Linked Matrix Factorization

Eric F. Lock[1]

with Michael J. O'Connell[2]

[1]University of Minnesota, Division of Biostatistics

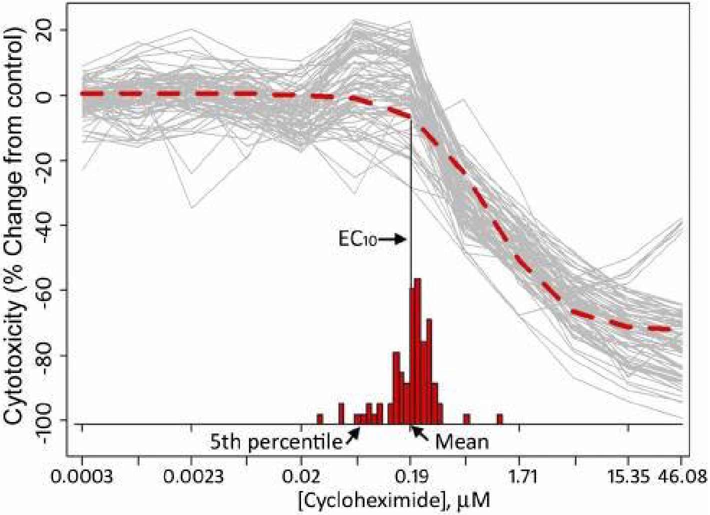
[2]Miami University, Department of Statistics

SDSS Seattle, 05/30/2019

Toxicity screening experiment

- Data for $1,086$ lymphoblastoid cell lines (1000 Genomes Project)
- 179 chemicals
- Collected by the Rusyn lab (UNC)
  - Initial analysis described in Abdo et al., 2015
  - Data available through Synapse DREAM challenge (Eduati et al., 2015)
- EC10 measured for each cell line $\times$ chemical pair
  - Lowest concentration with 10% cell death

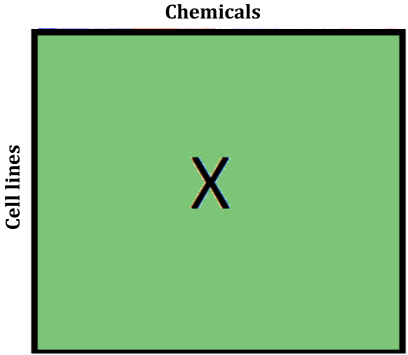
Cytotoxicity curves



(Ref: Lock et al., 2012)

