Linked Matrix and Tensor Decompositions

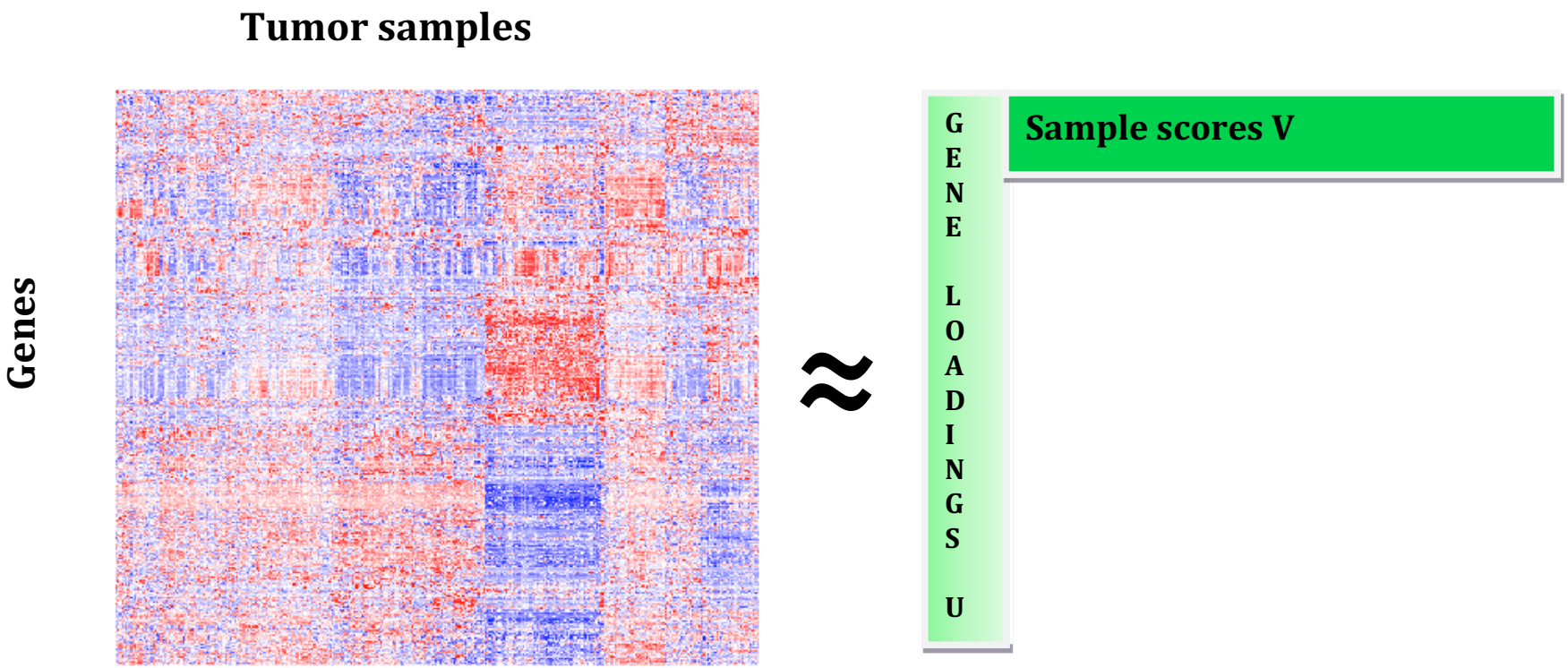
Eric F. Lock

University of Minnesota
Division of Biostatistics and Health Data Science

TRICAP, Alesund, Norway, 06/24/2025

Low-rank matrix approximation

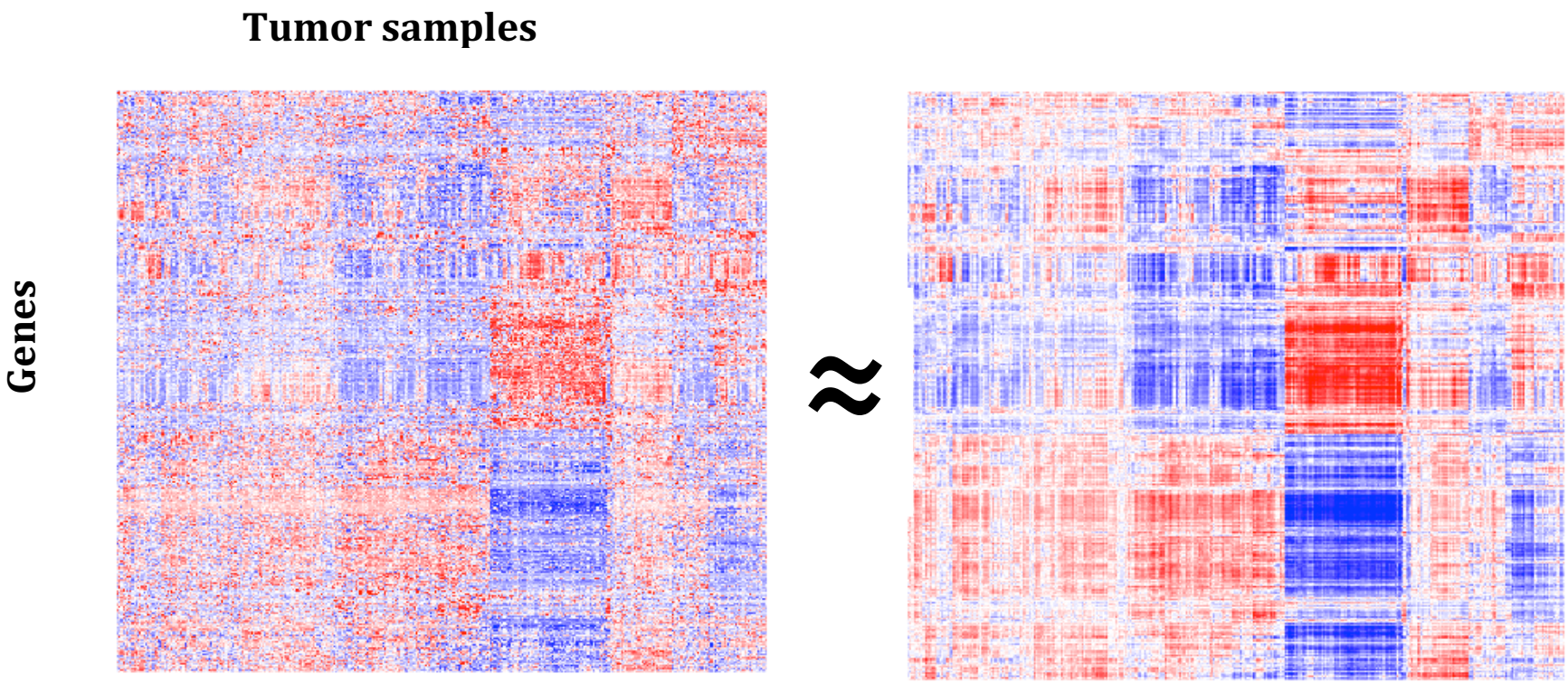
- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



- Low rank factorization: $X \approx UV$, $U : m \times r$, $V : r \times n$.

Low-rank matrix approximation

- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



- Low rank factorization: $X \approx UV$, $U : m \times r$, $V : r \times n$.

Low-rank matrix approximation

- $X = \mathbf{A} + E$ where rank($\mathbf{A}$)=r and noise E
- Singular value decomposition (SVD): $X = UDV^T$
  - D is diagonal with singular values $d_i = D[i, i]$
  - captures **signal** and noise:

$$U = [\mathbf{u_1}, \cdots, \mathbf{u_r}, u_{r+1}, \cdots]$$
$$D = \text{diag}(\mathbf{a_1} + e_1, \cdots, \mathbf{a_r} + e_r, e_{r+1}, \cdots)$$
$$V = [\mathbf{v_1}, \cdots, \mathbf{v_r}, v_{r+1}, \cdots]$$

Low-rank matrix approximation

- Approach 1: hard-thresholding
  - Minimize $\frac{1}{2}||X - \hat{X}||_F^2$ for $\text{rank}(\hat{X}) = r$
  - Then $\hat{X} = U\hat{D}V^T$ where $\hat{d}_i = \begin{cases} d_i \text{ for } i \leq r \\ 0 \text{ for } i > r \end{cases}$
  - Need to select r

- Over-fits! $\hat{D} = \text{diag}(\mathbf{a_1} + e_1, \cdots, \mathbf{a_r} + e_r, 0, \cdots)$

Low-rank matrix approximation

- Approach 2: soft-thresholding
  - Minimize $\frac{1}{2}||X - \hat{X}||_F^2 + \lambda||\hat{X}||_*$
  - $||\cdot||_*$ is the nuclear norm: $||\hat{X}||_* = \sum \hat{d}_i$
  - Then $\hat{X} = U\hat{D}V^T$ where $\hat{d}_i = max(d_i - \lambda, 0)$.
- Need to select λ
  - Assume $E : m \times n$ has iid $N(0, \sigma^2)$ entries
  - Then, largest singular value of $E \approx \sigma(\sqrt{m} + \sqrt{n})$
  - Set $\lambda = \hat{\sigma}(\sqrt{m} + \sqrt{n})$
- Over-shrinks! $\hat{D} = \text{diag}(\mathbf{a_1} + e_1 - \lambda, \cdots, \mathbf{a_r} + e_r - \lambda, 0, \cdots)$

Low-rank matrix approximation

- Minimizing $\frac{1}{2}||X - \hat{X}||_F^2 + \lambda||\hat{X}||_*$ is equivalent to

$$||X - \tilde{U}\tilde{V}||_F^2 + \lambda(||\tilde{U}||_F^2 + ||\tilde{V}||_F^2)$$

for $\tilde{U} : m \times r'$ $\tilde{V} : r' \times n$ and r' sufficiently large

- Posterior mode with $N(0, \sigma^2/\lambda)$ priors on $\tilde{U}$ and $\tilde{V}$

- More flexible model:

$$\tilde{U}[:, r] \sim N(\mathbf{0}, \sigma\tau_u^2[r]\mathbf{I}), \;\; \tilde{V}[:, r] \sim N(\mathbf{0}, \sigma\tau_v^2[r]\mathbf{I})$$

Low-rank matrix approximation

- Approach 3: Empirical variational Bayes
  - Approximate posterior with $q(\tilde{U}, \tilde{V}) = q_u(\tilde{U})q_v(\tilde{V})$
  - Minimize free energy

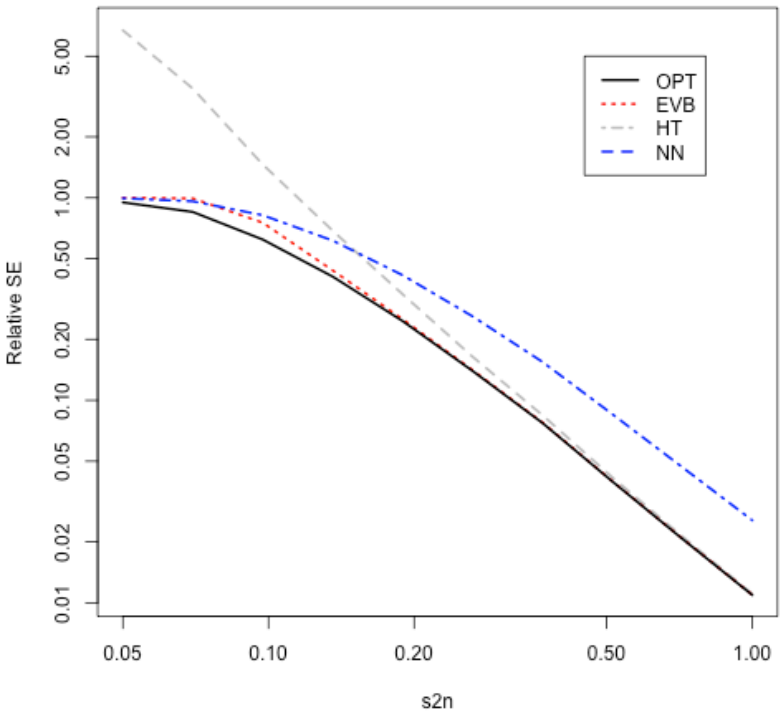
$$E_q \log \frac{q_u(\tilde{U})q_v(\tilde{V})}{p(X \mid \tilde{U}, \tilde{V}, \sigma)p(\tilde{U} \mid \tau_U)p(\tilde{V} \mid \tau_V)}$$

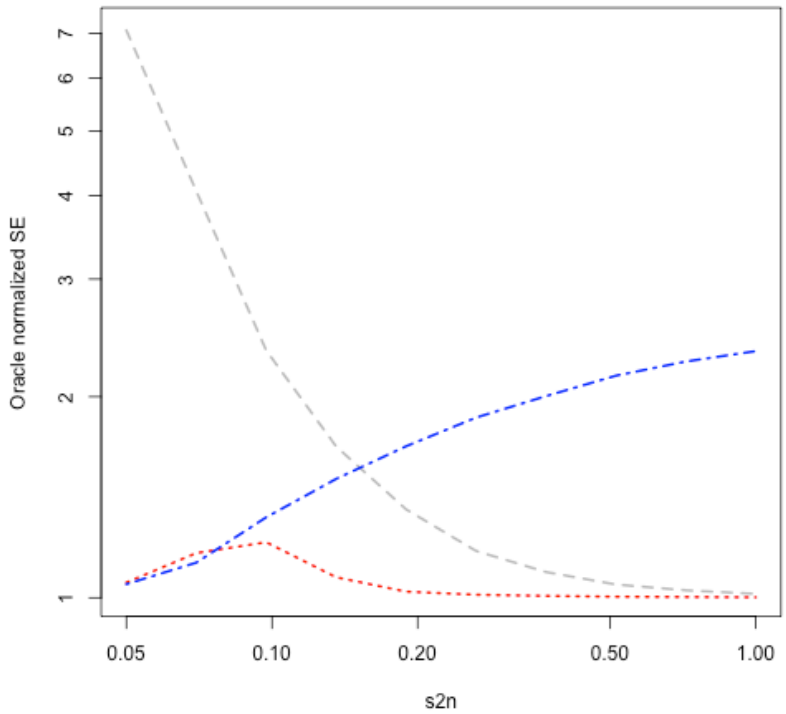
    over $\sigma, \tau_U, \tau_v, q_u, q_v$.
  - $\hat{X} = E_q(\tilde{U}\tilde{V}^T) = U\hat{D}V^T$ where $\hat{d}_i = f(d_i)$ for closed-form $f(\cdot)$

