Structured Matrix Completion with Applications to Genomic Data Integration

Aaron Jones

Duke University
BIOSTAT 900

October 14, 2016

Reference

Tianxi Cai, T. Tony Cai & Anru Zhang (2016) Structured Matrix Completion with Applications to Genomic Data Integration, Journal of the American Statistical Association, 111:514, 621-633, DOI: 10.1080/01621459.2015.1021005

Overview

Genomic Data Integration

- In genomics, often analyze data drawn from multiple studies/sources
  - E.g., combine separate studies conducted using different architecture
  - E.g., funding for NGS in a subset of patients, but SNP chip for the rest
  - E.g., may have other data for some patients (miRNA, methylation)
- Complete case analysis reduces power, and may bias associations
- The observed data are full rows (patients) and columns (loci) of the data matrix A; the missing data form a rectangular submatrix of A
  - Take advantage of the missingness structure to impute missing values

Structured Matrix Completion

- For $p_1 \times p_2$ matrix A, observe $m_1 < p_1$ rows and $m_2 < p_2$ columns:

$$A = \begin{array}{c} \begin{array}{cc} m_2 & p_2 - m_2 \end{array} \\ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & (A_{22}) \end{bmatrix} \end{array} \begin{array}{c} m_1 \\ p_1 - m_1 \end{array}$$

- Goal: fill in the missing block A_{22}, given fully observed A_{11}, A_{12}, A_{21}
- Problem: A_{22} could be anything, without some assumptions about A
- Solution: Assume A is approximately low-rank – sensible in genomics

Exact Low-Rank Matrix

Proposition 1: Suppose A is of rank r, the SVD of A_{11} is $A_{11} = U\Sigma V^T$, where $U \in \mathbb{R}^{p_1 \times r}, \Sigma \in \mathbb{R}^{r \times r}$, and $V \in \mathbb{R}^{p_2 \times r}$. If

$$\text{rank}(\begin{bmatrix} A_{11} & A_{12} \end{bmatrix}) = \text{rank}\left(\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}\right) = \text{rank}(A) = r,$$

then $\text{rank}(A_{11}) = r$ and A_{22} is exactly given by

$$A_{22} = A_{21}(A_{11})^{\dagger}A_{12} = A_{21}V(\Sigma)^{-1}U^T A_{12}.$$

- Simple, analytic solution, but $(A_{11})^{\dagger}$ is not continuous in A_{11}, so this method does not give approximate A_{22} for approximately low-rank A

Approximate Low-Rank Matrix

- Definition: A is approximately rank r if there is a significant gap between the rth and $(r+1)$th singular values, $\sigma_r(A)$ and $\sigma_{r+1}(A)$, and the tail $\left(\sum_{k \geq r+1} \sigma_k^q(A)\right)^{1/q}$ is small.
- Let $A = U\Sigma V$ be the SVD of an approximately low-rank matrix A and partition $U \in \mathbb{R}^{p_1 \times p_1}, \Sigma \in \mathbb{R}^{p_1 \times p_2}, V \in \mathbb{R}^{p_2 \times p_2}$ into blocks as

$$U = \begin{array}{c} \begin{array}{cc} r & p_1 - r \end{array} \\ \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{array}{c} m_1 \\ p_1 - m_1 \end{array} \end{array}$$

$$\Sigma = \begin{array}{c} \begin{array}{cc} r & p_2 - r \end{array} \\ \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{array}{c} r \\ p_1 - r \end{array} \end{array}$$

$$V = \begin{array}{c} \begin{array}{cc} r & p_2 - r \end{array} \\ \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} \begin{array}{c} m_2 \\ p_2 - m_2 \end{array} \end{array}$$

