Answers – Lecture 7
Undirected network reconstruction - part 3

Question 1
The set of edges of the conditional independence graph is given by: $\mathcal{E} = \{(1,4),(2,3),(2,4),(2,5),(4,5)\}$.

Question 2
One of the corollaries of the Inverse Variance Lemma states that the partial correlations can be obtained from the inverse of the covariance matrix by e.g.:

$$\rho(Y_1, Y_2 \,|\, Y_3) = \frac{-(\mathbf{\Sigma}^{-1})_{1,2}}{\sqrt{(\mathbf{\Sigma}^{-1})_{1,1}}\sqrt{(\mathbf{\Sigma}^{-1})_{2,2}}}.$$

First calculate the inverse covariance matrix is:

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} 0.3125 & -0.1250 & 0.0000 \\ -0.1250 & 0.5833 & 0.3333 \\ 0.0000 & 0.3333 & 0.3333 \end{pmatrix}.$$

One can now substitute the right elements in the expression for the partial correlation. Alternatively, one can standardize the inverse covariance matrix such that is has a unit diagonal. Hereto pre- and post-multiply it by a diagonal matrix with elements equal to reciprocal of the square root of diagonal elements of $\mathbf{\Sigma}^{-1}$: The partial correlation matrix thus equals:

$$\begin{pmatrix} (0.3125)^{-1/2} & 0.0000 & 0.0000 \\ 0.0000 & (0.5833)^{-1/2} & 0.0000 \\ 0.0000 & 0.0000 & (0.3333)^{-1/2} \end{pmatrix}.$$

After the standardization, only the off-diagonal elements need to be multiplied by -1. This eventually gives the partial correlation matrix:

$$\begin{pmatrix} 1.00000 & 0.2927700 & 0.0000000 \\ 0.29277 & 1.0000000 & -0.7559289 \\ 0.00000 & -0.7559289 & 1.0000000 \end{pmatrix}.$$

The graph with (in)dependencies can be read off this matrix. If it contains a zero, the multivariate normal distribution factorizes with respect to these variates (corresponding to the zero). The factorization means conditional independence. The graph thus has edge set: $\mathcal{E} = \{(1,2),(2,3)\}$.

Note: the standardization does not affect the presence of a zero. Nor does the multiplication by -1. Hence, the conditional independencies could have been concluded from the inverse covariance matrix.

Question 3

Question 3a)

Specify the parameters of the trivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the regression equations relating the genes. Directly given as $(\boldsymbol{\mu})_{1,1} = E(Y_1) = 1$ and $(\boldsymbol{\Sigma})_{1,1} = \mathrm{Var}(Y_1) = 4$. Consider $(\boldsymbol{\Sigma})_{1,2} = \mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Y_1, \frac{1}{2}Y_1 - \frac{7}{2} + \varepsilon_2) = \mathrm{Cov}(Y_1, \frac{1}{2}Y_1) + \mathrm{Cov}(Y_1, -\frac{7}{2}) + \mathrm{Cov}(Y_1, \varepsilon_2) = \frac{1}{2}\mathrm{Cov}(Y_1, Y_1) = 2$. Similarly, e.g. $(\boldsymbol{\Sigma})_{2,2} = \mathrm{Cov}(Y_2, Y_2) = \mathrm{Cov}(\frac{1}{2}Y_1 - \frac{7}{2} + \varepsilon_2, \frac{1}{2}Y_1 - \frac{7}{2} + \varepsilon_2) = \ldots$.

Alternatively, the theorem of Koller and Friedman (lecture notes, section on multivariate normal distribution), specifies the distribution of a random variable Y defined as $Y = \beta_0 + \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\varepsilon}$, with $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$. This theorem thus enables us to calculate the (unconditional) mean and variance of Y_2, as well as $\mathrm{Cov}(Y_1, Y_2)$. And consequently, that of Y_3. An example is detailed in the lecture notes. Here the resulting multivariate normal distribution equals:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix}, \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{pmatrix}\right)$$

Question 3b)

See question 2.