- No tuning parameters
- Just right! $\quad \hat{D} \approx \text{diag}(\mathbf{a_1}, \cdots, \mathbf{a_r}, 0, \cdots)$

Low-rank matrix approximation

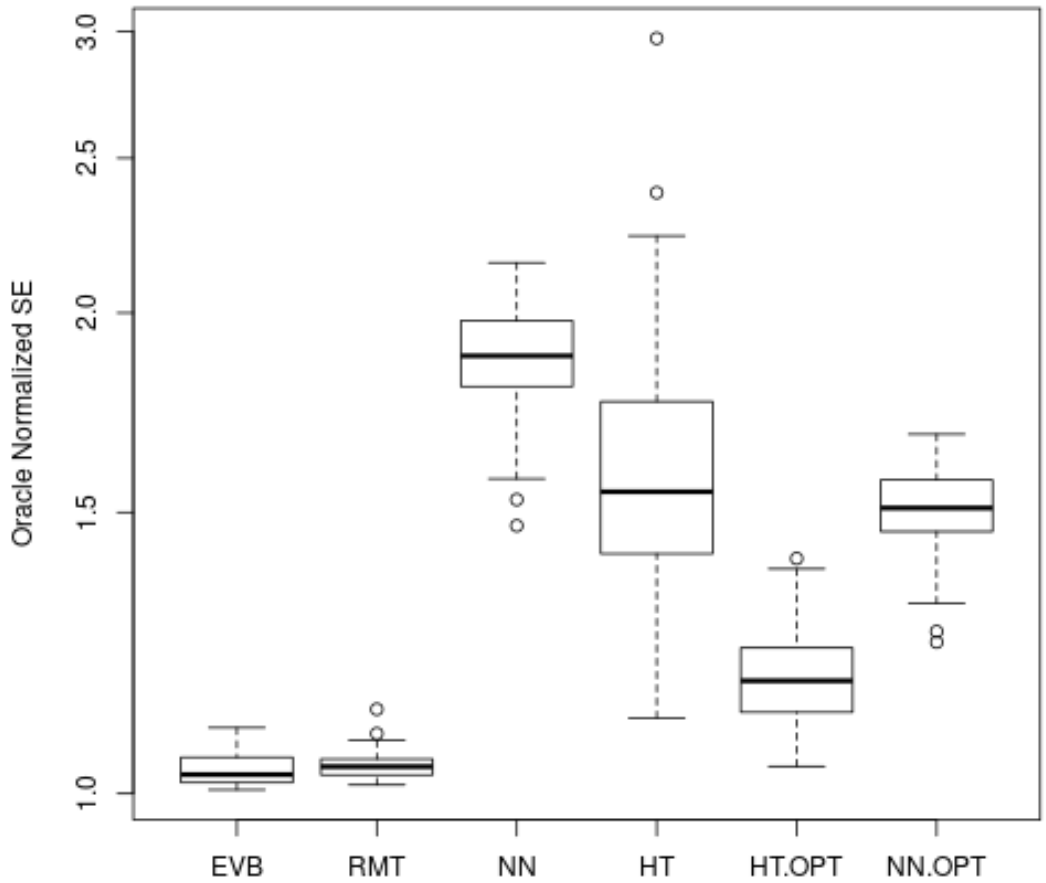
- Simulation: $m = 1000$, $n = 100$, $R = 10$, varying s2n.
- Relative squared error (SE) and oracle normalized SE





Low-rank matrix approximation

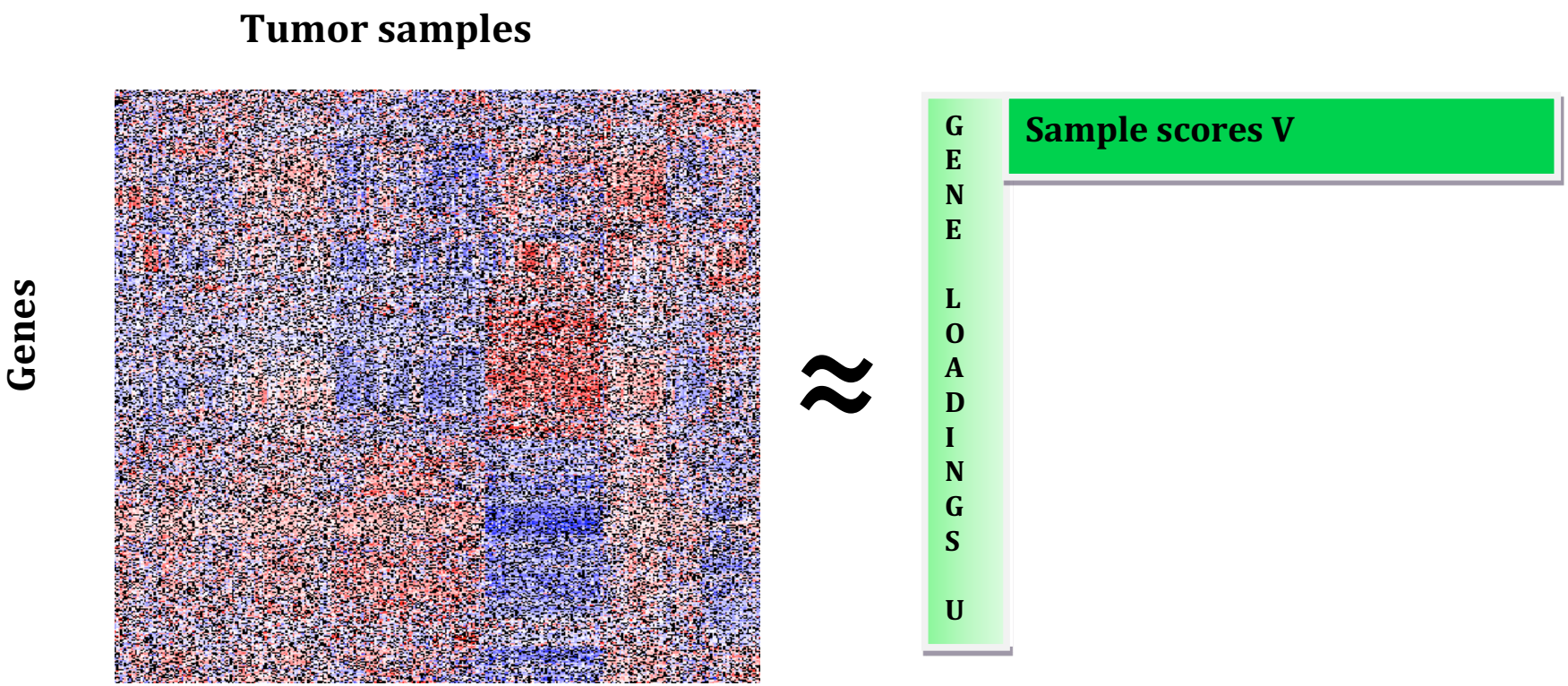
- Oracle normalized SE



RMT: Shabalin and Nobel, 2010.

Matrix factorization: missing data

- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples

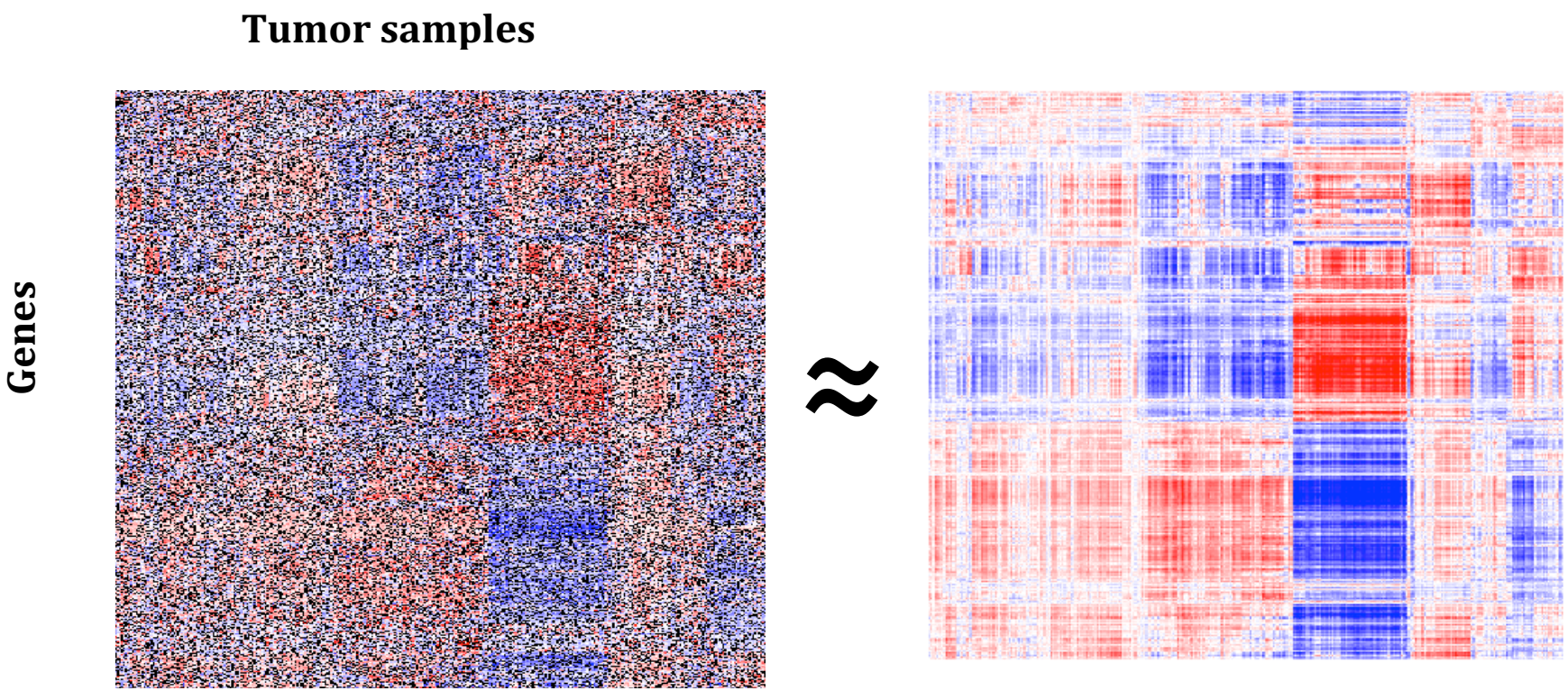


Matrix factorization: missing data

- Missing data $X_{\text{miss}} = \{X[i,j] : (i,j) \in \mathcal{M}\}$
- Minimize free energy over $\sigma, \tau_U, \tau_v, q_u, q_v$, and X_{miss}.
- EM-type approach:
  - Initialize X_{miss}
  - Update τ_U, τ_v, q_u, q_v given X_{miss}
  - Update $X_{\text{miss}} = E_q X_{\text{miss}}$
  - Repeat until convergence