Approximate Low-Rank Matrix

$$
\begin{aligned}
A &= U\Sigma V^T \\
&= \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_{11}^T & V_{21}^T \\ V_{12}^T & V_{22}^T \end{bmatrix} \\
&= \begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix} \Sigma_1 \begin{bmatrix} V_{11}^T & V_{21}^T \end{bmatrix} + \begin{bmatrix} U_{12} \\ U_{22} \end{bmatrix} \Sigma_2 \begin{bmatrix} V_{12}^T & V_{22}^T \end{bmatrix} \\
&= \begin{bmatrix} U_{11}\Sigma_1 V_{11}^T & U_{11}\Sigma_1 V_{21}^T \\ U_{21}\Sigma_1 V_{11}^T & U_{21}\Sigma_1 V_{21}^T \end{bmatrix} + \begin{bmatrix} U_{12}\Sigma_2 V_{12}^T & U_{12}\Sigma_2 V_{22}^T \\ U_{22}\Sigma_2 V_{12}^T & U_{22}\Sigma_2 V_{22}^T \end{bmatrix} \\
&= A_{\max(r)} + A_{-\max(r)},
\end{aligned}
$$

where $A_{\max(r)}$ is a rank-r approximation to A with the largest r singular values, and $A_{-\max(r)}$ has small singular values. Then by Proposition 1:

$$U_{21}\Sigma_1 V_{21}^T = (U_{21}\Sigma_1 V_{11}^T)(U_{11}\Sigma_1 V_{11}^T)^{-1}(U_{11}\Sigma_1 V_{21}^T)$$

Known "Rank" r

- Define the notation $M_{\bullet k} := \begin{bmatrix} M_{1k} \\ M_{2k} \end{bmatrix}$ and $M_{k\bullet} := \begin{bmatrix} M_{k1} & M_{k2} \end{bmatrix}$
- When r is known, we can estimate A_{22} by estimating $U_{\bullet 1}$ and $V_{\bullet 1}$ using the r principal components of $A_{\bullet 1}$ and $A_{1\bullet}$ as described below:

Algorithm 1 Structured Matrix Completion with a Known "Rank" r

1. Input: $A_{11} \in \mathbb{R}^{m_1 \times m_2}, A_{12} \in \mathbb{R}^{(p_1 - m_1) \times m_2}, A_{21} \in \mathbb{R}^{m_1 \times (p_2 - m_2)}$
2. Calculate the SVD of $A_{\bullet 1} = U^{(1)} \Sigma^{(1)} V^{(1)T}, A_{1\bullet} = U^{(2)} \Sigma^{(2)} V^{(2)T}$
3. Estimate the column space of U_{11} and V_{11} by $\hat{M} = U^{(2)}_{[,1:r]}, \hat{N} = V^{(1)}_{[,1:r]}$
4. Estimate A_{22} as $\hat{A}_{22} = A_{21} \hat{N} (\hat{M}^T A_{11} \hat{N})^{-1} \hat{M}^T A_{12}$

- Problem: Algorithm 1 assumes r is known, but r is generally unknown
- Solution: First estimate r with some $\hat{r}$, then run Algorithm 1 using $\hat{r}$

