Lecture 3
Dimension Reduction for Multi-Omics Data

April 18, 2023

Instructor: **Jack Pattee**

Outline

Dimension Reduction

- Principal components analysis

Multi-omics Dimension Reduction

- Stacked PCA
- Joint and Individual Variation Explained

Method Comparison

- Clustering
- Association
- Prediction

Dimension Reduction

- Dimension reduction seeks to explain the variation in a set of 'independent' or predictor variables using a set of derived features.
  - Often in the case of high-dimensional data.
- Considered a form of unsupervised machine learning; i.e., it does not consider an outcome variable.
- Many uses for dimension reduction
  - Clustering: grouping subjects or variables in the dataset.
  - Feature engineering: generating a set of derived features for subsequent use, i.e., in a predictive modeling application.
  - Visualization of high dimensional data.
  - Etc.

Principal Components Analysis

- A popular method for dimension reduction is principal components analysis (PCA).
- PCA projects the dataset into a new vector space with an orthonormal basis.
- If we consider our features to be encoded in the *n*-sample by *p*-feature matrix **X**, we can think of the principal components representation as closely related to the singular value decomposition of **X**.

Singular Value Decomposition

- According to SVD, any $n \times p$ matrix $\mathbf{X}$ can be decomposed into:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

  - $\mathbf{U}$ is an $n \times n$ matrix where columns are orthonormal.
  - $\boldsymbol{\Sigma}$ is an $n \times p$ rectangular diagonal matrix with entries equal to the square root of the eigenvalues of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$.
  - $\mathbf{V}$ is a $p \times p$ matrix with orthonormal columns.
- Note: $\boldsymbol{X}^T\boldsymbol{X} = \cdots = \boldsymbol{V}^T\boldsymbol{\Sigma}^2\boldsymbol{V}$, which satisfies the conditions for an eigendecomposition.
  - Thus, columns of $\mathbf{V}$ are equivalent to eigenvectors of the covariance matrix of $\mathbf{X}$
- Define $\boldsymbol{W}^T = \boldsymbol{\Sigma}\boldsymbol{V}^T$
- This gives us the PCA decomposition: $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}$.

PCA

Consider representation $\boldsymbol{T = XW}$.

- **T** is the principal component representation of our data, i.e., an $n \times p$ matrix where our original data has been projected into an orthonormal basis space. Each column is a principal component.
- **T** has two special properties:
  - Each column of **T** is orthogonal to all other columns.
  - Each column of **T** explains more variation than all subsequent columns and less variation than all preceding columns.
- **W** is the *loading matrix* of the principal component representation.

$$PC1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p$$
$$PC2 = w_{21}X_1 + w_{22}X_2 + \cdots + w_{2p}X_p$$
$$PC3 = w_{31}X_1 + w_{32}X_2 + \cdots + w_{3p}X_p$$

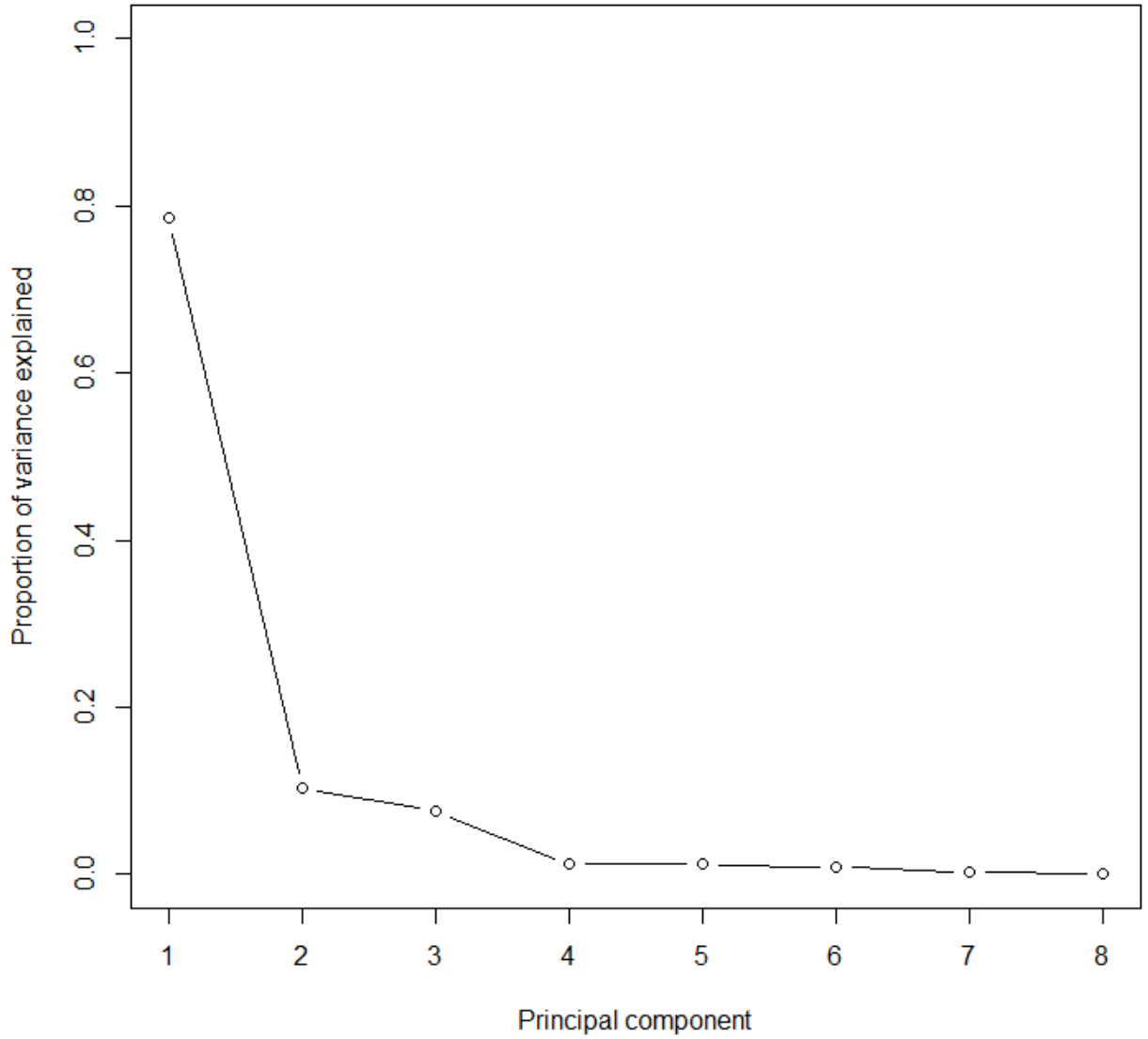
PCA

- If we want to generate a 1-dimensional representation of our data **X** in such a way that preserves the most variation in our data, this is done by projecting the data onto the first eigenvector of the covariance matrix.
  - i.e., $\boldsymbol{XW_1}$ has the largest sample variance among all (normalized) linear combinations of the columns of **X**.
- We can reduce the number of eigenvectors we use to reconstruct our data, thus projecting the data onto only those top few eigenvectors of the covariance matrix:

