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

Problem Set 3

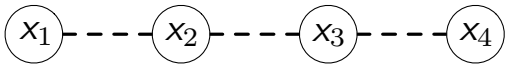
Issued: Thursday, Sept. 27, 2018 **Due:** Thursday, Oct. 11, 2018

Problem 3.1

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* $\mathbf{\Lambda} = [\Lambda_{ij}]$. 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:



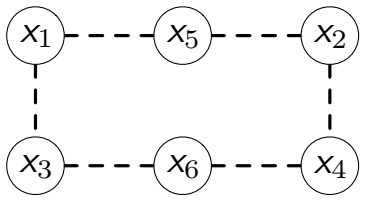
(a) List all conditional and unconditional independencies implied by the covariance graph.

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 3.2

Consider a random vector $\mathbf{x}$ made up of two subvectors $\mathbf{x}_1$ and $\mathbf{x}_2$ (i.e., $\mathbf{x} = [\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}}]^{\mathrm{T}}$), with information matrix and potential vector

$$\begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} \quad , \quad \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} \tag{1}$$

(where, of course, $\mathbf{J}_{11}$ and $\mathbf{J}_{22}$ are symmetric, and $\mathbf{J}_{21} = \mathbf{J}_{12}^{\mathrm{T}}$). If $\mathbf{J}_{12} = \mathbf{0}$, then the joint distribution for $\mathbf{x}_1$ and $\mathbf{x}_2$ factors and we easily see that $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent and that we can directly read off the marginal distribution for either of them. For example, in this case the information parameters for $\mathbf{x}_1$ are simply $\mathbf{J}_{11}$ and $\mathbf{h}_1$ (i.e., $\mathbf{x}_1 \sim \mathcal{N}^{-1}(\mathbf{h}_1, \mathbf{J}_{11})$). However, if $\mathbf{J}_{12} \neq \mathbf{0}$, then there is some work to be done to determine the information parameters for the marginal distribution for $\mathbf{x}_1$. Very importantly, and as you'll show in this problem, finding those information parameters is very closely related to computations that are likely familiar to you but from a very different context (namely solving simultaneous equations). Specifically, we obtain these information parameters by *Gaussian elimination*.

$$\mathbf{x}_1 \sim \mathcal{N}^{-1}(\mathbf{h}_a, \mathbf{J}_a)$$
$$\mathbf{J}_a = \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21} \quad , \quad \mathbf{h}_a = \mathbf{h}_1 - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{h}_2 \tag{2}$$

The operation involved in computing $\mathbf{J}_a$ is often referred to as the *Schur complement* formula, an operation that is central to Gaussian elimination. Now, we'll get at this answer in two different ways.

(a) Since $\mathbf{J}_{12} \neq \mathbf{0}$, we can't write the joint density as a product of the density for $\mathbf{x}_1$ and that for $\mathbf{x}_2$. However, if we can perform an invertible linear transformation into a new set of variables, in which we keep the components of $\mathbf{x}_1$ unchanged, maybe we can expose that marginal

density. That is, suppose we consider linear transformations of the form

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{z} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 + \mathbf{A}\mathbf{x}_1 \end{bmatrix}$$

Intuitively, what we would like to do is to subtract from $\mathbf{x}_2$ just enough of $\mathbf{x}_1$ to leave the difference uncorrelated with $\mathbf{x}_1$ (and hence independent by joint Gaussianity). Show that the right choice of $\mathbf{A}$ is $\mathbf{J}_{22}^{-1}\mathbf{J}_{21}$, that with this choice $\mathbf{x}_1$ and $\mathbf{z}$ are independent, and that the marginal distribution for $\mathbf{x}_1$ is as indicated above in eq. (2).

