Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2018

Problem Set 8

Issued: Thursday, Nov. 29, 2018 **Due:** Thursday, Dec. 6, 2018

Problem 8.1

In this problem, we try to learn undirected graph parameters for joint Gaussian distributions. Consider a joint Gaussian distribution over $x = [x_1, x_2, \ldots, x_6]$, as shown in Figure 1. Each node can be a Gaussian vector.

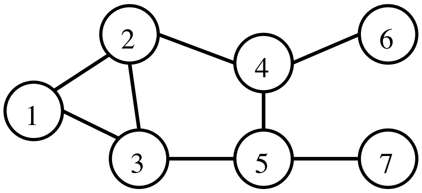

Figure 1:

(a) Suppose that you observe K i.i.d. samples $x^{(1)}, \ldots, x^{(K)}$. Provide the Maximum Likelihood estimator for the covariance matrix $\hat{\Sigma}$ and the mean $\hat{\mu}$.

(b) However, in order to do inference on the graphical model, you need to learn the information matrix J. In this example, assume that you are interested in estimating the block $J_{123,123}$ corresponding to the variables x_1, x_2, x_3.
One approach is to invert the estimated covariance matrix $\hat{\Sigma}$ to obtain an estimation of $\hat{J}$. Another approach is described below. Assume that you have a script of loopy Gaussian BP algorithm, which inputs an information matrix $\widetilde{J}$ for a potentially loopy graph and outputs all the messages $\widetilde{J}_{i \to j}$. Use this script to get an estimation of $\hat{J}_{123,123}$.
(hint: Specify the input and output to the script, and express the estimator $\hat{J}_{123,123}$ in terms of the output as well as $\hat{\Sigma}$)

(c) Comment on the complexity and the accuracy of the two approaches to learn Gaussian graphical models in (b).
(*(hint: in general, if a matrix A is sparse, the inverse A^{-1} will not have the sparsity pattern. Moreover, A^{-1} may not be sparse at all.)*)

Problem 8.2

In this problem, we consider the relationship between logistic regression and parameter learning in Ising models. Suppose we have Ising model $p(\cdot)$ over $\{-1,1\}^n$ given by:

$$p(x) \propto \exp(\sum_{i,j} \theta_{ij} x_i x_j)$$

We will refer to the collection of interaction parameters $(\theta_{ij})_{i,j\in[n]}$ as Θ and the collection $(\theta_{ij})_{j\in[n]}$ as Θ_i for every $i \in [n]$. We suppose that the underlying undirected graphical model G is known but the values of the parameter θ_{ij} are unknown (that is, $\{(i,j) : \theta_{i,j} \neq 0\}$). Let $X \sim p(\cdot)$. Given $i \in [n]$, we denote N_i to be the set of neighbours of i in the graph G

(a) Write down the conditional distribution of X_i conditioned on X_{N_i} and give an expression for $\mathbb{E}\left[X_i | X_{N_i}\right]$

We now consider logistic regression. Consider random variables $(W, Z) \in \mathbb{R}^d \times \{-1, 1\}$, such that:

$$\mathbb{P}(Z = 1|W) = \frac{e^{\beta^\intercal W}}{1 + e^{\beta^\intercal W}} := g(W;\beta) \tag{1}$$

for some parameter $\beta \in \mathbb{R}^d$. Logistic regression is the problem of estimating the parameter β for marginal arbitrary distributions of W. Suppose we have an algorithm to perform logistic regression given i.i.d samples from the model and obtain $\hat{\beta}$ such that

$$\mathbb{P}(Z = 1|W) = g(W;\hat{\beta})\,.$$

holds exactly with probability 1. We call this algorithm 'A'.

(b) Give an example of random variables (W, Z) such that $\hat{\beta} = \beta$ is not necessarily true. (consider the case when of the co-ordinates of W are linearly dependent)

(c) Show that we can use algorithm A to estimate Θ_i by $\hat{\Theta}_i$. What is the possible challenge in reconstructing Θ using the collection $\hat{\Theta}_i$?

To resolve this issue, we show that when $(W, Z) := (X_{N_i}, X_i)$, we must have $\beta = \hat{\beta}$ (that is, $\Theta_i = \hat{\Theta}_i$). Henceforth, we fix i and let $|N_i| =: d_i$.

(d) Let $t \in \{-1, 1\}^{d_i}$ be arbitrary. Show that $\mathbb{P}(X_{N_i} = t) > 0$.

Hint : Don't overthink.

(e) Recall the definition of $g(;)$ from Equation (1). Let $\hat{\Theta}_i$ be the reconstruction from part (c). Show that for every $j \in N_i$,

$$\mathbb{E}\left[g(X_{N_i}; 2\Theta_i)X_j\right] = \mathbb{E}\left[g(X_{N_i}; 2\hat{\Theta}_i)X_j\right] .$$

Hint : Consider $\mathbb{E}X_iX_j$ and consider the right conditional expectation.

