Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.438 Algorithms For Inference
Fall 2014

Problem Set 3

Issued: Thursday, September 25, 2014 **Due:** Thursday, October 2, 2014

Suggested Reading: Lecture notes 6–7

Problem 3.1

Let $x_1, x_2, \ldots, x_n$ denote a collection of jointly Gaussian random variables with information matrix $\mathbf{J}$. (Assume that $x_1, x_2, \ldots, x_n$ are nondegenerate, i.e., the covariance matrix is invertible so that $\mathbf{J}$ is well-defined.) $\mathbf{J}$ naturally induces a graph, namely the undirected graph formed by including edges between only those pairs of variables x_i, x_j for which $\mathbf{J}_{i,j} \neq 0$. By the Hammersley-Clifford Theorem, we know that $p_{x_1,x_2,\ldots,x_n}$ is Markov with respect to this graph.

Recall from lecture that given a distribution p_x, an undirected graph $\mathcal{G}$ is said to be a *perfect map* for p_x if p_x is Markov with respect to $\mathcal{G}$, *and* p_x does not satisfy any conditional independence statements that $\mathcal{G}$ does not imply.

(a) For this part only, assume that $\mathbf{J}_{1,2} = 0$ and that $n = 5$.

TRUE or FALSE? Under the previous assumptions, x_1 and x_2 are independent conditioned on x_3, x_4.

If TRUE, provide a proof. If FALSE, provide a counterexample.

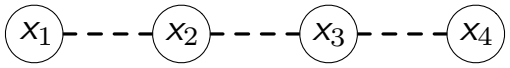
(b) Construct an example matrix $\mathbf{J}$ such that the associated graph is not a perfect map for $p_{x_1,x_2,\ldots,x_n}$ (you are free to choose whatever value of n you want). Your answer should include a proof that your construction works.

(c) (**Practice**) Prove that if the graph that we obtain from $\mathbf{J}$ is a tree, then this tree is always a perfect map for $p_{x_1,x_2,\ldots,x_n}$.

Hint: Show that if $\mathbf{J}$ is a tree, then the matrix $\mathbf{\Lambda} = \mathbf{J}^{-1}$ cannot have any zero entries. Then, use this property to prove the claim. Note that the two steps suggested in the hint are completely independent of each other.

Problem 3.2 (Practice)

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:



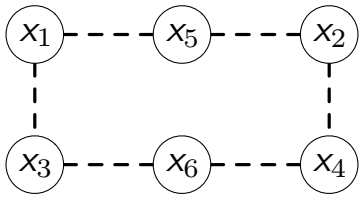
(a) List all conditional and unconditional independencies implied by the covariance graph.

(b) Draw a four-node undirected graphical model that is a minimal I-map of 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:



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

(d) 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 3.3
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 state

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

(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:

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

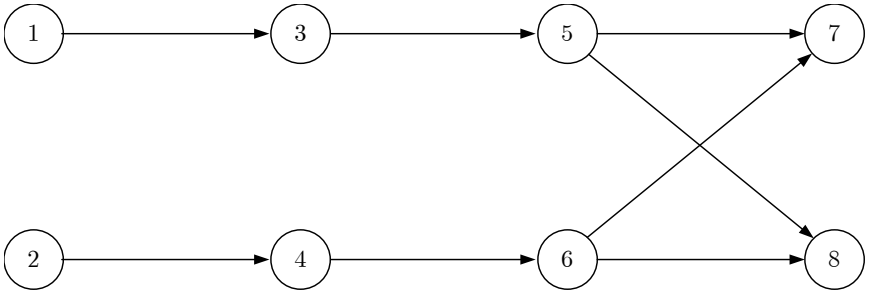
(i) Show that $\mathbf{x}_2$ and $\mathbf{x}_3$ are conditionally independent given $\mathbf{x}_1$ (refer to Problem 1.2 (b)).

(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 3.4

Consider the directed graph shown in the following figure.



(a) What is the corresponding moral graph?

(b) What is the reconstituted graph that results from invoking the UndirectedGraphEliminate algorithm (see Jordan Ch. 3) on the moral graph with the ordering $(8, 7, 6, 5, 4, 3, 2, 1)$?