Hint: $\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{A} & \mathbf{I} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A} & \mathbf{I} \end{bmatrix}$

(b) The information parameterization only implicitly specifies the mean of a Gaussian vector — i.e., we need to solve the equation $\mathbf{J}\mathbf{m} = \mathbf{h}$ to determine the mean. Consider again the case in which $\mathbf{x} = [\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}}]^{\mathrm{T}}$ has information parameterization as given in (1), and let $\mathbf{m}_1$, and $\mathbf{m}_2$ denote the means of $\mathbf{x}_1$, and $\mathbf{x}_2$, respectively. Set up the equations to be solved for these means from the information parameterization, eliminate $\mathbf{m}_2$, and show that what you are left with are precisely the equations

$$\mathbf{J}_a\mathbf{m}_1 = \mathbf{h}_a \ .$$

(c) As should be clear from this Gaussian elimination interpretation, in more complex situations — when $\mathbf{x}$ is composed of more than two component subvectors, we can, in principle perform Gaussian elimination in any order we like. For example, if $\mathbf{x}$ consists of three subvectors $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$, we can equivalently eliminate $\mathbf{x}_3$ first (obtaining the joint marginal for $\mathbf{x}_1$, and $\mathbf{x}_2$) and then eliminate $\mathbf{x}_2$, or we can eliminate $\mathbf{x}_2$ and $\mathbf{x}_3$ in the opposite order, or we can eliminate $\mathbf{x}_2$ and $\mathbf{x}_3$ simultaneously (viewing them together as a single, larger subvector). One case in which things are particularly simple is the case in which there is very special and important structure in the interdependencies of these three subvectors. Specifically, suppose that the information parameterization of $[\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}}, \mathbf{x}_3^{\mathrm{T}}]^{\mathrm{T}}$ has the following form:

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} & \mathbf{J}_{13} \\ \mathbf{J}_{21} & \mathbf{J}_{22} & \mathbf{0} \\ \mathbf{J}_{31} & \mathbf{0} & \mathbf{J}_{33} \end{pmatrix} \quad , \quad \mathbf{h} = \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \mathbf{h}_3 \end{pmatrix}$$

(i) Show that $\mathbf{x}_2$ and $\mathbf{x}_3$ are conditionally independent given $\mathbf{x}_1$.
Hint: Recall that if $(x||y)|z$, then $p(x, y, z) = h(x, z)g(y, z)$.

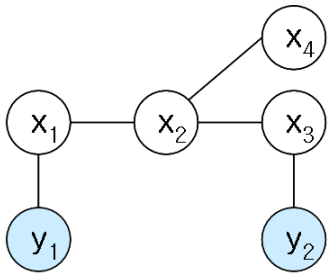
(ii) Show that the marginal distribution for $\mathbf{x}_1$ has additive form, in that the influences of $\mathbf{x}_2$ and $\mathbf{x}_3$ individually on $\mathbf{x}_1$ (as in eq. (2)) are simply added together to get their combined influence. That is,

$$\begin{aligned}
\mathbf{x}_1 &\sim \mathcal{N}^{-1}(\mathbf{h}_b, \mathbf{J}_b) \\
\mathbf{J}_b &= \mathbf{J}_{11} - (\mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21} + \mathbf{J}_{13}\mathbf{J}_{33}^{-1}\mathbf{J}_{31}) \\
\mathbf{h}_b &= \mathbf{h}_1 - (\mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{h}_2 + \mathbf{J}_{13}\mathbf{J}_{33}^{-1}\mathbf{h}_3)
\end{aligned}$$

Problem 3.3
Let $\mathbf{x} \sim \mathcal{N}^{-1}(\mathbf{h}_\mathbf{x}, \mathbf{J}_\mathbf{x})$, and $\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{v}$, where $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$.

(a) Find the potential vector and the information matrix of $p(\mathbf{x}, \mathbf{y})$ and $p(\mathbf{x}|\mathbf{y})$.

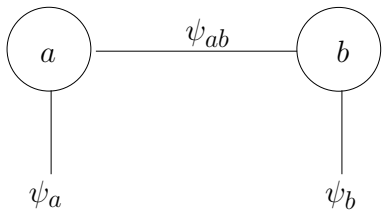
(b) Consider the following Gaussian graphical model:



Let $y_1 = x_1 + v_1$, $y_2 = x_3 + v_2$, and $\mathbf{R} = I$. Find $\mathbf{C}$. Represent messages $h_{x_3 \to x_2}$ and $J_{x_3 \to x_2}$ in terms of the elements of $\mathbf{h}_\mathbf{x}$ and $\mathbf{J}_\mathbf{x}$.

