PROBABILISTIC GRAPHICAL MODELS - EE 717
Volkan Cevher and Matthias Seeger
Teaching assistants: Hemant Tyagi & Young Jun Ko

Fall 2011

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Homework 4

Assigned: 17/10/2011.
Due: 28/10/2011.

Exercise 1. Expectation-Maximization for Bayesian Networks
In this exercise, you will derive general EM equations in order to learn the CPTs of a Bayesian network over discrete variables, given data cases that are partially incomplete. The data is such that variable values can be missing in some cases, but not in others. Partially incomplete data is a problem that occurs often in practice. In the absence of a proper treatment of missing values (integrating them out), people throw away incomplete cases or set missing entries to default values. The first is wasteful, and the second is just wrong.
The Bayesian network has j variables, x_j has parents $\boldsymbol{x}_{\pi_j}$ (note that π_j could be empty). Let $C_j := \{j\} \cup \pi_j$. The model is

$$P(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{j=1}^{J} P(x_j|\boldsymbol{x}_{\pi_j}, \boldsymbol{\theta}_j),$$

where $\boldsymbol{\theta}_j$ parameterizes the CPT of $P(x_j|\boldsymbol{x}_{\pi_j}, \boldsymbol{\theta}_j)$. There are n data cases $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)}$. Be careful to keep apart $\boldsymbol{x}_{C_j}$ (which is a variable) from $\boldsymbol{x}^{(i)}_{C_j}$ (which is an instantiation of this variable). These cases are not complete: they come with indexes $O_i \subset \{1, \dots, J\}$, $H_i := \{1, \dots, J\} \setminus O_i$, meaning that $\boldsymbol{x}^{(i)}_{O_i}$ is observed, but $\boldsymbol{x}^{(i)}_{H_i}$ is not. From these, define indexes local to each clique by $O_{i,j} := O_i \cap C_j$, $H_{i,j} := H_i \cap C_j$. Before you move on, make sure you understand this notation. For example, $\boldsymbol{x}_{C_j} = (x_j, \boldsymbol{x}_{\pi_j})$, $\boldsymbol{x}^{(i)}_{C_j} = (\boldsymbol{x}^{(i)}_{H_{i,j}}, \boldsymbol{x}^{(i)}_{O_{i,j}})$. Write down a small example for yourself if necessary.

1. [**3 points**] In the E step, we fix $\boldsymbol{\theta}$, and have to compute probabilities $Q_i(\boldsymbol{x}_{C_j})$. Recall that it is enough to compute marginal E-step distributions, instead of a joint $Q_i(\boldsymbol{x})$, because the M-step criterion decouples additively over the cliques:

   $$\begin{aligned}\sum\nolimits_i \mathrm{E}_{Q_i}[\log P(\boldsymbol{x}|\boldsymbol{\theta})] &= \sum\nolimits_i \mathrm{E}_{Q_i}\left[\sum\nolimits_j \log P(x_j|\boldsymbol{x}_{\pi_j}, \boldsymbol{\theta}_j)\right] \\ &= \sum\nolimits_{i,j} \mathrm{E}_{Q_i(\boldsymbol{x}_{C_j})}\left[\log P(x_j|\boldsymbol{x}_{\pi_j}, \boldsymbol{\theta}_j)\right].\end{aligned}$$

   Show how to compute $Q_i(\boldsymbol{x}_{C_j})$ for $i = 1, \dots, n$, $j = 1, \dots, J$.
   Hint: You'll find $Q_i(\boldsymbol{x}_{C_j}) = Q_i(\boldsymbol{x}_{H_{i,j}}|\boldsymbol{x}^{(i)}_{O_i})\mathrm{I}_{\{\boldsymbol{x}_{O_{i,j}}=\boldsymbol{x}^{(i)}_{O_{i,j}}\}}$. What is $Q_i(\boldsymbol{x}_{H_{i,j}}|\boldsymbol{x}^{(i)}_{O_i})$? You can assume to have an algorithm ready for computing posterior marginals, such as in your last recent programming exercise, but what is the model (nodes, potentials) you call it with?