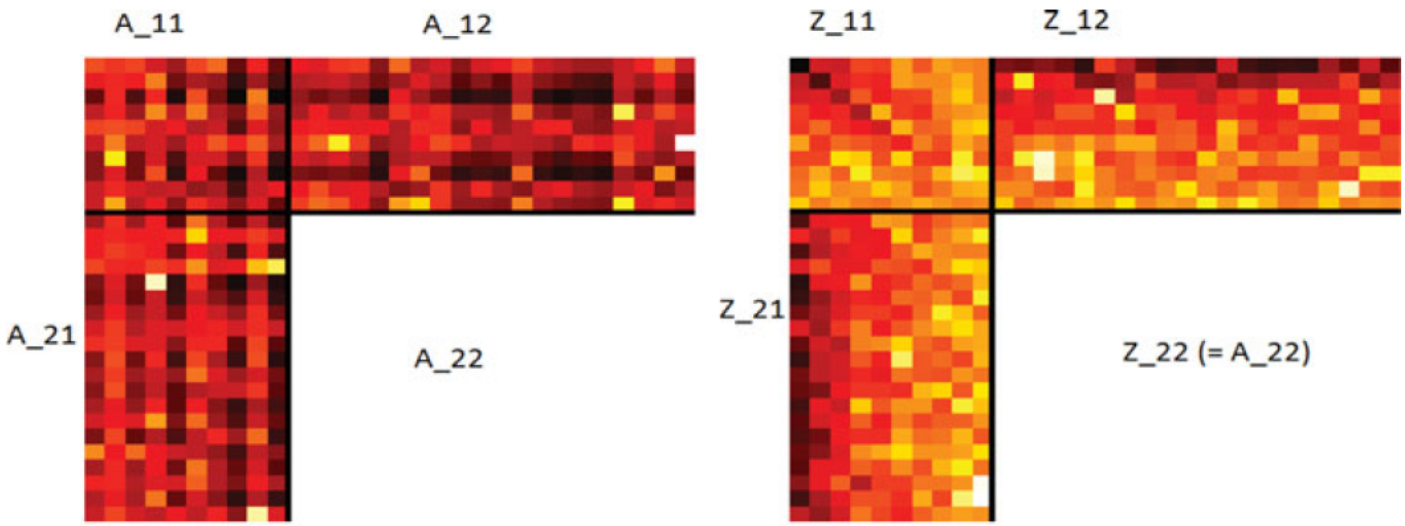
Unknown "Rank" r

The algorithm to recover A_{22} when r is unknown has three steps:

1. Rotate $A_{\bullet 1}$ and $A_{1\bullet}$ by SVD to move significant factors to the front:

$$A_{\bullet 1} = U^{(1)}\Sigma^{(1)}V^{(1)T}, A_{1\bullet} = U^{(2)}\Sigma^{(2)}V^{(2)T}$$

$$\Longrightarrow Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} = \begin{bmatrix} U^{(2)T}A_{11}V^{(1)} & U^{(2)T}A_{12} \\ A_{21}V^{(1)} & A_{22} \end{bmatrix}$$

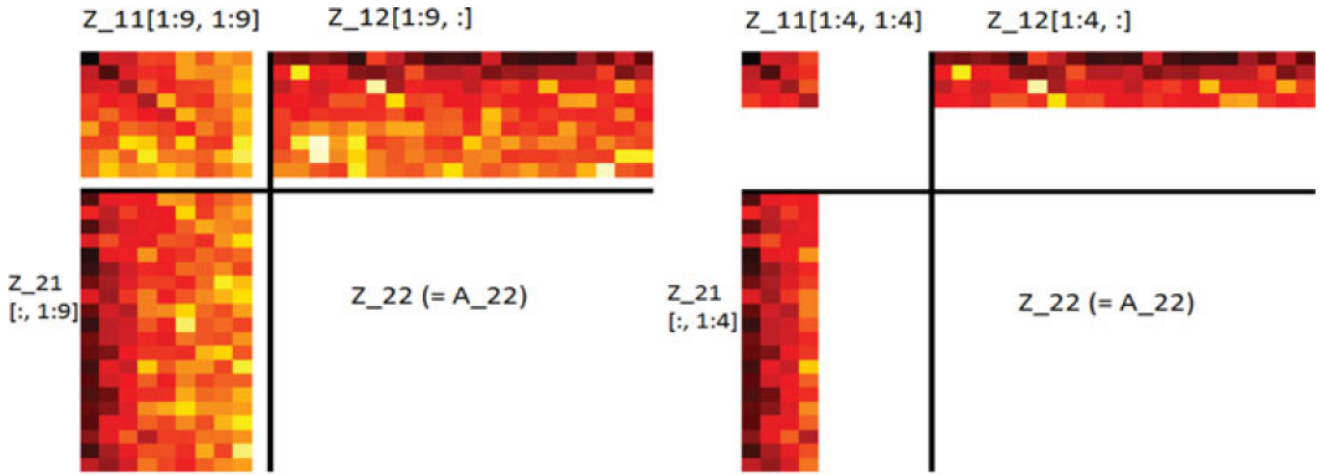


(a) heatmap of block-wise A (b) heatmap of block-wise Z after rotation

Figure 1. Illustrative example with $A \in \mathbb{R}^{30\times 30}$, $m_1 = m_2 = 10$. (A darker block corresponds to larger magnitude.)

