Exam Statistical Genomics 2017/2018

Date: Friday, April 6, 2018
Time: 9:00 - 12:00
Place: BB 5161.0293
Progress code: WISG-09

Rules to follow:

- This is a closed book exam. Consultation of books and notes is not permitted.
- Do not forget to write your name and student number onto each paper sheet.
- There are 4 exercises, and the numbers of points per exercise are indicated within boxes. You can reach $p = 90$ points and the exam grade will be computed as follows:

$$\text{grade} := \frac{10 + p}{10}$$

- I wish you success with the completion of the exam!

EXAM STARTS ON NEXT PAGE

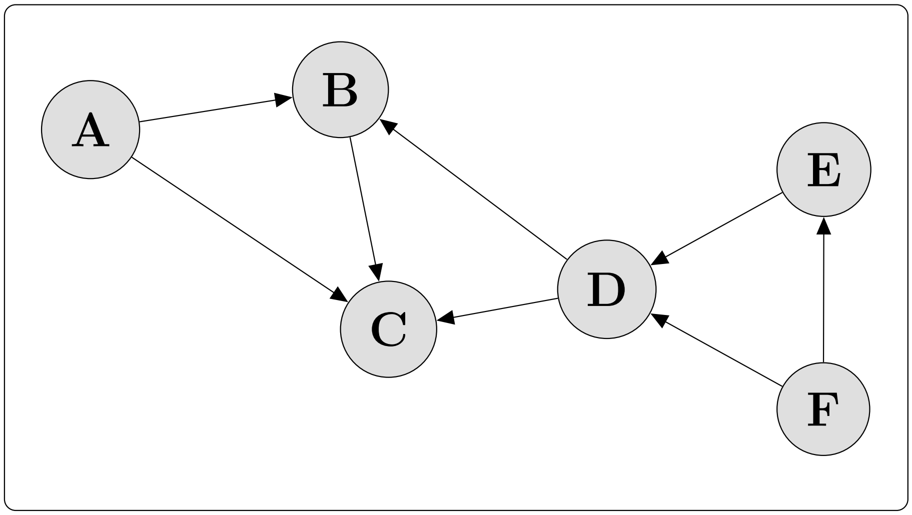

Figure 1: **The DAG for exercise no. 1.**

1. **Bayesian networks and directed acyclic graphs.** 30
Consider the directed acyclic graph (DAG) shown in Figure 1.

   (a) 3 Give the ancestor matrix of the graph.

   (b) 3 How many neighbour graphs can be reached by the 3 single edge operations?

   (c) 3 Give the CPDAG of the DAG.

   (d) 3 How many graphs are in the equivalence class defined by the CPDAG.

   (e) 3 Is there a DAG with the same skeleton but without any v-structures? If so, give an example. If not, give an explanation why that is impossible.

   (f) 3 Give a DAG with the same skeleton but in whose CPDAG the edge $F \to E$ is directed (compelled).

   (g) 3 Give the Markov Blanket of node D.

   (h) 3 List all paths (trails) from node A to node E, and indicate for each path whether it is open or blocked.

   (i) 3 List all open paths from node A to node E when conditional on $Z = \{C\}$.

   (j) 3 In Bayesian networks the joint distribution can be factorized into a product of local conditional distributions. Use this factorization to show that $P(A, B, C|D, E, F) = P(A, B, C|D)$. You can assume that all nodes are discrete binary variables.

2. **Structure MCMC sampling.** $\boxed{25}$
Consider a Bayesian network with $n = 2$ nodes X_1 and X_2. There are then three possible directed acyclic graphs (DAGs): $\mathbf{G_1}$: '$X_1 \rightarrow X_2$', $\mathbf{G_2}$: '$X_1 \leftarrow X_2$', and the empty graph without edges $\mathbf{G_3}$: '$X_1 \quad X_2$'. The structure MCMC sampling scheme defines a Markov Chain whose state space S is the set of those three DAGs: $S = \{\mathbf{G_1}, \mathbf{G_2}, \mathbf{G_3}\}$ The graph prior distribution is: $P(\mathbf{G_1}) = 0.4$, $P(\mathbf{G_2}) = 0.2$, and $P(\mathbf{G_3}) = 0.4$. The marginal likelihoods are: $P(data|\mathbf{G_1}) = 20a$, $P(data|\mathbf{G_2}) = 20a$ and $P(data|\mathbf{G_3}) = a$, where $a \in \mathbb{R}^+$.

