Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2020

Problem Set 1

Issued: Thursday, Sept. 3, 2020 **Due:** Tuesday, Sept. 22, 2020

Problem 1.1 (practice)
The random variables x and y are conditionally independent given z if

$$p_{x,y|z}(x, y|z) = p_{x|z}(x|z)p_{y|z}(y|z).$$

We denote this by $x \perp\!\!\!\perp y|z$.

(a) Show that for any random variables x, y, z, we have $x \perp\!\!\!\perp y|z$ if and only if the joint distribution for the three variables factors in the following form:

$$p_{x,y,z}(x, y, z) = h(x, z)g(y, z) \ .$$

(b) Construct an example of three binary random variables x, y and z that are pairwise independent but not mutually independent, then draw an undirected graphical model with the fewest number of edges that is consistent with this distribution. Show that there is no other undirected graphical model with fewer number of edges.

Problem 1.2 (practice)
Let $x_1, \ldots, x_N$ be random variables with joint distribution $p_{x_1,\ldots,x_N}(x_1, \ldots, x_N)$. Let $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D} \subset \{1, \ldots, N\}$ be disjoint subsets. Show that the following statements hold:

(a) Symmetry: If $x_\mathcal{A} \perp\!\!\!\perp x_\mathcal{B} \mid x_\mathcal{C}$, then $x_\mathcal{B} \perp\!\!\!\perp x_\mathcal{A} \mid x_\mathcal{C}$.

(b) Decomposition: If $x_\mathcal{A} \perp\!\!\!\perp x_{\mathcal{B}\cup\mathcal{D}} \mid x_\mathcal{C}$, then $x_\mathcal{A} \perp\!\!\!\perp x_\mathcal{B} \mid x_\mathcal{C}$ and $x_\mathcal{A} \perp\!\!\!\perp x_\mathcal{D} \mid x_\mathcal{C}$.

(c) Weak Union: If $x_\mathcal{A} \perp\!\!\!\perp x_{\mathcal{B}\cup\mathcal{D}} \mid x_\mathcal{C}$, then $x_\mathcal{A} \perp\!\!\!\perp x_\mathcal{B} \mid x_{\mathcal{C}\cup\mathcal{D}}$.

(d) Contraction: If $x_\mathcal{A} \perp\!\!\!\perp x_\mathcal{B} \mid x_\mathcal{D}$ and $x_\mathcal{A} \perp\!\!\!\perp x_\mathcal{C} \mid x_{\mathcal{B}\cup\mathcal{D}}$, then $x_\mathcal{A} \perp\!\!\!\perp x_{\mathcal{B}\cup\mathcal{C}} \mid x_\mathcal{D}$.

Problem 1.3

(a) Consider the graph shown in Figure 1.3-1. Determine whether the following statements are true or false and briefly justify why.

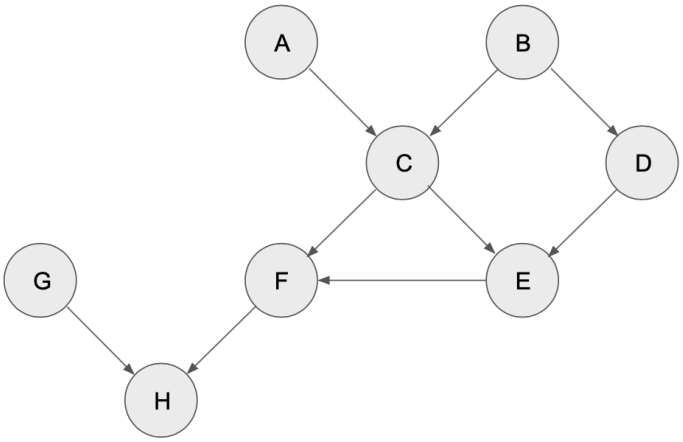


Figure 1.3-1

(i) $B \perp\!\!\!\perp G \mid A$

(ii) $C \perp\!\!\!\perp D \mid F$

(iii) $C \perp\!\!\!\perp D \mid A$

(iv) $H \perp\!\!\!\perp B \mid C, F$

(b) Two DAGs on N nodes are said to be *I-equivalent* (or *Markov equivalent*) if any distribution on $X_1, \ldots, X_N$ that can be factorized over one of these DAGs can also be factorized over the other. Draw all DAGs that are I-equivalent to the DAG in Figure 1.3-1.

(c) The skeleton of a DAG $\mathcal{G} = (V, E)$ is an undirected graph over the same node set V that contains an undirected edge for every edge in $\mathcal{G}$. Informally, by removing the arrow directions of a DAG you obtain its skeleton.