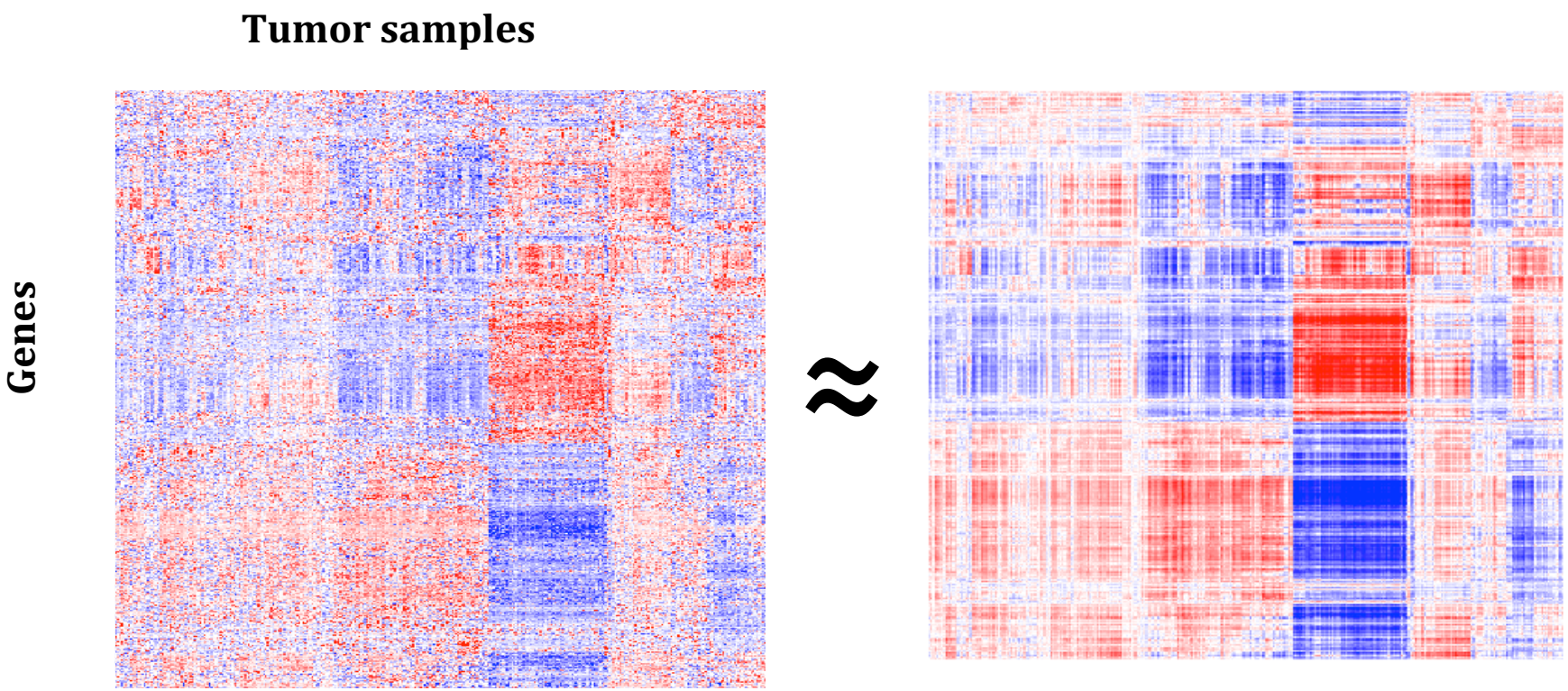
Matrix factorization: missing data

- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples

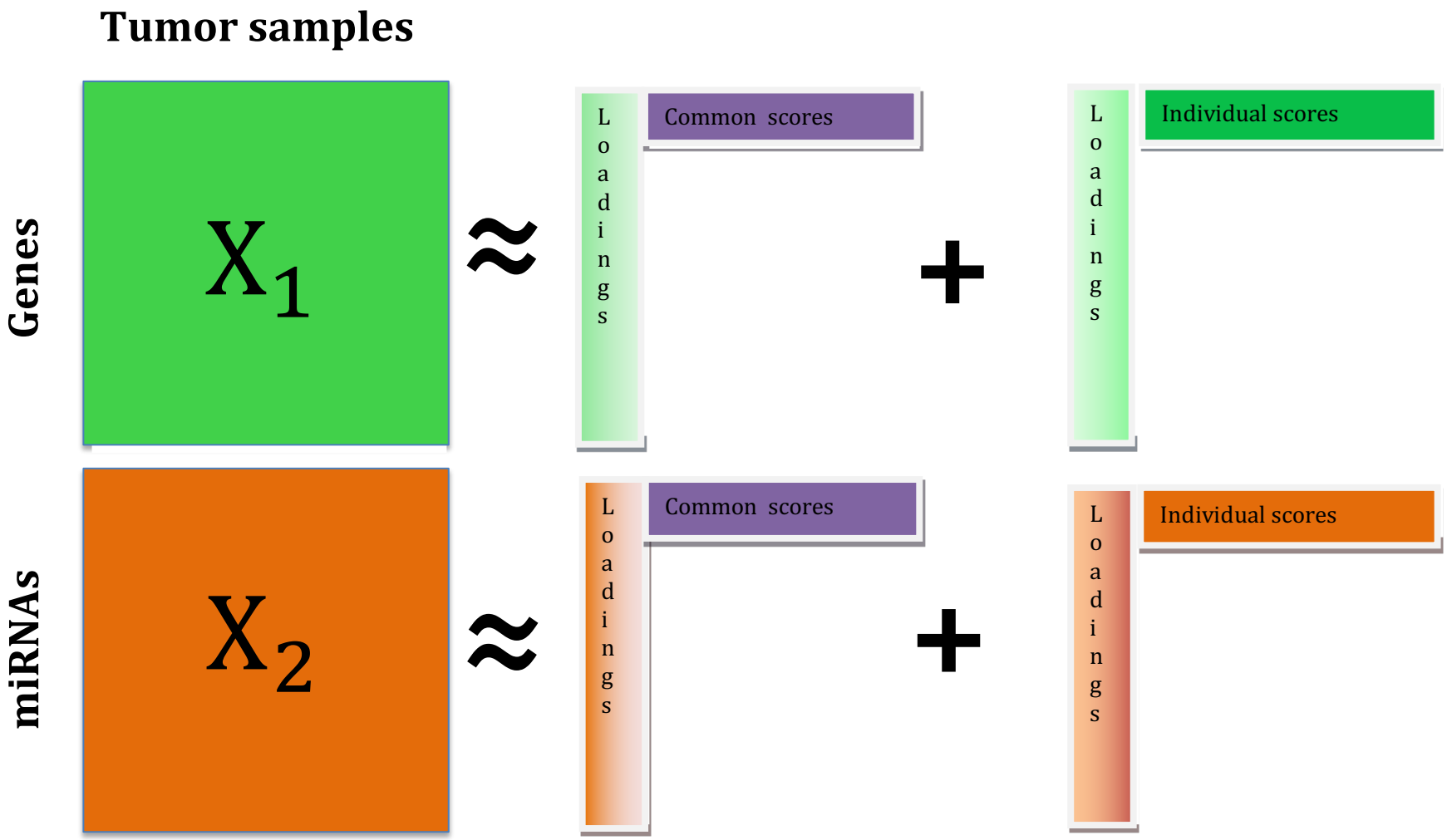


Matrix factorization: missing data

- Gene expression matrix $X : m \times n$
  - m genes for n breast cancer tumor samples



Vertically linked data: joint and individual factorization



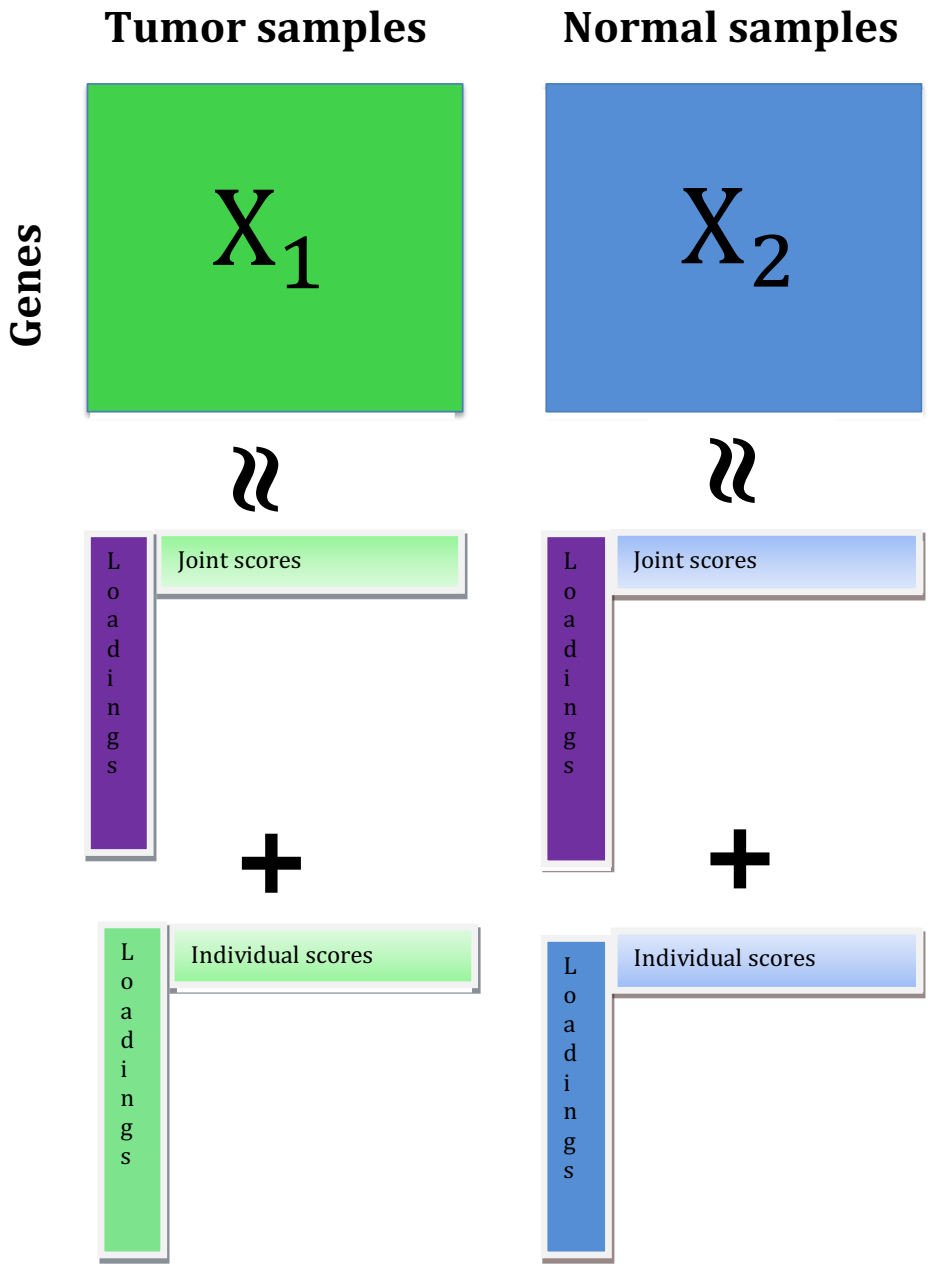
Joint + individual factorization methods

- JIVE [Lock et al., 2013]
  - "Joint and Individual Variation Explained"
- DISCO-SCA [Van Deun et al.,, 2013]
- AJIVE [Feng, Jiang, Hannig and Marron, 2018]
- SLIDE [Gaynanova and Li, 2018]
- GIPCA [Zhu, Li, Lock, 2020]

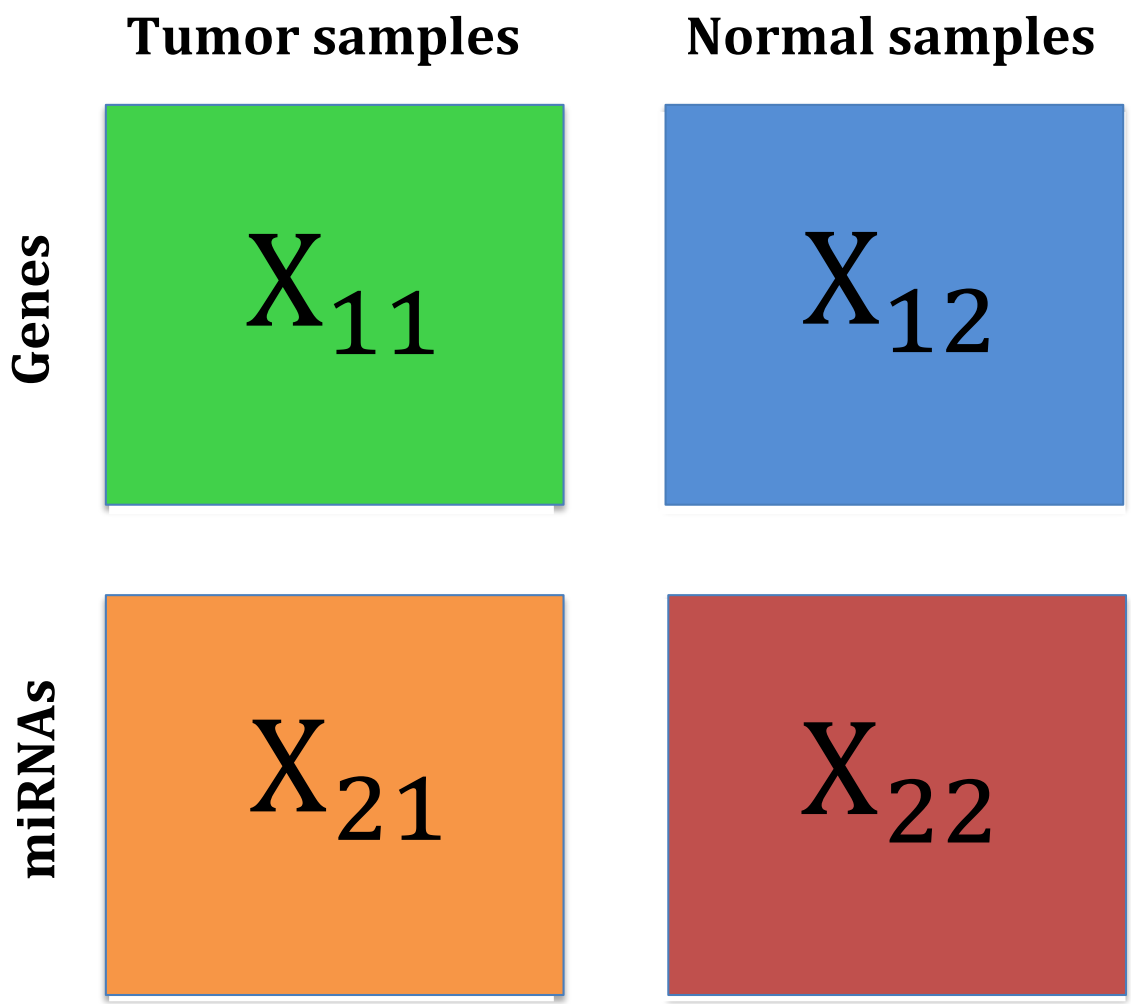
⋮

- The bi-factor method [Holzinger and Swineford, 1937]

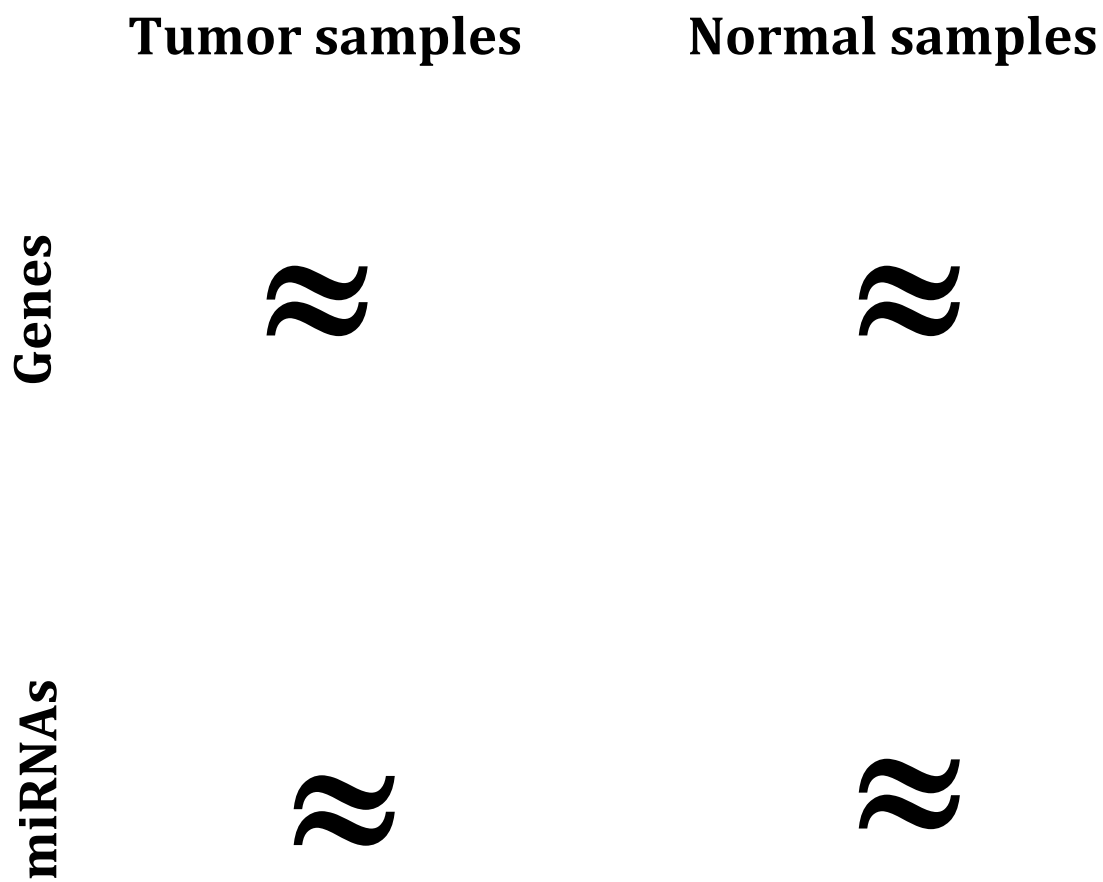
Horizontally linked data: Joint and individual