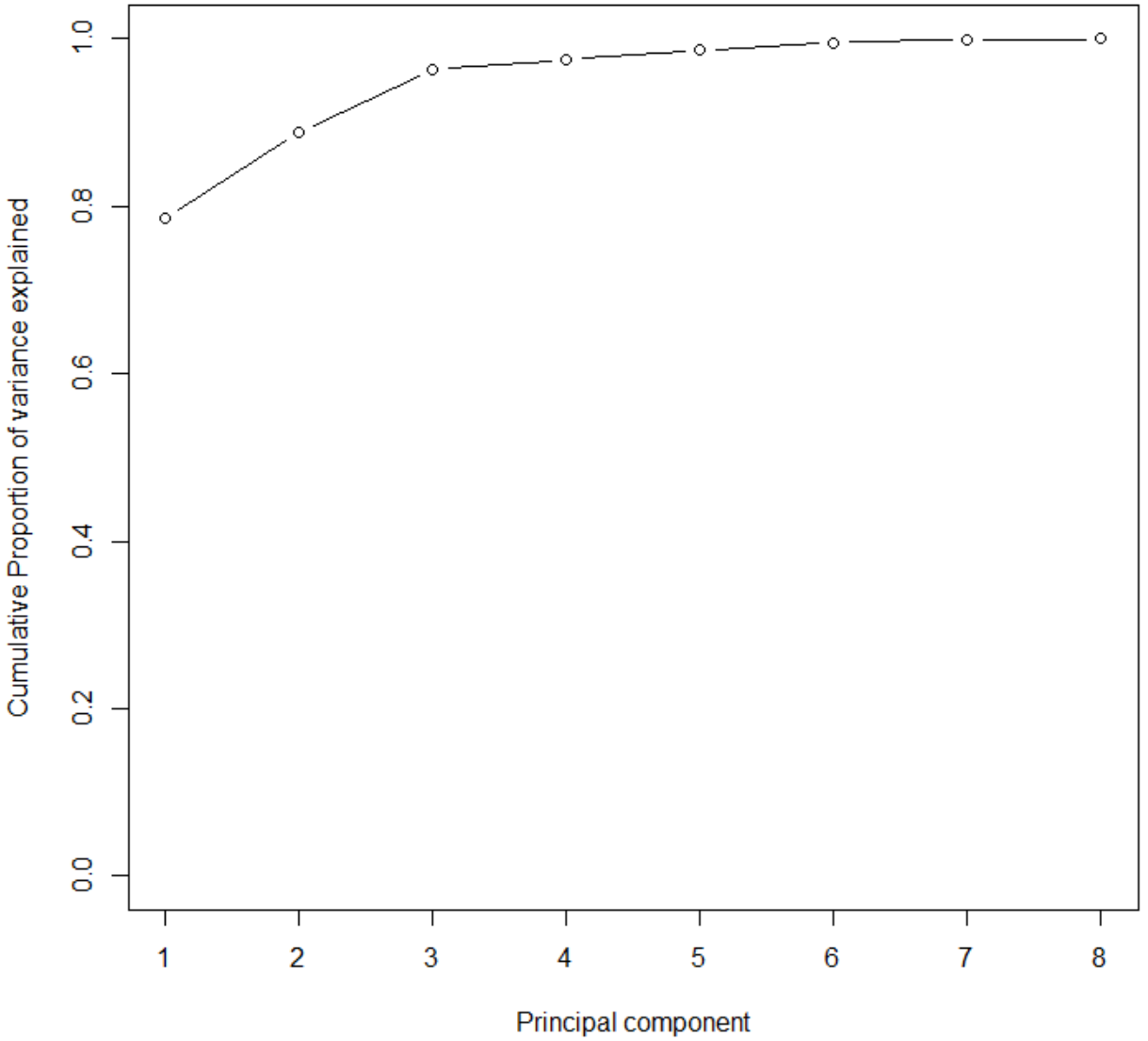
$$\boldsymbol{T}_L = \boldsymbol{XW}_L$$

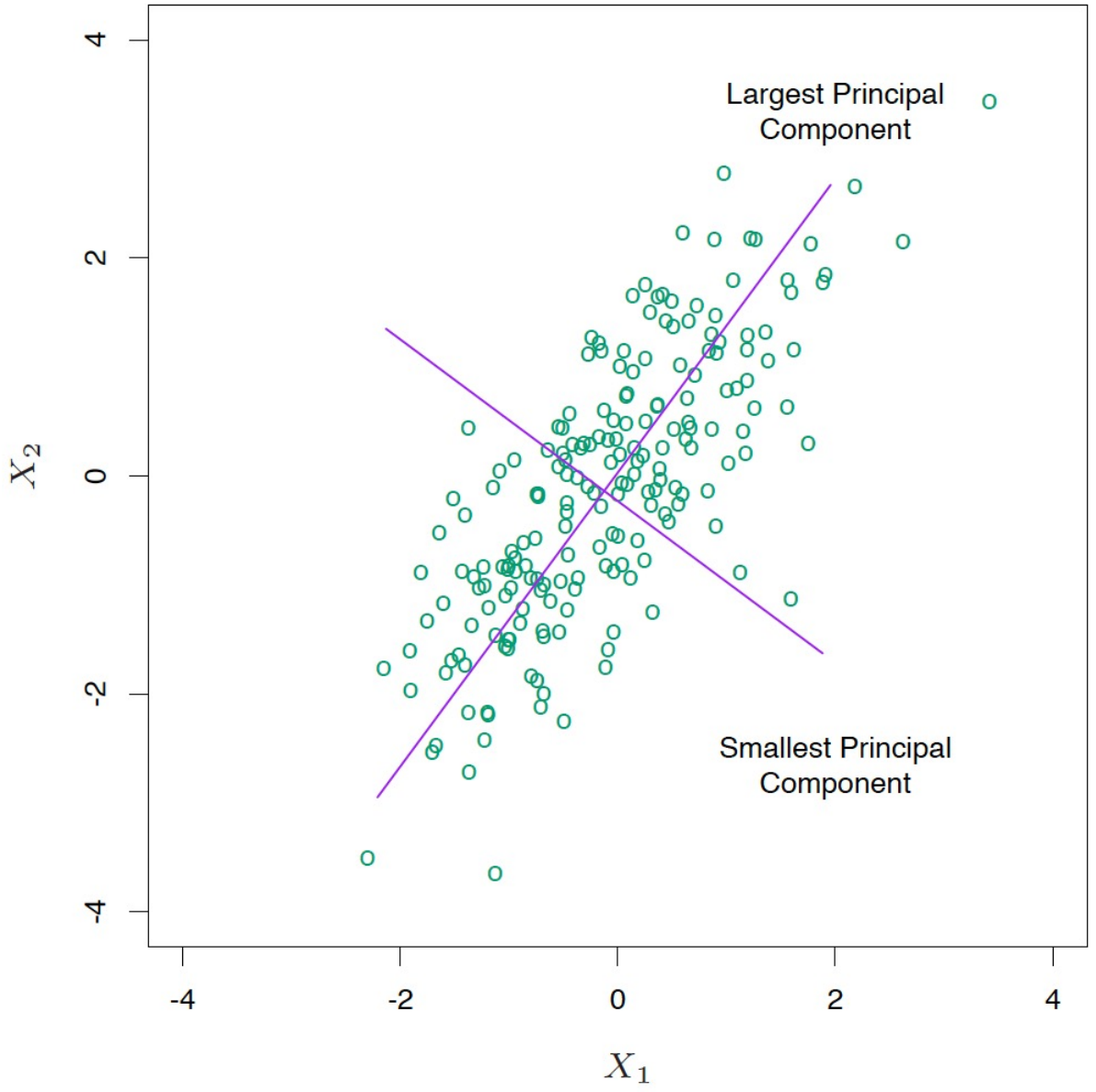
Where $\boldsymbol{W}_L$ retains only the first L eigenvectors.