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two DAGs over the same node set. Show the following statements:

(i) $\mathcal{G}_1$ and $\mathcal{G}_2$ having the same skeleton is a necessary, but not sufficient condition for $\mathcal{G}_1$ and $\mathcal{G}_2$ being I-equivalent.

(ii) $\mathcal{G}_1$ and $\mathcal{G}_2$ are I-equivalent if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same skeleton and the same set of immoralities (i.e., induced subgraphs of the form $X \rightarrow Z \leftarrow Y$).

In your solution, you may want to use the following concept:

- Let V be all nodes in a graphical model excluding a node i. A *Markov blanket* of node i is a set of nodes $V_0 \subseteq V$ such that $X_i \perp\!\!\!\perp X_{V \setminus V_0} \mid X_{V_0}$, i.e., X_{V_0} contains all the information necessary to infer X_i.

(d) The following questions are open-ended and focus on the similarities of different graphs in a Markov equivalence class.

(i) (optional) Develop a representation of a Markov equivalence class by a unique graph. Note: this graph may contain undirected edges.

(ii) (optional) Describe a set of moves between DAGs with which you can move from any member of a Markov equivalence class to any other member in the Markov equivalence class without leaving the Markov equivalence class.

To solve (ii), you may want to use the following concept:

- An edge $i \to j$ is *covered* if $\pi_i = \pi_j \setminus \{i\}$, where π_i denotes the set of parents of node i.

Problem 1.4

The Hammersley-Clifford theorem gives us a canonical method to turn an undirected graph into a factor graph. Specifically, assume that we are given a graph $\mathcal{G}$ and a strictly positive distribution P that is Markov with respect to $\mathcal{G}$. Then, Hammersley-Clifford tells us that P can be written as a product of potential functions for each maximal clique. Thus, to turn an undirected graph into a factor graph, we can simply define a factor node for each maximal clique in the graph. Figure 1.4-1 depicts an example graph and the associated factor graph.

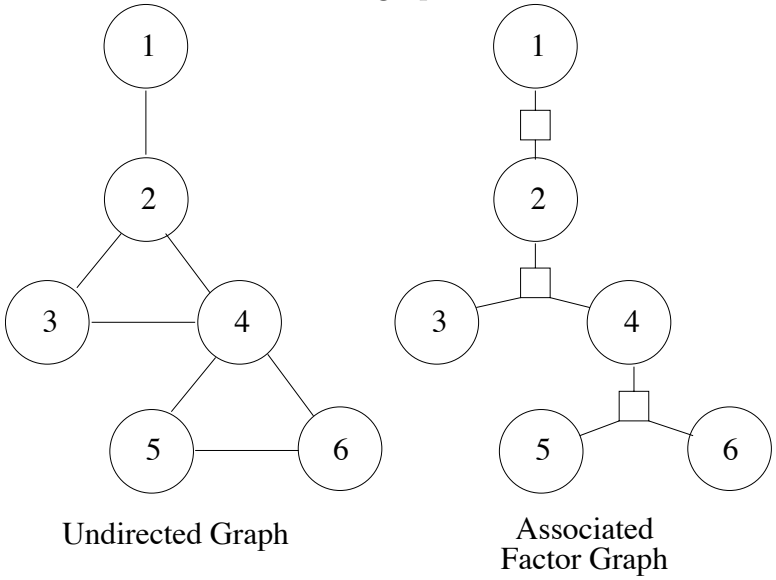


Figure 1.4-1

(a) Show that the factor graph produced by the canonical construction can be exponentially larger than the original graph. Specifically, show that there exists a constant $c > 1$ such that for all sufficiently large n, there exists an undirected

graph with n vertices such that the associated factor graph has at least c^n vertices.

(b) Show that there is a polynomial time algorithm, whose run time is no greater than a polynomial function of the number of vertices n, that given as input an undirected graph $\mathcal{G}$, determines whether or not the factor graph associated to $\mathcal{G}$ is a tree.

Note that the naive algorithm that computes the factor graph and then checks if it is a tree is not polynomial time, because from part (a) we know that just writing down the factor graph can take exponential time.

In your solution, you may want to use the following concepts and results:

- A graph is *chordal* if any cycle of length 4 or more nodes has a chord, which is an edge joining two nodes that are not adjacent in the cycle. You can use the fact that testing the chordality of a graph can be done in linear time (for fun, try to come up with the algorithm).
- We use K_4 to denote a clique containing 4 nodes, and say a graph contains K_4 if it has a clique of 4 nodes. We also use $\widetilde{K}_4$ to denote the graph generated from deleting one edge from K_4, and say a graph contains $\widetilde{K}_4$ if it has such a subgraph. You can use the fact that testing whether a graph contains $\widetilde{K}_4$ can be done in polynomial time (e.g. a brute-force search would take as little time as $\binom{n}{4} = \mathcal{O}(n^4)$).

