Simulation Results (2/13/17)

Model

We assume that our dataset, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is generated via the following model,

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{A} + \varepsilon, \tag{1}$$

where $\boldsymbol{Z} \in \mathbb{R}^{n \times k}$ represents the "latent" data where each sample (row) represents the expression of a gene cluster, and $\boldsymbol{A} \in \{0, 1\}^{k \times d}$ is the assignment matrix where each column only has one non-zero entry. That is, we are assuming a hard clustering where every gene belongs to only one cluster.

We assume each row of $\boldsymbol{Z}$ is drawn iid from $N(0, \boldsymbol{\Sigma})$, where we are interested in estimated the correlation matrix based on the latent covariance matrix $\boldsymbol{\Sigma}$. Also, $\varepsilon \in \mathbb{R}^n$ is drawn iid from $N(0, \sigma^2 \boldsymbol{I})$.

Algorithms

To be filled.

- Black: clustering based on CORD in https://arxiv.org/pdf/1508.01939.pdf where we choose the appropriate tuning parameter to achieve K clusters.
- Red: spectral clustering method in Section 6 of https://arxiv.org/pdf/1606.05100.pdf with a set K.
- Green: hierarchical clustering where we use single-linkage clustering and prune the tree at K (which we set beforehand).
- Blue: stochastic block model where we first estimate a graph of d nodes based on the significant correlations (using the Fisher Z-transform), and then use spectral clustering on the adjacency matrix with a set K.

In all methods above, we average the respective columns of $\boldsymbol{X}$ to estimate the latent covariance matrix, $\boldsymbol{\Sigma}$.

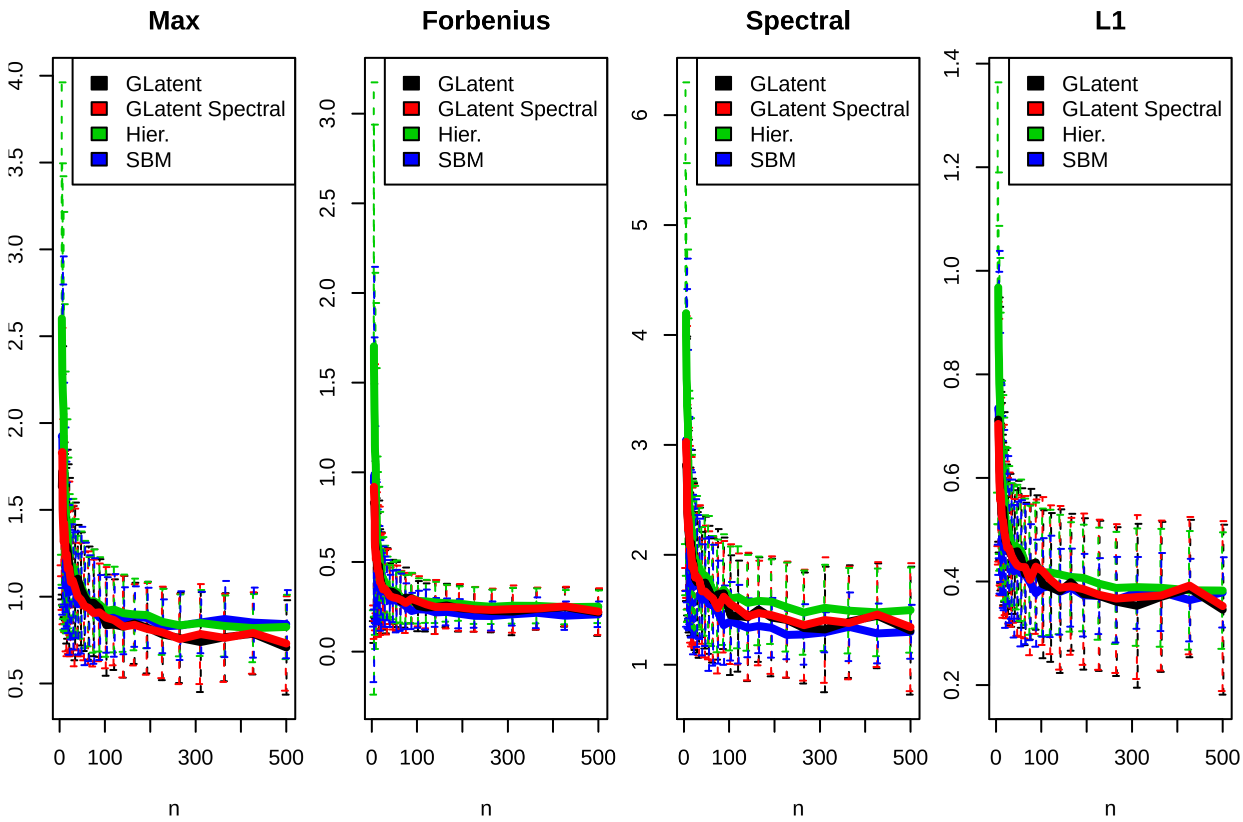
Simulation Results

Model 1 (Standard) Correctly specified

We have $K = 4$, and each cluster has 6 genes.