Unknown "Rank" r

2. If A were exactly rank-r, the $[r+1, \ldots, m_1]$ rows and $[r+1, \ldots, m_2]$ columns of Z would be zero, but they are nonzero (yet small) due to the perturbation $A_{-\max(r)}$. So, since we want $A_{\max(r)}$, the best rank-r approximation to A, ignore these rows/columns and use the first r.
   - However, r is unknown, so estimate it by the largest $\hat{r}$ such that $Z_{11,[1:\hat{r},1:\hat{r}]}$ is nonsingular and $\sigma_1(Z_{21,[1:\hat{r},1:\hat{r}]} Z^{-1}_{11,[1:\hat{r},1:\hat{r}]}) \leq 2\sqrt{\frac{p_1}{m_1}}$.
3. As before, estimate A_{22} as $\hat{A}_{22} = \hat{Z}_{22} = Z_{21,[,1:\hat{r}]} Z^{-1}_{11,[1:\hat{r},1:\hat{r}]} Z_{12,[1:\hat{r},]}$



(a) Intermediate step when $\hat{r} = 9$ (b) Identify the position to truncate at $\hat{r} = 4$

Unknown "Rank" r

Algorithm 2 Structured Matrix Completion with an Unknown "Rank" r

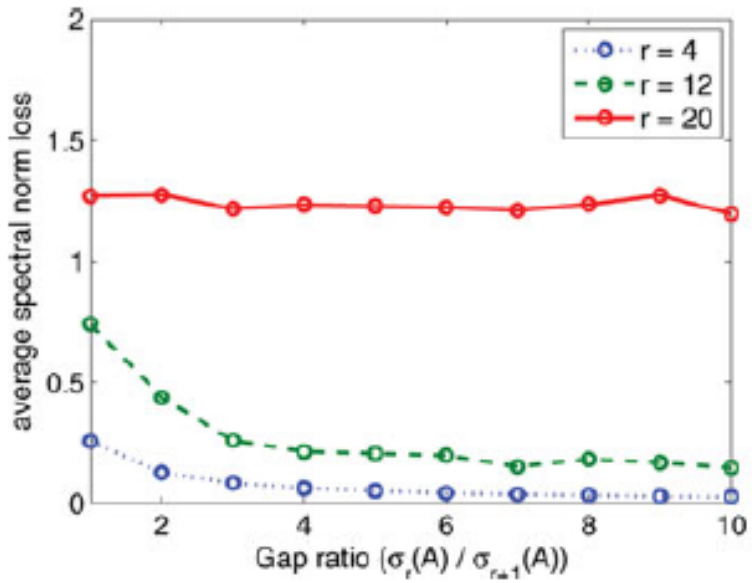
1. Input: $A_{11} \in \mathbb{R}^{m_1 \times m_2}, A_{12} \in \mathbb{R}^{(p_1 - m_1) \times m_2}, A_{21} \in \mathbb{R}^{m_1 \times (p_2 - m_2)}$, thresholding level T_R (or T_C)
2. Calculate the SVD of $A_{\bullet 1} = U^{(1)} \Sigma^{(1)} V^{(1)T}, A_{1\bullet} = U^{(2)} \Sigma^{(2)} V^{(2)T}$
3. Calculate $Z_{11} = U^{(2)T} A_{11} V^{(1)}, Z_{12} = U^{(2)T} A_{12}, Z_{21} = A_{21} V^{(1)}$
4. Estimate the column space of U_{11} and V_{11} by $\hat{M} = U^{(2)}_{[,1:r]}, \hat{N} = V^{(1)}_{[,1:r]}$
5. For $s = \min(m_1, m_2), \ldots, 2, 1$:
   - Calculate $D_{R,s} = Z_{21,[,1:s]} Z^{-1}_{11,[1:s,1:s]}$ (or $D_{C,s} = Z^{-1}_{11,[1:s,1:s]} Z_{12,[1:s,]}$)
   - If $Z_{11,[1:s,1:s]}$ is not singular and $\sigma_1(D_{R,s}) \leq T_R$ (or $\sigma_1(D_{C,s}) \leq T_C$):
     - $\hat{r} = s$
6. If $\hat{r}$ is still unassigned, then $\hat{r} = 0$
7. Estimate A_{22} as $\hat{A}_{22} = \hat{Z}_{22} = Z_{21,[,1:\hat{r}]} Z^{-1}_{11,[1:\hat{r},1:\hat{r}]} Z_{12,[1:\hat{r},]}$

