dcsr-Omega

dscr-Omega (Dynamic Statistical Comparisons of estimation of precision matrix Ω)

- Motivation
- methods
- dscr simulation

Motivation

In last decade, the estimation of covariance matrix and its inverse become important in statistical analysis, for example, multiple comparison, principal component analysis, linear discriminant analysis and graphical model.

Motivation

Multiple comparison of gene expression level between two tissues.

$$\hat{\beta} = \bar{X}_1 - \bar{X}_2$$

There are correlation across genes.

$$\hat{\beta} \sim N(\beta, \Sigma_{P \times P})$$

where β is the true value and Σ is the covariance matrix under the Gaussian assumption. We need to estimate the $\Omega = \Sigma^{-1}$ to get the likelihood of β.

Motivation (Optional)

In Gaussian graphical model

$$(i,j) \in Edges \Leftrightarrow \Omega_{ij} \neq 0$$

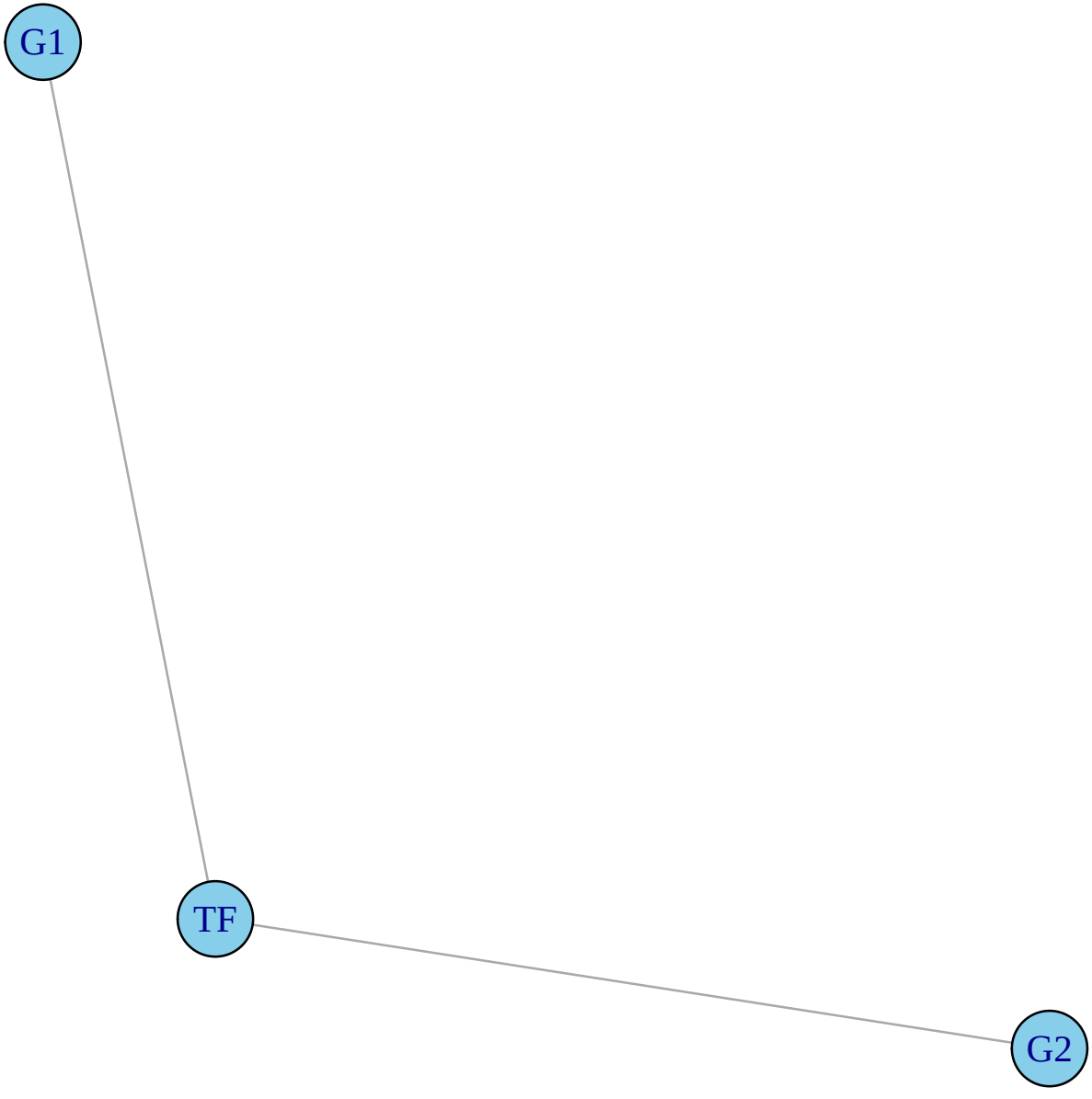
$$(i,j) \notin Edges \Leftrightarrow \Omega_{ij} = 0$$