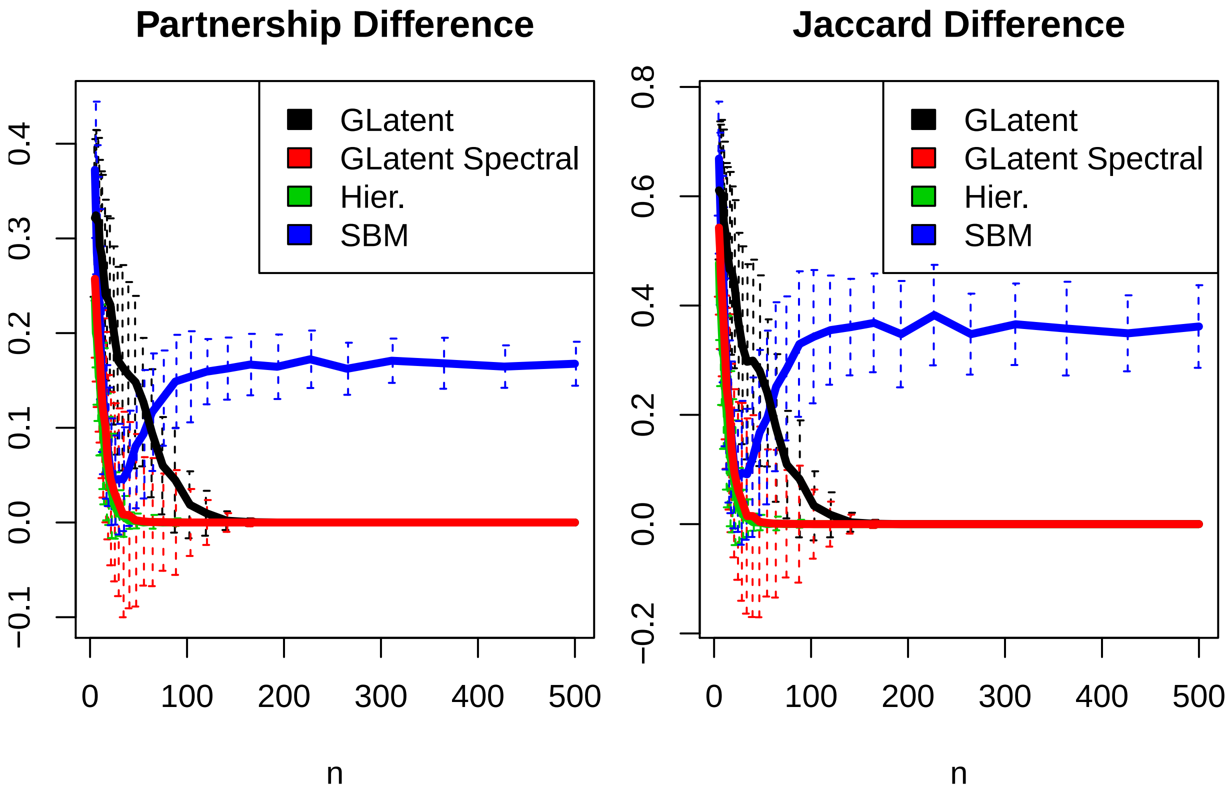
Estimation of Latent Covariance Matrix

We show the "Max", "Forbenius", "Spectral", and "L1" errors between the estimated latent covariance matrix and the population covariance matrix below, from left to right.