Question 3c)

In the lecture notes (Section Regression vs. partial correlation), the relation between the partial correlations and the regression coefficients is detailed. For instance, in the regression equation $Y_1 \,|\, Y_2, Y_3 = \beta_{12}Y_2 + \beta_{13}Y_3 + \varepsilon_1$, we wish to specify β_{12} and β_{13}. Then: $\beta_{12} = \rho_{12.3}\sqrt{\sigma^{(22)}/\sigma^{(11)}}$, where $\sigma^{(22)} = (\Sigma^{-1})_{22}$. Thus: $\beta_{21} = 0.2927700\sqrt{0.5833/0.3125} \approx 0.40$. Doing this for all βs, we obtain the following system of regression equations:

$$\begin{aligned} Y_1 &= 0.4Y_2 + \varepsilon_1 \\ Y_2 &= 0.2142857Y_1 - 0.5714286Y_3 + \varepsilon_2 \\ Y_3 &= -Y_2 + \varepsilon_3 \end{aligned}$$

Question 4

Question 4a

The set of edges of the conditional independence graph is given by: $\mathcal{E} = \{(1,2), (1,3), (2,3)\}$. This is a fully connected graph.

Question 4b

From the regression equations we define:

$$\mathbf{B} = \begin{pmatrix} 0.000000 & -0.2407139 & 0.7232975 \\ -1.175004 & 0.0000000 & 1.0925333 \\ 1.075472 & 0.3327965 & 0.0000000 \end{pmatrix}.$$

When $\mathbf{Y} = (Y_1, Y_2, Y_3)^T$, the (estimated) system of regression equations can than be written as: $\mathbf{Y} = \mathbf{BY} + \boldsymbol{\varepsilon}$. The partial correlations are then given by $\text{sign}(\beta_{j_1,j_2})\sqrt{\beta_{j_1,j_2}\beta_{j_2,j_1}}$. The partial correlation matrix thus equals:

$$\mathbf{R} = \begin{pmatrix} 1.00000000 & -0.5318268 & 0.8819786 \\ -0.5318268 & 1.0000000 & 0.6029853 \\ 0.6029853 & 0.6029853 & 1.00000000 \end{pmatrix}.$$

Question 4c
H_0: $\rho = 0$. Under H_0, the sample partial correlation is then distributed as: $\mathcal{N}(0, 1/(n-2-k))$, where k is the number of variables conditioned on. Here $k = 2$. The standard deviation is thus (approximately) equal to 0.1020621. For the normal distribution, it is well known that 95% of the data falls into the interval two standard deviations in either direction from the mean. The partial correlation that statistically significant deviate from zero are:

$$\begin{pmatrix} NA & < 0.05 & < 0.05 \\ < 0.05 & NA & < 0.05 \\ < 0.05 & < 0.05 & NA \end{pmatrix}.$$

Hence, all partial correlation differ significantly from zero.

Question 4d
From the residual errors from regression equations are exactly the (square root of the) inverse variances. Define the diagonal matrix $\mathbf{U}$ with : $\text{diag}(\mathbf{U}) = (1/0.6742129, 1/1.4895880, 1/0.8221256)$. The (estimate of the) covariance matrix is now given by:

$$(\mathbf{URU})^{-1} = \begin{pmatrix} 2.0465 & 0.0000 & 2.2010 \\ 0.0000 & 3.4866 & 1.1603 \\ 2.2010 & 1.1603 & 3.4291 \end{pmatrix}.$$

Question 5
Question 5a
First find the joint distribution of Y_1 and Y_2. From the exercise we know $(Y_1, Y_2)^T = \mathbf{B}\,(Y_1, Y_2)^T + (\varepsilon_1, \varepsilon_2)$, where:

$$\mathbf{B} = \begin{pmatrix} 0.0 & -1.0 \\ -0.5 & 0.0 \end{pmatrix}.$$

The (partial) correlation is then given by $\text{sign}(\beta_{1,2})\sqrt{\beta_{1,2}\beta_{2,1}} = -\frac{1}{2}\sqrt{2}$. On the other hand, the conditional variances are exactly the inverse variances. Define the diagonal matrix $\mathbf{U}$ with : $\text{diag}(\mathbf{U}) = (\frac{1}{2}\sqrt{2}, \frac{1}{2})$. The (estimate of the) covariance matrix is now given by:

$$(\mathbf{URU})^{-1} = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

So far, we thus have:

$$\text{Cov}\left(\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}\right) = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