Conditionally independent

$$i \perp j | all\ others \Leftrightarrow \Omega_{ij} = 0$$

$$i \perp j | all\ others \Leftrightarrow (i,j) \notin Edges$$

gene network



Difficulty (problems never sleep)

Sample covariance

$$\Sigma_n = \frac{1}{n} X^T X$$

When P (number of genes) larger than n (sample size)

- Σ_n is not stable
- Σ_n is not invertible
- number of parameters is larger than sample size

Reparameterization

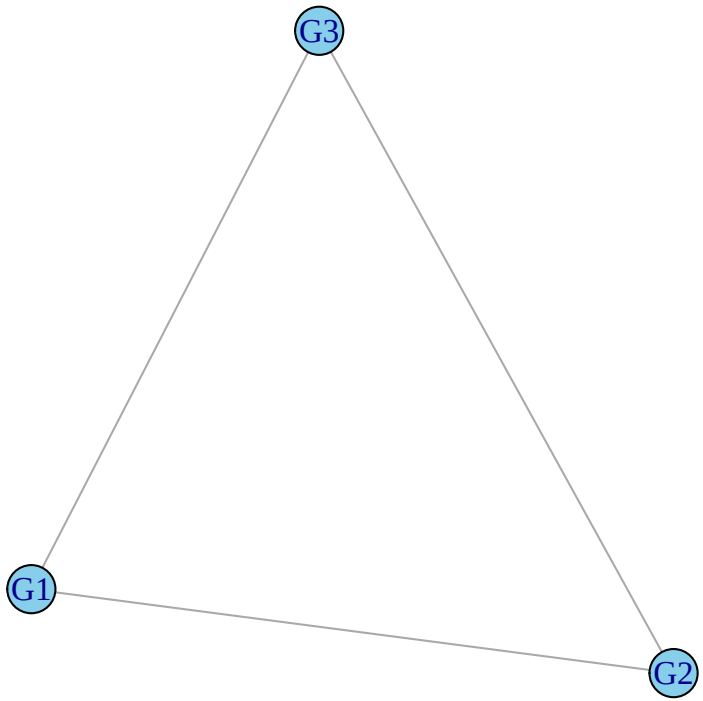
- sparse assumption on covariance matrix: Σ
- sparse assumption on precision matrix: Σ^{-1}
- diagonal matrix + low rank matrix

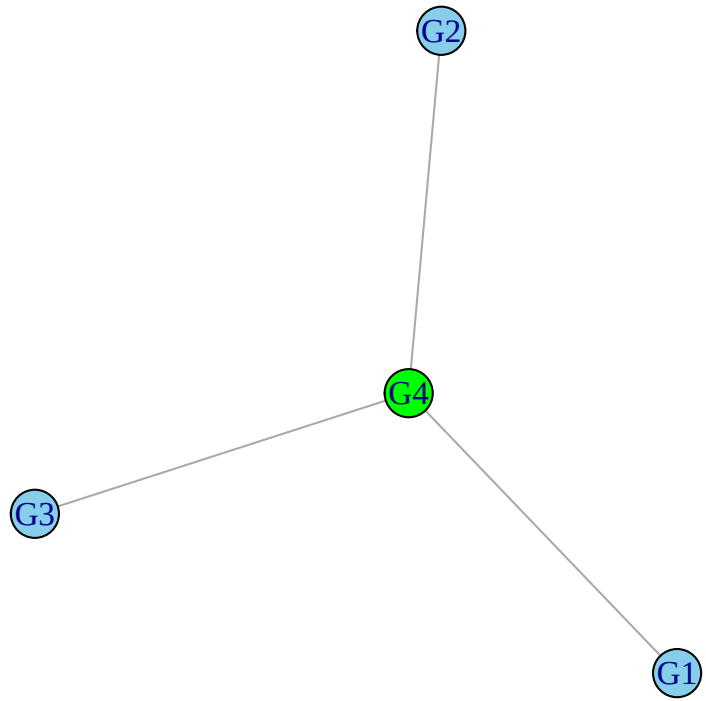
Sparse Estimation

- Glasso: Graphical Lasso
- Clime: Constrained L_1 Minimization
- Tiger: Tuning-Insensitive Approach for Optimally Estimating Gaussian Graphical Models
- PCS: Partial Correlation Screening

(problems couldn't fall asleep again)

- inverse of sparse matrix is no longer sparse
- sparsity changes after adding or deleting genes for precision matrix





Factor Model

$$X_{n\times P} = F_{n\times K} L_{K\times P} + E_{n\times P}$$

where X is gene expression level matrix, F is factor matrix, L is loading matrix and $E_{.i} \sim N(0, \Psi)$

For individual i:

$$X_{i1} = L_{11}F_{i1} + L_{21}F_{i2} + \cdots + L_{K1}F_{iK} + E_{i1}$$

$$X_{i2} = L_{12}F_{i1} + L_{22}F_{i2} + \cdots + L_{K2}F_{iK} + E_{i2}$$

$$X_{iP} = L_{1P}F_{i1} + L_{2P}F_{i2} + \cdots + L_{KP}F_{iK} + E_{iP}$$

F_{ik} stands for influence of factor k on individual i, $L_{k,j}$ stands for influence of factor k on gene j.