Proportion of total variance explained by each PC

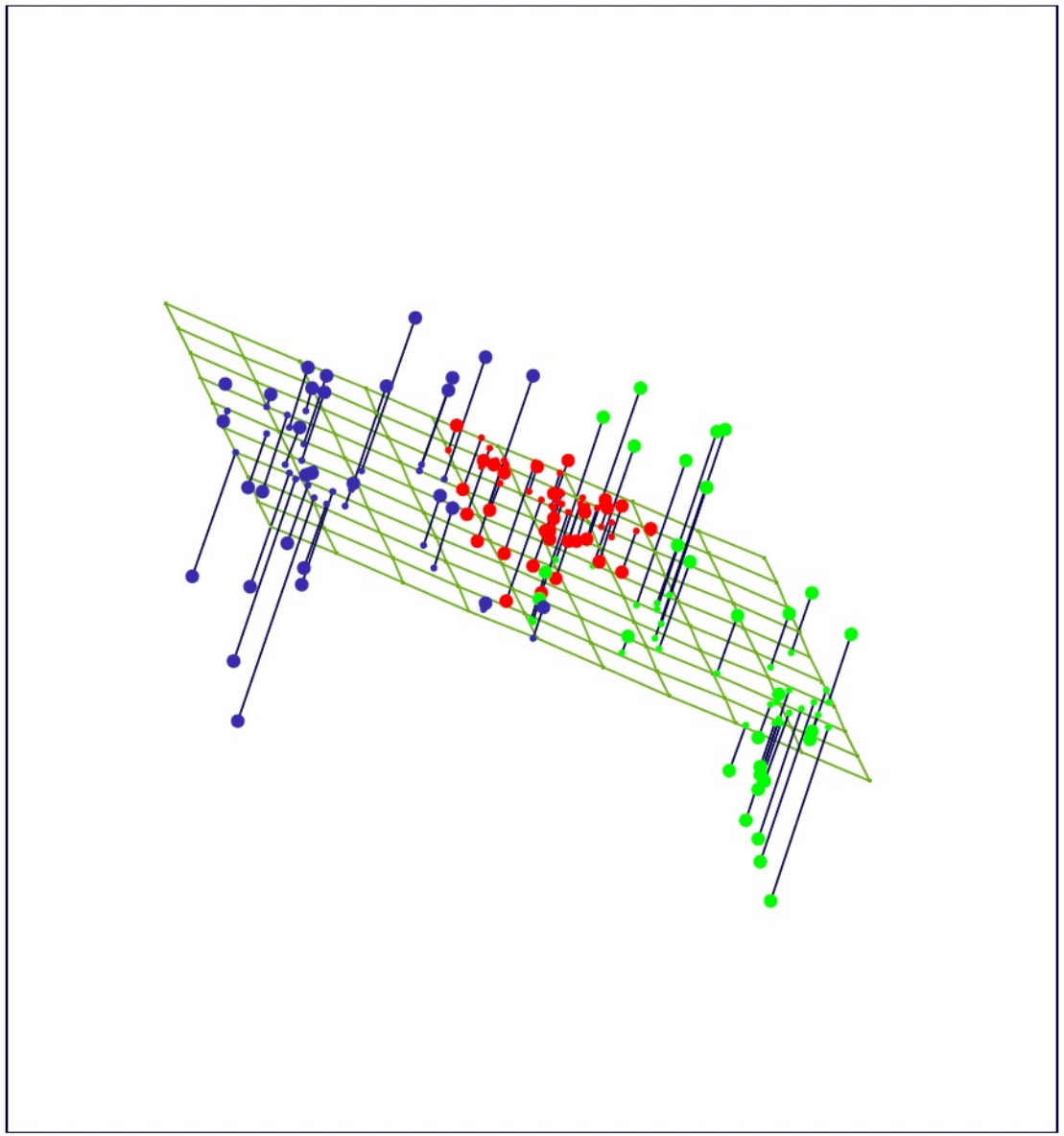


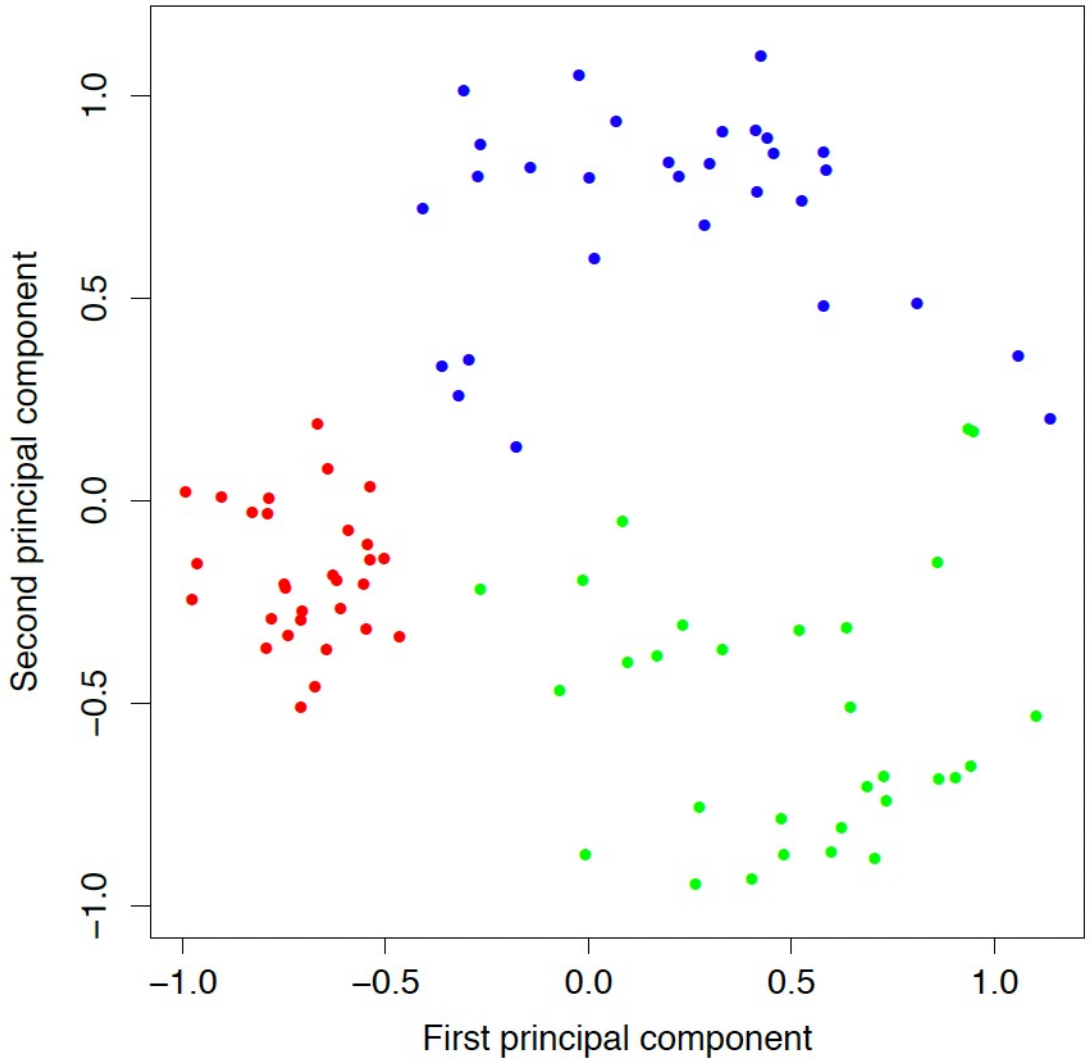
Cumulative proportion of variance explained