It rests to apply the theorem of Koller and Friedman (lecture notes, section on multivariate normal distribution) to find the marginal mean and variance of Y_3 and its covariance with Y_1 and Y_2.

$$\text{Cov}\left(\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}\right) = \begin{pmatrix} 1 & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{3}{2} \end{pmatrix}.$$

Question 5b
The inverse covariance matrix is given by:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 3 & 3 & -1 \\ 3 & 5 & -1 \\ -1 & -1 & 1 \end{pmatrix}.$$

There are no zero-elements: hence, none of the partial correlation is zero.

Question 6
In the lecture notes an explicit expression for the inverse of a 2×2 block matrix is given in terms of elements of the original 2×2 block matrix. From this we obtain:

$$\boldsymbol{\Omega}_{12} = -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$$

Now clearly, if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, then $\boldsymbol{\Omega}_{12} = \mathbf{0}$.

Question 7
Question 7a
The set of edges of the conditional independence graph is given by: $\mathcal{E} = \{(1,2),(2,3),(2,4)\}$. This is a so-called star graph.

Question 7b
From the regression equations we define:

$$\mathbf{B} = \begin{pmatrix} 0.00000 & -0.48543 & 0.03006 & -0.03074 \\ -0.34517 & 0.00000 & -0.65077 & -0.36067 \\ 0.02160 & -0.65772 & 0.00000 & -0.02214 \\ -0.01594 & -0.26307 & -0.01598 & 0.00000 \end{pmatrix}.$$

When $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)^T$, the (estimated) system of regression equations can than be written as: $\mathbf{Y} = \mathbf{B}\mathbf{Y} + \boldsymbol{\varepsilon}$. The partial correlations are then given by $\text{sign}(\beta_{j_1,j_2})\sqrt{\beta_{j_1,j_2}\beta_{j_2,j_1}}$. The partial correlation matrix thus equals:

$$\mathbf{R} = \begin{pmatrix} 1.00000000 & -0.4093357 & 0.02547856 & -0.02213551 \\ -0.40933566 & 1.0000000 & -0.65423406 & -0.30803198 \\ 0.02547856 & -0.6542341 & 1.00000000 & -0.01880467 \\ -0.02213551 & -0.3080320 & -0.01880467 & 1.00000000 \end{pmatrix}.$$

Question 7c
H_0: $\rho = 0$. Under H_0, the sample partial correlation is then distributed as: $\mathcal{N}(0, 1/(n-2-k))$, where k is the number of variables conditioned on. Here $k = 2$. The standard deviation is thus (approximately) equal to 0.03168621. For the normal distribution, it is well known that 95% of the data falls into the interval two standard deviations in either direction from the mean. The partial correlation that statistically significant deviate from zero are:

$$\begin{pmatrix} NA & < 0.05 & > 0.05 & > 0.05 \\ < 0.05 & NA & < 0.05 & < 0.05 \\ > 0.05 & < 0.05 & NA & > 0.05 \\ > 0.05 & < 0.05 & > 0.05 & NA \end{pmatrix}.$$

Hence, only partial correlations between Y_2 and some other gene, conditioned on the remaining genes differ significantly from zero.

Question 7d
From the residual errors from regression equations are exactly the (square root of the) inverse variances. Define the diagonal matrix $\mathbf{U}$ with : $\text{diag}(\mathbf{U}) = (1/0.7074, 1/0.5965, 1/0.5995, 1/0.5095)$. The (estimate of the) covariance matrix is now given by:

$$(\mathbf{URU})^{-1} = \begin{pmatrix} 0.7869958 & -0.5743491 & 0.3918300 & 0.1322914 \\ -0.5743491 & 1.1536689 & -0.7649498 & -0.2821249 \\ 0.3918300 & -0.7649498 & 0.8672393 & 0.1811383 \\ 0.1322914 & -0.2821249 & 0.1811383 & 0.3287839 \end{pmatrix}.$$