Estimation of the Clusterings

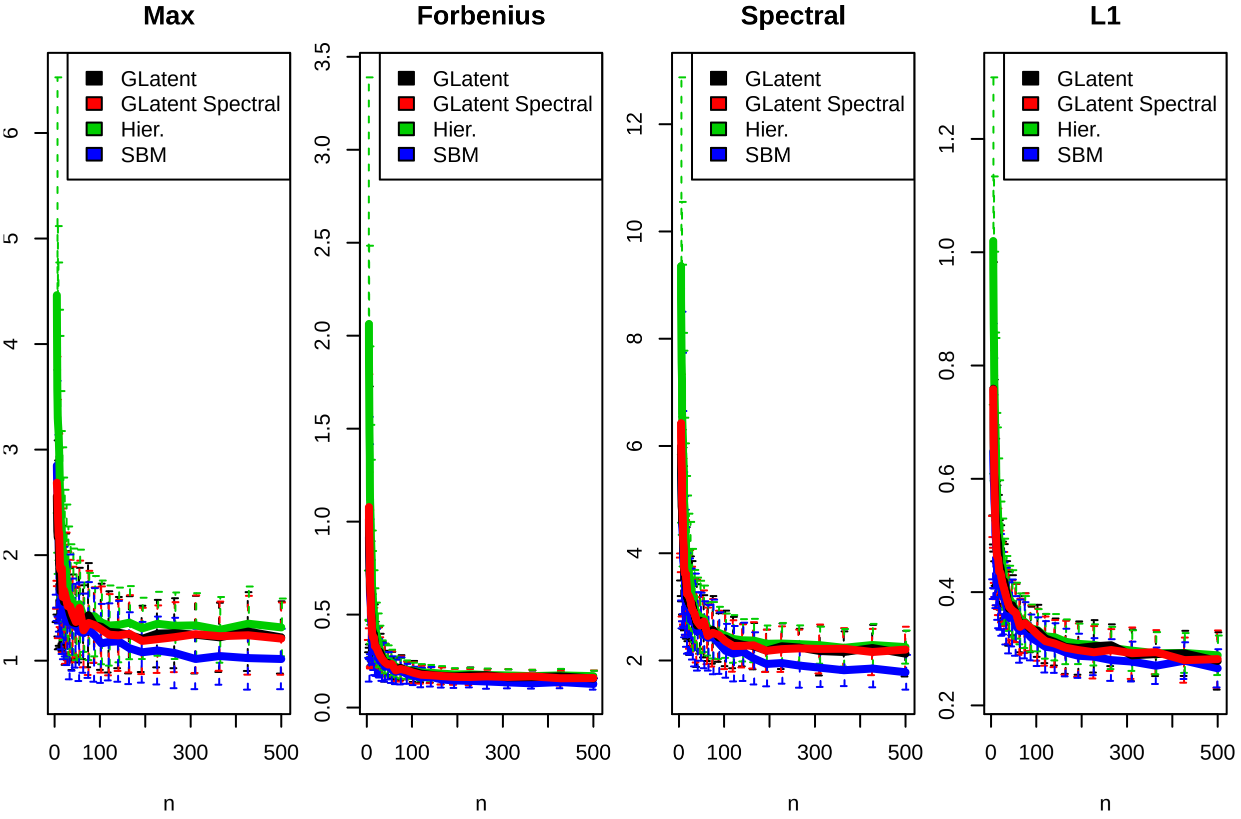
We show the partnership difference and the Jaccard difference in the clustering assignments, on the left and right.



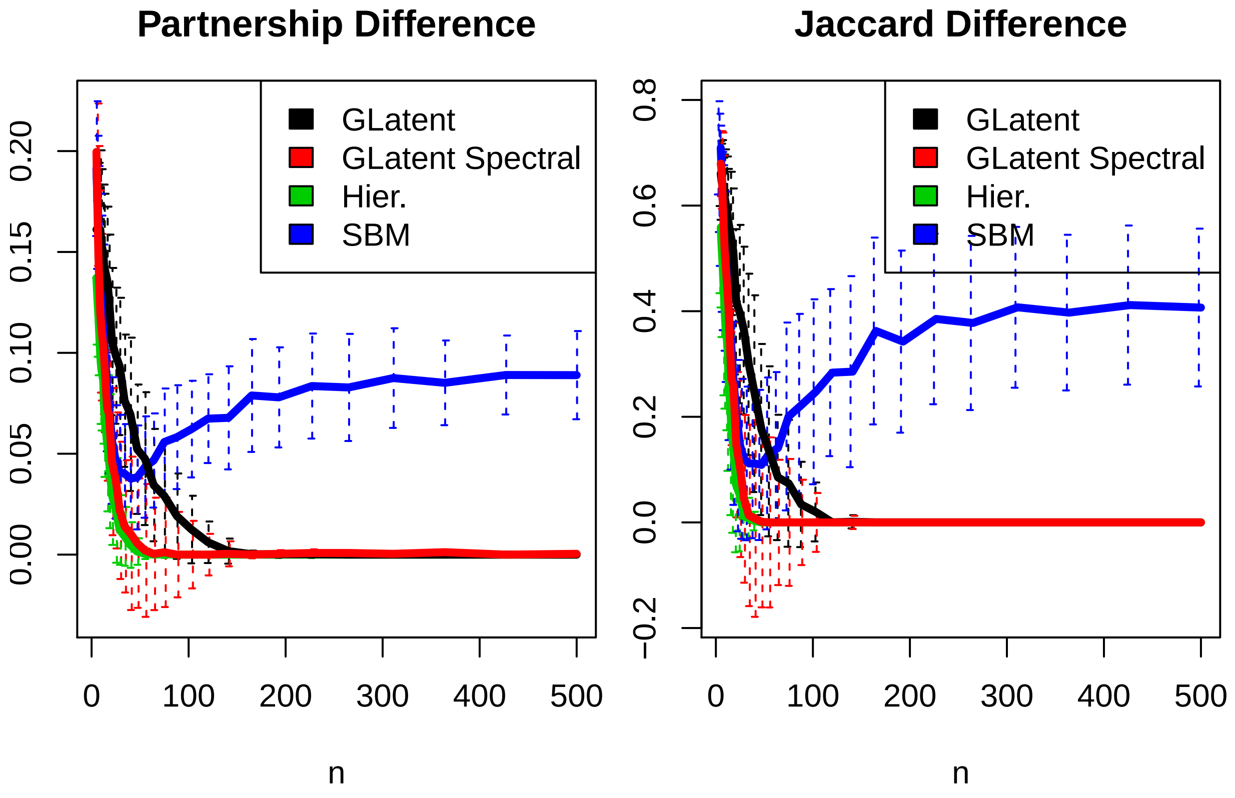
Model 2 (Fragment: many small clusters) Correctly specified