(f) Use parts (d) and (e) to show that $\Theta_i = \hat{\Theta}_i$

Hint: Let $f : \mathbb{R} \to \mathbb{R}$ be continuous, $f(x) \geq 0\ \forall x$ and there exists a unique $x^* \in \mathbb{R}$ such that $f(x^*) = 0$,. For any real valued random variable Γ, if $\mathbb{E}f(\Gamma) = 0$ then $\Gamma = x^*$ almost surely. For instance, if $\mathbb{E}\Gamma^2 = 0$, then $\Gamma = 0$ almost surely.

Problem 8.3

Assume we are given a causal DAG G. In this problem, we will find a type of adjustment set for X_i and one of its descendant X_j, that can be used instead of the parents of X_i, that is, X_{π_i} (e.g., if some of the parents can't be measured). We say that a subset of variables X_A satisfies the *backdoor criterion* relative to X_i and X_j if:

(i) no node in A is a descendant of i

(ii) A blocks every path between i and j that contains an edge into i (i.e., the *backdoor* paths, hence the name)

Assume that X_A satisfies the backdoor criterion. We will prove that X_A is an adjustment set for X_i and X_j.

(a) Show from d-separation that $X_i \perp\!\!\!\perp X_A \mid X_{\pi_i}$

(b) Show from d-separation that $X_j \perp\!\!\!\perp X_{\pi_i} \mid X_i, X_A$

(c) Show that X_A is an adjustment set for X_i and X_j. *(Hint: start with the fact that the parent set is an adjustment set)*

(d) Consider the graph in Fig. 2. Which subsets of $\{X_4, X_5, X_6\}$ satisfy the backdoor criterion relative to X_i and X_j?

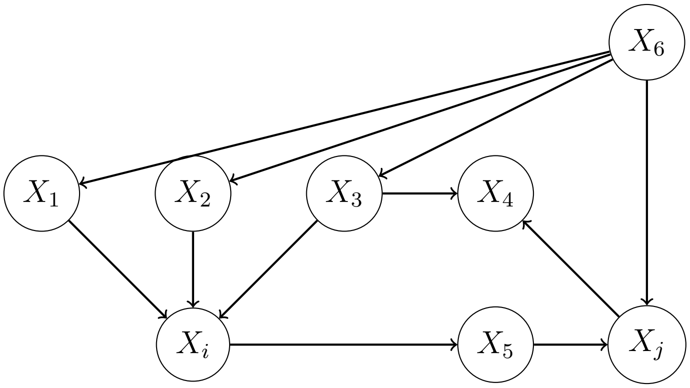

Figure 2: A graph

Problem 8.4

To health specialists studying infant mortality, there is a consensus that the mortality rate of infants (i.e., the probability that the infant dies in the first month, represented by the variable $D \in \{0,1\}$) is directly influenced by a combination of:

- $S \in \{0,1\}$, 1 if the the mother smokes cigarettes and 0 if she doesn't
- $W \in \{\ell, h\}$, whether the birth weight of the infant is low (ℓ) or high (h), and
- $O \in \{0,1\}$, the presence (1) or absence (0) of other factors (e.g., malnutrition of the mother)

Furthermore, it is known that both smoking and the other factors have an effect on birth weight. No other causal relationships between the variables are believed to be true.

(a) Draw a causal DAG relating W, S, O, and D that represents the causal beliefs of health specialists.

(b) Determine if the following statements are true or false in the causal DAG you drew in (a):

(i) $P_{W|do(S=s)} = P_{W|S}(w|s)$

(i) $P_{D|do(S=s)} = P_{D|S}(d|s)$

(ii) $S \perp\!\!\!\perp O \mid W = \ell$

(c) The following statements are widely regarded as fact in the medical community. Express them in terms of inequalities between probability distributions (possibly conditioninal and/or interventional):

(i) For the average infant (i.e., no additional information is known about the infant), smoking causes an increased probability of death.

(ii) A mother's smoking causes an increase in the probability that an infant has low birth weight.

(iii) An infant with a low birth weight has a higher mortality rate.

(iv) When looking at children with only low birth weight, it has be observed that having a smoking mother ($S = 1$) actually *decreases* the mortality rate of the child.

This has been called the *Birth Weight Paradox.* As a doctor, if you know that a baby will have low birth weight (say, from an ultrasound scan), should you tell the would-be mother that her infant is more likely to survive if she starts smoking? That is, is it true that

$$P_{D|W,do(S=1)}(1|\ell) < P_{D|W,do(S=0)}(1|\ell) \tag{2}$$

(d) **(Optional)** Provide a counterexample to (2) in terms of conditional probabilities for your causal DAG from (a) that match the facts from (c).