   (a) $\boxed{10}$ Compute the 3-by-3 transition matrix T of the Markov Chain when only single edge additions and deletions are implemented (no single edge reversals).

   (b) $\boxed{10}$ Compute the 3-by-3 transition matrix T of the Markov Chain when all three single edge operations (additions, deletions and reversals) are used.

   (c) $\boxed{5}$ Give the stationary distribution(s) of the two Markov chains in (a) and (b).

3. **Gaussian Bayesian networks.** $\boxed{20}$
Consider three random variables X_1, X_2, and X_3, which are in the following regression relationships to each other:

$$\begin{aligned} X_1 &= 2 + \epsilon_1 \\ X_2 &= (-1) \cdot X_1 + \epsilon_2 \\ X_3 &= 2 \cdot X_2 + \epsilon_3 \end{aligned}$$

where ϵ_1, ϵ_2, and ϵ_3 are independently standard Gaussian $N(0,1)$ distributed random variables. This can be interpreted as a Gaussian Bayesian network with the directed acyclic graph: '$X_1 \rightarrow X_2 \rightarrow X_3$'.

   (a) The 3-dimensional random vector $\mathbf{X} := (X_1, X_2, X_3)^T$ is multivariate Gaussian distributed. Give its expectation vector and its covariance matrix. $\boxed{10}$

   (b) The graph is equivalent to the graph '$X_1 \leftarrow X_2 \leftarrow X_3$'. Give the regression equations for the latter graph. $\boxed{10}$

<u>HINTS</u>: For part (a), recall that for random variables X, Y and Z:

- $Cov(X,Y) = Cov(Y,X)$
- $Cov(X,X) = Var(X)$
- $Cov(X, Y+Z) = Cov(X,Y) + Cov(X,Z)$
- $Cov(c, X) = 0$ for $c \in \mathbb{R}$

For part (b), recall that for a vector $(X_1, \ldots, X_n)^T$ with a multivariate Gaussian distribution:

- $E[X_i|X_j = a] = E[X_i] + \frac{Cov(X_i,X_j)}{Var(X_j)} \cdot (a - E[X_j])$
- $Var(X_i|X_j = a) = \left(1 - \frac{Cov(X_i,X_j)^2}{Var(X_i) \cdot Var(X_j)}\right) \cdot Var(X_i)$

4. **Hidden Markov model.** 15

Consider a set of six binary random Mariables $\{C_1, C_2, C_3, E_1, E_2, E_3\}$ and the following probabilistic relationships:

$$\begin{aligned} p(C_1 = 1) &= 0.5 \\ p(C_1 = 2) &= 0.5 \end{aligned}$$

and for $t = 2, 3$:

$$\begin{aligned} p(C_t = 1|C_{t-1} = 1) &= 0.8 \\ p(C_t = 2|C_{t-1} = 1) &= 0.2 \\ p(C_t = 1|C_{t-1} = 2) &= 0.3 \\ p(C_t = 2|C_{t-1} = 2) &= 0.7 \end{aligned}$$

Moreover, for $t = 1, 2, 3$:

$$\begin{aligned} P(E_t = 1|C_t = 1) &= 0.1 \\ P(E_t = 2|C_t = 1) &= 0.9 \\ P(E_t = 1|C_t = 2) &= 0.9 \\ P(E_t = 2|C_t = 2) &= 0.1 \end{aligned}$$

(a) Visualize the relationships between the six variables through a directed acyclic graph (DAG), and factorize the joint distribution into a product of local conditional distributions. 5

(b) Compute the following two conditional probabilities: 10

- $P(E_2 = 1|C_1 = 1, C_2 = 1, C_3 = 1, E_1 = 1, E_3 = 1)$
- $P(E_2 = 1|C_1 = 1, C_3 = 1, E_1 = 1, E_3 = 1)$

SOLUTION EXERCISE 1:

For notational convenience, identify: $A = 1$, $B = 2$, $C = 3$, $D = 4$, $E = 5$, and $F = 6$.