Elements of Statistical Learning (2nd Ed.), Hastie, Tibshirani, and Friedman 2009, Chapter 3





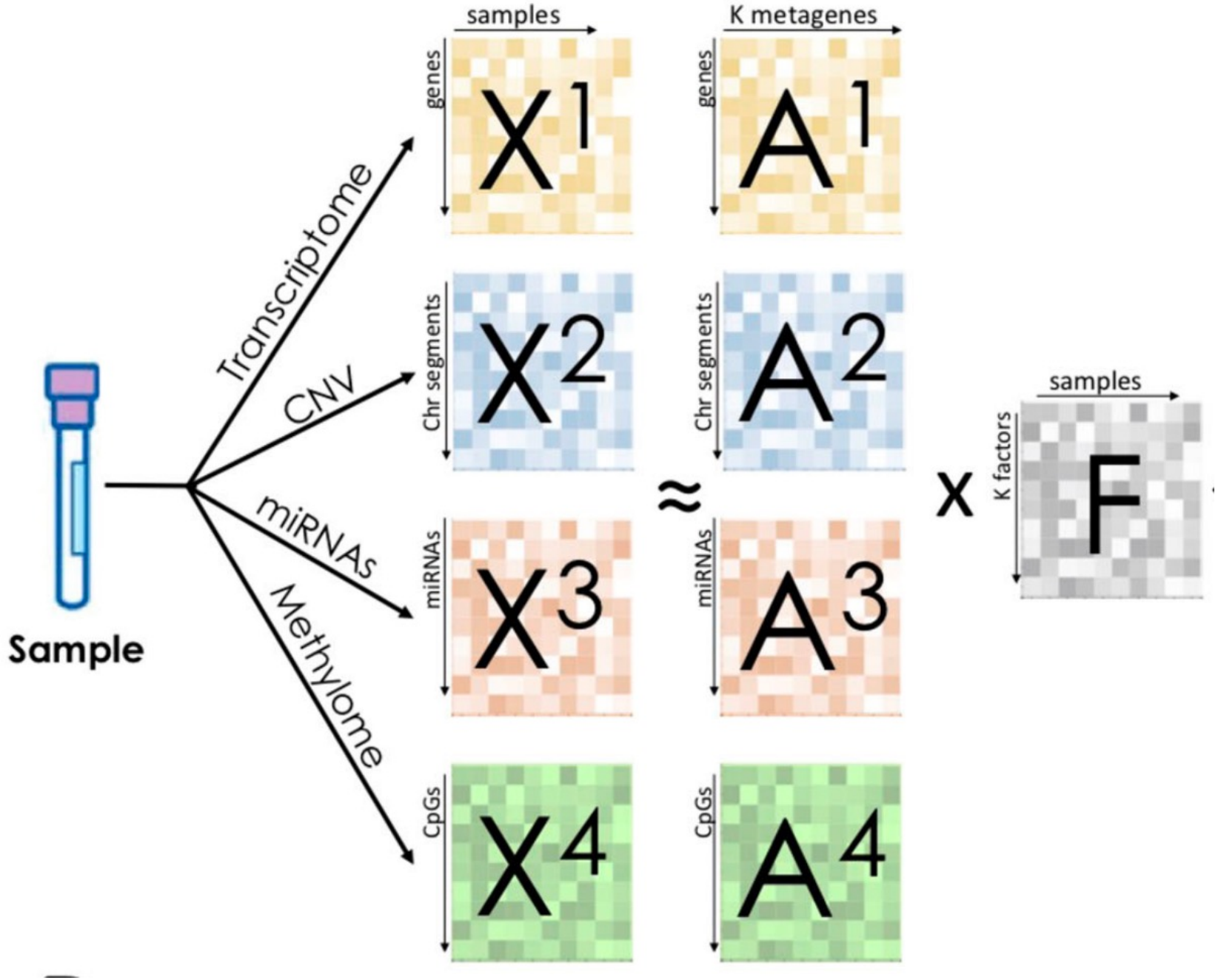
Elements of Statistical Learning (2nd Ed.), Hastie, Tibshirani, and Friedman 2009, Chapter 14

Dimension Reduction for Multi-omics

- According to Cantini et al: “Multiomics integrative approaches should be able to capture not only signals shared by all omics data, but also those emerging from the complementarity of various omics data”.
- Want a way to model variation shared among datatypes and unique to each datatype.
- Informatively combining information from multiple datatypes allows for more precise analysis.

Multi-omics

- Consider the multi-omics context. We have n individuals where we have measured omic values for some set of k omics types. So, say, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$, for $k \geq 2$, where $\boldsymbol{X}_i$ is $p_i \times n$.
  - Variables are rows, subjects are columns.
  - p_i can vary widely according to study and data type; transcriptomic data may have $p_i \sim 20{,}000$, whereas miRNA may have $p_i \sim 1{,}000$
- We want a dimension-reduced representation of the multi-omic data.



Cantini et al, Nature Communications 2021

Multi-omics

- Rows of the factor matrix **F** are called ‘factors’.
- Columns the weight matrix are called ‘metagenes’.
- Factors are projections onto the sample space, whereas metagenes are projections onto the omics space.
- The factor matrix can be used to cluster samples, whereas columns of weight matrix can be used to extract markers (i.e., selecting the top ranked genes).

