Exercises – Lecture 6
Undirected network reconstruction - part 2

Question 1 *(without R)*
A small experiment has been conducted to assess the effects of (the expression levels of) genes 1 and 2, denoted by X_1 and X_2, on (the expression levels of) a third gene, denoted Y. The experimental design (the settings of X_1 and X_2) and the outcome (Y) are given in the table below. In the experiment the actual expression levels of gene 1 and 2 are not actually measured. They are either knocked-out or enhanced. Knock-out, encoded by a `-1` in the table below, ensure no mRNA can be formed. Enhanced, encoded by a `1` in the table below, means that the transcription of the gene has been stimulated. The expression levels of Y have measured by means of PCR and are given as such below.

Observation	X_1	X_2	Y
1	-1	-1	2.4
2	-1	1	0.4
3	1	-1	3.2
4	1	1	1.6

a) Give the linear regression model (with intercept) that relates X_1 and X_2 to Y, including the assumptions accompanying the model.

b) Estimate all (!) parameters of the regression model from the experimental data given in the table above.

Question 2 *(without R)*
Consider a three-gene pathway. The expression levels of genes A and B, denoted Y_A and Y_B, are distributed as $\mathcal{N}(0, 1)$ and $\mathcal{N}(2, \frac{1}{100})$. Gene C is regulated by genes A and B. Their relation is given by: $Y_C = \frac{1}{2} + \frac{3}{10} Y_A - \frac{3}{10} Y_B + \varepsilon_C$, with $\varepsilon_C \sim \mathcal{N}(0, \frac{1}{100})$. Furthermore, assume Y_A, Y_B, and ε_C are independent.

a) Calculate the correlation between the (expression levels of) gene C and that of both its regulators. Can you explain the difference between the two correlations (in relation to their regression coefficients, which both equal $\frac{3}{10}$)?

b) Draw a (marginal) independence graph relating the three genes in accordance with the correlation analysis. Hereto apply a cut-off of 0.3: only correlations larger than 0.3 in an absolute sense are considered to truly represent an association and yield an edge.

Data on the expression levels of the three genes has been obtained from a thousand cells. Below you find the R results of regression the expression levels of gene C on that of the other two genes.

```
Coefficients :
                    Estimate   Std.Error   t − value    Pr(> |t|)
(Intercept)         0.440580    0.063999       6.884   1.03e − 11   * * *
Y_A                 0.300915    0.003122      96.374   < 2e − 16    * * *
Y_B                −0.271230    0.032010      −8.473   < 2e − 16    * * *

---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error:  0.09929 on 997 degrees of freedom
Multiple R-squared:  0.9033, Adjusted R-squared:  0.9031
F-statistic:  4654 on 2 and 997 DF, p-value:  < 2.2e-16
```

c) Draw the conditional independence graph relating the three genes in accordance with the regression analysis above.

d) Explain the differences between the two graphs obtained under parts *b)* and *c)*.

Question 3 *(without R)*
Consider a three-gene pathway. The expression levels of genes A and B, denoted Y_A and Y_B, are independent and normally distributed. Gene C is regulated by genes A and B. Their relation is given by:

$$Y_C \quad = \quad 2 - \frac{3}{2} Y_B - \frac{1}{2} Y_A Y_B + \varepsilon_C, \tag{1}$$

with $\varepsilon_C \sim \mathcal{N}(0, 1)$. Furthermore, assume Y_A, Y_B, and ε_B are independent. How would you interpret the third term ($-\frac{1}{2} Y_A Y_B$) on the right-hand side of the equality sign in Formula (1)?

Question 4 *(without R)*
In the models considered, the expression levels of one gene are affected by that of other genes in a linear fashion. That is, a gene's expression levels are a linear combination of expression levels of other genes (plus some noise).

a) Assume a standard trivariate normal random variable $\mathbf{Y}$. Define the matrix $\mathbf{A}$:

$$\mathbf{A} \quad = \quad \begin{pmatrix} 0 & -2 & -3 \\ -1 & 0 & 1 \\ -1 & 1 & 2 \end{pmatrix}.$$

Determine the covariance matrix of the new trivariate random variable $\mathbf{Z} = \mathbf{A}\mathbf{Y}$, a linear transformation of $\mathbf{Y}$.

b) Now assume the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{Y}$ is:

$$\mathbf{\Sigma} \quad = \quad \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

Determine the covariance matrix of the new trivariate random variable $\mathbf{Z} = \mathbf{A}\mathbf{Y}$, a linear transformation of $\mathbf{Y}$.

Question 5 *(without R)*
The random variable $\mathbf{Z} = (\mathbf{Y}^T, \mathbf{X}^T)^T$, with $\mathbf{Y} = (Y_1, Y_2, Y_3)^T$ and $\mathbf{X} = (X_1, X_2)^T$, is normally distributed with a zero mean and a covariance matrix equal to:

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & 0 & 0 & 1 & -1 \\ 0 & 3 & 1 & 0 & -1 \\ 0 & 1 & 3 & 0 & 1 \\ 1 & 0 & 0 & 4 & -2 \\ -1 & -1 & 1 & -2 & 4 \end{pmatrix}$$

a) Give the distribution of $\mathbf{Y}$ conditional on $\mathbf{X}$.
b) Give the distribution of $\mathbf{X}$ conditional on $\mathbf{Y}$.

Question 6* *(without R)*

a) An investigator wants to use multiple regression to predict a variable, Y, from two other variables, X_1 and X_2. He/She proposes forming a new variable $X_3 = X_1 + X_2$ and using multiple regression to predict Y from the three X variables. Show that he/she will run into problems when applying multiple regression. *Hint:* recall $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, and try to calculate this.

b) In the traditional regression setting the number of observations n on the response and the explanatory variables exceeds the number of explanatory variables p. Show now that, if $p > n$, you run into problems when trying to estimate the regression parameters. *Hint:* recall $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, and try to calculate this.

Question 7 *(without R)*
A medical researcher wishes to explain the differences of expression levels of gene A between three cell types by means of a linear regression model. The expression levels of gene A have been measured twice for each cell type. Give a parametrization of the design matrix $\mathbf{X}$ that he/she might want to use.

Question 8 *(with R)*
Consider a three-gene pathway. Data has been obtained on the expression levels of all three genes in 31 samples. Develop a model relating the expression levels of gene C on one hand to that of its suspected regulators (gene A and B) on the other. Among others, generate scatterplots, fit regression models, and investigate the residuals.

	gene A	gene B	gene C
1	3.18	8.54	3.20
2	3.43	7.70	3.39
3	3.29	7.53	3.15
4	5.77	8.38	5.98
5	3.96	7.37	3.49
6	4.11	9.39	3.94
7	2.51	5.51	2.46
8	2.44	5.85	2.46
9	4.35	8.04	4.56
10	3.11	7.03	2.99
11	3.64	7.53	3.22
12	3.96	9.22	4.50
13	4.42	8.38	4.31
14	4.53	7.37	4.71
15	5.06	7.03	4.88
16	5.17	7.87	5.26
17	4.25	6.86	3.79
18	3.43	7.70	3.42
19	3.54	6.52	3.42
20	5.53	8.54	6.36
21	6.70	9.56	8.23
22	3.22	8.88	3.28
23	5.74	8.38	6.61
24	3.29	6.02	2.92
25	4.28	5.68	3.73
26	2.33	6.69	2.46
27	3.33	8.38	3.53
28	3.36	7.53	3.29
29	5.77	8.38	6.03
30	5.60	8.71	6.39
31	3.40	8.21	3.66