Problem 1.5

A *perfect map* for a distribution P is a graph $\mathcal{G}$ (directed or undirected) such that P satisfies a conditional independence relationship if and only if this conditional independence relationship is implied by $\mathcal{G}$, i.e., $\mathcal{I}(P) = \mathcal{I}(\mathcal{G})$ (using the notation from class).

(a) Construct a distribution which has no perfect map. Specifically, no DAG or undirected graph should be a perfect map for the distribution.

(b) Construct a DAG and a distribution such that the DAG is a perfect map for the distribution, but no undirected graph is a perfect map for the distribution.

(c) Do the reverse of part (b), i.e., construct an undirected graph and a distribution such that the undirected graph is a perfect map for the distribution, but no DAG is a perfect map for the distribution.

(d) (optional) Prove that for any undirected graph $\mathcal{G}$, there exists some probability distribution P such that $\mathcal{G}$ is a perfect map for P.

(*Hint:* Start with a collection of independent random variables $y_1, y_2, \ldots, y_N$, where N may be much larger than n, the number of vertices of $\mathcal{G}$. Then, for each vertex of $\mathcal{G}$, try to associate a cleverly chosen subset of the y_i.)

(e) (optional) Prove that for any DAG $\mathcal{G}$, there exists some probability distribution P such that $\mathcal{G}$ is a perfect map for P.

(f) (optional) This part is concerned with the opposite question to parts (b) and (c) – when can a distribution P have a perfect map that is a DAG and a perfect map that is an undirected graph? Two equivalent answers to this question are stated below:

(i) For an arbitrary undirected graph $\mathcal{G}$, there exists a directed graph that implies exactly the same conditional independencies as $\mathcal{G}$ if and only if $\mathcal{G}$ is chordal.

(ii) For an arbitrary DAG $\mathcal{G}$, there exists an undirected graph that implies exactly the same conditional independencies as $\mathcal{G}$ if and only if moralizing $\mathcal{G}$ does not add any edges.

Prove either statement (i) or (ii), i.e., you can prove whichever version you find easier to prove.

(g) Parts (b) and (c) show that there are distributions for which directed graphs can be perfect maps but for which no undirected graphs can be perfect maps, and vice versa. From parts (d) and (e), we also know that no DAGs or undirected graphs are "useless", i.e., there is always some distribution for which a given graph is a perfect map.

Part (f) characterizes when a distribution can have a DAG and an undirected graph as perfect map. So, the final cases to consider are the following:

(i) Can two different undirected graphs be perfect maps for the same distribution?

(ii) Can two different DAGs be perfect maps for the same distribution?

Show that the answer to part (i) is no, but that the answer to part (ii) is yes. Thus, we can throw away some DAGs without changing the set of distributions that have DAGs as perfect maps, but throwing away *any* undirected graph will shrink the set of distributions that have undirected graphs as perfect maps.

Problem 1.6 (practice)

Let there be K different coins, each with different biases, $c_1, \ldots, c_K \in [0, 1]$. The k^{th} coin comes up heads with probability c_k, and tails with probability $1 - c_k$. Let

$t \in \{1, 2, \ldots K\}$ be a random variable having the mass function $p_t(\cdot)$. Then define the random variable $x \in \{0, 1\}$ to be 1 if the t^{th} coin comes up heads, and 0 if the t^{th} coin comes up tails. In other words, to generate a sample of x, we first sample t, then toss coin number t, and set x equal to the indicator function of coin t coming up heads.

(a) Write down an undirected graphical model description of the above described mixture distribution involving variables x and t. Provide the diagram of the undirected graphical model in addition to the potential functions.

In addition, we are given K (known) N-dimensional vectors $\boldsymbol{\theta}^1, \ldots \boldsymbol{\theta}^K \in \mathbb{R}^N$. If $t = k$, we generate random variables $\mathbf{y} = (y_1, \ldots, y_N) \in \{0, 1\}^N$ according to

$$p_{\mathbf{y}|t}(\mathbf{y}|k) \propto \exp\big(\sum_{i=1}^{N} \theta_i^k y_i\big), \tag{1}$$

for $\mathbf{y} = (y_1, \ldots, y_N) \in \{0, 1\}^N$, where we used the notation $\boldsymbol{\theta}^k = (\theta_1^k, \ldots, \theta_N^k)$.