We have $K = 8$, and each cluster has 3 genes.

Estimation of Latent Covariance Matrix



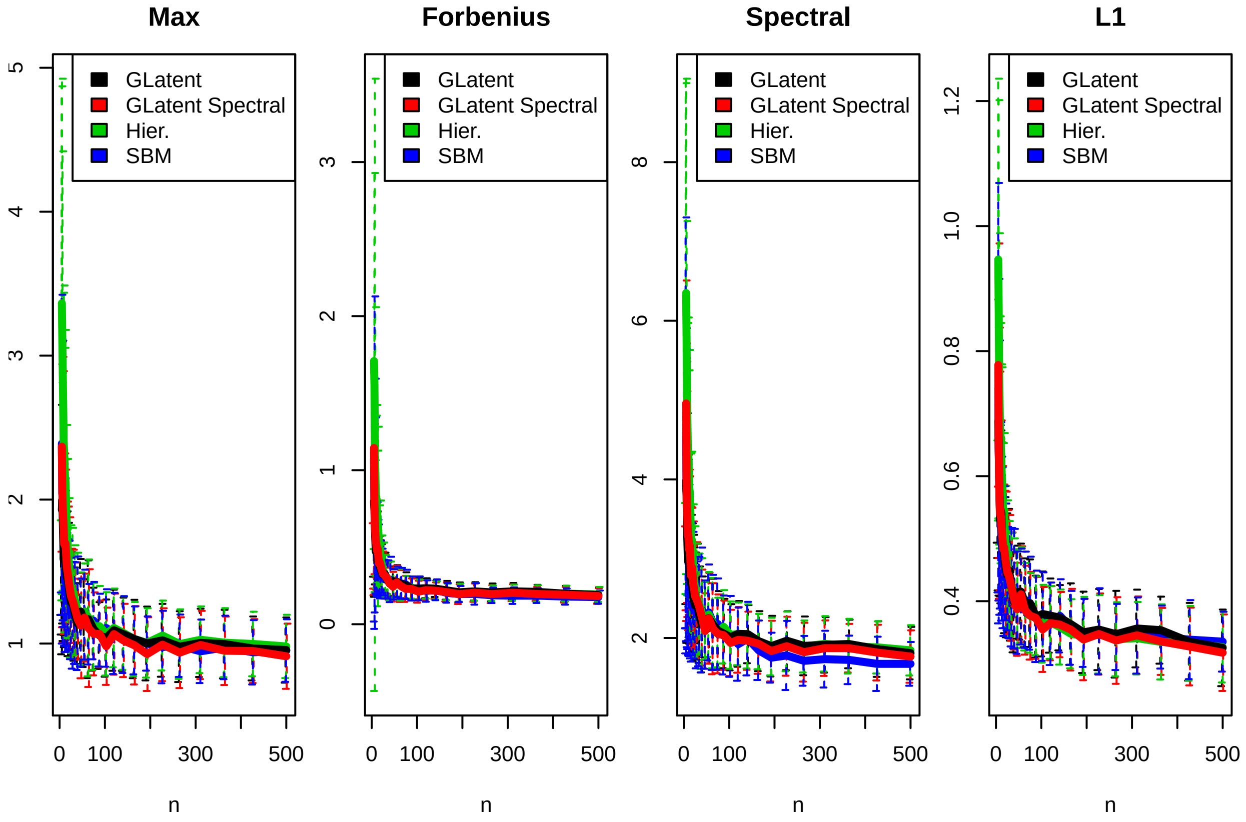
Estimation of the Clusterings



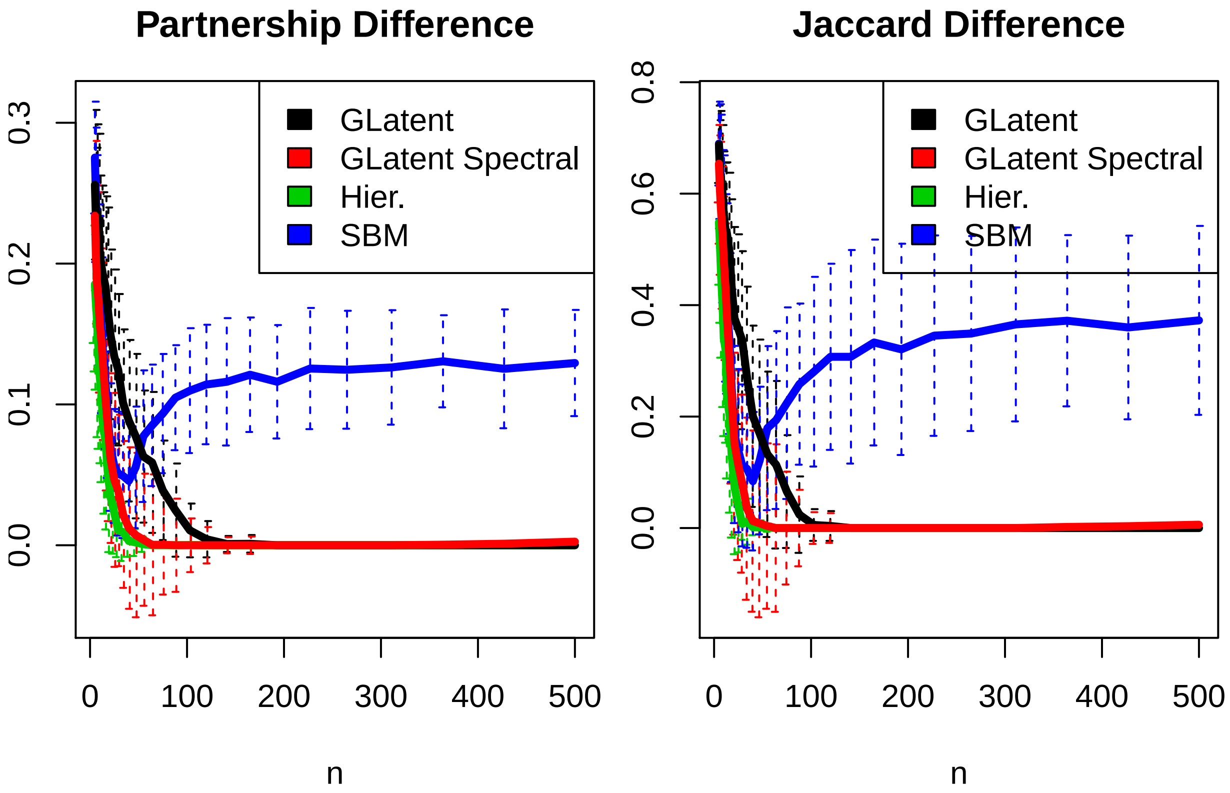
Model 3 (Blockwise: many clusters of different sizes) Correctly specified

We have $K = 6$, where the first cluster has 2 genes, the second cluster has 3 genes, and so on. There are a total of 27 genes this way.