Part (a): Ancestor matrix is:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Part (b):

- By edge deletions: 8
- By edge reversals: 5
- By edge additions: 10

Answer: By single edge operations 23 neighbour graphs can be reached.

Part (c): The CPDAG is shown in Figure 2.

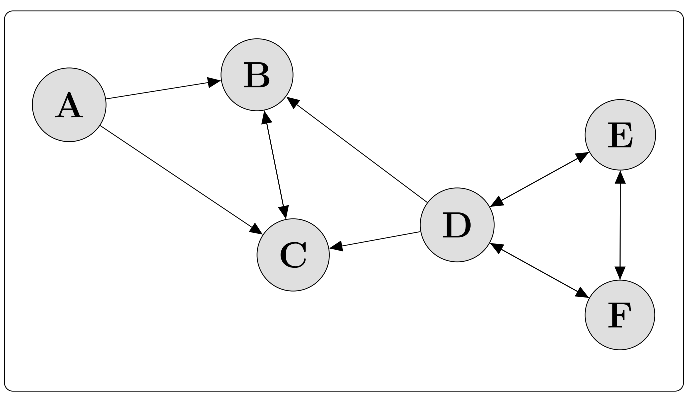

Figure 2: **Exercise 1(c): CPDAG. Reversible edges are represented as bi-directional.**

Part (d): There could be up to $2^4 = 16$ graphs in the equivalence class, as 4 edges in the CPDAG are reversible. However, only 12 of them actually belong to the same equivalence class. The 4 disqualified graphs have no additional v-structures, but invalid cycles.

Part (e): Yes, there is one. See Figure 3.

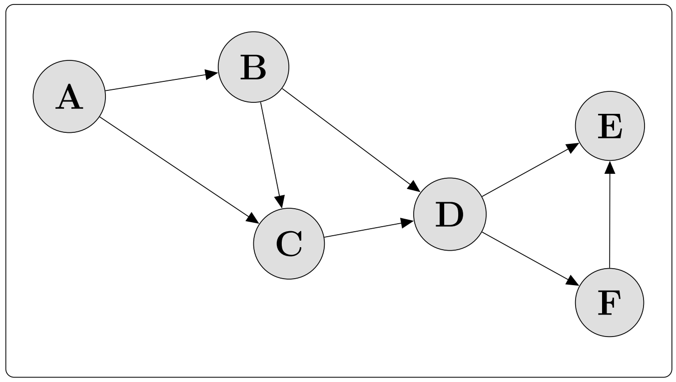

Figure 3: **Exercise 1(e): A DAG with the same skeleton but without v-structures.**

Part (f): Such a DAG can be found in Figure 4.

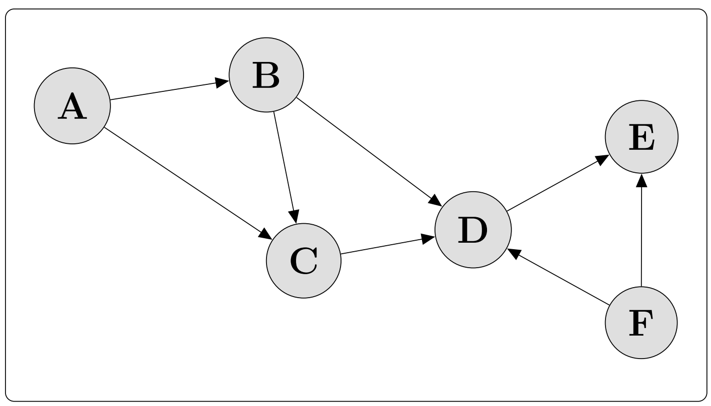

Figure 4: **Exercise 1(f): A DAG with the same skeleton and in whose CPDAG the edge from F to E is directed.**

Part (g): $\text{MB}(D) = \{A, B, C, E, F\}$, i.e. the 5 other nodes.