Multi-omics

- Consider ‘stacking’ the k matrices:

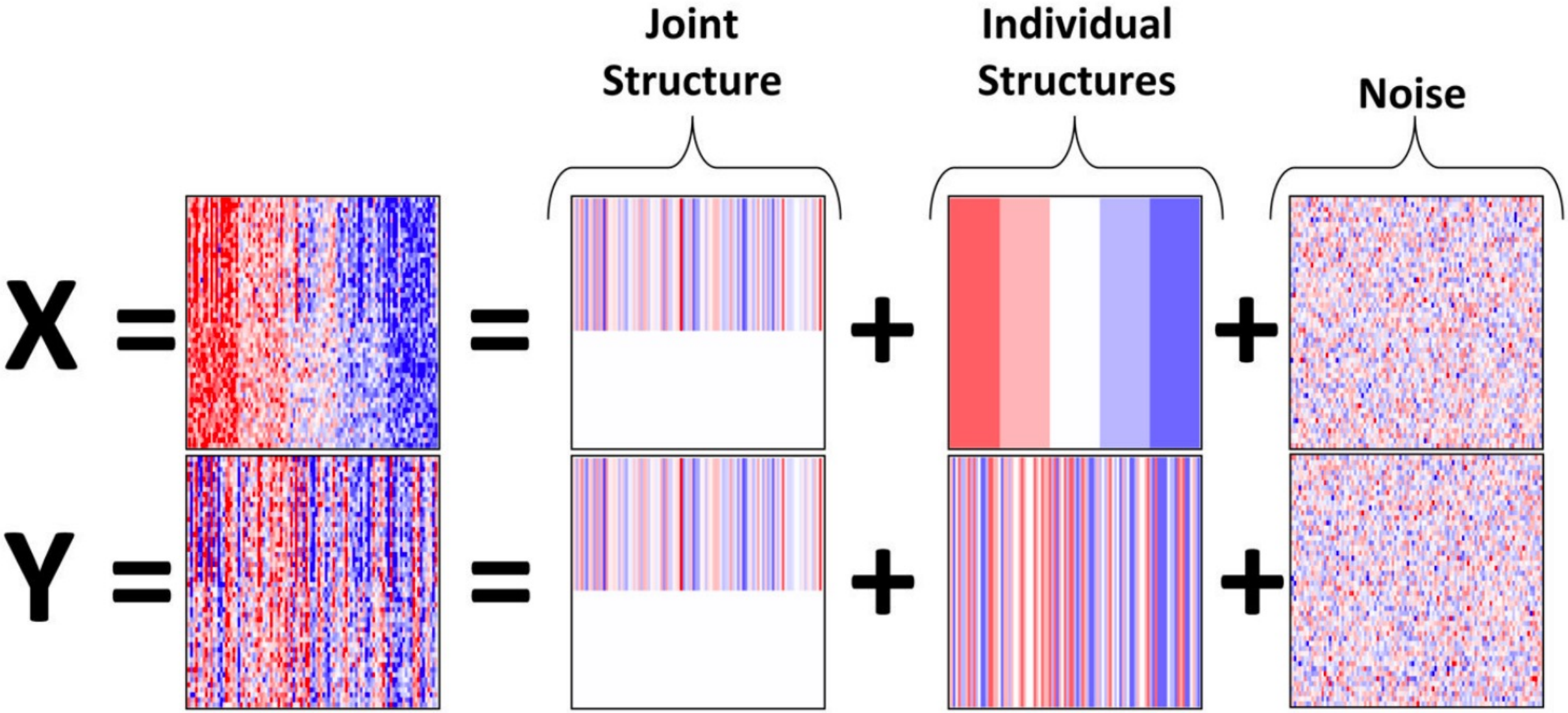
$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_k \end{pmatrix}$$

- Could simply perform PCA on the stacked matrix **X**: this is termed ‘consensus PCA’ (Wold, Kettaneh, and Tjessem, 1996). Consensus PCA is leveraged by popular ‘iCluster’ method (Shen, Olshen, and Ladanyi, 2009).
- However, such an approach cannot measure the so-called ‘joint’ and ‘individual’ structure.
  - Broadly: this is the distinction between whether or not a dimension reduction method estimates omics-specific factors.
- For example: could be certain patterns unique to miRNA and transcriptomic data, and certain patterns shared between the two.

Joint and Individual Variation Explained

- One example of a method that estimates both individual and joint variation is JIVE (Lock et al, 2013).
- Idea is to decompose **X** into joint and individual variation, where matrix $\boldsymbol{J}_i$ encodes the contribution of omics datatype *I* to the joint structure and $\boldsymbol{A}_i$ encodes the individual variation.
- Error terms indicate that we use a reduced representation (i.e., limit the number of ranks used in the reconstruction).

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



Lock et al, Ann Appl Stat 2013

JIVE

- JIVE can be conceptualized as a form of multidimensional PCA.
- **X** has an SVD decomposition, and thus a PCA representation **X=WU**.
- Thus, the rank r joint matrix **J** can be written as the product of a $p \times r$ loading matrix **W** and an $r \times n$ factor matrix **U**:

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_1 \\ \vdots \\ \boldsymbol{W}_k \end{pmatrix}$$

Where each $\boldsymbol{W}_i$ gives the loadings of the joint structure corresponding to datatype i.

- The rank r individual structure matrix $\boldsymbol{A}_i$ for $\boldsymbol{X}_i$ can be written as $\boldsymbol{V}_i\boldsymbol{U}_i$, where $\boldsymbol{V}_i$ is a $p_i \times r_i$ loading matrix and $\boldsymbol{U}_i$ is an $r_i \times n$ factor matrix.

JIVE

- This gives us the full model as follows:

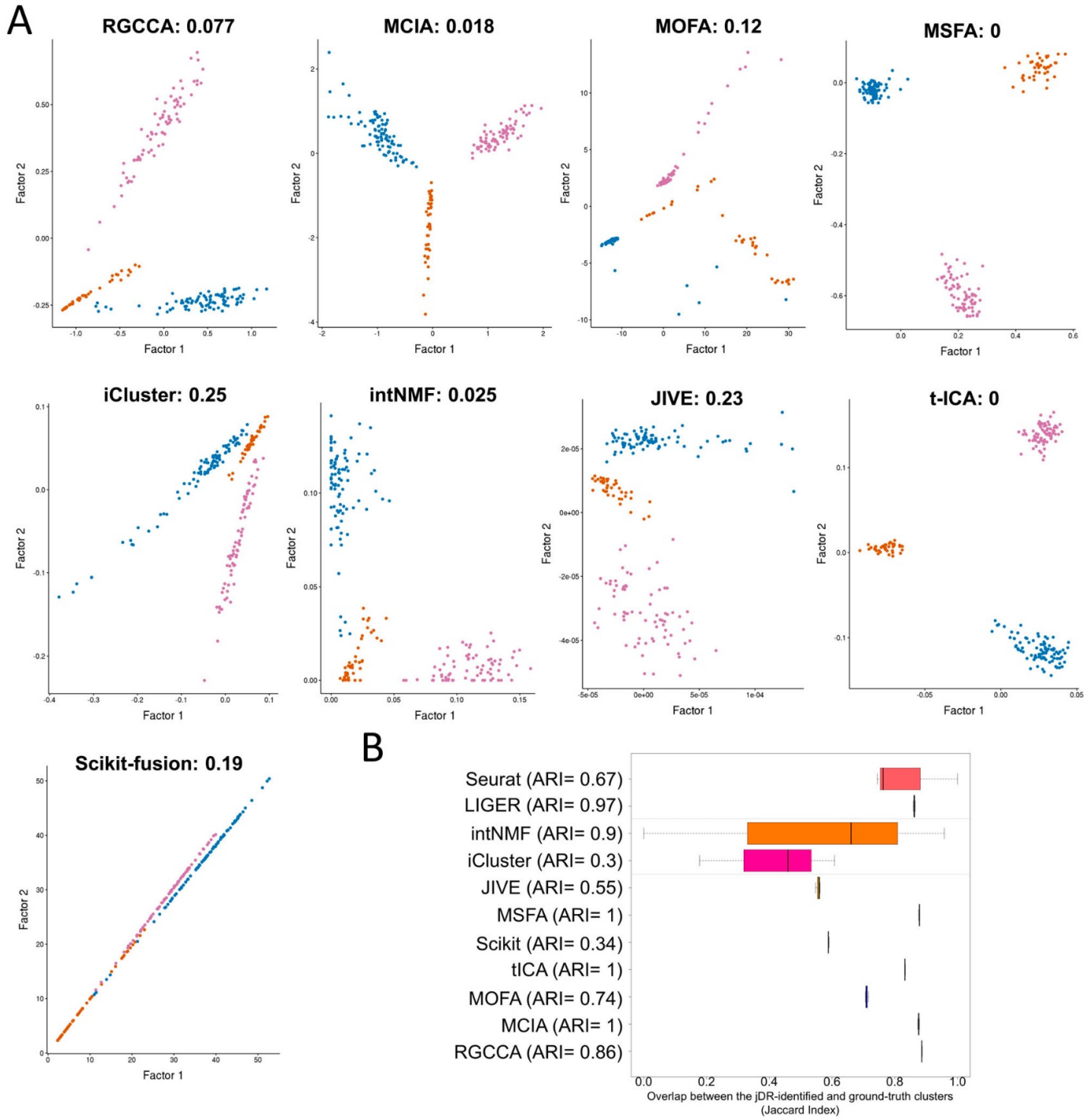
$$\boldsymbol{X}_1 = \boldsymbol{W}_1\boldsymbol{U} + \boldsymbol{V}_1\boldsymbol{U}_1 + \boldsymbol{R}_1$$
$$\vdots$$
$$\boldsymbol{X}_k = \boldsymbol{W}_k\boldsymbol{U} + \boldsymbol{V}_k\boldsymbol{U}_k + \boldsymbol{R}_k$$

Applications

- Now that we have a multi-omic dimension reduction, what can we do with it?
  1. Cluster samples.
  2. Predict clinical outcome.
  3. Associate factors with clinical variables of interest (age, gender).
- Review paper (Cantini et al) compares nine methods: intNMF, JIVE, MCIA, RGCCA, iCluster, MOFA, tICA, MSFA, and data fusion.

Single-Cell Clustering

- Cantini et al used factors derived from multi-omic dimension reduction methods to cluster samples in single cell data.
  - These methods were originally designed for bulk data, but characterizing their performance on single-cell data is of interest.
- Two methods, intNMF and iCluster, intrinsically perform clustering; for the other seven methods, k-means consensus clustering was applied to the factor matrix.
- Two ‘omics types: scRNA-seq and scATAC-seq, which measure gene expression and chromatin accessibility, respectively.
- Applied to 206 cells from three cancer cell lines.

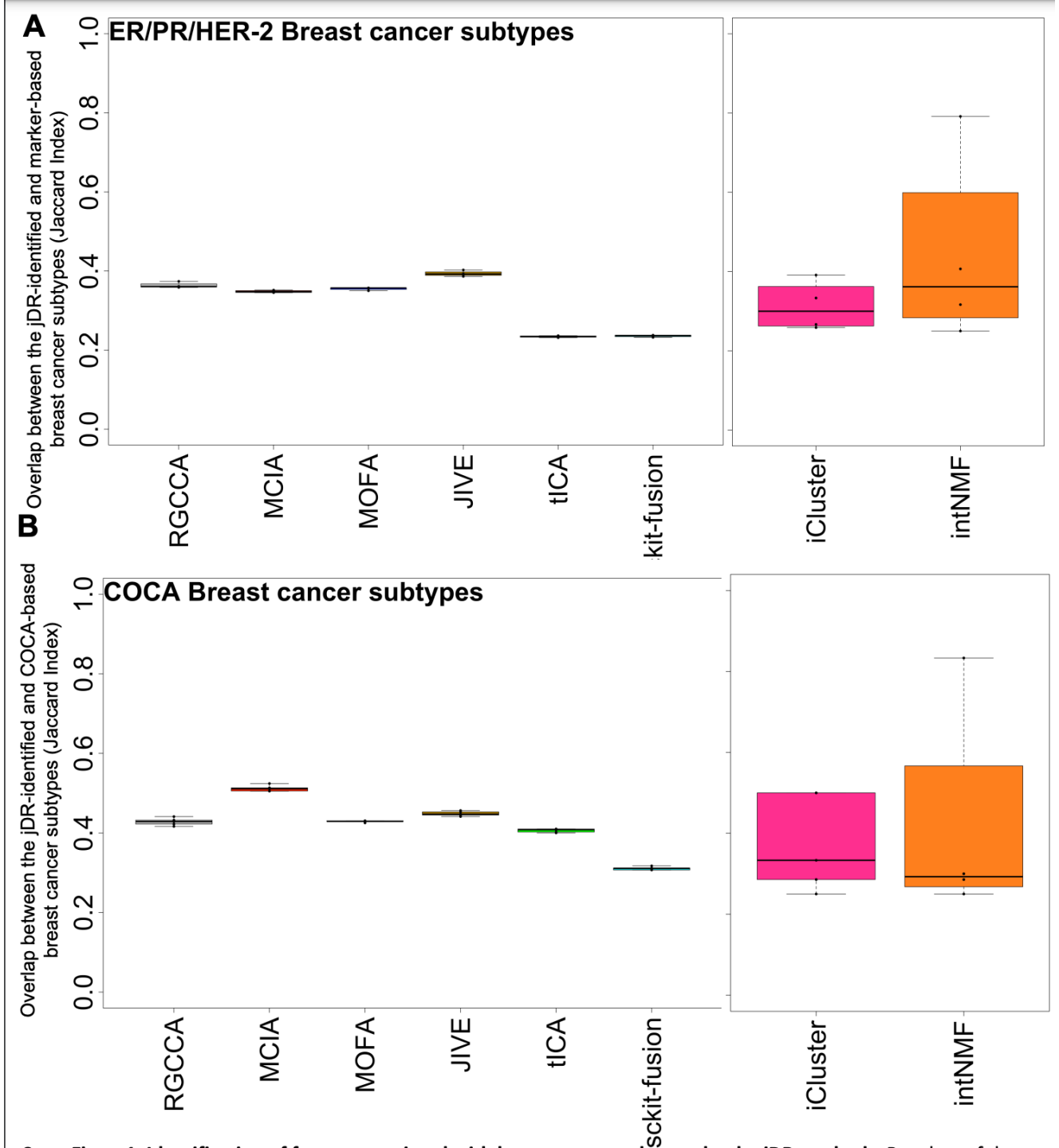

A
RGCCA: 0.077
MCIA: 0.018
MOFA: 0.12
MSFA: 0
iCluster: 0.25
intNMF: 0.025
JIVE: 0.23
t-ICA: 0
Scikit-fusion: 0.19
Factor 1
Factor 2
B
Seurat (ARI= 0.67)
LIGER (ARI= 0.97)
intNMF (ARI= 0.9)
iCluster (ARI= 0.3)
JIVE (ARI= 0.55)
MSFA (ARI= 1)
Scikit (ARI= 0.34)
tICA (ARI= 1)
MOFA (ARI= 0.74)
MCIA (ARI= 1)
RGCCA (ARI= 0.86)
Overlap between the jDR-identified and ground-truth clusters
(Jaccard Index)

Application to TCGA Data

- Analyzed data from The Cancer Genome Atlas on ten different cancer types.
- Omics data types were gene expression, DNA methylation and miRNA expression.