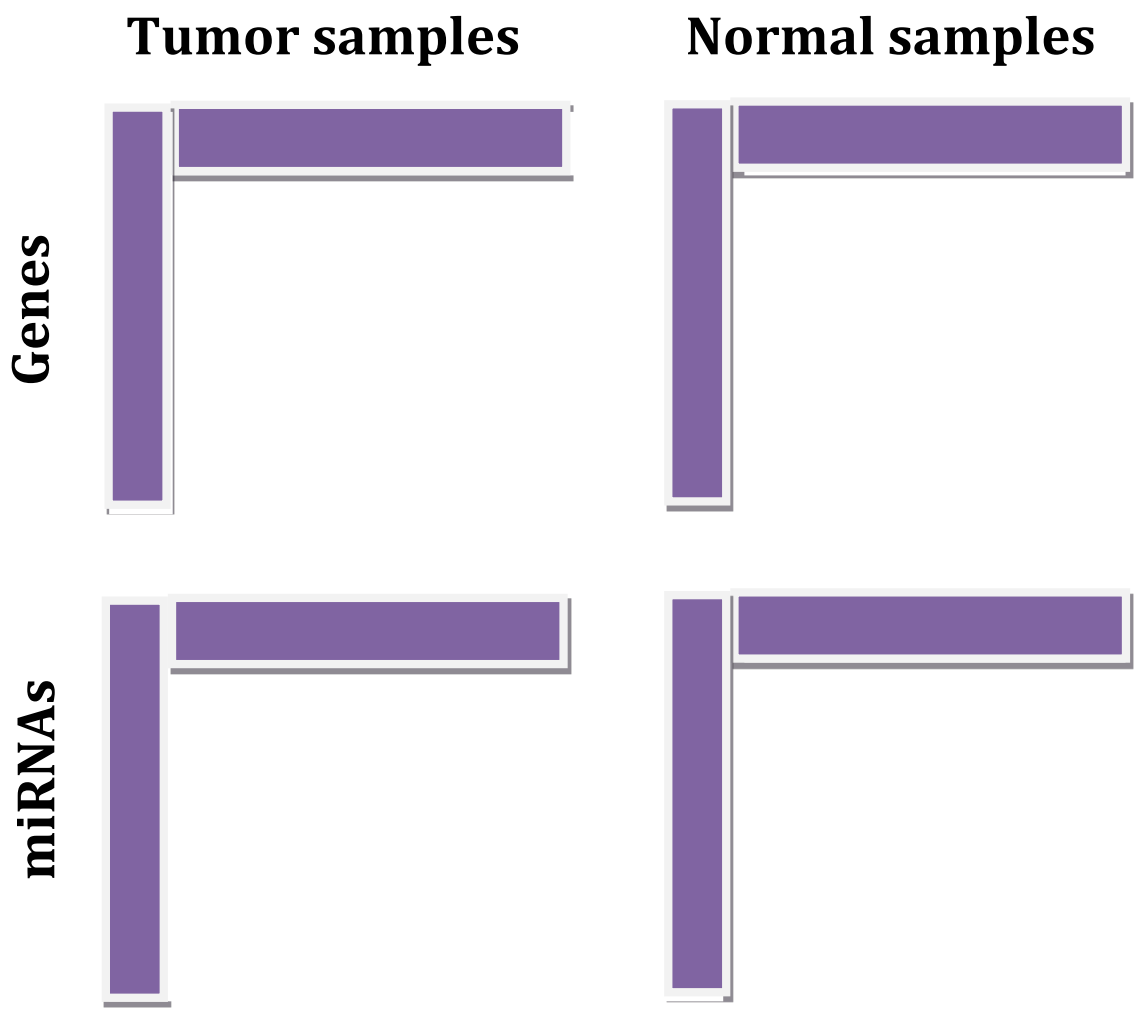
Bidimensionally linked data: BIDIFAC



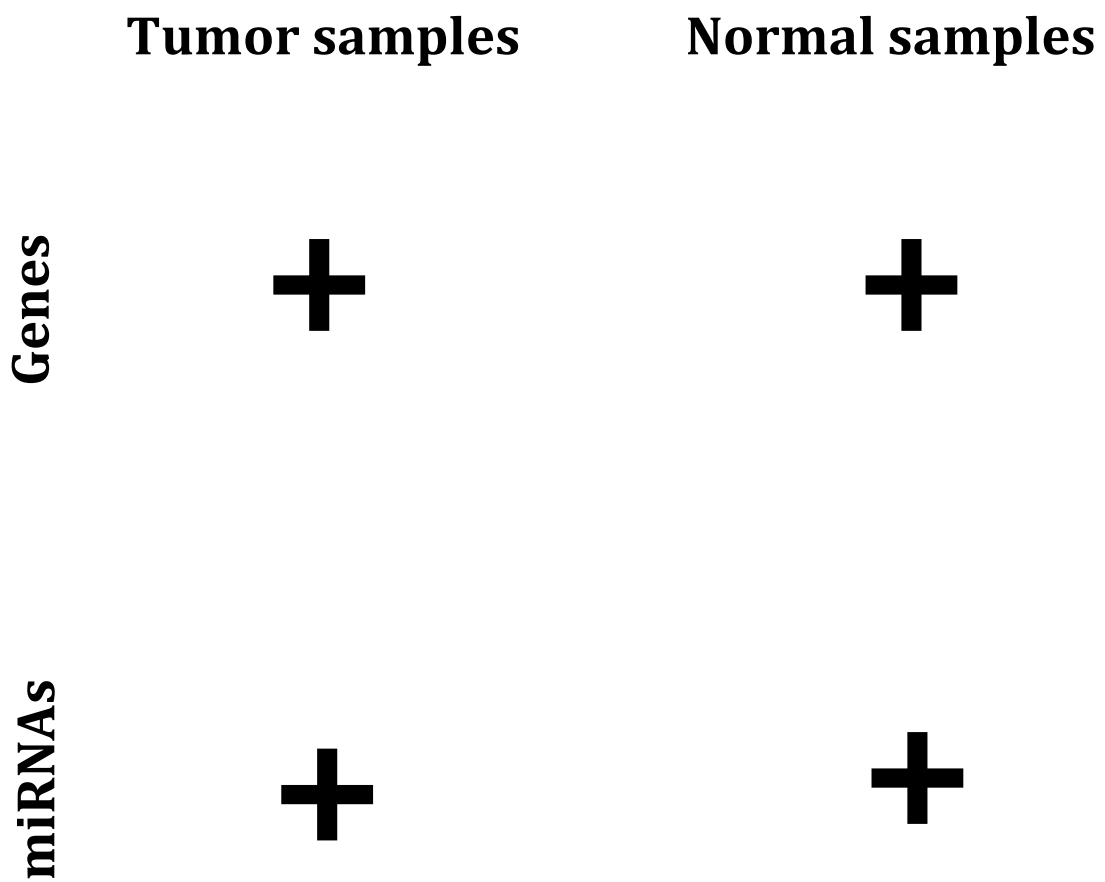
Bidimensionally linked data: BIDIFAC



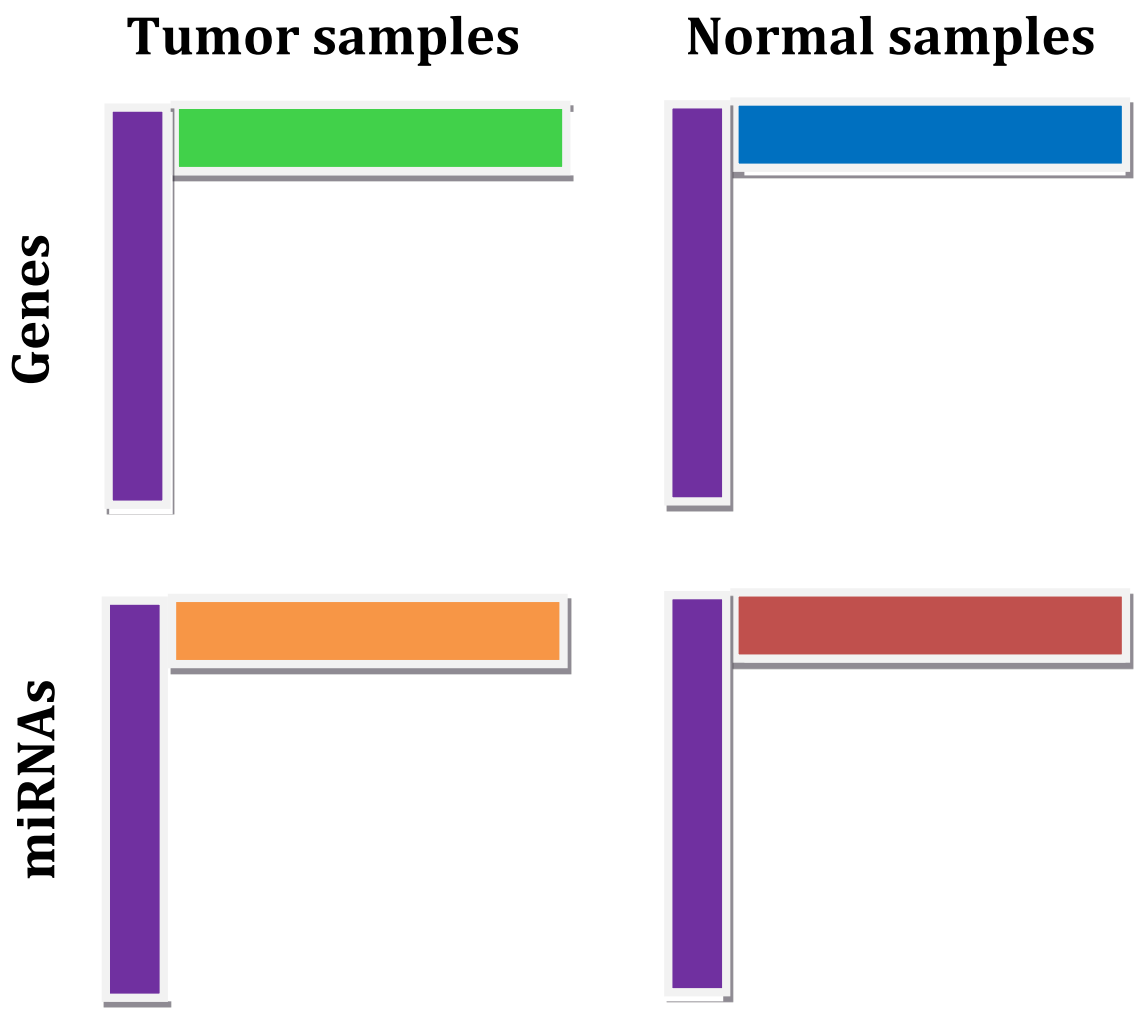
Bidimensionally linked data: BIDIFAC



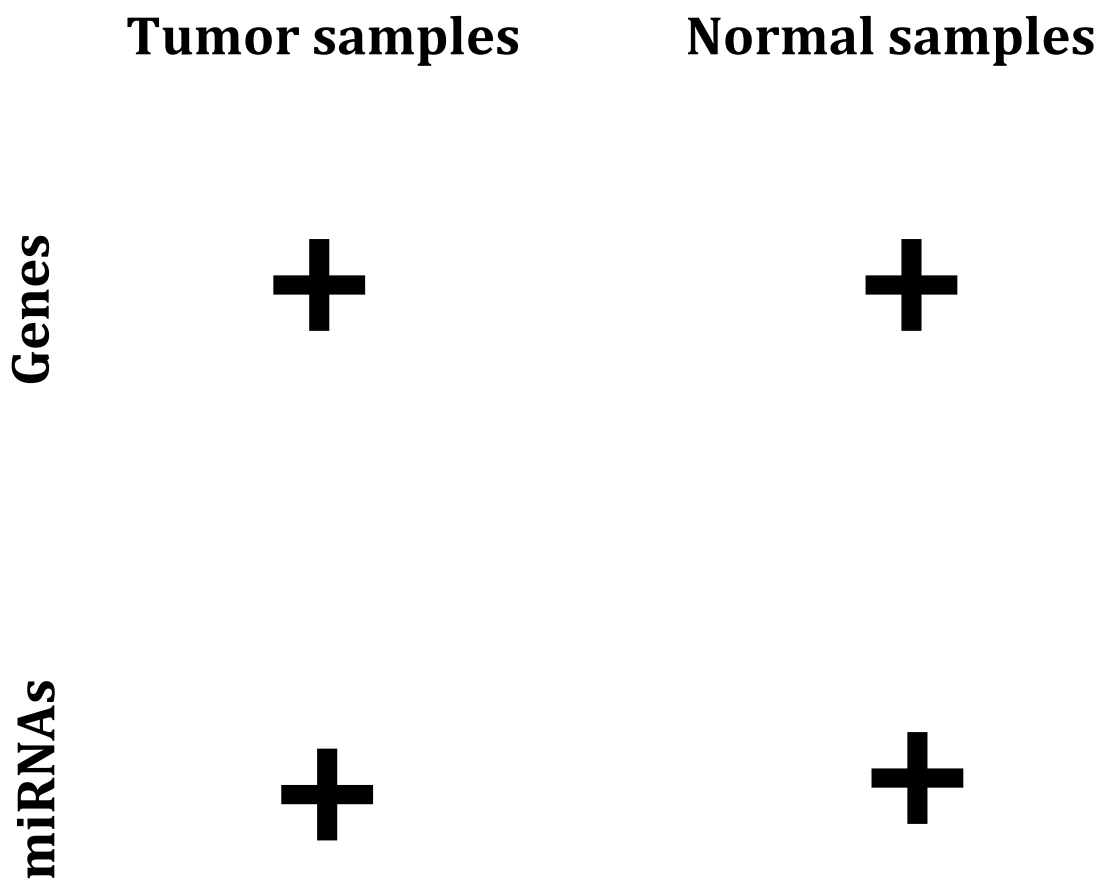
Bidimensionally linked data: BIDIFAC



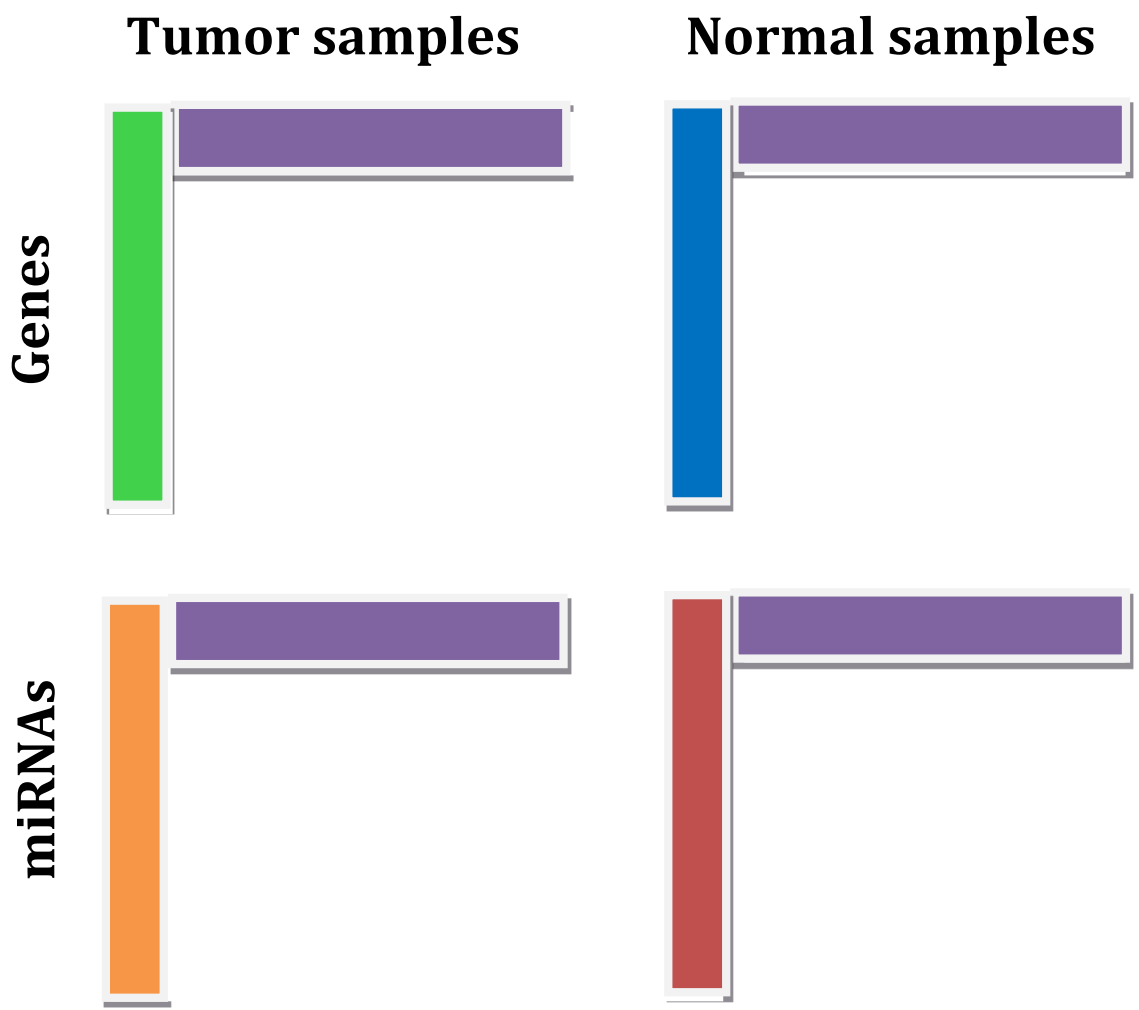
Bidimensionally linked data: BIDIFAC



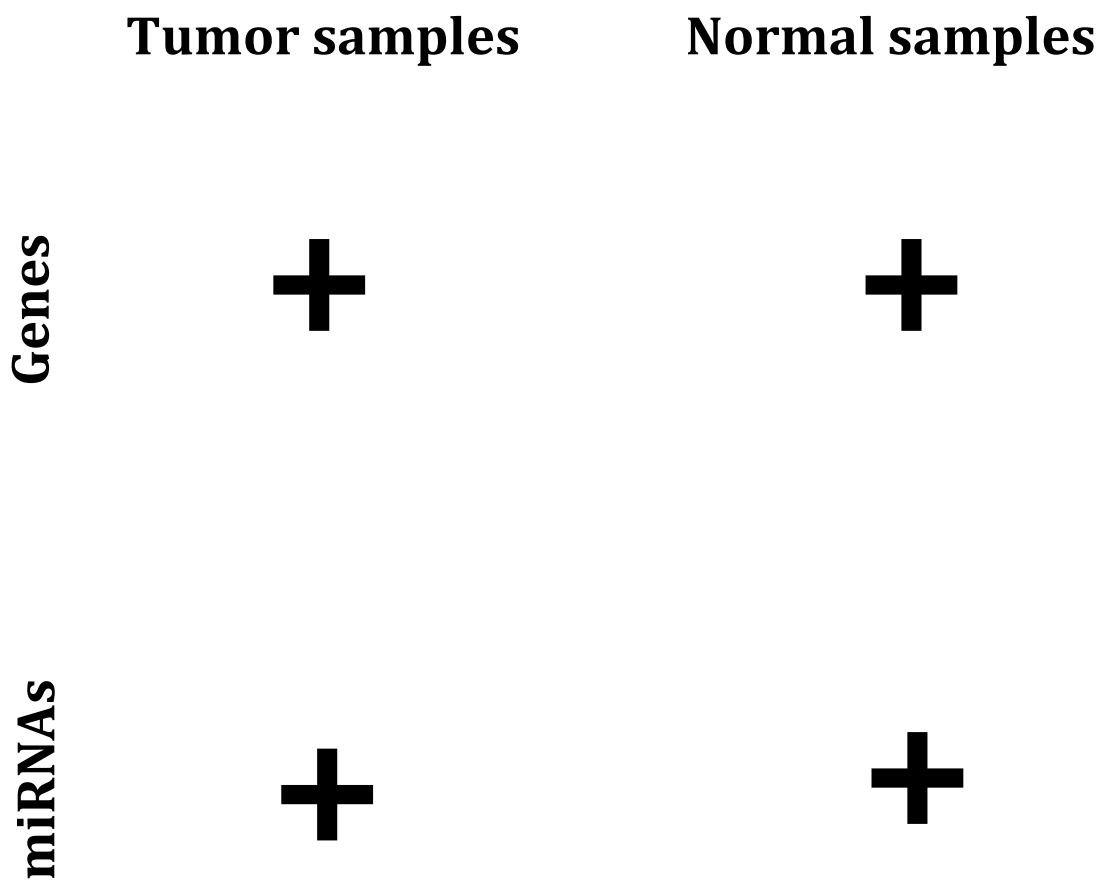
Bidimensionally linked data: BIDIFAC



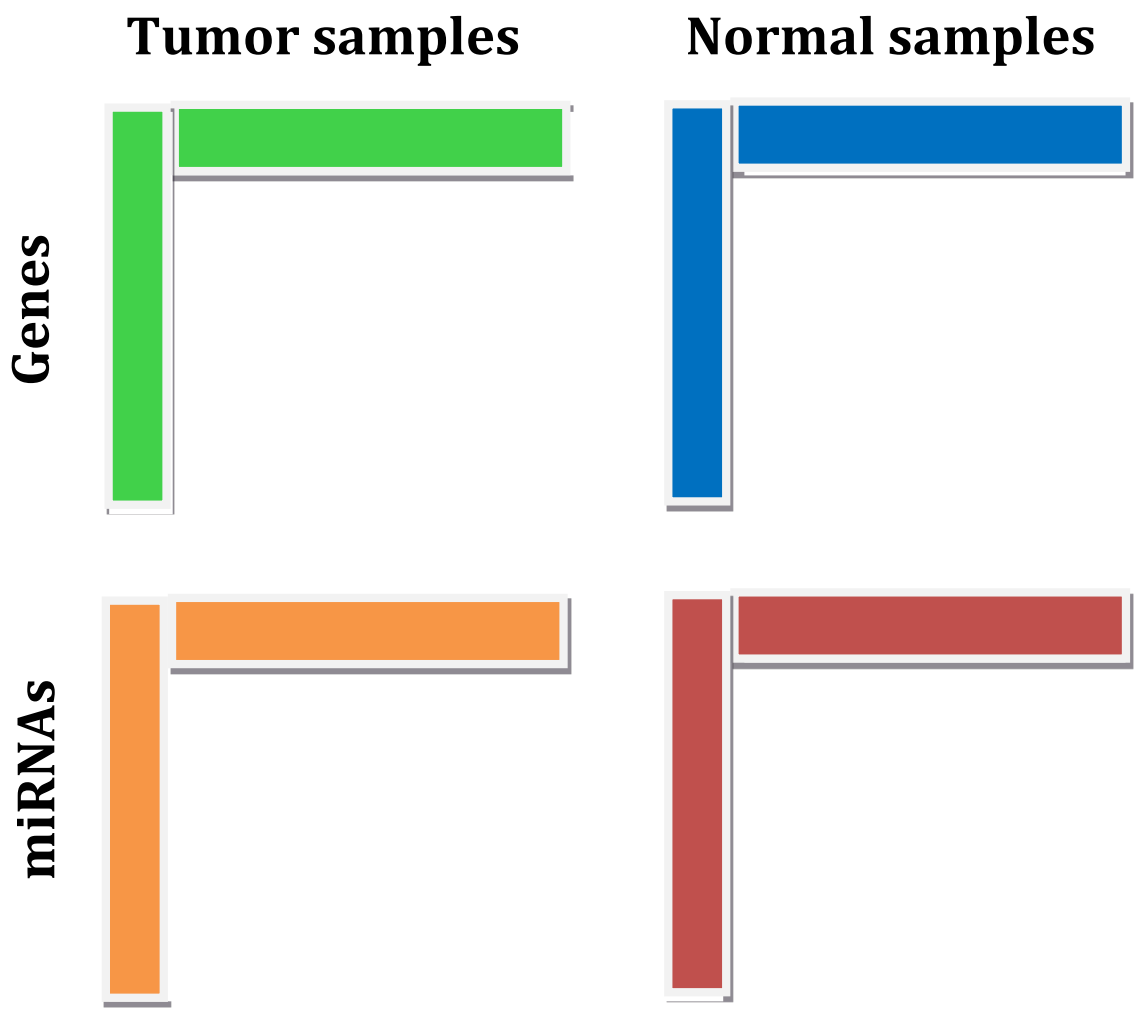
Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC



Bidimensionally linked data: BIDIFAC



BIDIFAC: general framework

- Linked matrices $\{X_{ij} : m_i \times n_j \mid i = 1, \ldots, I,\ j = 1, \ldots, J\}$:

$$X_{\cdot\cdot} = \begin{bmatrix} X_{11} & \ldots & X_{1J} \\ \vdots & \ddots & \vdots \\ X_{I1} & \ldots & X_{IJ} \end{bmatrix}$$