Estimation of Latent Covariance Matrix



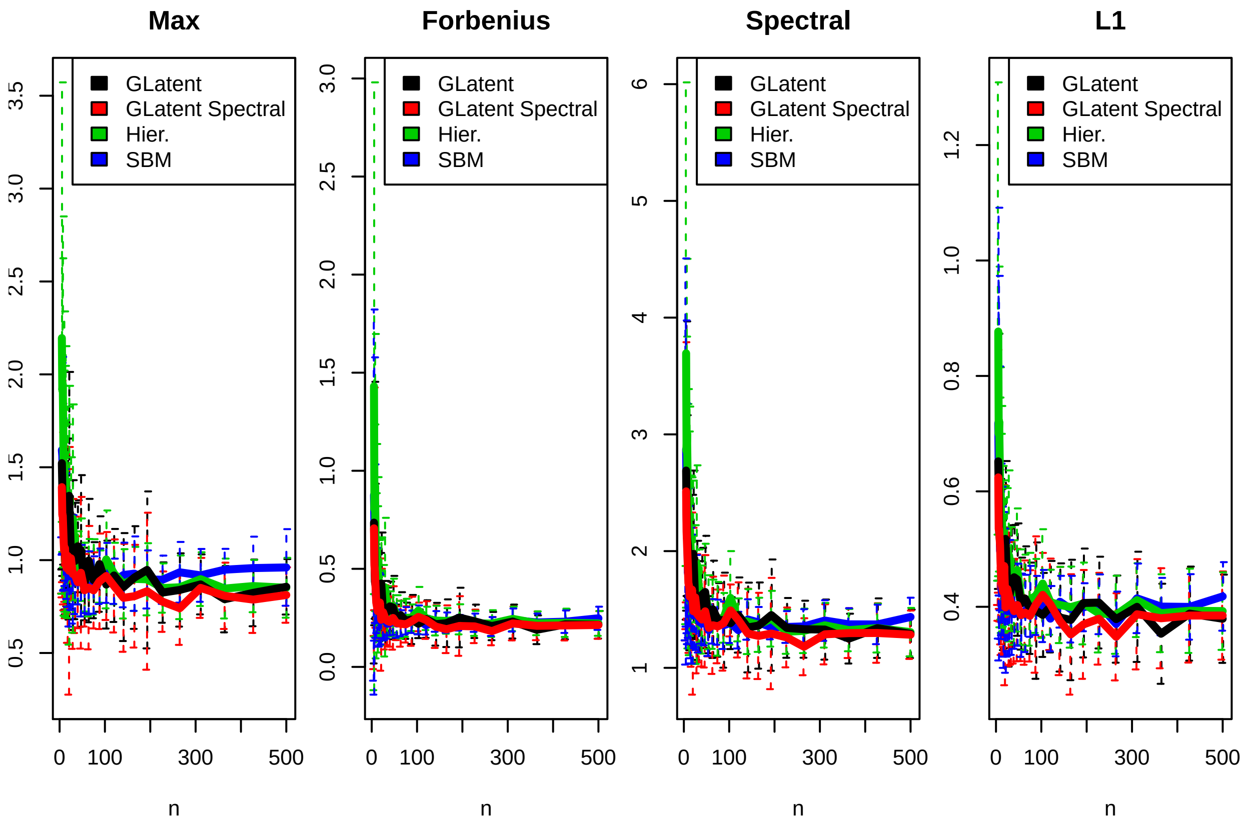
Estimation of the Clusterings



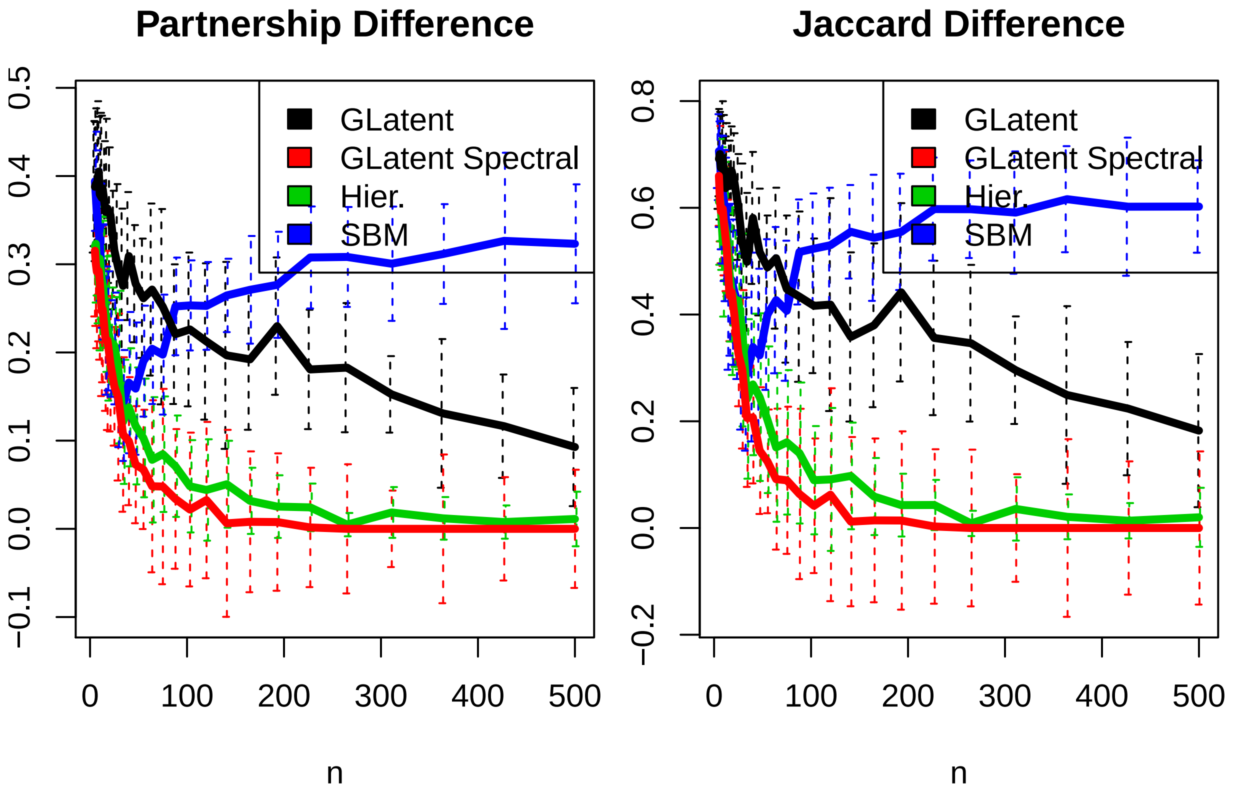
Model 4 (Mixed) Misspecified

We have $K = 4$, and each cluster has 6 genes. However, we construct the $\boldsymbol{A}$ matrix in the following way. First set $\boldsymbol{A} \in \mathbb{R}^{K \times d}$ to be a $\{0, 