2. [**4 points**] In the M step, whose goal is the CPT update $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}'$, we fix all Q-step distributions Q_i (consider them independent of $\boldsymbol{\theta}'$). Let

   $$n_j(\boldsymbol{x}_{C_j}) := \sum\nolimits_{i=1}^{n} Q_i(\boldsymbol{x}_{C_j}).$$

Show how to compute the new CPTs $P(x_j|\boldsymbol{x}_{\pi_j}, \boldsymbol{\theta}'_j)$ in terms of the counts $n_j(\boldsymbol{x}_{C_j})$. *Hint*: Write the M-step criterion $\sum_i \mathrm{E}_{Q_i}[\log P(\boldsymbol{x}|\boldsymbol{\theta})]$ in the form $\sum_i \sum_j \sum_{\boldsymbol{x}_{C_j}}$. Reorder this (finite) summation, so that $\sum_i$ is innermost. Use the fact (shown in the lecture) that

$$\underset{P(x_j|\boldsymbol{x}_{\pi_j})}{\operatorname{argmax}} \sum_{x_j} \alpha_j(\boldsymbol{x}_{C_j}) \log P(x_j|\boldsymbol{x}_{\pi_j}) = \frac{\alpha_j(\boldsymbol{x}_{C_j})}{\sum_x \alpha_j(\boldsymbol{x}_{\pi_j}, x_j = x)},$$

if $\alpha_j(\boldsymbol{x}_{C_j}) \geq 0$ for all $\boldsymbol{x}_{C_j}$, where the maximization is over all distributions $P(x_j|\boldsymbol{x}_{\pi_j})$ over x_j.

3. [**4 points**] You may have noted a problem with the M step update. What happens if $n_j(\boldsymbol{x}_{C_j}) = 0$ for some j, $\boldsymbol{x}_{\pi_j}$, and all values x_j? Before you move on, make sure you understand when this happens, using a small example if necessary: if for some assignment $\boldsymbol{x}_{\pi_j}$, in every case $\boldsymbol{x}^{(i)}$, there is some $k_i \in \pi_j \cap O_i$ (observed and among parents of j) such that $x_{k_i} \neq x_{k_i}^{(i)}$. If your network contains large cliques (some variables have many parents), and/or many data values are missing, this is not unlikely to happen. For example, think about consumer preference data ("people who bought this book, also looked at ...").
A common remedy is *smoothing*: instead of computing the M-step update in terms of $n_j(\boldsymbol{x}_{C_j})$, we use $n_j(\boldsymbol{x}_{C_j}) + \alpha_j$ for some constants $\alpha_j > 0$. Then, all counts are positive. In this exercise, you'll find an interpretation of smoothing as maximum a posteriori (MAP) estimation (discussed in the lecture).
The *Dirichlet* distribution is a distribution over distributions:

$$\mathcal{D}(\boldsymbol{p}|\boldsymbol{\beta}) \propto \prod_{k=1}^{K} p_k^{\beta_k - 1}, \quad \beta_k > 0,$$

over all $\boldsymbol{p}$ such that $p_k \geq 0$ and $\sum_k p_k = 1$. Define a prior distribution

$$P(\boldsymbol{\theta}) \propto \prod_{j=1}^{J} \prod_{\boldsymbol{x}_{\pi_j}} \mathcal{D}(P(x_j|\boldsymbol{x}_{\pi_j}, \boldsymbol{\theta}_j) \,|\, (\alpha_j + 1)\mathbf{1}).$$

Here, $(\alpha_j + 1)\mathbf{1}$ denotes a vector with constant coefficients equal to $\alpha_j + 1$. Instead of maximizing the log likelihood $\log P(\{\boldsymbol{x}^{(i)}\}|\boldsymbol{\theta})$, we are interested in maximizing the log posterior:

$$\max_{\boldsymbol{\theta}} \log P(\{\boldsymbol{x}^{(i)}\}|\boldsymbol{\theta}) + \log P(\boldsymbol{\theta}).$$

Show that the smoothed EM variant just mentioned can be seen as a method for finding a posterior mode.
Hint: $\log P(\boldsymbol{\theta})$ does not depend on the data, whether observed or latent, so the E-step remains unchanged, and the M-step criterion is the same as above plus $\log P(\boldsymbol{\theta})$. Use the form you derived above (with $\sum_i$ innermost) and marry it with the $\log P(\boldsymbol{\theta})$ terms.

