multi-source factorization models
  - CIFA (Z. Guoxo et al., 2014)
  - Bayesian joint analysis (P. Ray et al., 2014)
  - JINMF (Yang & Michailidis, 2015)

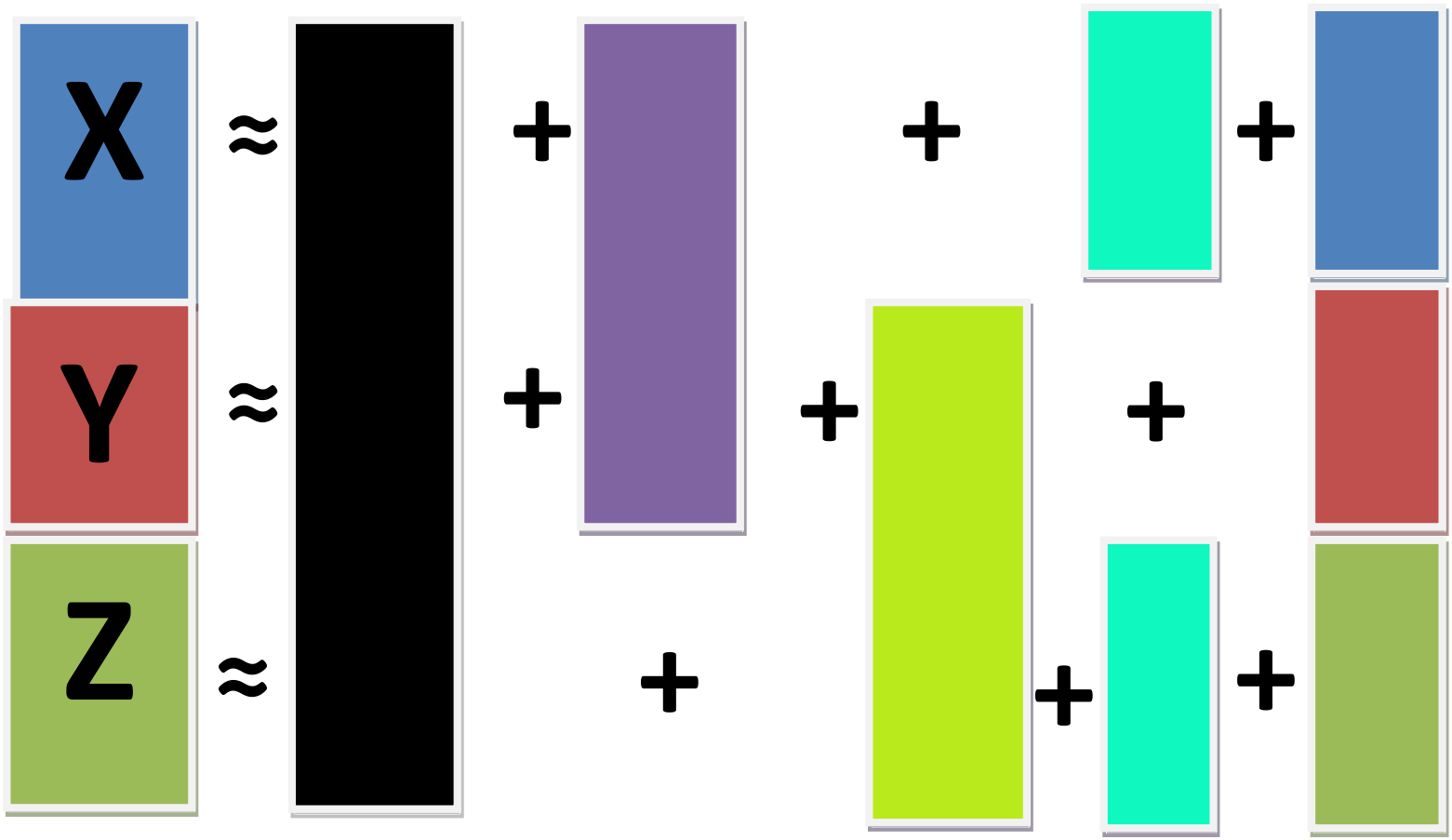
Future work: Factorial JIVE

- More than two datasets (standard JIVE):



Future work: Factorial JIVE

- Factorial model:



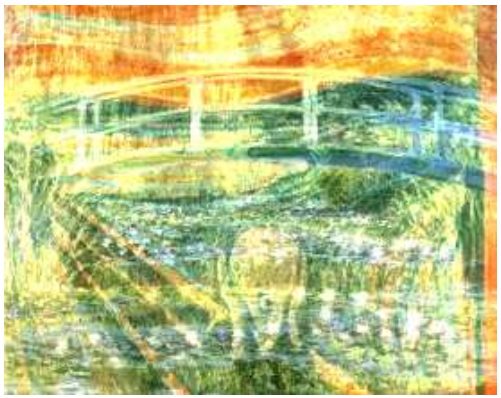
Future work: higher-order arrays

- JIVE and BCC apply to collection of 2D arrays
- One dimension in common
  - Same columns (samples) different rows (variables)
  - Same rows (variables) different columns (samples)
- What if both dimensions are common?
- What about higher-order arrays?