(c) Now assume that we have an additional measurement $y_3 = x_3 + v_3$, where v_3 is a zero-mean Gaussian variable with variance 1 and is independent from all other variables. Represent messages $h_{x_3 \to x_2}$ and $J_{x_3 \to x_2}$ in terms of the elements of $\mathbf{h}_\mathbf{x}$ and $\mathbf{J}_\mathbf{x}$.

Problem 3.4
Consider the following 2-node undirected graphical model:



where the variables x_a and x_b are binary, and the compatibility functions are given by:

$$\psi_a(0) = \psi_a(1) = \psi_b(0) = \psi_b(1) = 1 \tag{3}$$

$$\psi_{ab}(0,0) = \psi_{ab}(1,1) = 1 \quad , \quad \psi_{ab}(1,0) = \psi_{ab}(0,1) = 10 \tag{4}$$

(a) Compute the max-marginals for each variable, and show that there is no unique maximizing value for each of the variables. Explain why independently choosing the maximizing values for each of the variables does not lead to the maximum of the joint distribution.

(b) In general, we shall define *edge* max-marginals $\bar{p}_{ij}(x_i, x_j) = \max_{\mathbf{x} \setminus \{x_i, x_j\}} p_{\mathbf{x}}(\mathbf{x})$ for every edge (i, j) in a tree.

Choose one of the maximizing values at one node, say node "a". Show that in order to maximize the joint probability, you can use the edge max-marginal to determine the value of the other node.

Computational Exercise 1: Belief Propagation

In this exercise, you will implement the parallel sum-product algorithm and test it on two datasets. To make it easy to use the same algorithm for both datasets, you should make no assumptions about the graph structure, but you can assume that the variables are binary and there are only nodewise and pairwise potentials, i.e. there are no potentials involving more than three nodes. For example, you may want to write a function with signature

$$\texttt{nodes} \times \texttt{edges} \times \texttt{node\_potentials} \times \texttt{edge\_potentials} \to \texttt{marginals}$$

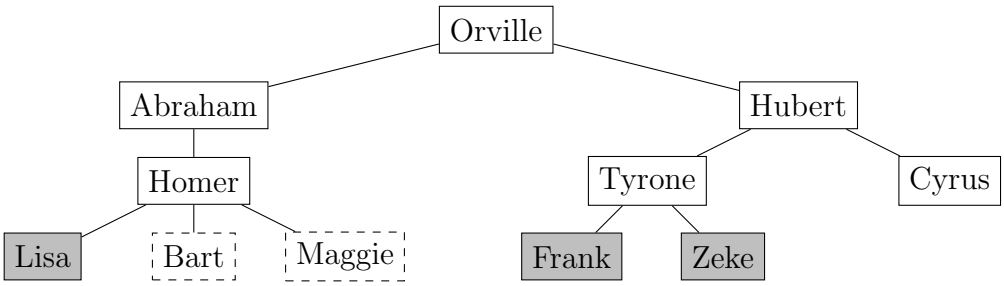
However, the implementation details, including the programming language, are up to you: we will only read your code and we will not run it.

Instead, your submission for this exercise should be a LaTeX document including your answers, any accompanying figures, and a readable picture of your code.

(a) **Genetics**

You have been hired by a new genome sequencing company, 23-Know-Me, to develop a new product. Families should be able to input a family tree and known genetic traits from some members of the family, and receive the probabilities that other members of the family have those traits. These probabilities are to be computed on the basis of proprietary data gathered by 23-Know-Me about the co-occurence of traits between parents and the children.

