dscr-Omega

In many cases, we need estimation of precision matrix to get calculate the likelihood. Since in the real data, we always meet this situation that sample size is around few hundreds and much smaller than number of varibles which is around ten to hundred of thousands. We need assumptions to make constraint on the number of unknown parameters to estimate. There are two possible assumptions:

- sparse assumption on precision matrix: Ω
- diagonal matrix + low rank matrix

Sparse estimation:

- Glasso
$$min : logdet(\Omega) - trace(\Sigma_n \Omega) - \rho||\Omega||_1$$
and then convert to a lasso problem for each row
- Clime
$$min : ||\Omega||_1$$
subject to
$$||\Sigma_n \Omega - I||_\infty$$
also use sparse estimation of each row and combine them together
- PCS & Tiger
$$X_j = \sum_{i \neq j} \beta_i X_i + V_j$$
$$\beta_i = -\Omega_{ij}/\Omega_{jj};\ V_j \sim N(0, \frac{1}{\Omega_{jj}})$$
$i, j = 1 \cdots P$, X is the n by P data matrix. PCS screening these partial correlations, Tiger minimize the sum square of prediction error with sparsity constraint.

One of the limitation of sparse assumption of precision matrix is that the marginal precision matrix is not sparse when the joint precision matrix is sparse

$$(\Sigma_O)^{-1} = \Omega_O - \Omega_{OH}\Omega_H^{-1}\Omega_{HO}$$

where O stand for observed, H means hidden, and the subscript means the corresponding partition of the Ω matrix.

Instead of starting from the precision matrix itself with sparsity assumption, we can start from the data and try to learn the data well, and then turn to the precision matrix. We first put the data into a low-dimension space which hopefully could capture all the variation information in this low-dimensional space in order to estimate the covariance matrix and precision matrix. We use factor model which is easy to extend in dimension to model the expression level data.

$$X_{n \times P} = F_{n \times K} L_{K \times P} + E_{n \times P} \tag{1}$$

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

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.

The estimation of covariance matrix is

$$\hat{\Sigma} = \Psi + L^T \Lambda_F L \tag{2}$$

where $\Lambda_F = \frac{1}{n} F^T F$

For the precision matrix:

$$\begin{aligned} \hat{\Omega} &= \hat{\Sigma}^{-1} && (3) \\ &= \Psi^{-1} - \Psi^{-1} L^T (\Lambda_F + L\Psi^{-1}L^T)^{-1} L\Psi^{-1} && (4) \end{aligned}$$

- Both $\hat{\Sigma}$ and $\hat{\Omega}$ are diagonal plus low rank.
- Inverse $\hat{\Sigma}$ is not computationally expensive.
- Model is statistically interpretable.

Comparison criterion:

All the scores we use are not depend on the true value.

Likelihood:

$$logdet(\Omega) - trace(\Sigma_n \Omega) = -E_p log(p/q) + C$$

F-norm of error:

$$\frac{\partial}{\partial \Omega} E_p ||\nabla logp - \nabla logq||_2^2 = \frac{1}{2}\Sigma_n \Omega + \frac{1}{2}\Omega \Sigma_n - I$$

prediction error

$$\sum_j (X_j - (\sum_{i \neq j} -\Omega_{ij}/\Omega_{jj} X_i))^2$$

this is for one observation, we have n observation, so just add them up.