0.7\}$ matrix where each column only has one non-zero entry. Then, for each column, randomly add 0.3 to one of the K possible columns. Essentially, this means most genes are 70% in one cluster, 30% to be in another.

Estimation of Latent Covariance Matrix



Estimation of the Clusterings



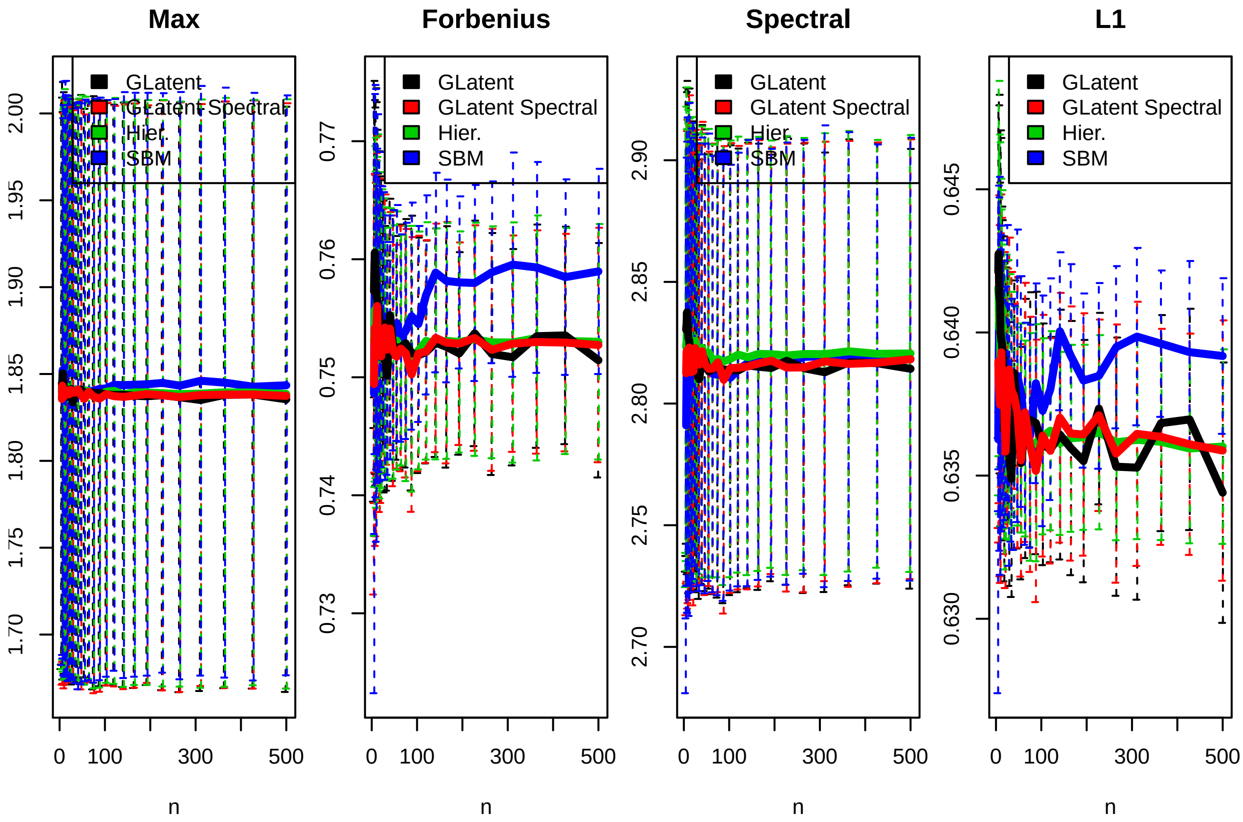
Model 5 (Monotone Transformation) Misspecified