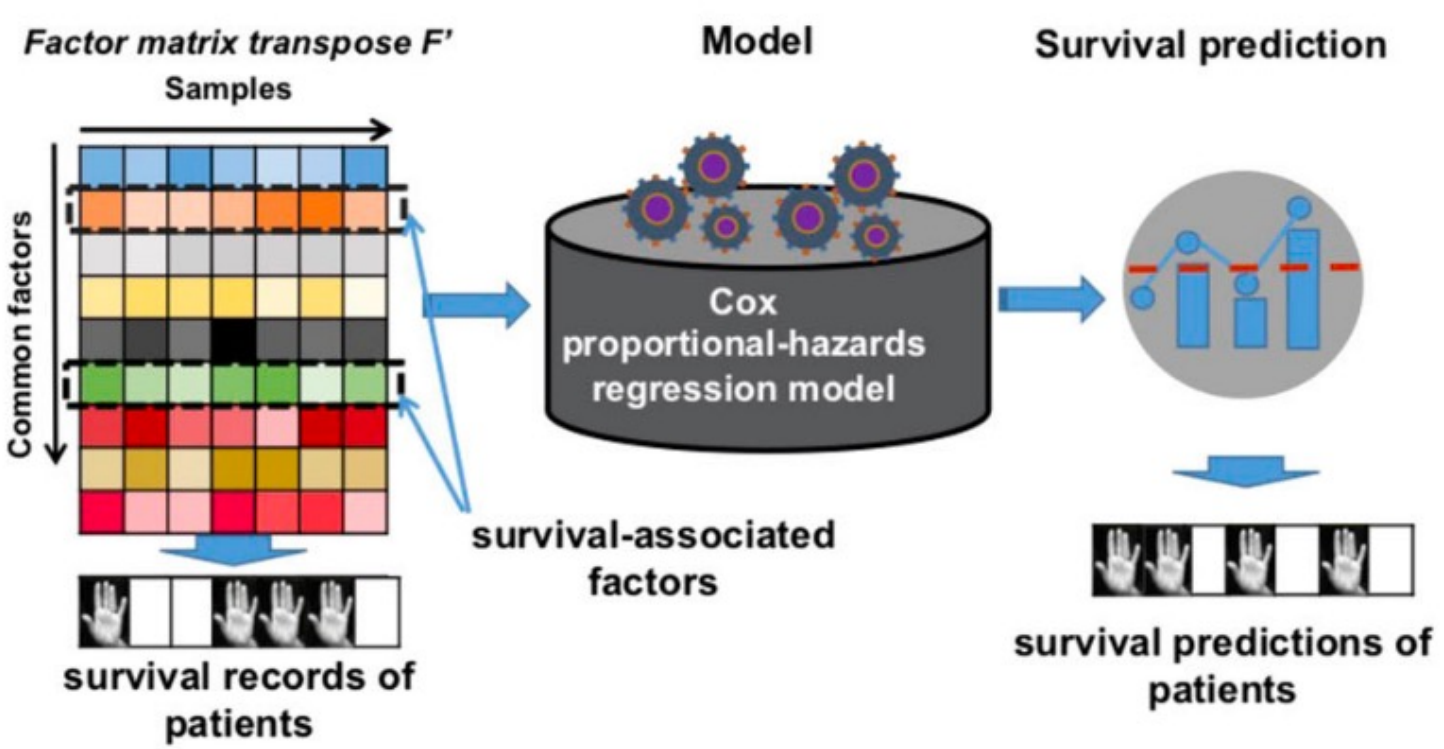
TCGA Clustering

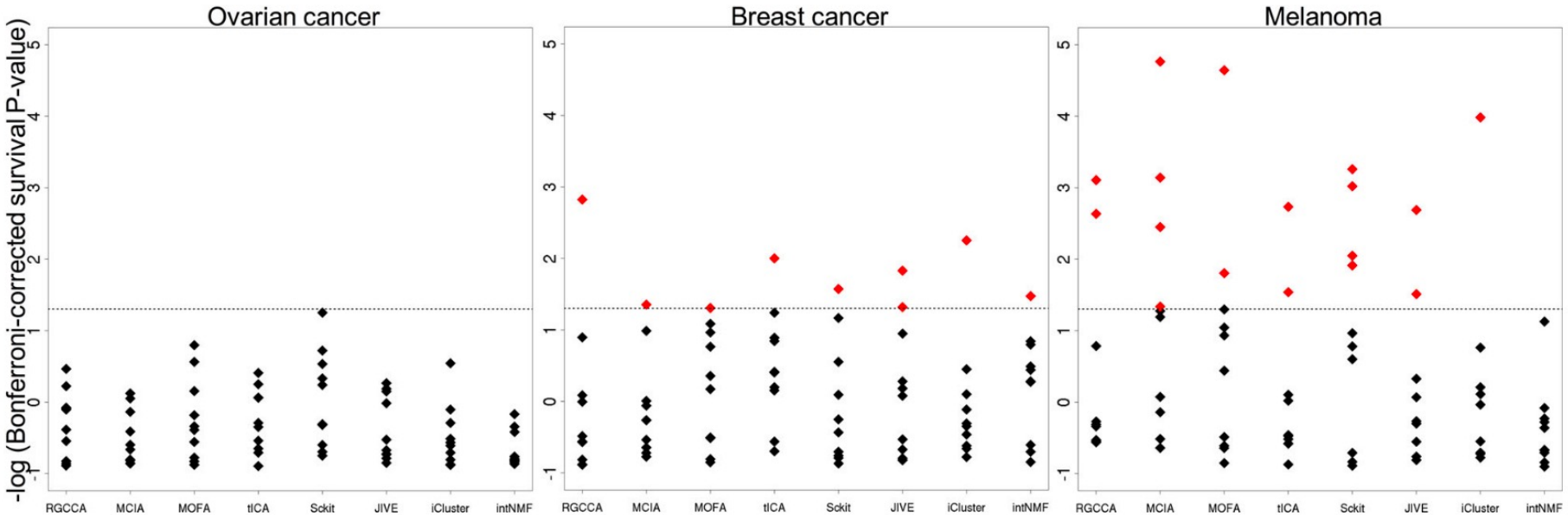
- Used clinical subtyping to assess clustering performance of jDR methods.
  - Note: these cannot be considered a 'ground truth' assessment.
- For BRCA, compared clustering to two subtypes: ER/PR/HER-2, and COCA classification.
  - COCA: integrative subtyping obtained by separately clustering different 'omics types and performing a consensus clustering on the shared results.
- For methods that do not automatically perform clustering, conducted k-means using the first four derived factors.
- JIVE has best performance based on clinical subtyping.