Part (h): There are 8 paths between A and E, and they are all blocked:

- $A \to B \leftarrow D \leftarrow E$, blocked
- $A \to B \leftarrow D \leftarrow F \to E$, blocked
- $A \to B \to C \leftarrow D \leftarrow E$, blocked
- $A \to B \to C \leftarrow D \leftarrow F \to E$, blocked
- $A \to C \leftarrow B \leftarrow D \leftarrow E$, blocked
- $A \to C \leftarrow B \leftarrow D \leftarrow F \to E$, blocked
- $A \to C \leftarrow D \leftarrow E$, blocked
- $A \to C \leftarrow D \leftarrow F \to E$, blocked

Part (i): Conditional on $Z = \{C\}$, all the 8 blocked paths become open paths.
Recall that a collider can be 'opened' by conditioning on the node itself or on one of its descendants. Here, C is a descendant of B.

- $A \to B \leftarrow D \leftarrow E$, open
- $A \to B \leftarrow D \leftarrow F \to E$, open
- $A \to B \to C \leftarrow D \leftarrow E$, open
- $A \to B \to C \leftarrow D \leftarrow F \to E$, open
- $A \to C \leftarrow B \leftarrow D \leftarrow E$, open
- $A \to C \leftarrow B \leftarrow D \leftarrow F \to E$, open
- $A \to C \leftarrow D \leftarrow E$, open
- $A \to C \leftarrow D \leftarrow F \to E$, open

Part (j): For the given graph we have:

$$P(A, B, C, D, E, F) = P(A) \cdot P(B|A, D) \cdot P(C|A, B, D) \cdot P(D|E, F) \cdot P(E|F) \cdot P(F)$$

And the marginal distribution of D, E, and F is:

$$\begin{aligned}
P(D, E, F) &= \sum_a \sum_b \sum_c P(A = a, B = b, C = c, D, E, F) \\
&= \sum_a \sum_b \sum_c P(A = a) \cdot P(B = b|A = a, D) \cdot P(C = c|A = a, B = b, D) \cdot P(D|E, F) \cdot P(E|F) \cdot P(F) \\
&= P(D|E, F) \cdot P(E|F) \cdot P(F) \sum_a \sum_b \sum_c P(A = a) \cdot P(B = b|A = a, D) \cdot P(C = c|A = a, B = b, D) \\
&= P(D|E, F) \cdot P(E|F) \cdot P(F) \sum_a P(A = a) \sum_b P(B = b|A = a, D) \sum_c P(C = c|A = a, B = b, D) \\
&= P(D|E, F) \cdot P(E|F) \cdot P(F)
\end{aligned}$$

It follows:

$$P(A, B, C|D, E, F) \quad = \quad \frac{P(A, B, C, D, E, F)}{P(D, E, F)} = P(A) \cdot P(B|A, D) \cdot P(C|A, B, D)$$

And the expression on the right does not depend on E and F.

SOLUTION EXERCISE 2:

For notational convenience, identify: $\mathbf{G_1}$ with 1, $\mathbf{G_2}$ with 2, and $\mathbf{G_3}$ with 3.

The proposal probabilities can then be arranged in a 3-by-3 matrix $\mathbf{Q}$. The element $\mathbf{Q}_{i,j}$ is the probability for proposing a move from $\mathbf{G_i}$ to $\mathbf{G_j}$. The Metropolis-Hastings acceptance probability $\mathbf{A}_{i,j}$ for the move from $\mathbf{G_i}$ to $\mathbf{G_j}$ is given by:

$$\mathbf{A}_{i,j} := A(\mathbf{G_i} \to \mathbf{G_j}) = \min\{1, \frac{p(data|\mathbf{G_j})}{p(data|\mathbf{G_i})} \cdot \frac{p(\mathbf{G_j})}{p(\mathbf{G_i})} \cdot \frac{\mathbf{Q}_{j,i}}{\mathbf{Q}_{i,j}}\}$$