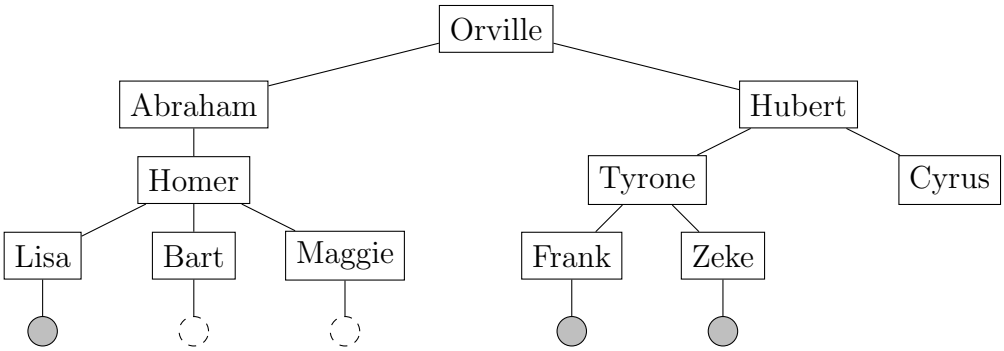
In this problem, you will consider a single binary trait, A, with $A = 1$ indicating a pollen allergy and $A = 0$ indicating no pollen allergy. 23-Know-Me wants you to test your algorithm on the following family tree:



The family members in the dashed boxes have been observed not to have a pollen allergy, while the family members in the grey boxes have been observed to have a pollen allergy.

(i) Assume that for each parent-child pair (i, j), the compatibility function is $\psi_{ij}(A_i, A_j) = \alpha \mathbb{1}_{A_i = A_j} + \mathbb{1}_{A_i \neq A_j}$, with $\alpha > 1$. Find the marginal probability that each family member has a pollen allergy before conditioning on the observed variables and after conditioning on the observed variables, for $\alpha = \{2, 4, 6, 8, 10\}$.

Now assume that the observations of pollen allergies are unreliable: there is an indicator O_i for each family member i that has taken a test for pollen allergies, which is related to the truth A_i by $p(O_i|A_i) = \beta \mathbb{1}_{A_i=O_i} + (1-\beta)\mathbb{1}_{A_i \neq O_i}$, with $\beta \in (.5, 1]$. We represent this using the following graph, where circles indicate test results for pollen allergies, with grey indicating $O_i = 1$ and dashes indicating $O_i = 0$.



(ii) Find the new marginal probability for each A_i after conditioning on the observed variables, for $\alpha = \{2, 4, 6, 8, 10\}$ and $\beta = \{0.8, 0.9\}$.

(b) **Image Denoising** As an avid photographer, you've been taking 1,000s of pictures with your retro black-and-white camera. Unfortunately, it seems the camera is a little too retro: many of your pictures have unwanted noise, like the picture below:

Determined to have clean pictures, you decide to use your inference skills to remove the noise. You think that the true image can be modeled by a ferromagnetic Ising model: neighboring pixels are more likely to have the same value than opposite values, i.e. $\psi_{ij}(Z_i, Z_j) = \alpha \mathbb{1}_{Z_i = Z_j} + \mathbb{1}_{Z_i \neq Z_j}$, with $\alpha > 1$. You don't think there's a bias toward pixels being black or white, so you assume that $\phi_i(Z_i) \equiv 1$.

After reading up about this problem on BinaryPhotography.com, you find that a good model of the noise-generating process is that each pixel independently flips color with probability $\beta \in [0, .5)$, i.e. $p(X_i|Z_i) = (1-\beta)\mathbb{1}_{X_i = Z_i} + \beta \mathbb{1}_{X_i \neq Z_i}$.

With a model in hand, you're ready to de-noise your images by finding the marginal distributions over each pixel. Luckily, for your day job at 23-Know-Me, you have an implementation of the parallel sum-product algorithm on hand. You know it's only supposed to work on trees, but you decide to try it on your problem anyways.

(i) Draw the graphical model corresponding to a 3x3 image with noise as described above, both before observing the noise and after observing the noise. Specify the node and edge potentials for both models.

For this part, you may draw the model by hand and insert the image.

You can find the above image on the class Stellar site. You will run your implementation of parallel sum-product on the graphical model described above, using the values of each pixel of the image as your observed value.

To help report the performance of your algorithm as the parameters vary, write a function that converts the nodewise marginals into a *marginal image*, a grayscale image where the value of each pixel is equal to the probability that the true pixel is black.

(ii) Fixing $\beta = .1$, show the marginal images for $\alpha = \{2, 5, 10\}$.

(iii) Fixing $\alpha = 2$, show the marginal images for $\beta = \{0.05, 0.2, 0.4\}$