Exercise 2. ITERATIVE PROPORTIONAL FITTING
Recall the iterative proportional fitting (IPF) algorithm from the lecture, a coordinate-ascent algorithm for maximum-likelihood learning of log-linear (undirected) Markov random fields. In this exercise, you will show that this algorithm converges to the unique global solution. Recall the setup:

$$P(\boldsymbol{x}) = Z^{-1} e^{\sum_j \boldsymbol{\theta}_j^T \boldsymbol{f}_j(\boldsymbol{x}_{C_j})}, \quad \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J).$$

Given some completely observed data, we obtain empirical moments

$$\boldsymbol{f}_j(\tilde{\boldsymbol{x}}_{C_j}) := n^{-1} \sum_{i=1}^{n} \boldsymbol{f}_j(\boldsymbol{x}_{C_j}^{(i)}),$$

where $\boldsymbol{x}^{(1)}$, …, $\boldsymbol{x}^{(n)}$ are the data cases, and the negative log likelihood (normalized by $1/n$) is

$$\phi(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta}) - \sum_{j=1}^{J} \boldsymbol{\theta}_j^T \boldsymbol{f}_j(\tilde{\boldsymbol{x}}_{C_j}).$$

$\phi(\boldsymbol{\theta})$ is a jointly convex function, which means that $(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J) \mapsto \phi(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J)$ is convex. Make sure you understand that joint convexity implies that ϕ is convex w.r.t. any subset of its arguments (the converse is not true in general: there are functions $f(\boldsymbol{\theta})$ which are convex w.r.t. to any single $\boldsymbol{\theta}_j$, but are not jointly convex). For jointly convex functions, under mild additional assumptions that we'll ignore here, coordinate descent algorithms provably converge to the unique minimum. Such algorithms are iterative. In each iteration, pick some j and update $\boldsymbol{\theta} \to \boldsymbol{\theta}'$ such that

$$\boldsymbol{\theta}'_k = \boldsymbol{\theta}_k,\ k \neq j, \quad \boldsymbol{\theta}'_j = \mathrm{argmax}\, \phi(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{j-1}, (\cdot), \boldsymbol{\theta}_{j+1}, \dots, \boldsymbol{\theta}_J).$$

The schedule for picking j must be such that each component is picked infinitely often.

1. **[3 points]** Show that IPF is a coordinate descent algorithm on the negative log likelihood $\phi(\boldsymbol{\theta})$.
   Hint: Recall $\nabla_{\boldsymbol{\theta}_j} \log Z$ from the lecture.

When you get into the situation of having to learn a log-linear MRF from data, remember that while IPF is a simple algorithm to write down and code, it can be terribly slow to converge in many cases. For many models and simple implementations, the cost of obtaining all marginals is not much higher than that of obtaining marginals at one clique only (for example, remember the difference between node elimination and belief propagation on trees), and the former computation gets you the complete gradient $\nabla_{\boldsymbol{\theta}} \phi$, which you can feed into off-the-shelf nonlinear optimization code (nonlinear conjugate gradients or Quasi-Newton).

Exercise 3. Completing the Square

In this exercise, you'll work with "completing the square" once more, in order to rapidly determine Gaussian conditional distributions such as the posterior. Suppose we have two variables $\boldsymbol{x}$, $\boldsymbol{y}$, with

$$P(\boldsymbol{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Psi}), \quad P(\boldsymbol{y}|\boldsymbol{x}) = N(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{\Gamma}).$$

This is very important, and I noticed some of you have problems with it. So, please go through this exercise carefully, do it as hinted, and check whether you are comfortable with it. Spend the time to understand why every single step is taken. The world is full of ML people who'll tell you "ah, but it's *just* linear algebra", but then spend five pages with 25 typos for simple Gaussian problems, which is not only embarrassing, but also limits their horizon considerably in terms of the scope of problems they can address.

1. **[3 points]** Work out the posterior $P(\boldsymbol{x}|\boldsymbol{y})$ in a self-contained, yet short way. *Do not use Bayes rule as it stands.*
   Hint: Understand that the joint distribution $P(\boldsymbol{x}, \boldsymbol{y})$ is Gaussian, so that the conditional $P(\boldsymbol{x}|\boldsymbol{y})$ must be Gaussian as well. All you need to determine, is $\boldsymbol{\mu}_{x|y} = \mathrm{E}[\boldsymbol{x}|\boldsymbol{y}]$ and $\boldsymbol{\Sigma}_{x|y} = \mathrm{Cov}[\boldsymbol{x}|\boldsymbol{y}]$. Next, write $P(\boldsymbol{x}, \boldsymbol{y}) = C_1 e^{-(1/2)q(\boldsymbol{x},\boldsymbol{y})}$, where C_1 does not depend on $\boldsymbol{x}$, $\boldsymbol{y}$. Understand that $q(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{y}|\boldsymbol{x}) + q(\boldsymbol{x})$, where $q(\boldsymbol{y}|\boldsymbol{x})$ comes from $P(\boldsymbol{y}|\boldsymbol{x})$, $q(\boldsymbol{x})$ from $P(\boldsymbol{x})$, but just as well $q(\boldsymbol{x}, \boldsymbol{y}) = q(\boldsymbol{x}|\boldsymbol{y}) + q(\boldsymbol{y})$. Also note that $q(\boldsymbol{y}|\boldsymbol{x})$ and $q(\boldsymbol{x})$ come in the centered form you expect from the Gaussian density form:

$$q(\boldsymbol{x}) = (\boldsymbol{x} - \underbrace{\boldsymbol{\mu}}_{\text{Mean}})^T(\underbrace{\boldsymbol{\Psi}}_{\text{Cov}})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \quad q(\boldsymbol{y}|\boldsymbol{x}) = (\boldsymbol{y} - \boldsymbol{W}\boldsymbol{x})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{x}).$$

Importantly, understand that if you massage $q(\boldsymbol{x}, \boldsymbol{y})$ into the form $q(\boldsymbol{x}|\boldsymbol{y}) + q(\boldsymbol{y})$, so that $q(\boldsymbol{x}|\boldsymbol{y})$ has such a centered form as well, you can just read of $\boldsymbol{\mu}_{x|y}$ and $\boldsymbol{\Sigma}_{x|y}$. This is the basis for "completing the square".
Here, we are after $P(\boldsymbol{x}|\boldsymbol{y})$, so write out $q(\boldsymbol{x}, \boldsymbol{y})$, and look at it *as a function of* $\boldsymbol{x}$ *only*: there is a quadratic part $(\boldsymbol{x}^T(\dots)\boldsymbol{x})$ and a linear part $(\boldsymbol{x}^T(\dots))$. This must be $q(\boldsymbol{x}|\boldsymbol{y})$ up to a constant part, because $q(\boldsymbol{y})$ does not depend on $\boldsymbol{x}$. Determining this constant part, or equivalently bringing $q(\boldsymbol{x}|\boldsymbol{y})$ into centered form, is what is called "completing the square". In order to figure this out, it's easiest to work backwards: write down what we want to get:

$$q(\boldsymbol{x}|\boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{\mu}_{x|y})^T \boldsymbol{\Sigma}_{x|y}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_{x|y}) = \boldsymbol{x}^T \boldsymbol{\Sigma}_{x|y}^{-1} \boldsymbol{x} - 2\boldsymbol{x}^T \boldsymbol{\Sigma}_{x|y}^{-1} \boldsymbol{\mu}_{x|y} + \boldsymbol{\mu}_{x|y}^T \boldsymbol{\Sigma}_{x|y}^{-1} \boldsymbol{\mu}_{x|y}.$$

Therefore, if

$$q(\boldsymbol{x}, \boldsymbol{y}) = \underbrace{\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}}_{\text{quad. part}} - \underbrace{2\boldsymbol{x}^T \boldsymbol{b}}_{\text{lin. part}} + C_2,$$

C_2 independent of $\boldsymbol{x}$, by matching the expressions, you see directly that $\boldsymbol{\Sigma}_{x|y} = \boldsymbol{M}^{-1}$, and $\boldsymbol{\mu}_{x|y} = \boldsymbol{M}^{-1}\boldsymbol{b}$. Once you understood these steps, it is best to memorize the recipe:

- Write down joint $P(\boldsymbol{x}, \boldsymbol{y})$, or better $-2 \log P(\boldsymbol{x}, \boldsymbol{y})$, dropping terms independent of $\boldsymbol{x}$, $\boldsymbol{y}$
- Identify quadratic part $\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}$. At this point, you already have $\mathrm{Cov}[\boldsymbol{x}|\boldsymbol{y}] = \boldsymbol{M}^{-1}$
- Identify linear part $-2\boldsymbol{x}^T \boldsymbol{b}$. In order to complete the square, write it as $-2\boldsymbol{x}^T \boldsymbol{M}(\boldsymbol{M}^{-1}\boldsymbol{b})$. At this point, you already have $\mathrm{E}[\boldsymbol{x}|\boldsymbol{y}] = \boldsymbol{M}^{-1}\boldsymbol{b}$

You should now be able to go back to the "completing the square" exercise two sheets earlier, where the goal (translated into notation here) was to collect the leftover $q(\boldsymbol{y})$ after stripping away $q(\boldsymbol{x}|\boldsymbol{y})$ (*including its constant part* $\boldsymbol{b}^T \boldsymbol{M}^{-1} \boldsymbol{b}$), and match it against the known form of $P(\boldsymbol{y})$.