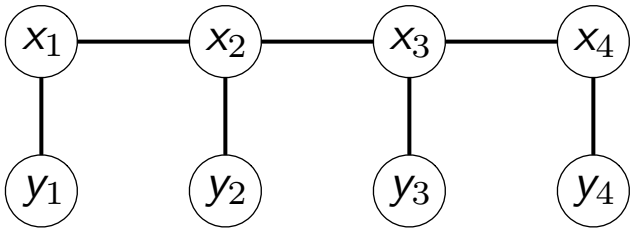
(c) What is the reconstituted graph that results from invoking the UndirectedGraphEliminate algorithm on the moral graph with the ordering $(8, 5, 6, 7, 4, 3, 2, 1)$?

(d) Suppose you wish to use the Eliminate algorithm to calculate $p_{x_1|x_8}(x_1|x_8)$. (Suppose that each x_i is binary and that the local conditionals do not exhibit any special symmetries.) What elimination ordering is optimal? Why?

Problem 3.5 (Practice)

In this problem, we consider the elimination algorithm for inference and its computational complexity.

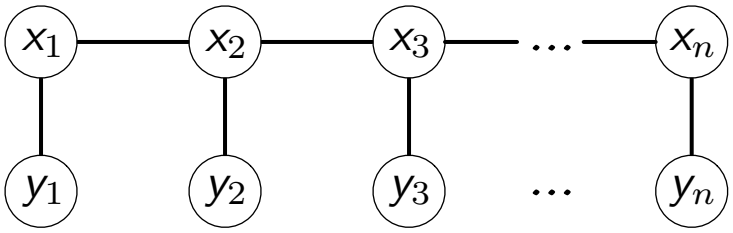
(a) Consider the following graph on 8 nodes.



Draw the reconstituted graph induced by the elimination ordering

$$(x_4, y_4, y_3, x_3, x_1, x_2, y_2, y_1).$$

(b) Now consider a graph on $2n$ nodes as drawn in the following figure, in which every random variable takes on values from a finite alphabet of size k. (That is, $\forall i \in \{1, \ldots, n\},\ x_i \in \mathcal{X},\ y_i \in \mathcal{Y}$ and $|\mathcal{X}| = |\mathcal{Y}| = k$.)

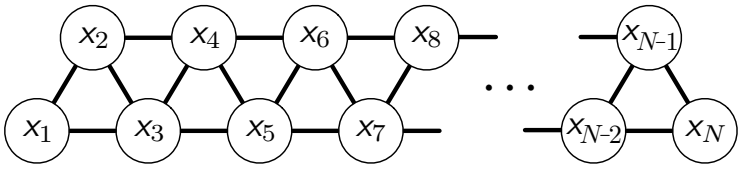


Describe an elimination ordering that requires the least computation and one that requires the most computation. Determine the asymptotic time complexity of the algorithm for each of these orderings with respect to k and n.

(c) Give an example of an undirected graph on n nodes such that the maximal clique size is constant with respect to n (i.e., the maximal clique size does *not* depend on n), but where the computation time required to perform elimination with any ordering is proportional to $k^{\alpha n}$, where k is the size of the alphabet. Specify the value of α for your example. Note that depending on the elimination ordering, your graph may have different values of α. However, your graph should be such that α is lower-bounded by some positive constant across all elimination orderings.

Problem 3.6

Consider the following undirected graphical model over N discrete random variables:



We define clique potentials $\psi_{i,i+1,i+2}(x_i, x_{i+1}, x_{i+2})$ for each triplet of random variables (x_i, x_{i+1}, x_{i+2}) for $i \in \{1, \ldots, N-2\}$.

(a) We run the elimination algorithm on this graph with the ordering $(x_1, x_2, \ldots, x_N)$. Give the list of active potentials just before eliminating node i.

Recall that the list of active potentials contains all potentials (including intermediate computations) which have not yet been processed.

(b) In a message-passing scheme for marginalization on this graph, each node $i \in \{1, \ldots, N-1\}$ sends a forward message τ_i as follows:

$$\tau_1(x_2, x_3) = \sum_{x_1} f_a(\psi_{1,2,3}(x_1, x_2, x_3)),$$

$$\tau_i(x_{i+1}, x_{i+2}) = \sum_{x_i} f_b(\tau_{i-1}(x_i, x_{i+1}),\, \psi_{i,i+1,i+2}(x_i, x_{i+1}, x_{i+2})), \quad i \in \{2, \ldots, N-2\},$$

$$\tau_{N-1}(x_N) = \sum_{x_{N-1}} f_c(\tau_{N-2}(x_{N-1}, x_N)).$$

Determine functions f_a, f_b, and f_c so that $p_{x_n}(x_N) \propto \tau_{N-1}(x_N)$.