We divide the variation of gene expression into variation common genes and variation specific to gene.

Sparse Factor Model

- dense factor loading: batch effect, latent population structure
- sparse factor loading: transcript factors, biological covariates

Estimate Precision Matrix

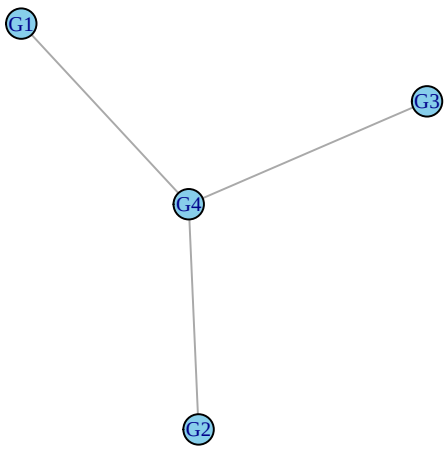
- covariance matrix
$$\hat{\Sigma} = \Psi + L^T \Lambda_F L$$
where $\Lambda_F = \frac{1}{n} F^T F$
- precision matrix
$$\hat{\Omega} = \hat{\Sigma}^{-1}$$
- Both $\hat{\Sigma}$ and $\hat{\Omega}$ are diagonal plus low rank.
- Inverse $\hat{\Sigma}$ is not computationally expensive.
- Model is statistically interpretable.