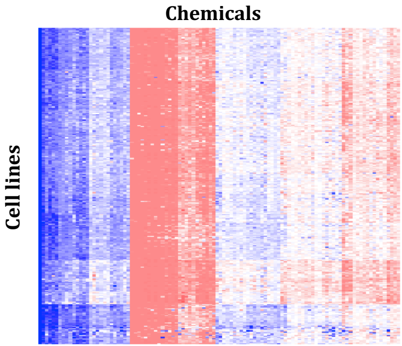
Toxicity matrix

- X : 1086×179 of log(EC10) values



Toxicity matrix

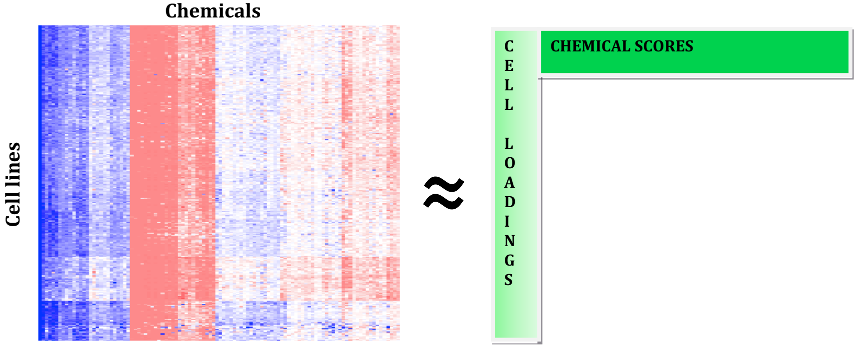
- X : 1086×179 of log(EC10) values



- Heatmap: red = more toxic, blue = less toxic

Toxicity matrix

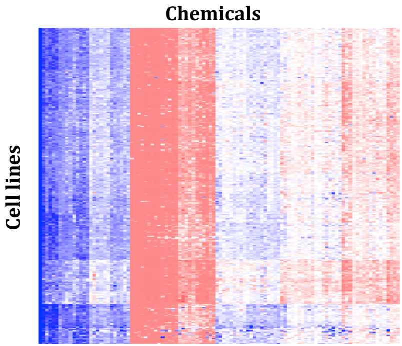
- X : 1086×179 of log(EC10) values



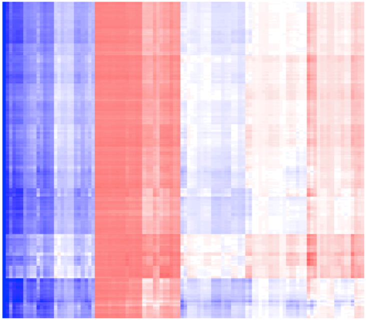
- Low rank factorization: $X \approx UV$, $U : 1086 \times r$, $V : r \times 179$.

Toxicity matrix (rank 3 approximation)

- X : 1086×179 of log(EC10) values



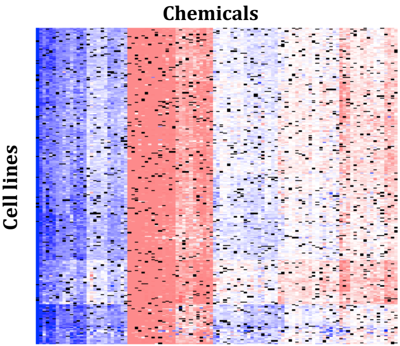
$\approx$



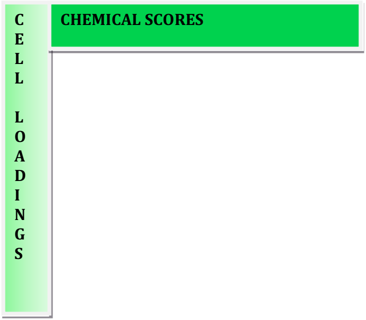
- Low rank factorization: $X \approx UV$, $U : 1086 \times 3$, $V : 3 \times 179$.

Toxicity matrix (5% missing values)

- X : 1086×179 of log(EC10) values



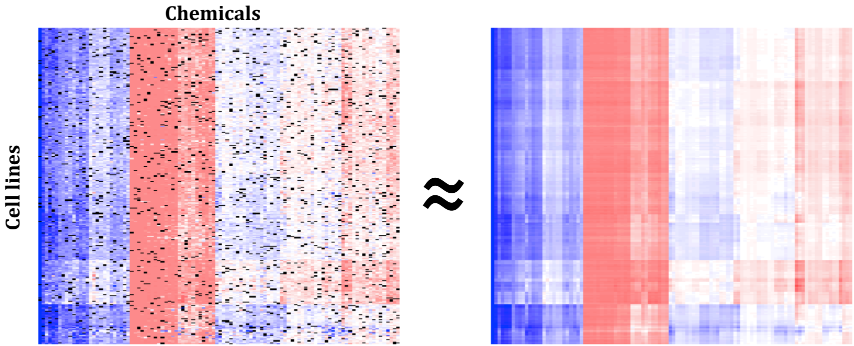
$\approx$



- Low rank factorization: $X \approx UV$, $U : 1086 \times r$, $V : r \times 179$.

Toxicity matrix (5% missing values)

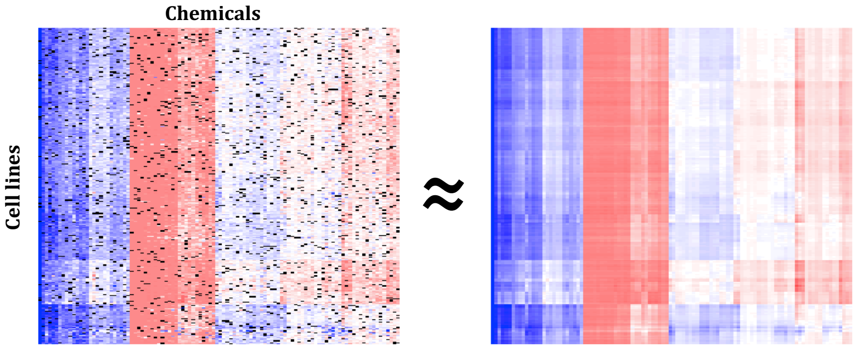
- X : 1086×179 of log(EC10) values



- Low rank factorization: $X \approx UV$, $U : 1086 \times 3$, $V : 3 \times 179$.