We have $K = 4$, and each cluster has 6 genes. This is the same as Model 1, but instead of (1), we instead have

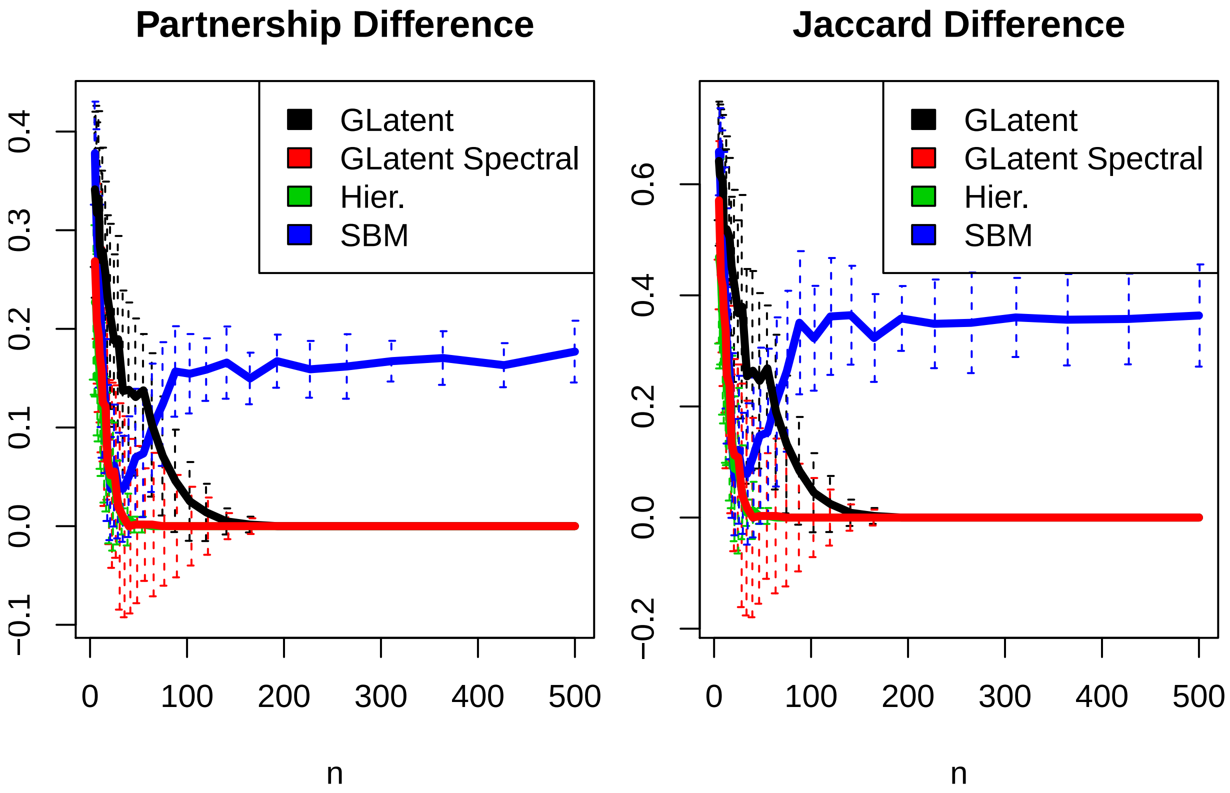
$$\boldsymbol{X} = \text{logistic}(\boldsymbol{Z}\boldsymbol{A}) + \varepsilon,$$

where logistic$(\cdot)$ is the logistic function that we apply entry-wise.

Estimation of Latent Covariance Matrix



Estimation of the Clusterings


Partnership Difference
Jaccard Difference
GLatent
GLatent Spectral
Hier.
SBM
−0.1
0.0
0.1
0.2
0.3
0.4
−0.2
0.0
0.2
0.4
0.6
0
100
200
300
400
500
n