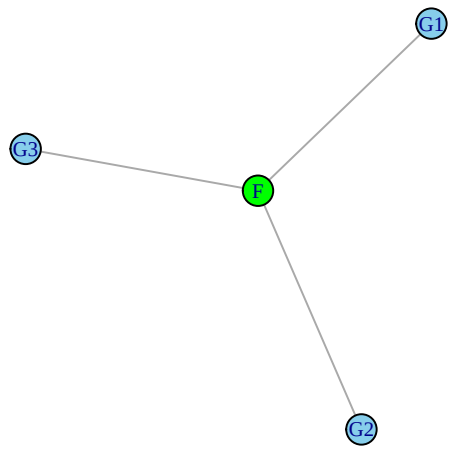
Interpretation

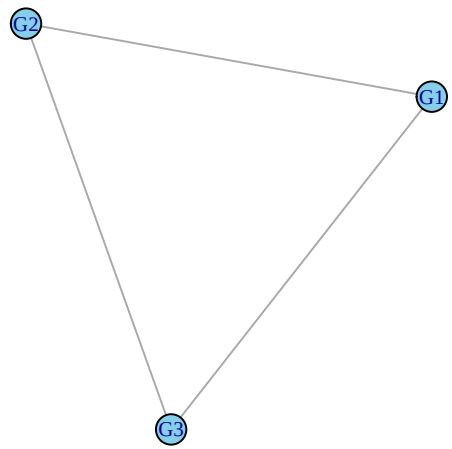
plot 1: sparse network

plot 2: factor model

plot 3: network of factor model in plot 2



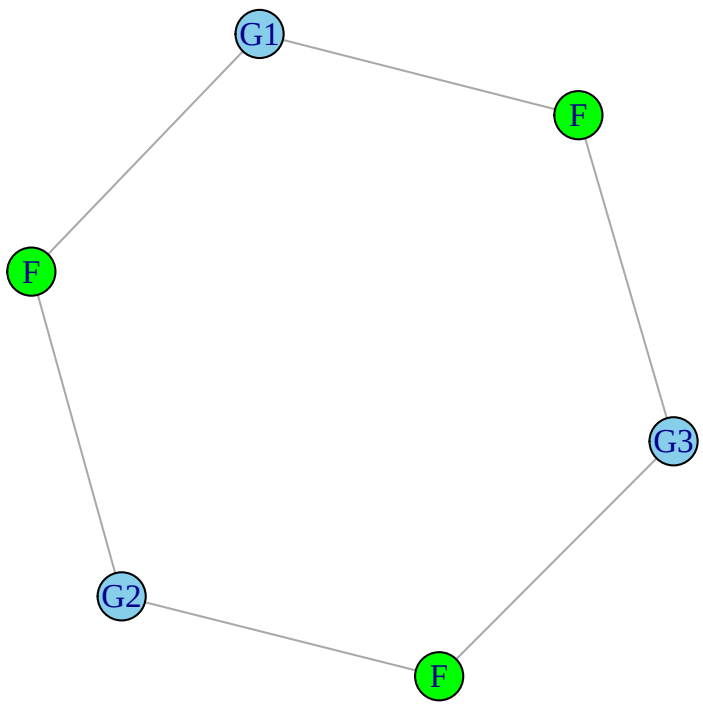


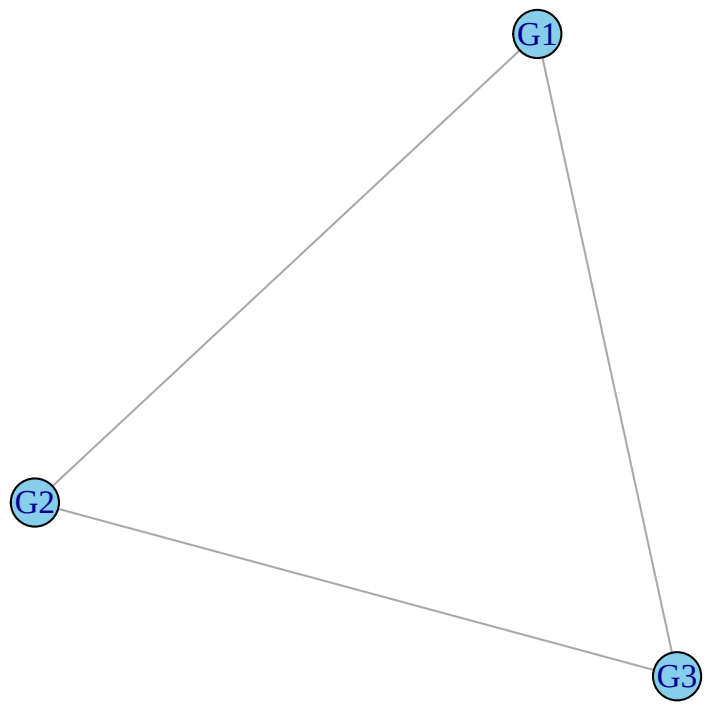


Interpretation

Sparse loadings:

```
##       [,1] [,2] [,3]
## [1,]    1    1    0
## [2,]    1    0    1
## [3,]    0    1    1
```





DSCR package (Dynamic Statistical Comparisons)

- Good artist copy, great artist steal. So, the following are from Matthew's github page for dscr

Dynamic statistical comparisons (DSC) are an attempt to change the way that researchers perform statistical comparisons of methods. When a new statistical method is developed, it is almost inevitable that it will be useful to compare it to other methods for tackling the same problem. However, the way these comparisons are currently (usually) done is suboptimal in so many ways. First, comparisons are usually performed by the research group that developed one of the methods, which almost inevitably favors that method. Furthermore, performing these kinds of comparisons is incredibly time-consuming, requiring careful familiarization with software implementing the methods, and the creation of pipelines and scripts for running and comparing them.

DSCR package (money in the bank)

And in fast-moving fields new methods or software updates appear so frequently that comparisons are out of date before they even appear. In summary, the current system results in a large amount of wasted effort, with multiple groups performing redundant and sub-optimal comparisons. A DSC is a public Internet repository that allows methods to be compared with one another in a reproducible and easily-extensible way. . . . We envisage DSC as complementing, rather than replacing, the standard publication model: papers introducing a new method will "deposit"" the method in the DSC repo, in the same way that scientific data are deposited in data repositories

DSCR

- scenarios: generate different scenarios for comparison
- data maker: take scenario names, provide input and meta
  generate data in different scenario, split the data into training set (input) and test set (meta)
- methods: take the input, provide output
  use the training set to estimate parameters
- score: take output and meta, provide scores as criterion for comparison
  use estimated parameters from training set and data from test set calculate score
- run dcs: addscenario(), addmethod(), addscore(), run_dsc()

No way to cheat

- test new method
- try new data

DSCR-Omega

In our application:

- scenarios: identity, diagonal, toeplitz, lung data in GTEX data
- methods: all mentioned above
- scores: likelihood, prediction error, measure of difference: $\Sigma\Omega - I$
  use $\hat{\Omega}$ learned from training set calculate the likelihood, prediction error and measure of difference: $\Sigma\Omega - I$ based on the data from test set

Futher work

- 80% genes but not all coregulate, it's hard to tell whether the loadings are sparse or dense.
  We need a more flexible shrinkage method to estimate the loadings.
- only a subgroup of observations or all of them have a certain latent factor.
  We also need a flexible shrinkage method on factors to capture the structure in samples.
- The solution of those questions above comes to the upcoming ASH-SFA.
- To be continued.