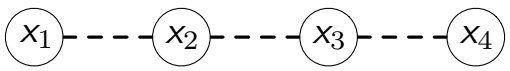
(b) Write down an undirected graphical model description of the above described mixture distribution involving varables $x, t, y_1, \ldots y_N$. Provide the diagram of the undirected graphical model in addition to the potential functions.

Problem 1.7

Let $\mathbf{x} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$ denote a collection of jointly Gaussian random variables with information matrix $\mathbf{J} = [J_{ij}]$. Recall that we can form the corresponding undirected graphical model by including edges between only those pairs of variables x_i, x_j for which $J_{ij} \neq 0$.

In this problem, we consider a graph induced by the sparsity pattern of the *covariance matrix* $\boldsymbol{\Lambda} = [\Lambda_{ij}]$. That is, we form an undirected graph by including edges between only those pairs of variables x_i, x_j for which $\Lambda_{ij} \neq 0$. The edges are drawn in dashed lines, and this graph is called a *covariance graph*.

Consider the following covariance graph:



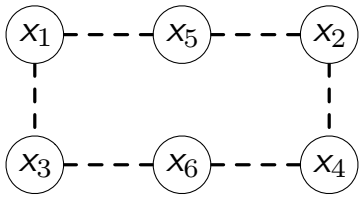
(a) List all conditional and unconditional independencies implied by the covariance graph.

For the remainder of the problem, you may find useful the following results on an arbitrary random vector $\mathbf{y}$ partitioned into two subvectors $\mathbf{y}_1$ and $\mathbf{y}_2$ (i.e., $\mathbf{y} = [\mathbf{y}_1^{\mathrm{T}}, \mathbf{y}_2^{\mathrm{T}}]^{\mathrm{T}}$), with information matrix and covariance matrix

$$\begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix}, \begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} \end{bmatrix}.$$

Specifically, the conditional distribution $p_{\mathbf{y}_1|\mathbf{y}_2}(\mathbf{y}_1|\mathbf{y}_2)$ has information matrix $\mathbf{J}_{11}$ and covariance matrix $\mathbf{\Lambda}_{11} - \mathbf{\Lambda}_{12}\mathbf{\Lambda}_{22}^{-1}\mathbf{\Lambda}_{21}$. The marginal distribution $p_{\mathbf{y}_1}(\mathbf{y}_1)$ has information matrix $\mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}$ and covariance matrix $\mathbf{\Lambda}_{11}$.

Consider the following covariance graph:



(b) Draw a covariance graph with the fewest possible (dashed) edges for p_{x_1,x_2,x_3,x_4}.

(c) Draw a covariance graph with the fewest possible (dashed) edges for $p_{x_1,x_2,x_3,x_4|x_5,x_6}$.

Problem 1.8

(a) Suppose we have n i.i.d samples drawn from the joint distribution of random variables $X_1, X_2, ..., X_p, Y$ and want to predict Y as a linear combination of X_i for $1 \leq i \leq p$. We can write this as the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Y}$ is an $n \times 1$ vector of observations of Y, $\mathbf{X}$ is an $n \times p$ matrix of observations of $X_1, X_2, ..., X_p$, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of error terms, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. We will assume that $\boldsymbol{\varepsilon} \sim \mathtt{N}(\mathbf{0}, \sigma^2\mathbf{I})$, i.e. the error vector consists of i.i.d Gaussian random variables with mean 0 and constant variance. We want to estimate the unknown parameter $\boldsymbol{\beta}$.

(i) Determine the log-likelihood function $\ell(\boldsymbol{\beta}; \mathbf{Y})$, in terms of $\mathbf{X}, \mathbf{Y}, \sigma$.

(ii) We are interested in finding the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, where

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}; \mathbf{Y})$$

Determine the maximum likelihood estimate for $\boldsymbol{\beta}$. How does this relate to the least-squares estimate of $\boldsymbol{\beta}$? Recall the least-squares estimate satisfies

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

(b) Let $\mathbf{x}$ be a p-dimensional random variable with joint Gaussian distribution $\mathbf{x} \sim \mathtt{N}(\boldsymbol{\mu}, \mathbf{\Lambda})$ where $\boldsymbol{\mu}$ is an $p \times 1$ mean vector, and $\mathbf{\Lambda}$ is a $p \times p$ covariance matrix.

Suppose we have data $\mathcal{D} = \{x_1, ..., x_n\}$ of n i.i.d samples of $\mathbf{x}$, i.e., each x_i is a $p \times 1$ vector. Find the ML estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$.