Part (a) When only single edge additions and delitions are allowed, we have:

$$\mathbf{Q} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

and the four required acceptance probabilities are:

$$\begin{aligned}
\mathbf{A}_{1,3} &= \min\{1, \frac{a}{20a} \cdot \frac{0.4}{0.4} \cdot \frac{0.5}{1}\} = \frac{1}{40} = 0.025 \\
\mathbf{A}_{2,3} &= \min\{1, \frac{a}{20a} \cdot \frac{0.4}{0.2} \cdot \frac{0.5}{1}\} = \frac{1}{20} = 0.05 \\
\mathbf{A}_{3,1} &= \min\{1, \frac{20a}{a} \cdot \frac{0.4}{0.4} \cdot \frac{1}{0.5}\} = 1 \\
\mathbf{A}_{3,2} &= \min\{1, \frac{20a}{a} \cdot \frac{0.2}{0.4} \cdot \frac{1}{0.5}\} = 1
\end{aligned}$$

For $i \neq j$ we have the transition probabilities: $\mathbf{T}_{i,j} = \mathbf{Q}_{i,j} \cdot \mathbf{A}_{i,j}$, and for the diagonal elements we then compute: $\mathbf{T}_{i,i} = 1 - \sum_{j \neq i} \mathbf{T}_{i,j} \quad (i = 1, 2, 3)$. This way, we compute the elements of the transition matrix $\mathbf{T}$:

$$\mathbf{T} = \begin{pmatrix} 0.975 & 0 & 0.025 \\ 0 & 0.95 & 0.05 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

(b) When all three single edge operations are allowed, we have:

$$\mathbf{Q} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

so that the Hastings-ratio is always equal to 1.
The six required acceptance probabilities are then:

$$\begin{aligned}
\mathbf{A}_{1,3} &= \min\{1, \frac{a}{20a} \cdot \frac{0.4}{0.4}\} = 0.05 \\
\mathbf{A}_{2,3} &= \min\{1, \frac{a}{20a} \cdot \frac{0.4}{0.2}\} = 0.1 \\
\mathbf{A}_{3,1} &= \min\{1, \frac{20a}{a} \cdot \frac{0.4}{0.4}\} = 1 \\
\mathbf{A}_{3,2} &= \min\{1, \frac{20a}{a} \cdot \frac{0.2}{0.4}\} = 1 \\
\mathbf{A}_{1,2} &= \min\{1, \frac{20a}{20a} \cdot \frac{0.2}{0.4}\} = 0.5 \\
\mathbf{A}_{2,1} &= \min\{1, \frac{20a}{20a} \cdot \frac{0.4}{0.2}\} = 1
\end{aligned}$$

Again we use for $i \neq j$: $\mathbf{T}_{i,j} = \mathbf{Q}_{i,j} \cdot \mathbf{A}_{i,j}$. And for $i = 1,2,3$: $\mathbf{T}_{i,i} = 1 - \sum_{j \neq i} \mathbf{T}_{i,j}$.

The transition matrix $\mathbf{T}$ is then:

$$\mathbf{T} = \begin{pmatrix} 0.725 & 0.25 & 0.025 \\ 0.5 & 0.45 & 0.05 \\ 0.5 & 0.5 & 0 \end{pmatrix}$$

(c) For both Markov Chains it is guaranteed that they will have the posterior distribution as stationary distribution. Therefore, we have to compute the posterior distribution.

Normalization constant:

$$\begin{aligned} P(data) &= \sum_{i=1}^{3} p(data|\mathbf{G_i}) \cdot p(\mathbf{G_i}) \\ &= 20a \cdot 0.4 + 20a \cdot 0.2 + a \cdot 0.4 \\ &= 12.4a \end{aligned}$$

For the posterior probabilities we use:

$$P(\mathbf{G_i}|data) = \frac{p(data|\mathbf{G_i}) \cdot p(\mathbf{G_i})}{p(data)}$$

This way, we get the same stationary distribution for both Markov Chains, namely:

$$\begin{aligned} P(\mathbf{G_1}|data) &= \frac{20a \cdot 0.4}{12.4a} = \frac{8}{12.4} \approx 0.645 \\ P(\mathbf{G_2}|data) &= \frac{20a \cdot 0.2}{12.4a} = \frac{4}{12.4} \approx 0.323 \\ P(\mathbf{G_2}|data) &= \frac{a \cdot 0.4}{12.4a} = \frac{0.4}{12.4} \approx 0.032 \end{aligned}$$

SOLUTION EXERCISE 3:

Part (a) Compute the marginal distributions:

- $X_1 = 2 + \epsilon_1$ implies that $X_1 \sim N(2, 1)$
- $X_2 = -X_1 + \epsilon_2$ implies that $X_2 \sim N(-2, 2)$, as X_1 and ϵ_2 have independent Gaussian distributions.
- $X_3 = 2X_2 + \epsilon_3$ implies that $X_3 \sim N(-4, 9)$, as X_2 and ϵ_3 have independent Gaussian distributions.