Toxicity matrix (5% missing values)

- X : 1086×179 of log(EC10) values

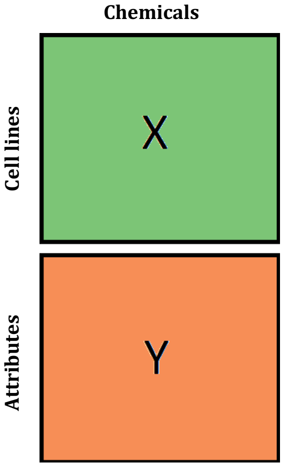


- Low rank factorization: $X \approx UV$, $U : 1086 \times 3$, $V : 3 \times 179$.

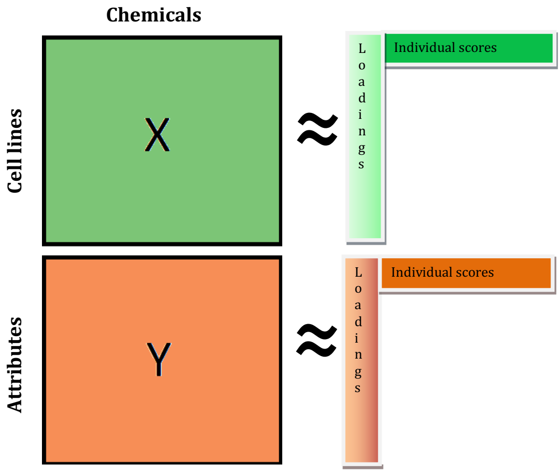
Chemical attributes

- Also have 9432 quantitative attributes for each chemical
  - 160 descriptors using Chemistry Development Kit (CDK)
  - $9,272$ descriptors using Simplex representation of molecular structure (SIRMS)
- Linked data matrices:
  - X : 1086×179 of log(EC10) values
  - Y : 9272×179 of chemical attributes

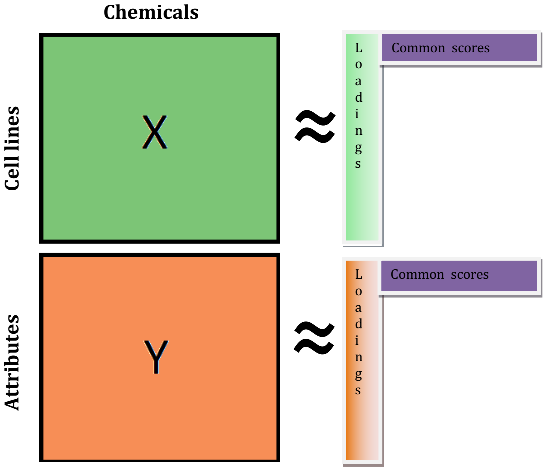
Vertically linked data



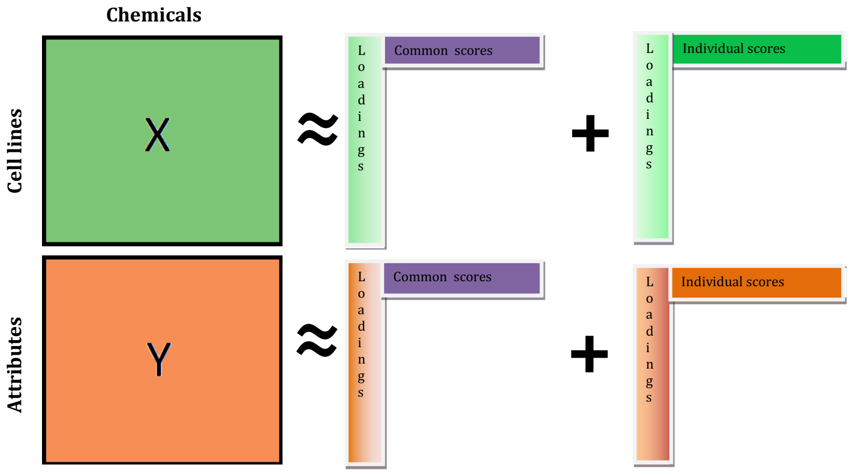
Vertically linked data: separate factorizations