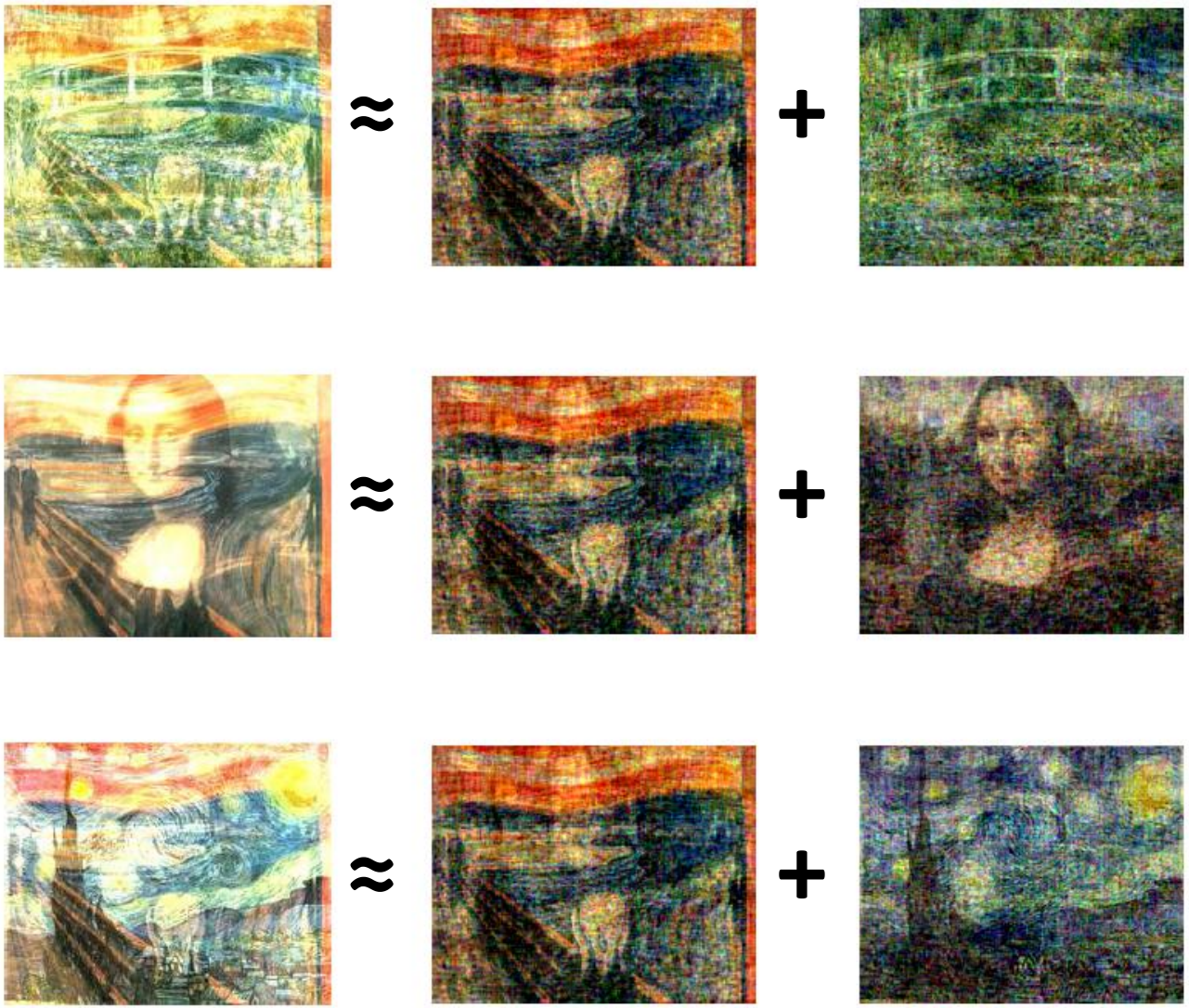
Future work: higher-order arrays

- JIVE and BCC apply to collection of 2D arrays
- One dimension in common
  - Same columns (samples) different rows (variables)
  - Same rows (variables) different columns (samples)
- What if both dimensions are common?
- What about higher-order arrays?

Mixed Art



Mixed Art: Estimated decomposition



Mixed Art: Actual decomposition

Future work: higher-order arrays

- JIVE applies to a collection of 2D arrays
- One dimension in common
  - Same columns (samples) different rows (variables)
  - Same rows (variables) different columns (samples)
- What if both dimensions are common?
- What about higher-order arrays?

Future work: higher-order arrays

- Multiple higher-order arrays $\mathbb{X}_1, \mathbb{X}_2, \ldots$ for a single dataset.
- Some dimensions are shared, some aren't
- Example:
  - $\mathbb{X}_1$: fMRI tensor of order 5, $\mathbb{R}^{\mathbf{N} \times \mathbf{T} \times d_x \times d_y \times d_z}$

    $$\textbf{Samples} \times \textbf{Time} \times X \times Y \times Z$$

  - $\mathbb{X}_2$: Gene expression time course tensor of order 3

    $$\textbf{Samples} \times \textbf{Time} \times \text{Genes}$$

  - $\mathbb{X}_3$: Genotype data matrix $\mathbb{R}^{\mathbf{N} \times d_s}$

    $$\textbf{Samples} \times \text{SNPS}$$

- Goal: general integrative models for shared dimensions

Thank you!

- BCC reference
  - EF Lock and DB Dunson. **Bayesian Consensus Clustering**, *Bioinformatics*, 29 (20), 2013.
- JIVE reference
  - EF Lock, KA Hoadley, JS Marron, and AB Nobel. **Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types**. *Annals of Applied Statistics*, 7 (1), 2013.
- TCGA breast data reference
  - Cancer Genome Atlas Network. **Comprehensive molecular portraits of human breast tumours**.*Nature*, 490 (7418), 2012.
- Software for JIVE (Matlab) and BCC (R) is available at
  - `www.tc.umn.edu/~elock/software`