The expectation vector is $(2, -2, -4)^T$. The diagonal elements of the covariance matrix are: $\Sigma_{1,1} = 1$, $\Sigma_{2,2} = 2$, and $\Sigma_{3,3} = 9$. The non-diagonal elements of the covariance matrix are the covariances: $\Sigma_{i,j} = Cov(X_i, X_j)$ $(i \neq j)$.

$$\begin{aligned}
\Sigma_{1,2} &= Cov(X_1, X_2) = Cov(X_1, -X_1 + \epsilon_2) = Cov(2 + \epsilon_1, -2 - \epsilon_1 + \epsilon_2) = Cov(\epsilon_1, -\epsilon_1) \\
&= -1 \\
\Sigma_{1,3} &= Cov(X_1, X_3) = Cov(X_1, 2X_2 + \epsilon_3) = Cov(X_1, 2(-X_1 + \epsilon_2) + \epsilon_3) \\
&= Cov(X_1, -2X_1 + 2\epsilon_2 + \epsilon_3) = Cov(X_1, -2X_1) = -2Var(X_1) \\
&= -2 \\
\Sigma_{2,3} &= Cov(X_2, X_3) = Cov(X_2, 2X_2 + \epsilon_3) = Cov(X_2, 2X_2) = 2Var(X_2) \\
&= 4
\end{aligned}$$

Altogether, this yields

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 2 \\ -2 \\ -4 \end{pmatrix}, \begin{pmatrix} 1 & -1 & -2 \\ -1 & 2 & 4 \\ -2 & 4 & 9 \end{pmatrix}\right)$$

Part (b)

- The marginal distribution of X_3 is: $X_3 \sim N(-4, 9)$, we can write that as:

$$X_3 = -4 + \tilde{\epsilon_3} \quad \text{where} \quad \tilde{\epsilon_3} \sim N(0, 9)$$

- The conditional distribution of X_2 given $X_3 = a$, is:

$$X_2|(X_3 = a) \sim N\left(-2 + \frac{4}{9}(a + 4), (1 - \frac{16}{2 \cdot 9}) \cdot 2\right) = N(-\frac{2}{9} + \frac{4}{9}a, \frac{2}{9})$$

Thus, we have:

$$X_2 = -\frac{2}{9} + \frac{4}{9}X_3 + \tilde{\epsilon_2} \quad \text{where} \quad \tilde{\epsilon_2} \sim N(0, \frac{2}{9})$$

- The conditional distribution of X_1 given $X_2 = a$, is:

$$X_1|(X_2 = a) \sim N\left(2 + \frac{-1}{2}(a+2), (1 - \frac{1}{1 \cdot 2}) \cdot 1\right) = N(1 - \frac{1}{2}a, \frac{1}{2})$$

  Thus, we have:

$$X_1 = 1 - \frac{1}{2}X_2 + \tilde{\epsilon_1} \quad \text{where} \quad \tilde{\epsilon_1} \sim N(0, \frac{1}{2})$$

Summary: Regression relationships for DAG '$X_3 \to X_2 \to X_1$':

$$\begin{aligned} X_3 &= -4 + \tilde{\epsilon_3} \\ X_2 &= -\frac{2}{9} + \frac{4}{9}X_3 + \tilde{\epsilon_2} \\ X_1 &= 1 - \frac{1}{2}X_2 + \tilde{\epsilon_1} \end{aligned}$$

where $\tilde{\epsilon_3} \sim N(0, 9)$, $\tilde{\epsilon_2} \sim N(0, \frac{2}{9})$, and $\tilde{\epsilon_1} \sim N(0, \frac{1}{2})$.