Vertically linked data: joint factorization



Vertically linked data: JIVE factorization



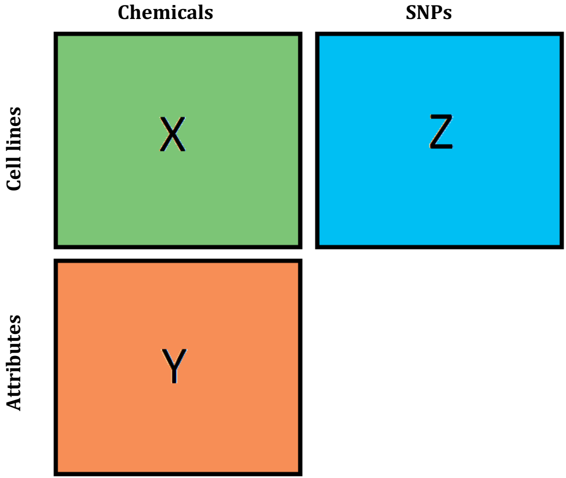
Joint + individual factorization methods

- JIVE [Lock, Hoadley, Marron, and Nobel, 2013]
- AJIVE [Feng, Jiang, Hannig and Marron, 2018]
- SLIDE [Gaynanova and Li, 2018]
- GIPCA [Zhu, Li, Lock, 2018]

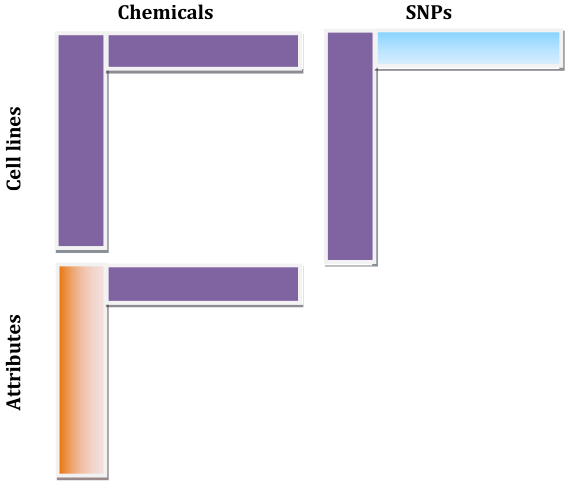
Genotype data

- Also have genotype data available for each cell line.
  - Single-nucleotide polymorphisms (SNPs) (minor allele count $0, 1, 2$)
- Linked data matrices after filtering:
  - X : 751×105 of log(EC10) values
  - Y : 105×105 of chemical attributes
  - Z : 751×441 of SNPs

Bidimensionally linked data



Bidimensionally linked data: joint factorization



Joint linked matrix factorization: model

- Approximation of rank r :

$$X = US_xV^T + E_x$$
$$Y = US_yV_y^T + E_y$$
$$Z = U_zS_zV^T + E_z$$

  - $U : 751 \times r$, $U_z : 441 \times r$
  - $V : 105 \times r$, $V_y : 105 \times r$
  - S_x, S_y, S_z are $r \times r$
  - E_x, E_y, E_z are error matrices (iid entries, mean 0)
- Identifiable if
  - Columns of each of U, U_z, V, V_y are orthonormal
  - S_x, S_y, S_z are diagonal

Joint linked matrix factorization: model

- Approximation of rank r :

$$X = US_xV^T + E_x$$
$$Y = US_yV_y^T + E_y$$
$$Z = U_zS_zV^T + E_z$$

  - $U : 751 \times r$, $U_z : 441 \times r$
  - $V : 105 \times r$, $V_y : 105 \times r$
  - S_x, S_y, S_z are $r \times r$
  - E_x, E_y, E_z are error matrices (iid entries, mean 0)
- Identifiable if
  - Columns of each of U, U_z, V, V_y are orthonormal
  - S_x, S_y, S_z are diagonal

Joint linked matrix factorization: estimation

- Minimize overall squared residuals (SSR):

$$||X - USV||_F^2 + ||Y - US_y V_y||_F^2 + ||Z - U_z S_z V||^2$$

  - Iteratively estimate each of

$$U, V, S, US_y, \text{ and } U_z S_z$$

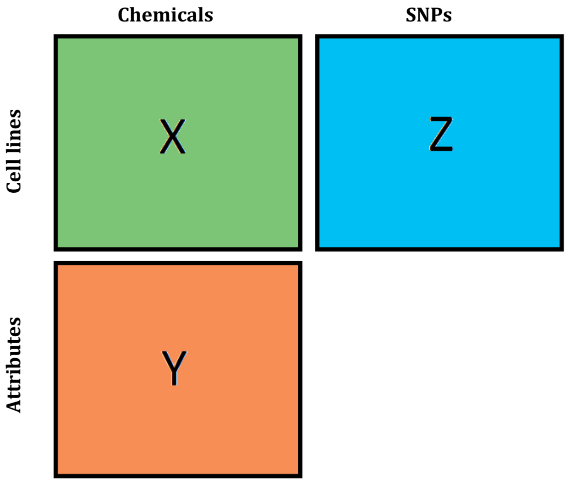
    to minimize SSR.