A
ER/PR/HER-2 Breast cancer subtypes
Overlap between the jDR-identified and marker-based breast cancer subtypes (Jaccard Index)
0.0
0.2
0.4
0.6
0.8
1.0
RGCCA
MCIA
MOFA
JIVE
tICA
kit-fusion
iCluster
intNMF
B
COCA Breast cancer subtypes
Overlap between the jDR-identified and COCA-based breast cancer subtypes (Jaccard Index)
0.0
0.2
0.4
0.6
0.8
1.0
RGCCA
MCIA
MOFA
JIVE
tICA
sckit-fusion
iCluster
intNMF

Predicting Survival

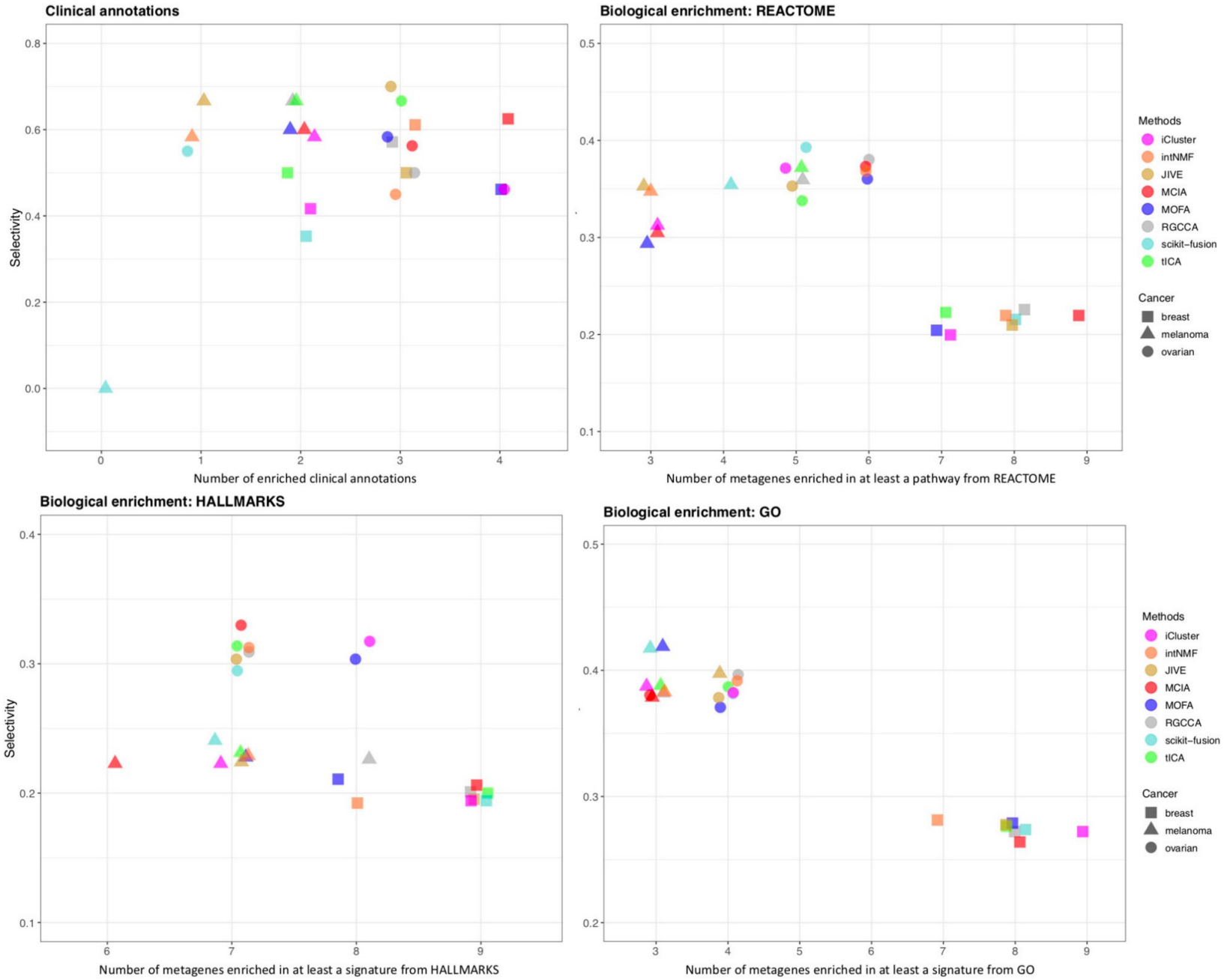
- Used first ten factors to predict survival via Cox regression.
- Examined the number of derived features associated with survival.
- Appears that the number of associated factors depends more on the cancer type than the dimension reduction approach.
- RGCCA, MCIA, and JIVE showed most promising results for those cancers where predicting survival was most difficult.


Factor matrix transpose F'
Samples
Common factors
survival records of patients
Model
Cox proportional-hazards regression model
survival-associated factors
Survival prediction
survival predictions of patients


Ovarian cancer
Breast cancer
Melanoma
-log (Bonferroni-corrected survival P-value)
RGCCA
MCIA
MOFA
tICA
Sckit
JIVE
iCluster
intNMF

Association with Clinical Factors

- Assess the association of dimension-reduced factors with clinical annotations.
  - May be able to determine if one factor is a proxy for some clinical feature.
- Performed analysis in TCGA data, using three omics datatypes.
- Investigated four variables: ‘age’, ‘days to new tumor’, ‘gender’, and ‘neo-adjuvant therapy administration’.
- Investigated selectivity of methods, i.e., ability to generate a one-to-one map between reduced factors and clinical annotations.
- RGCCA, MCIA, and MOFA are the best-performing algorithms overall.


Clinical annotations
Selectivity
Number of enriched clinical annotations
Biological enrichment: REACTOME
Number of metagenes enriched in at least a pathway from REACTOME
Biological enrichment: HALLMARKS
Selectivity
Number of metagenes enriched in at least a signature from HALLMARKS
Biological enrichment: GO
Number of metagenes enriched in at least a signature from GO
Methods
iCluster
intNMF
JIVE
MCIA
MOFA
RGCCA
scikit−fusion
tICA
Cancer
breast
melanoma
ovarian

References

- Elements of Statistical Learning (2nd Ed.), Hastie, Tibshirani, and Friedman 2009.
- Cantini, Laura, et al. "Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer." *Nature communications* 12.1 (2021): 124.
- Wold, Svante, Nouna Kettaneh, and Kjell Tjessem. "Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection." *Journal of chemometrics* 10.5-6 (1996): 463-482.
- Shen, Ronglai, et al. "Integrative subtype discovery in glioblastoma using iCluster." *PloS one* 7.4 (2012): e35236.
- Lock, Eric F., et al. "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types." *The annals of applied statistics* 7.1 (2013): 523.