SOLUTION EXERCISE 4:

Part (a) For the DAG see Figure 5. The implied factorization is

$$P(C_1, C_2, C_3, E_1, E_2, E_3) = P(C_1) \cdot P(C_2|C_1) \cdot P(C_3|C_2) \cdot P(E_1|C_1) \cdot P(E_2|C_2) \cdot P(E_3|C_3)$$

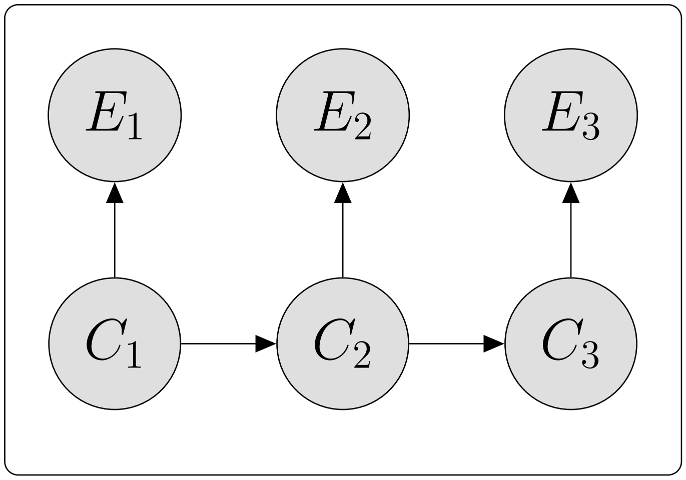

Figure 5: **Exercise 4(a): Graphical representation of the DAG.**

Part (b)

- From the DAG it can be seen:
  $P(E_2 = 1|C_1 = 1, C_2 = 1, C_3 = 1, E_1 = 1, E_3 = 1) = P(E_2 = 1|C_2 = 1) = 0.1$

- From the DAG it can be seen:

$$P(E_2 = 1|C_1 = 1, C_3 = 1, E_1 = 1, E_3 = 1) \quad = \quad P(E_2 = 1|C_1 = 1, C_3 = 1)$$

  and then

$$\begin{aligned}
P(E_2 = 1|C_1 = 1, C_3 = 1) &= \sum_{i=1}^{2} P(E_2 = 1, C_2 = i|C_1 = 1, C_3 = 1) \\
&= \sum_{i=1}^{2} P(E_2 = 1|C_1 = 1, C_2 = i, C_3 = 1) \cdot P(C_2 = i|C_1 = 1, C_3 = 1) \\
&= \sum_{i=1}^{2} P(E_2 = 1|C_2 = i) \cdot P(C_2 = i|C_1 = 1, C_3 = 1) \\
&= 0.1 \cdot \frac{0.8 \cdot 0.8}{0.8 \cdot 0.8 + 0.2 \cdot 0.3} + 0.9 \cdot \frac{0.2 \cdot 0.3}{0.8 \cdot 0.8 + 0.2 \cdot 0.3} \\
&= 0.1 \cdot \frac{0.64}{0.7} + 0.9 \cdot \frac{0.06}{0.7} \approx 0.169
\end{aligned}$$

  In the last step we have used:

$$
\begin{aligned}
P(C_2 = i|C_1 = 1, C_3 = 1) &= \frac{P(C_2 = i, C_1 = 1, C_3 = 1)}{P(C_1 = 1, C_3 = 1)} \\
&= \frac{P(C_3 = 1|C_2 = i) \cdot P(C_2 = i|C_1 = 1) \cdot P(C_1 = 1)}{\sum_{j=1}^{2} P(C_2 = j, C_1 = 1, C_3 = 1)} \\
&= \frac{P(C_3 = 1|C_2 = i) \cdot P(C_2 = i|C_1 = 1) \cdot P(C_1 = 1)}{\sum_{j=1}^{2} P(C_3 = 1|C_2 = j) \cdot P(C_2 = j|C_1 = 1) \cdot P(C_1 = 1)} \\
&= \frac{P(C_3 = 1|C_2 = i) \cdot P(C_2 = i|C_1 = 1)}{\sum_{j=1}^{2} P(C_3 = 1|C_2 = j) \cdot P(C_2 = j|C_1 = 1)}
\end{aligned}
$$