  - Proceed until convergence

Joint linked matrix factorization: scaling

- Center each matrix to have mean 0
  - Y: subtract mean from each attribute
  - Z: subtract mean from each gene
  - X: subtract overall mean for all EC10 values.
- Scale each matrix to have same total sum of squares.
  - $||X||_F^2 = ||Y||_F^2 = ||Z||_F^2$
  - Gives each dataset same total signal power

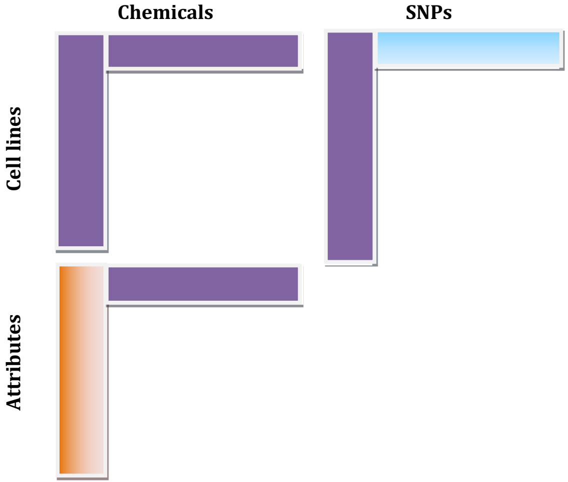
Joint and individual linked matrix factorization (LMF-JIVE)



Joint and individual linked matrix factorization (LMF-JIVE)

	Chemicals	**SNPs**
Cell lines	$\approx$	$\approx$
Attributes	$\approx$	

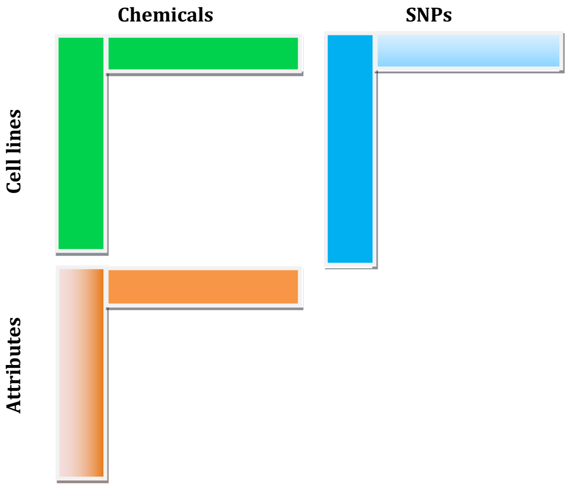
Joint and individual linked matrix factorization (LMF-JIVE)



Joint and individual linked matrix factorization (LMF-JIVE)

	Chemicals	**SNPs**
Cell lines	+	+
Attributes	+	

Joint and individual linked matrix factorization (LMF-JIVE)



LMF-JIVE: model

- Model decomposition:

$$
\begin{aligned}
X &= J_x + A_x + E_x \\
Y &= J_y + A_y + E_y \\
Z &= J_z + A_z + E_z
\end{aligned}
$$

where

$$J_x = U_J S_{Jx} V_J^T, \quad J_y = U_{Jy} S_{Jy} V_J^T, \quad J_z = U_J S_{Jz} V_{Jz}^T$$

and

$$A_x = U_{Ax} S_{Ax} V_{Ax}^T, \quad A_y = U_{Ay} S_{Ay} V_{Ay}^T, \quad A_z = U_{Az} S_{Az} V_{Az}^T$$

- $\text{rank}(J_x) = \text{rank}(J_y) = \text{rank}(J_z) = r$
- $\text{rank}(A_x) = r_x$, $\text{rank}(A_y) = r_y$, $\text{rank}(A_z) = r_z$