(c) Each node $i \in \{N, N-1, \ldots, 2\}$ also sends a backward message η_i as follows:

$$\begin{aligned}
\eta_N(x_{N-1}, x_{N-2}) &= \sum_{x_N} g_a(\psi_{N-2,N-1,N}(x_{N-2}, x_{N-1}, x_N)), \\
\eta_i(x_{i-1}, x_{i-2}) &= \sum_{x_i} g_b(\eta_{i+1}(x_i, x_{i-1}) \ , \ \psi_{i-2,i-1,i}(x_{i-2}, x_{i-1}, x_i)), \quad i \in \{N-1, \ldots, 3\}, \\
\eta_2(x_1) &= \sum_{x_2} g_c(\eta_3(x_2, x_1)).
\end{aligned}$$

Determine functions g_a, g_b, and g_c so that $p_{x_1}(x_1) \propto \eta_2(x_1)$.

(d) We compute the remaining marginal distributions from these messages as follows:

$$\begin{aligned}
p_{x_2} &\propto h_a(\tau_1 \ , \ \eta_3 \ , \ \eta_4), \\
p_{x_i} &\propto h_b(\tau_{i-2} \ , \ \tau_{i-1} \ , \ \eta_{i+1} \ , \ \eta_{i+2}), \qquad\qquad i \in \{3, \ldots, N-2\}, \\
p_{x_{N-1}} &\propto h_c(\tau_{N-3} \ , \ \tau_{N-2} \ , \ \eta_N).
\end{aligned}$$

Determine functions h_a, h_b, and h_c.

(e) Express, in $O(\cdot)$ notation, the minimal complexity of obtaining all singleton marginals via the algorithm we designed in parts (b)-(d). Express your answer in terms of the number of variables N and the alphabet size k. Assume that all N variables are defined over the same alphabet.

Problem 3.7 (Practice)

Let $w_i = (x_i, y_i)$, for $i \in \{1, 2, 3\}$ be a collection of random variables. In this problem, we use the notation $a \leftrightarrow b \leftrightarrow c$ to indicate that a, b, c form a Markov chain.

(a) The relation $w_1 \leftrightarrow w_2 \leftrightarrow w_3$ does not imply both $x_1 \leftrightarrow x_2 \leftrightarrow x_3$ and $y_1 \leftrightarrow y_2 \leftrightarrow y_3$. Show this by constructing an example in the form of a fully labeled undirected graph.

(b) The relations $x_1 \leftrightarrow x_2 \leftrightarrow x_3$ and $y_1 \leftrightarrow y_2 \leftrightarrow y_3$ together do not imply $w_1 \leftrightarrow w_2 \leftrightarrow w_3$. Show this by constructing an example in the form of a fully labeled undirected graph.

For parts (c) and (d) below, we also require that $x_1 \perp\!\!\!\perp y_1$.

(c) Does the statement in part (a) continue to hold? If so, provide a fully labeled undirected graph as an example; if not, explain why not.

(d) Does the statement in part (b) continue to hold? If so, provide a fully labeled undirected graph as an example; if not, explain why not.

Problem 3.8 (Practice)

This problem illustrates some limitations of graphical models in revealing probabilistic relationships among even three random variables.

(a) Let $x \leftrightarrow y \leftrightarrow z$ form a Markov chain such that x and z are independent. Show by example that each of the following is true:

(i) z need not be independent of y

(ii) x need not be independent of y

(iii) z need not be independent of (x, y)

(iv) x need not be independent of (y, z)

(b) Suppose x,y,z are are collection of random variables such that x and y are independent, and x and z are independent.

(i) Show by example that x need not be independent of (y, z)

(ii) Prove that if $x \leftrightarrow y \leftrightarrow z$ also form a Markov chain, then x is independent of (y, z).

6.438 Algorithms for Inference

Fall 2014