- Decompose $X_{\cdot\cdot}$ into structural *modules*:

$$X_{\cdot\cdot} = \sum_{k=1}^{K} S_{\cdot\cdot}^{(k)} + E_{\cdot\cdot},$$

where presence of submatrices $S_{ij}^{(k)}$ are determined by binary row indicators $R : I \times K$ and column indicators $C : J \times K$:

$$S_{ij}^{(k)} = \begin{cases} 0_{M_i \times N_j} & \text{if } R[i,k] = 0 \text{ or } C[j,k] = 0 \\ U_i^{(k)} V_j^{(k)} & \text{if } R[i,k] = 1 \text{ and } C[j,k] = 1 \end{cases}.$$

EV-BIDIFAC

- Model for $i = 1, \ldots I$ and $j = 1, \ldots, J$:

$$\mathbf{X}_{ij} = \sum_{k=1}^{K} \mathbf{U}_i^{(k)} \mathbf{V}_j^{(k)T} + \mathbf{E}_{ij} \text{ where } \mathbf{E}_{ij}[l, m] \overset{iid}{\sim} \text{Normal}(0, \sigma_{ij}^2),$$

$$\mathbf{U}_i^{(k)}[\cdot, r] = \mathbf{0} \text{ if } \mathbf{R}[i, k] = 0, \text{Normal}(\mathbf{0}, \sigma_{ij}\tau_u^2[k, r]\mathbf{I}) \text{ if } \mathbf{R}[i, k] = 1$$