Hint: You can use the following equalities that hold for any square matrix A and vector x, and where $tr\,[A]$ refers to the trace of A (the sum of its diagonal elements):

$$\frac{\partial}{\partial A} x^T A x = \frac{\partial}{\partial A} tr\left[x x^T A\right] = [x x^T]^T = x x^T$$

$$\frac{\partial}{\partial A} \log |A| = A^{-T}$$

Computational Problem 1

Consider a set of random variables X_1, X_2, X_3, Y which are jointly Gaussian and have a joint distribution following the conditional independencies of the DAG in Figure 1.8-1 below.

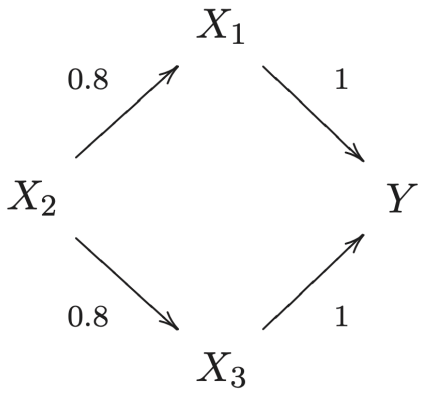


Figure 1.8-1

The graph should be interpreted as $X_2 = \varepsilon_2$, $X_1 = 0.8X_2 + \varepsilon_1$, and $X_3 = 0.8X_2 + \varepsilon_3$ where $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are mutually independent Gaussian random variables with mean zero and constant variance. Further, suppose that $\varepsilon_1, \varepsilon_2, \varepsilon_3$ have been defined such that X_1, X_2, X_3 all have variance 1 (this allows us to meaningfully compare their regression coefficients). We are interested in estimating the causal effect of X_i on Y for $i = 1, 2, 3$.

(a) (i) One method for explaining the effect of X_i on Y is to simply apply multiple linear regression $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ of X_i on Y, where ε is an independent random normal variable with mean 0 and variance 1. Determine the coefficients $\beta_1, \beta_2, \beta_3$ (Note: this should be relatively straightforward, using the specified interpretation of the graph). Which variable is most important?

(ii) Another method is to determine the causal effect of X_i on Y as the regression coefficient of X_i in the regression of Y on X_i and π_i, as dis-

cussed in recitation. This coefficient is given by $\beta_{i|\pi_i}$, where for any set $S \subseteq \{X_1, ..., X_n, Y\} \setminus \{X_i\}$,

$$\beta_{i|S} = \begin{cases} 0 & \text{if } Y \in S \\ \text{coefficient of } X_i \text{ in } Y \sim X_i + S & \text{if } Y \notin S \end{cases}$$

and π_i are the parents of X_i. Determine $\beta_{i|\pi_i}$ for $i = 1, 2, 3$. Under this analysis, which variable is the most important? Explain briefly why the results differ from that of (a)(i).

(b) On Stellar, in data.csv, you will find samples of randomly generated joint Gaussian variables corresponding to the conditional dependencies in the DAG in Figure 1.8-2 below. Each column represents a different variable, and each row corresponds to a sample drawn from the joint distribution. Each variable has been defined to have a variance of 1, similar to the previous part.

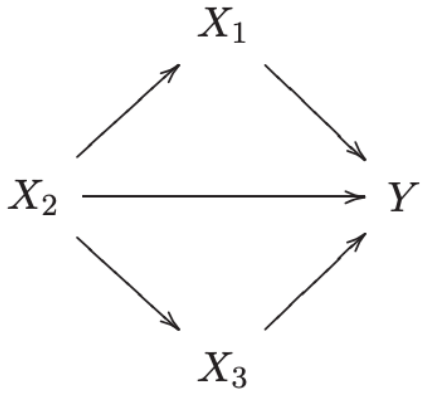


Figure 1.8-2

Find the maximum likelihood estimate of the regression coefficients of each X_i for $i = 1, 2, 3$ using both methods in (a)(i) and (a)(ii) and state the difference between the two results.

(c) (optional) Suppose we do not know the true underlying DAG, but we learn that X_1 is independent of X_3 given X_2. From this information, we can learn the Markov equivalence class (the set of all DAGs that encode the same set of conditional independencies) of the true DAG, but not the true DAG itself. In this case, how could we define the causal effect of X_i on Y?

If you're interested in further reading on this topic, you can read this paper which shows how we can compute the causal effect of policies like wearing masks on the spread of SARS-CoV-2.