Answers – Lecture 6
Undirected network reconstruction - part 2

Question 1
Question 1a)
Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ for observations $i = 1, \ldots, n$. The following assumptions should be included: $\varepsilon \ \mathcal{N}(0, \sigma^2)$ and $\mathrm{Cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) = 0$ if $i_1 \neq i_2$. Hence, identically and independently distributed errors following a normal with zero mean.

Question 1b)
The lecture notes give the estimator for the regression coefficients: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. In this $\mathbf{Y}^{\mathrm{T}} = (2.4, 0.4, 3.2, 1.6)^{\mathrm{T}}$ and

$$\mathbf{X} \quad = \quad \begin{pmatrix} 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix},$$

where the first column represents the intercept (or, a variable that has been kept constant during the experiment). Work out the linear algebra and obtain: $\hat{\beta}_0 = 1.9$, $\hat{\beta}_1 = 0.5$, $\hat{\beta}_2 = -0.9$.

Rests to estimate the variance of the error. Its maximum likelihood estimator is (refer lecture notes):

$$\hat{\sigma}^2_{ML} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})^2.$$

The ML estimator then equals $\hat{\sigma}^2_{ML} = 0.01$, and $\hat{\sigma}_{ML} = 0.1$. The unbiased (least squares estimator) would replace the factor $1/n$ by $1/(n-p)$ where p is the number of estimated parameters, and amounts to $\hat{\sigma}^2 = 0.04$, and $\hat{\sigma} = 0.2$.

Question 2
Question 2a)
$\mathrm{Var}(Y_C) = \frac{9}{100}\mathrm{Var}(Y_A) + \frac{9}{100}\mathrm{Var}(Y_B) + \mathrm{Var}(\varepsilon_C) = \frac{1009}{10000}$.
$\mathrm{Cov}(Y_A, Y_C) = \frac{3}{10}\mathrm{Cov}(Y_A, Y_A) = \frac{3}{10}$.
$\mathrm{Cov}(Y_B, Y_C) = -\frac{3}{10}\mathrm{Cov}(Y_B, Y_B) = -\frac{3}{1000}$.
$\mathrm{Cor}(Y_A, Y_C) = \frac{3}{10}/[\frac{1009}{10000}]^{1/2} \approx 0.944$.
$\mathrm{Cor}(Y_B, Y_C) = -\frac{3}{1000}/[\frac{1}{100}\frac{1009}{10000}]^{1/2} \approx -0.0944$.
The difference between the two correlation coefficients is due to the difference in the variance of expression levels of gene A and B.

Question 2b)
Networks: (regression) $Y_A \to Y_C \leftarrow Y_B$ and (correlation) $Y_A \to Y_C$. The regression analysis takes into account the variance of Y_A and Y_B. Recall: t-statistic is given by $\hat{\beta}/\hat{\sigma}(\hat{\beta})$. Further, $\mathrm{Var}(\hat{\beta}) = \sigma^2[\mathbf{X}^T\mathbf{X}]^{-1}$, where $\mathbf{X}$ is the design matrix and σ^2 the residual variance. The design matrix comprises of a column of ones (corresponding to the intercept) and the data of Y_A and Y_B (centered around zero). In particular when Y_A and Y_B are centered around zero, the variances of Y_A and Y_B are proportional to diagonal elements of $\mathbf{X}^T\mathbf{X}$. Hence, they appear in the t-statistic.

Question 3
This is so-called interaction term. This means that the effect of one variable, say Y_B, depends on that of another variable, here Y_A.

Question 4
Question 4a
From the lecture notes we know $\mathrm{Var}(\mathbf{Z}) = \mathrm{Cov}(\mathbf{Z},\mathbf{Z}) = \mathrm{Cov}(\mathbf{AY},\mathbf{AY}) = \mathbf{A}\mathrm{Cov}(\mathbf{Y},\mathbf{Y})\mathbf{A}^{\mathrm{T}} = \mathbf{A}\mathrm{Var}(\mathbf{Y})\mathbf{A}^T$. As $\mathrm{Var}(\mathbf{Y}) = \mathbf{I}$, we have $\mathrm{Var}(\mathbf{Z}) = \mathbf{A}\mathbf{A}^T$. Thus:

$$\mathrm{Var}(\mathbf{Z}) = \begin{pmatrix} 13 & -3 & -8 \\ -3 & 2 & 3 \\ -8 & 3 & 6 \end{pmatrix}.$$

Question 4b
Using the same reasoning as for Question 4a (now replacing $\mathrm{Var}(\mathbf{Y})$ by $\mathbf{\Sigma}$):

$$\mathrm{Var}(\mathbf{Z}) = \begin{pmatrix} 14 & -2 & -7 \\ -2 & 4 & 4 \\ -7 & 4 & 6 \end{pmatrix}.$$

Question 5

Question 5a
The exercise assumes that the joint distribution of $\mathbf{Y}$ and $\mathbf{X}$ is a multivariate normal:

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{XX} & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_{YY} \end{pmatrix} \right)$$

Then, in the lecture notes (section on the multivariate normal) the conditional distribution of $\mathbf{Y}|\mathbf{X}$ is given:

$$\mathbf{Y}|\mathbf{X} = \mathcal{N}(\boldsymbol{\mu}_Y + \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}(\mathbf{X} - \boldsymbol{\mu}_X), \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}).$$

The relevant submatrices are:

$$\mathbf{\Sigma}_{YY} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix}, \quad \mathbf{\Sigma}_{XX} = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}, \quad \mathbf{\Sigma}_{YX} = \mathbf{\Sigma}_{XY}^T = \begin{pmatrix} 1 & -1 \\ 0 & -1 \\ 0 & 1 \end{pmatrix}.$$

Substitute this in the equation above:

$$\mathbf{Y}|\mathbf{X} \;=\; \mathcal{N}\left(\frac{1}{6}\begin{pmatrix} 1 & -1 \\ -1 & -2 \\ 1 & 2 \end{pmatrix}\mathbf{X}, \frac{1}{6}\begin{pmatrix} 10 & -1 & 1 \\ -1 & 16 & 8 \\ 1 & 8 & 16 \end{pmatrix}\right).$$

Question 5b
Along similar lines as Question 5a, we obtain:

$$\mathbf{X}|\mathbf{Y} \;=\; \mathcal{N}\left(\frac{1}{2}\begin{pmatrix} 1 & 0 & 0 \\ -1 & -1 & 1 \end{pmatrix}\mathbf{Y}, \frac{1}{2}\begin{pmatrix} 7 & -3 \\ -3 & 5 \end{pmatrix}\right).$$

Question 6
Question 6a)
The design matrix $\mathbf{X}$ will be of rank lower than the number of variables, and so will $\mathbf{X}^T\mathbf{X}$ be (this follows from the identity $\text{rank}(\mathbf{X}^T\mathbf{X}) = \text{rank}(\mathbf{X}\mathbf{X}^T) = \text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^T)$ which only requires that $\mathbf{X}$ be a real matrix). Consequently, $\mathbf{X}^T\mathbf{X}$ is singular and its inverse does not exist. Try for a simple choice of X_1 and X_2.

Question 6b)
The design matrix $\mathbf{X}$ will be of rank lower than the number of variables, and so will $\mathbf{X}^T\mathbf{X}$ be. Consequently, $\mathbf{X}^T\mathbf{X}$ is singular and its inverse does not exist. Try for a small p and n.

Question 7
A possible choice for the design matrix could be:

$$\mathbf{X} \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Question 8
R-code to inserted.