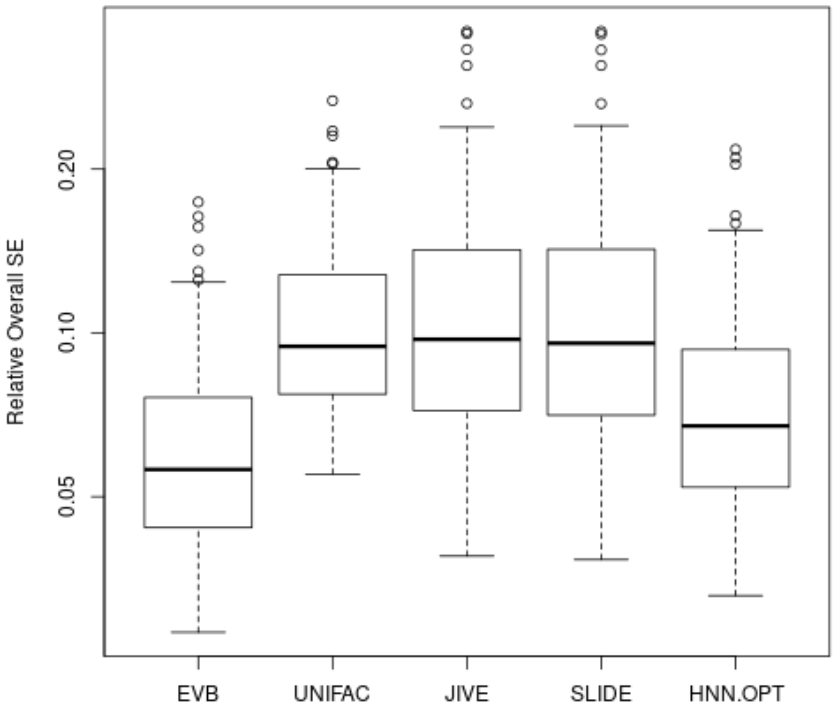
$$\mathbf{V}_j^{(k)}[\cdot, r] = \mathbf{0} \text{ if } \mathbf{C}[j, k] = 0, \text{Normal}(\mathbf{0}, \sigma_{ij}\tau_v^2[k, r]\mathbf{I}) \text{ if } \mathbf{C}[j, k] = 1$$

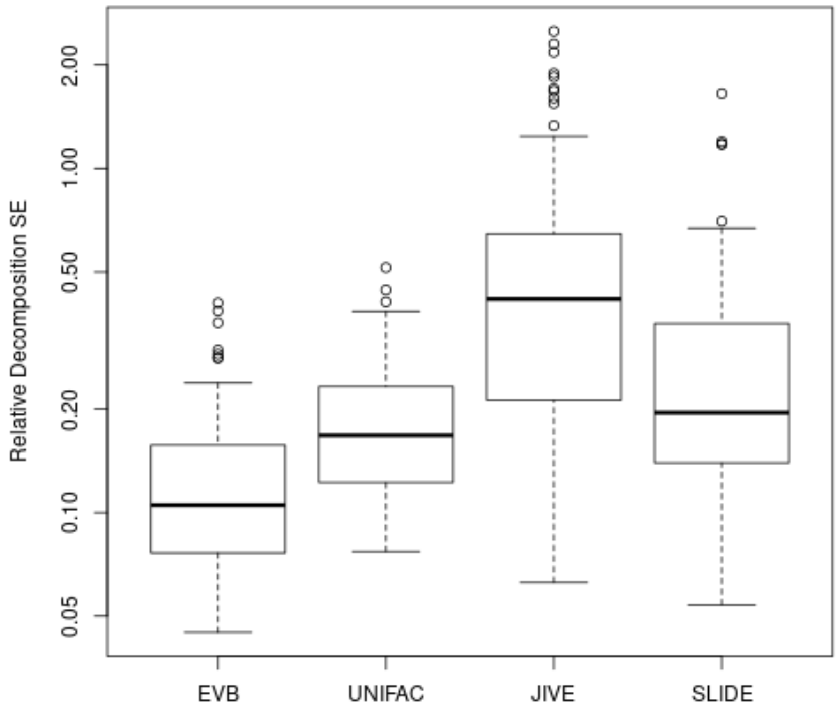
- Estimate under empirical variational Bayes framework

- Gives a unique decomposition under general conditions!

EV-BIDIFAC: Results

- Simulation: two linked matrices, shared & unshared structures.

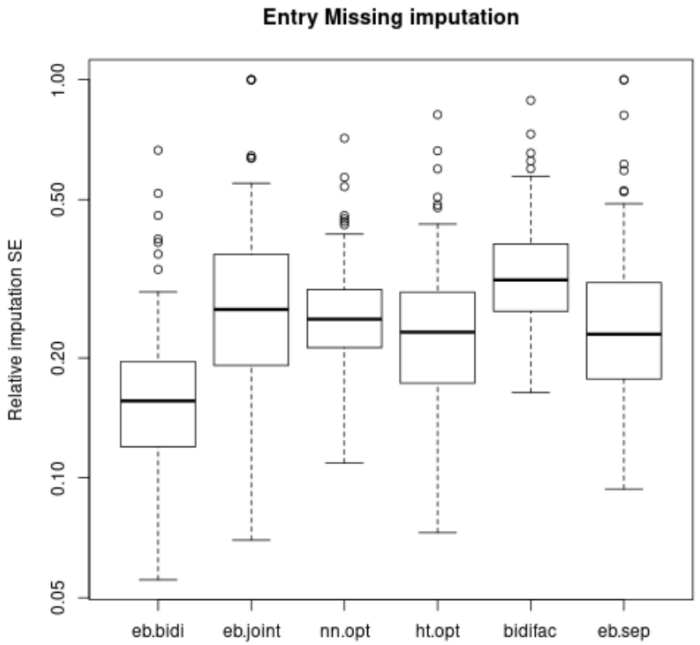


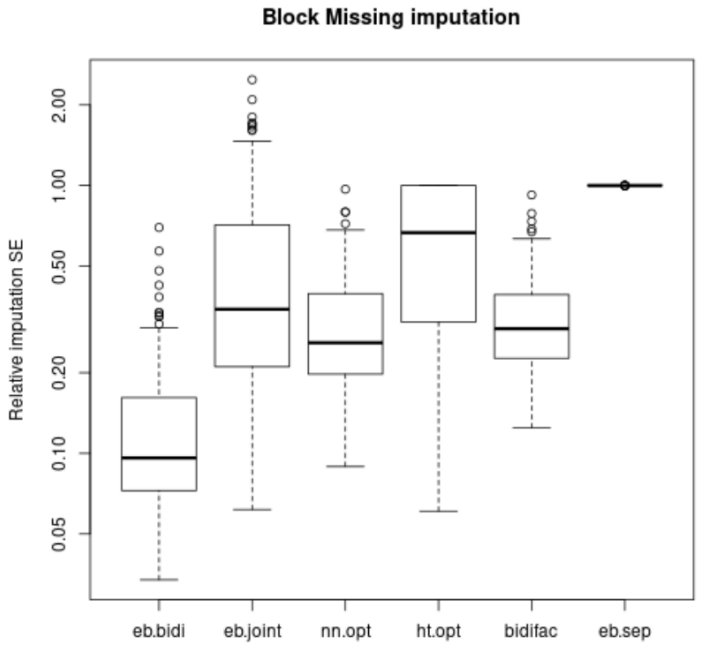


UNIFAC: Park and Lock, 2020
HNN: Yi, Wang, and Gaynanova, 2022

EV-BIDIFAC: Results

- Simulation: 2 x 2 linked matrices
- 5 modules with rank 2 structure, 4 with no structure
- 10% of entries and rows/columns missing

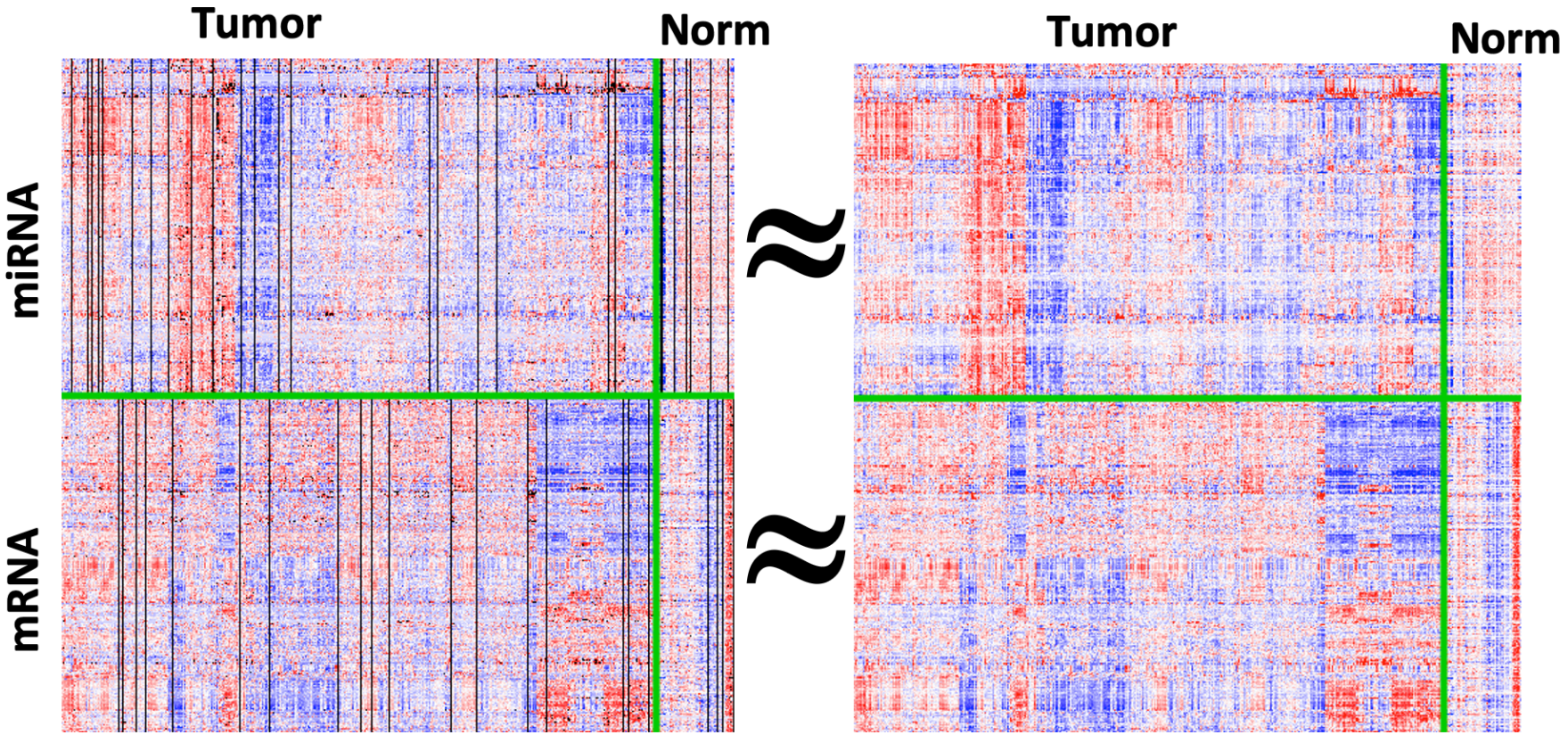




- $\hat{S}_{\cdot\cdot}^{(k)} = \mathbf{0}$ for modules with no true structure 400/400 times.

EV-BIDIFAC: Results

- Minimize free energy via EM for missing data



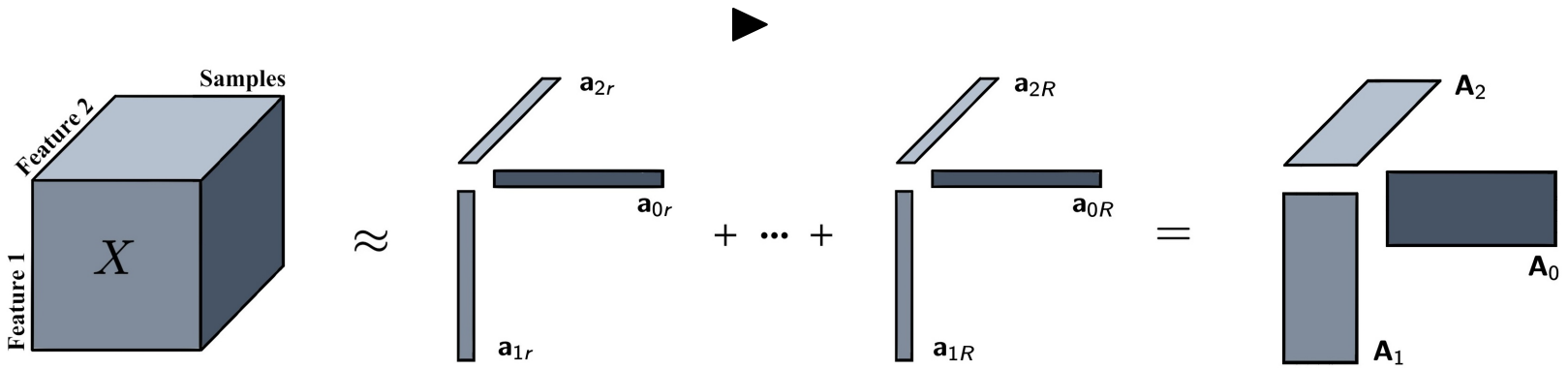
EV-BIDIFAC: BRCA data results

- Imputation accuracy for BRCA (held-out data):

Method	Entry-missing	Col-missing	Row-missing	Overall
EV-BIDIFAC	**0.389**	**0.756**	**0.903**	**0.682**
BIDIFAC	0.526	0.870	**0.909**	0.768
EB-SEP	**0.381**	1.00	1.00	0.79
EB-JOINT	0.510	1.01	1.67	1.06
NN-SEP	0.536	1.00	1.00	0.845
NN-JOINT	0.614	0.891	**0.881**	0.796
HT-SEP	0.459	1.00	1.03	0.831
HT-JOINT	0.520	4.50	13.1	6.05
KNN	0.646	1.26	1.01	0.974
StructureMC	-	0.977	1.01	-

Low-rank tensor factorization: Candecomp/Parafac (CP)

- Three-way tensor with dimensions $\mathbb{X} : I_1 \times I_2 \times I_3$.