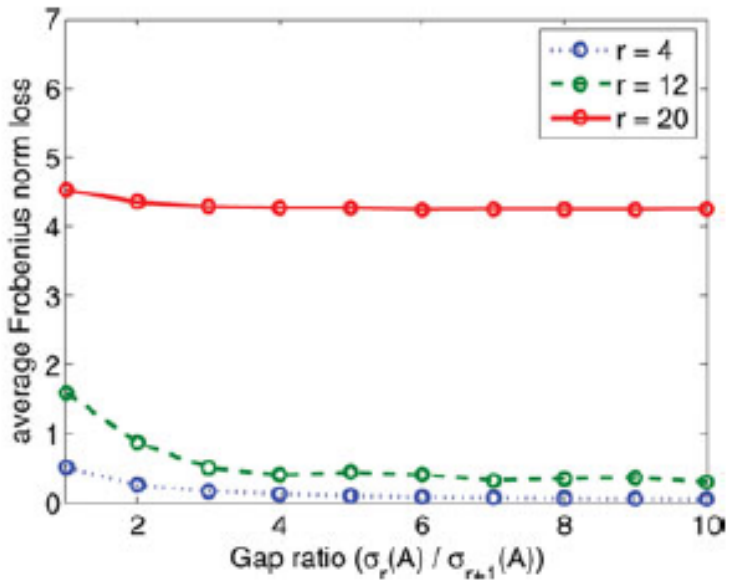
Theoretical Analysis

- The paper presents upper and lower bounds for the estimation errors of Algorithms 1 & 2, so the optimal rate of recovery can be given for certain classes of approximately low-rank matrices
- There are also probability bounds on the estimation errors for fixed A and random rows/columns observed, and also for random A

Simulation

- Fix $p_1 = p_2 = 1000, m_1 = m_2 = 50$
- Choose singular values as $\{1, \overset{r-2}{\ldots}, 1, g^{-1}1^{-1}, g^{-1}2^{-1}, \ldots\}$
- Vary gap ratio $g = 1, 2, \ldots, 10$, rank $r = 4, 12, 20$