LMF-JIVE: model

- Identifiability conditions:

$$(i)\ row(J_x) = row(J_y) \text{ and } col(J_x) = col(J_z)$$

$$(ii)\ row(A_x) \cap row(A_y) = \{\mathbf{0}\} \text{ and } col(A_x) \cap col(A_z) = \{\mathbf{0}\}$$

$$(iii)\ row(J_x) \cap row(A_x) = \{\mathbf{0}\} \text{ and } col(J_x) \cap col(A_x) = \{\mathbf{0}\}.$$

$$(iv)\ J_y A_y^T = 0_{m_2 \times m_2} \text{ and } J_z^T A_z = 0_{n_2 \times n_2}$$

LMF-JIVE: estimation

- Given ranks, minimize overall squared residuals (SSR):

$$||X - J_x - A_x||_F^2 + ||Y - J_y - A_y||_F^2 + ||Z - J_z - A_z||^2$$

  - Iteratively update all terms to minimize SSR
  - Proceed until convergence

- Post-hoc projections to ensure $J_y A_y^T = 0_{m_2 \times m_2}$ and $J_z^T A_z = 0_{n_2 \times n_2}$

LMF-JIVE: estimation

- Given ranks, minimize overall squared residuals (SSR):

$$||X - J_x - A_x||_F^2 + ||Y - J_y - A_y||_F^2 + ||Z - J_z - A_z||^2$$

  - Iteratively update all terms to minimize SSR
  - Proceed until convergence

- Post-hoc projections to ensure $J_y A_y^T = 0_{m_2 \times m_2}$ and $J_z^T A_z = 0_{n_2 \times n_2}$

Missing data imputation

- Algorithm to impute missing data in X:
  - *Initialize* missing entries to obtain the complete matrix $\hat{X}$.
  - (1) Estimate $\{J_x, A_x\}$ from LMF-JIVE on $\{\hat{X}, Y, Z\}$.
  - (2) Update missing entries in $\hat{X}$:
    $$\hat{X}[i,j] = \begin{cases} X_{ij} \text{ if } X_{ij} \text{ is observed} \\ J_x[i,j] + A_x[i,j] \text{ if } X_{ij} \text{ is missing.} \end{cases}$$
  - *Repeat* steps (1) and (2) until convergence.
- EM Algorithm under Gaussian error
- Allows for imputation of entire rows or columns of X

Rank selection: imputation cross-validation

- Randomly select values of X to hold out as missing

- Compute relative imputation error for given ranks

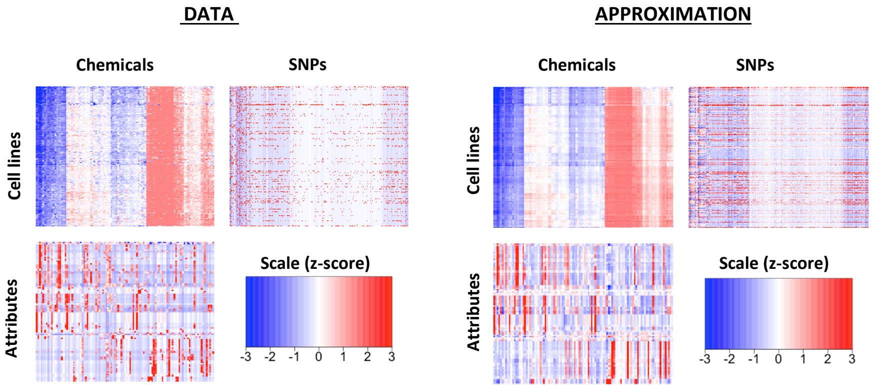
$$RSE = \frac{||\hat{X}[\text{missing values}] - X[\text{missing values}]||_F^2}{||X[\text{missing values}]||_F^2}$$

- Select ranks $\{r, r_x, r_y, r_z\}$ that minimize RSE

- Forward selection approach

LMF-JIVE: low-rank approximation

- Selected ranks: $r = 3, r_X = 4, r_Y = 2, r_Z = 6$



Imputation comparison

- Relative squared error (RSE) for imputed values

	LMF	SVD	softImpute	LMF-JIVE
Missing chemical and cell line	0.878	1.02	1.00	**0.854**
Missing chemical	0.898	1.02	1.00	**0.875**
Missing cell line	0.203	0.208	1.00	**0.201**
Missing entry	0.164	**0.112**	0.113	0.114

Thank you!

- Email: elock@umn.edu
- Slides: http://ericfrazerlock.com/Talks.html
- MJ O'Connell and EF Lock. Linked Matrix Factorization. *Biometrics*, doi: 10.1111/biom.13010, 2018.
- Code: https://github.com/lockEF/LMF