- $\mathbb{X} \approx \sum_{r=1}^{R} \mathbf{a}_{1r} \circ \mathbf{a}_{2r} \circ \mathbf{a}_{3r} = [\![\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3]\!]$.
  - Factor matrices $\mathbf{A}_n = [\mathbf{a}_{n1}, \mathbf{a}_{n2}, \ldots, \mathbf{a}_{nR}] : I_n \times R,\ n = 1, 2, 3.$
- Alternatively: $\mathbb{X} \approx \sum_{r=1}^{R} \lambda_r \cdot \tilde{\mathbf{a}}_{1r} \circ \tilde{\mathbf{a}}_{2r} \circ \tilde{\mathbf{a}}_{2r} = [\![\boldsymbol{\lambda}; \tilde{\mathbf{A}}_1, \tilde{\mathbf{A}}_2, \tilde{\mathbf{A}}_3]\!]$
  - Where $||\tilde{\mathbf{a}}_{nr}|| = 1$ for all n, r.

CP factorization: L_2 penalization

Theorem

For the CP approximation of an N-way tensor, a L_2 penalty on factor matrices $\mathbf{A}_i$, $i = 1, \ldots, n$ is equivalent to an $L_{2/N}$ penalty on weights $\boldsymbol{\lambda}$ for fixed rank R, i.e,

$$\min \left\| \mathbb{X} - [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n]\!] \right\|_F^2 + \sigma \sum_{i=1}^{N} \left\| \mathbf{A}_i \right\|_F^2 \tag{1}$$

$$= \min \left\| \mathbb{X} - [\![\boldsymbol{\lambda}; \tilde{\mathbf{A}}_1, \ldots, \tilde{\mathbf{A}}_n]\!] \right\|_F^2 + N\sigma \|\boldsymbol{\lambda}\|_{2/N}^{2/N}. \tag{2}$$

- Encourages rank sparsity; $\hat{\lambda}_r = 0$ for $\hat{R} < r \leq R$
- (1) is efficiently solved by penalized ALS
  - Iteratively update $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n$ with closed-form solution

Multiple linked tensor factorization: motivation

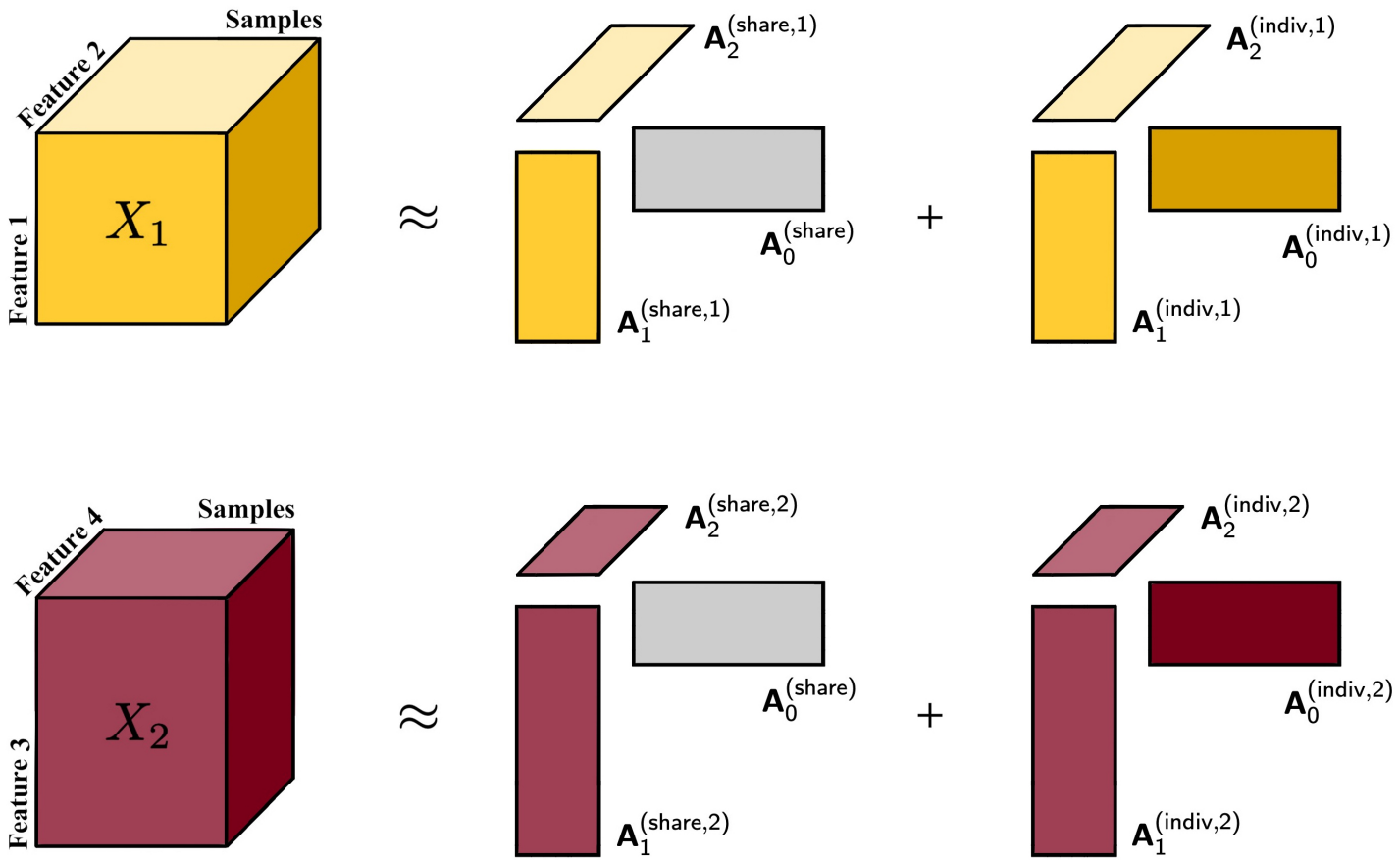
- Study on iron deficiency anemia in infant rhesus monkeys
- Sysmex Hematology Data
  $\mathbb{X}_1$: Infants $\times$ Age$\times$ Hematology Indices
  - Collected at 2 weeks, 2 months, 4 months, 6 months, and 8 months.
  - Includes 18 hematological indices (e.g., red blood cell indices, hemoglobin levels).
- Magnetic Resonance Imaging (MRI) Data
  $\mathbb{X}_2$: Infants$\times$ DTI Parameters $\times$ Brain Regions
  - Measurement from 14 brain regions.
  - Focused on four diffusion tensor imaging (DTI) parameters:
    - Axial diffusivity (AD)
    - Fractional anisotropy (FA)
    - Mean diffusivity (MD)
    - Radial diffusivity (RD)

Multiple linked tensor factorization: motivation

- Different forms of missingness
  - Entry-wise missing: RetHE missing at 2 weeks for an infant
  - Slice-wise missing: no 2-week hematology for an infant
  - Tensor-wise missing: no MRI data for an infant.
- Goals:
  - Approximate underlying signals of multiple tensors.
  - Identify shared and individual structures among the tensors
  - Impute missing data of various forms

Proposed Model: MULTIFAC

- Multiple linked tensor factorization (MULTIFAC) method:



General Framework

- Consider multiple tensors sharing the same sample set: $\{\mathbb{X}_k : I_0 \times I_1^{(k)} \times \cdots \times I_{N_k}^{(k)} | k = 1, \ldots, K\}$.
- Minimize the following objective function

$$\sum_{k=1}^{K} \left\| \mathbb{X}_k - [\![\mathbf{A}_0, \mathbf{A}_1^{(k)}, \ldots, \mathbf{A}_{N_k}^{(k)}]\!] \right\|_F^2 + \sigma \left(\left\| \mathbf{A}_0 \right\|_F^2 + \sum_{k=1}^{K} \sum_{i=1}^{N_k} \left\| \boldsymbol{A}_i^{(k)} \right\|_F^2 \right)$$

- Iteratively update $\mathbf{A}_0$ and each $\mathbf{A}_i^{(k)}$ with closed form solution.
- "Maximum" overall rank R: $\mathbf{A}_0 : N \times R$, $\mathbf{A}_i^{(k)} : I_i^{(k)} \times R$.

General Framework

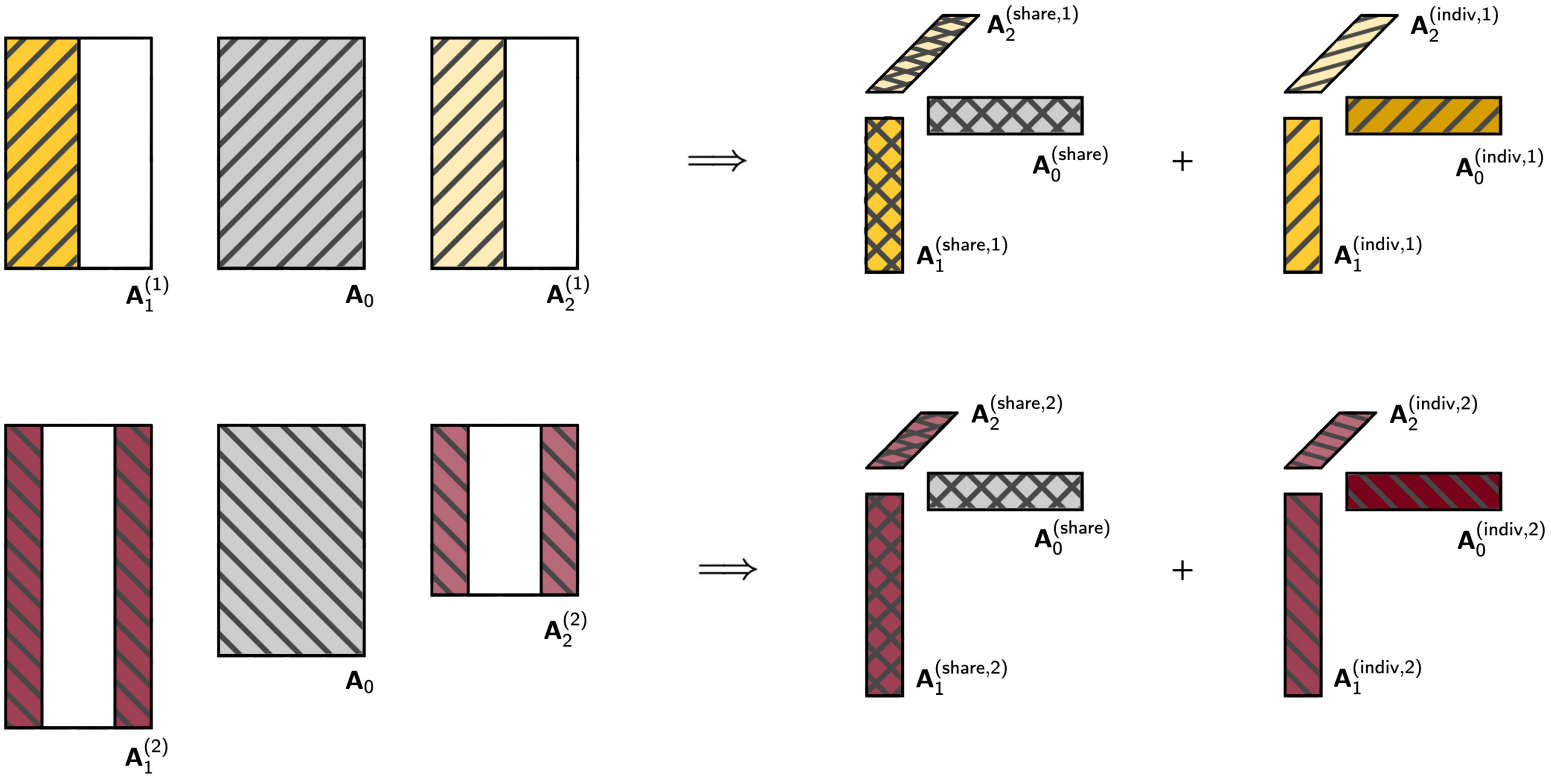
- Equivalent to the following objective:

$$\sum_{k=1}^{K} \left\| \mathbb{X}_k - [\![\boldsymbol{\lambda}_0 * \boldsymbol{\lambda}_{(0)k}; \tilde{\mathbf{A}}_0, \tilde{\mathbf{A}}_1^{(k)}, \ldots, \tilde{\mathbf{A}}_{N_k}^{(k)}]\!] \right\|_F^2$$
$$+ \sigma \left(\|\boldsymbol{\lambda}_0\|_2^2 + \sum_{k=1}^{K} N_k \|\boldsymbol{\lambda}_{(0)k}\|_{2/N_k}^{2/N_k} \right)$$

- Induces overall rank sparsity
- Identifies shared and unshared low-rank signal

Rank Sparsity Property for MULTIFAC

- Rank sparsity property produces individual structures.



Missing data imputation

- Consider missing entries and ‘tensor-wise” missing
- Initialize missing values, e.g., via mean-imputation
- Update missing values with current structure at each iteration

Model selection: cross-validation

- Hold out entries or tensor slices as "missing"
- Optimize imputation accuracy
- Step 1: Selection of tensor ranks:
  - Start with a large overall rank R
  - Grid search on the penalty factor σ.
  - Fix shared and unshared ranks at optimal value
- Step 2: Fine tuning of σ.
  - Fixed ranks R_{shared}, R_{Indiv1}, R_{Indiv2}
  - Adjust σ to optimize imputation accuracy

Related work

- Coupled matrix/tensor factorization[1]
  - Focused on integrating tensors with matrices
  - MULTIFAC tensor must be at least 3-way to identify shared and unshared signals
- Structured data fusion[2]
  - Very general framework
  - Requires specified tensor ranks (shared, unshared) and penalties
  - Implemented in Matlab package TensorLab
  - Optimized using non-linear least squares

[1] Acar, Kolda, Dunlavy 2011; and others.
[2] Sorber, Van Barel, De Lathauwer, 2015

Simulation Study

- Two tensors $X_1 : 100 \times 100 \times 4$ and $X_2 : 100 \times 40 \times 10 \times 3$.
  - Factor matrices generated from standard normal distribution.
  - Rank for shared structure: 2, rank for individual structures: 3.
  - Signal-to-noise ratio: 1/3, 1, 3.
- Key measurement: related squared error (RSE) = norm of residual/ norm of true signal.
- Complete data decomposition + missing data imputation.
- Compare to TensorLab with true ranks

Simulation Results

Table: Performance metrics for decomposing complete tensors.