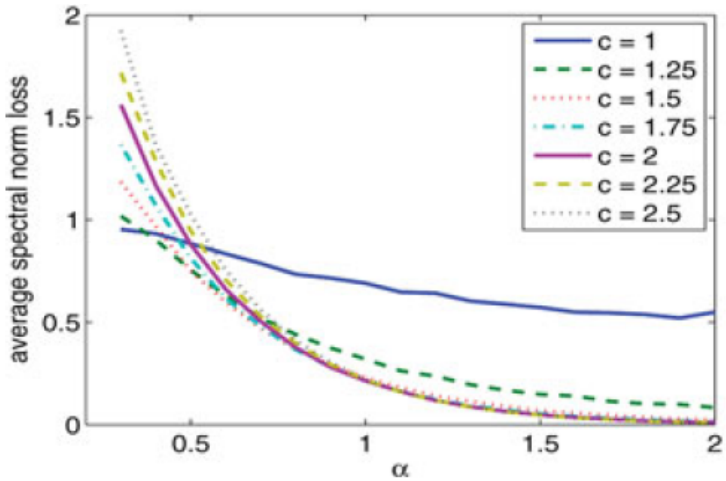


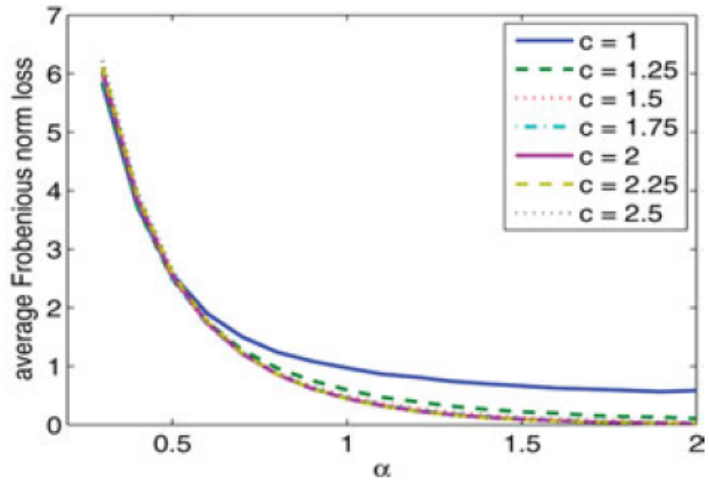


- Algorithm improves as r gets smaller and $g = \dfrac{\sigma_r(A)}{\sigma_{r+1}(A)}$ gets larger

Simulation

- Fix $p_1 = p_2 = 1000, m_1 = m_2 = 50$
- Choose singular values as $\{j^{-\alpha} : j = 1, 2, \ldots, \min(p_1, p_2)\}$
- Vary α between 0.3 and 2, and $T_R = c\sqrt{\frac{p1}{m1}}$ for c between 1 and 6