Method	SNR	RSE^1_{full}	RSE^2_{full}	RSE^1_{share}	RSE^2_{share}	RSE^1_{Indiv}	RSE^2_{Indiv}
MULTIFAC	3	0.113 (0.025)	0.046 (0.003)	0.162 (0.187)	0.112 (0.227)	0.156 (0.095)	0.116 (0.233)
Tensorlab		0.25 (0.145)	0.212 (0.127)	1.327 (2.371)	0.975 (0.524)	1.028 (1.573)	0.814 (0.409)
MULTIFAC	1	0.169 (0.02)	0.083 (0.02)	0.235 (0.224)	0.144 (0.213)	0.232 (0.143)	0.161 (0.229)
Tensorlab		0.295 (0.117)	0.24 (0.13)	1.254 (1.211)	1.509 (3.305)	1.051 (1.093)	1.363 (2.941)
MULTIFAC	1/3	0.274 (0.028)	0.151 (0.041)	0.379 (0.239)	0.293 (0.305)	0.392 (0.207)	0.31 (0.276)
Tensorlab		0.361 (0.094)	0.262 (0.127)	1.378 (1.584)	0.903 (0.632)	1.141 (1.168)	0.761 (0.5)

Simulation Results

Table: Imputation performance for tensors with missing data.

Tensor	SNR	$RSE_{observe}$	$RSE_{missing}$	$RSE_{entry\text{-}wise}$	$RSE_{tensor\text{-}wise}$
X_1	3	0.126 (0.032)	0.543 (0.095)	0.129 (0.033)	0.262 (0.206)
X_2		0.049 (0.004)	0.525 (0.117)	0.049 (0.004)	0.232 (0.252)
X_1	1	0.174 (0.016)	0.564 (0.093)	0.179 (0.017)	0.327 (0.221)
X_2		0.084 (0.018)	0.531 (0.112)	0.086 (0.019)	0.281 (0.24)
X_1	1/3	0.286 (0.033)	0.602 (0.086)	0.292 (0.036)	0.447 (0.227)
X_2		0.164 (0.051)	0.561 (0.112)	0.166 (0.053)	0.39 (0.232)

Data Analysis: Iron Deficiency Data

- Sysmex Hematology Data X_1 : Infants $\times$ Age$\times$ Hematology Indices
  - Collected at 2 weeks, 2 months, 4 months, 6 months, and 8 months.
  - Includes 18 hematological indices (e.g., red blood cell indices, hemoglobin levels).
- Magnetic Resonance Imaging (MRI) Data X_2 : Infants$\times$ DTI Parameters $\times$ Brain Regions
  - Measurement from 14 brain regions.
  - Focused on four diffusion tensor imaging (DTI) parameters:
    - Axial diffusivity (AD)
    - Fractional anisotropy (FA)
    - Mean diffusivity (MD)
    - Radial diffusivity (RD)

Data Analysis: Iron Deficiency Data

Proportion of variance explained & (rank):

Dataset, Structure	Total	Shared	Individual
Hematology (X_1)	0.757 (13)	0.244 (6)	0.451 (7)
MRI (X_2)	0.708 (16)	0.373 (6)	0.311 (10)

Data Analysis: Iron Deficiency Data

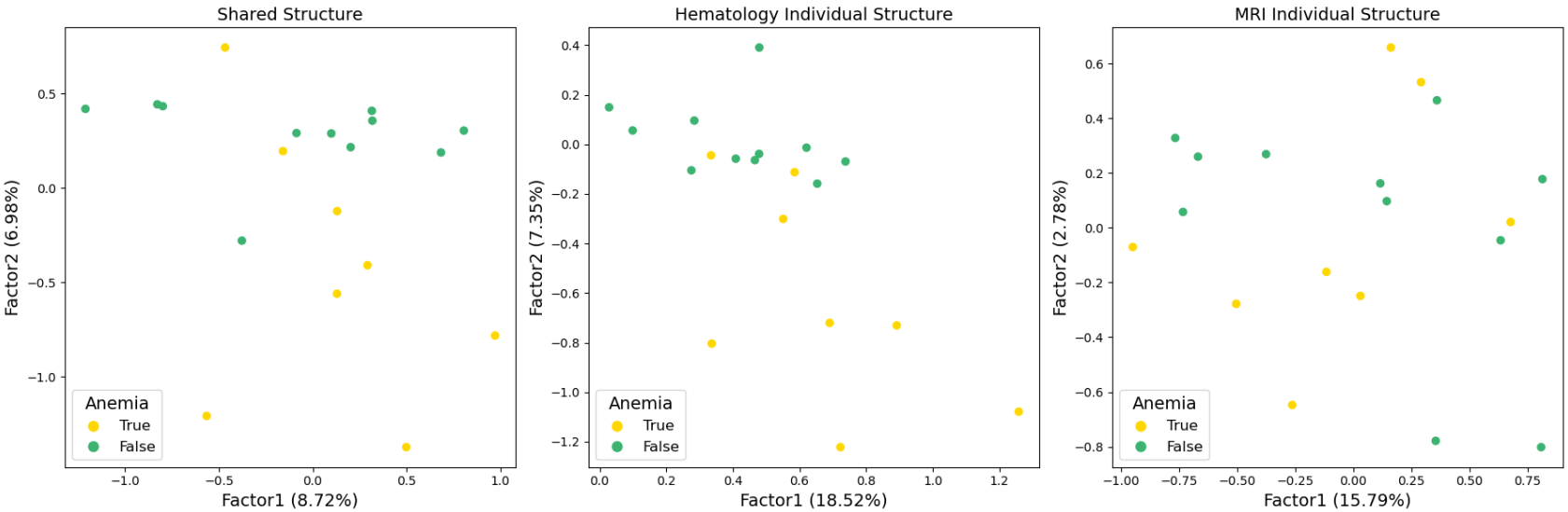


Figure: Top 2 Sample Loadings in Different Structures

Data Analysis: Iron Deficiency Data

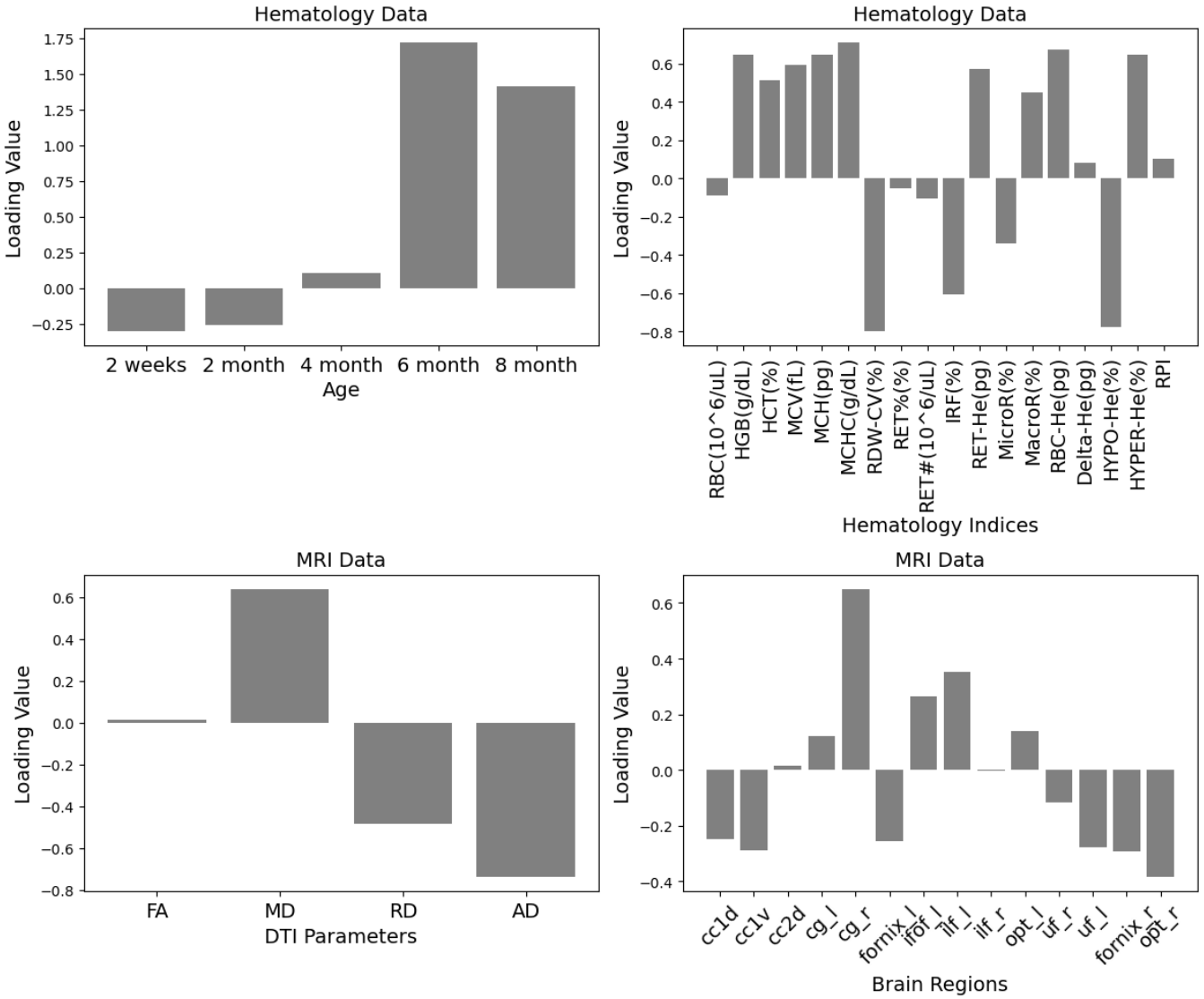


Figure: Loadings for the top shared component

Data Analysis: Iron Deficiency Data

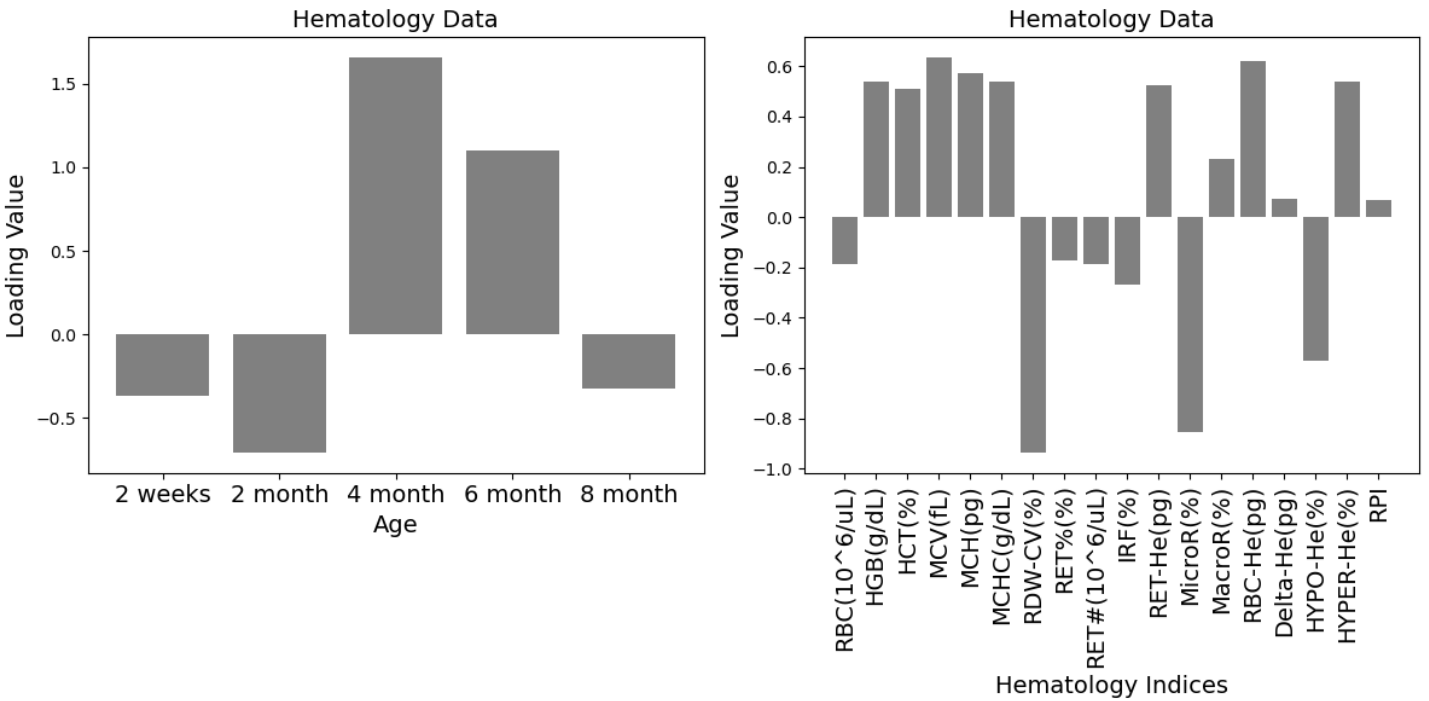


Figure: Loadings for the second hematology component

Future work

- Extending EVB approach to linked tensor context
- Extending EVB approach to reduced rank regressions
  - Multiple augmented reduced rank regression [Wang and Lock, 2024]
- Fully Bayesian inference based on EVB solution
  - Useful for multiple imputation and propagating uncertainty
  - Bayesian simultaneous factorization [Samorodnitsky, Wendt and Lock, 2024]

Thank you!

- Support: NIGMS grant R01-GM130622

- **EV-BIDIFAC**: EF Lock. Empirical Bayes Linked Matrix Decomposition. *Machine Learning*, 113 (10): 7451-7477.
  - R Code: https://github.com/lockEF/bidifac

- **MULTIFAC**: Zhiyu Kang, R Rao and EF Lock. Multiple Linked Tensor Factorization. *arXiv*: 2502.20286, 2025.
  - Python Code: https://github.com/lockEF/bidifac