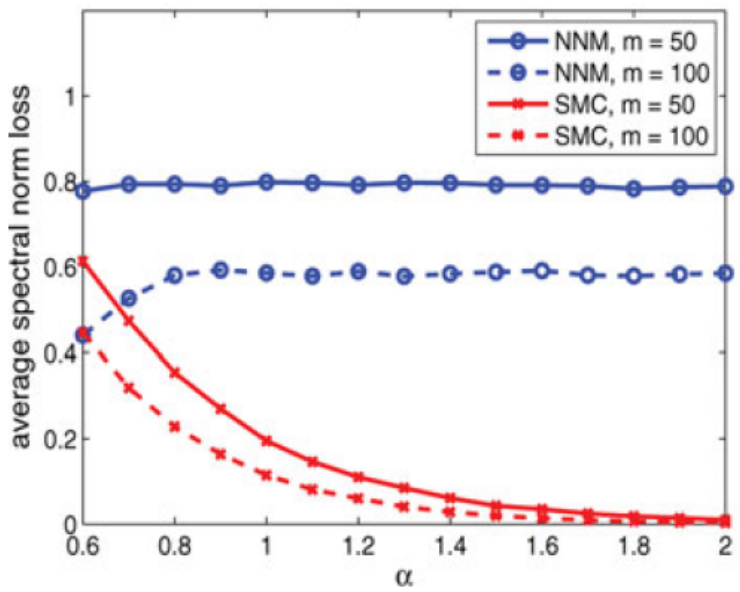


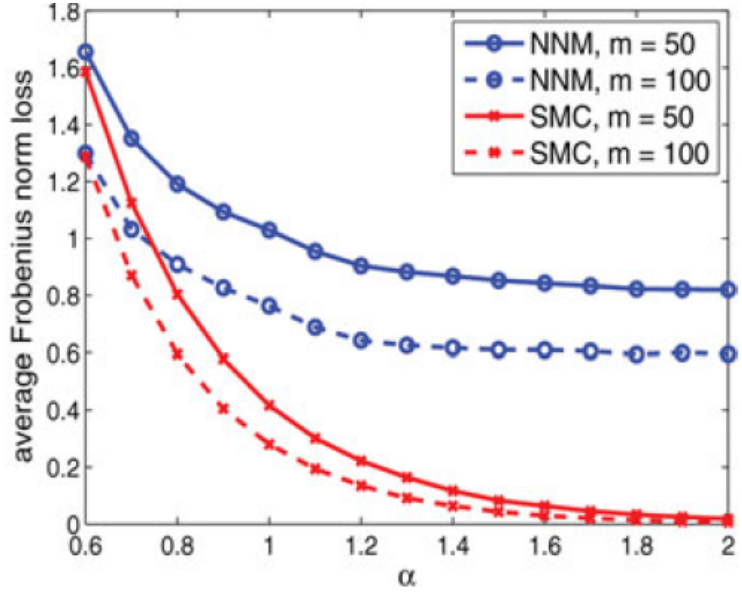


- Algorithm does well if α is not too small and improves as α gets larger
- The paper identifies $c = 2$ as the recommended optimal value

Simulation

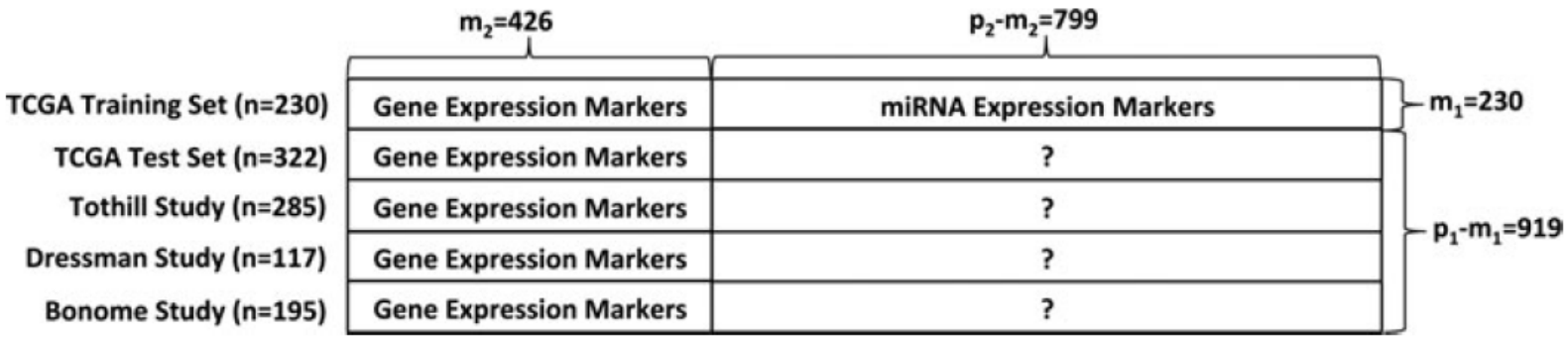
- Fix $p_1 = p_2 = 1000$
- Choose singular values as $\{j^{-\alpha} : j = 1, 2, \ldots, \min(p_1, p_2)\}$
- Vary α between 0.6 and 2, and $m_1 = m_2 = 50$ or 100
- Compare SMC to constrained nuclear norm minimization (NNM)





- SMC outperforms NNM in approximately low-rank matrices with rectangular missingness

Application



(a) Integrated analysis with imputed miRNA versus single study with observed miRNA

	logHR			SE			*p*-Value		
	Ori.	SMC	NNM	Ori.	SMC	NNM	Ori.	SMC	NNM
$\mathcal{G}$	0.067	0.143	0.168	0.041	0.034	0.028	0.104	0.000	
miRNA_1^{PC}	− 0.012	− 0.019	− 0.013	0.009	0.006	0.012	0.218	0.001	0.283
miRNA_2^{PC}	0.023	0.018	− 0.005	0.014	0.009	0.014	0.092	0.039	0.725

- Imputing the missing miRNA expression levels reduces the standard errors and increases power
- Adding the imputed miRNA significantly improves the predictive ability of the model