Stat521A Spring 2009: homework 3

1 Pairwise independence does not imply mutual independence

We say that two random variables are **pairwise independent** if

$$p(X_2|X_1) = p(X_2) \tag{1}$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{2}$$

We say that n random variables are **mutually independent** if

$$p(X_i|X_S) = p(X_i) \quad \forall \subseteq \{1, \ldots, n\} \setminus \{i\} \tag{3}$$

and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{4}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example. Hint: consider $C = \text{xor}(A, B)$ for independent binary variables A and B.

2 Conditional independence iff joint factorizes

In the book we said $X \perp Y|Z$ iff

$$p(x, y, z) = p(x|z)p(y|z) \tag{5}$$

for all x, y, z such that $p(z) > 0$. Now prove the following alternative definition: $X \perp Y|Z$ iff there exist function g and h such that

$$p(x, y|z) = g(x, z)h(y, z) \tag{6}$$

for all x, y, z such that $p(z) > 0$.

3 Moralization does not introduce new independence statements

Recall that the process of moralizing a DAG means connecting together all "unmarried" parents that share a common child, and then dropping all the arrows. Let M be the moralization of DAG G. Show that $CI(M) \subseteq CI(G)$, where CI are the set of conditional independence statements implied by the model.

4 Gaussian DAGs vs Gaussian MRFs

Consider a Gaussian DAG in which each CPD has the form

$$X_j = \mu_j + \sum_{k \in \pi_j} w_{jk}(X_k - \mu_k) + \sqrt{v_j} Z_j \tag{7}$$

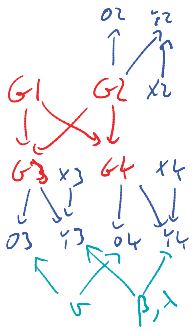


Figure 1: A DGM of a small family tree. G1 and G2 are the parents of G3 and G4.

Let $\mu_j = 0$ and $v_j = 1$ for all j. Define the matrix

$$\mathbf{T} = \begin{pmatrix} 1 & & & & \\ -w_{21} & 1 & & & \\ -w_{32} & -w_{31} & 1 & & \\ \vdots & & & \ddots & \\ -w_{d1} & -w_{d2} & \ldots & -w_{d,d-1} & 1 \end{pmatrix}$$

Let $\mathbf{M} = \mathbf{T}^T$ be upper triangular. Then, as we showed in lecture 5, we have $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = \mathbf{M}\mathbf{M}^T$ (since $\mathbf{D} = \text{diag}(v_j) = \mathbf{I}$). Note that $\mathbf{M}_{:,j}$ are the weights into node j, and $\mathbf{M}_{i,:}$ are the weights out of node i.
Prove or disprove the following statement:

$$M_{i,j} = 0 \quad \forall j > i \quad \implies \Omega_{i,j} = 0 \tag{8}$$

5 EM for family-based genetic association

Consider the DGM in Figure 1. G_i represents the true but unknown genotype of person i, and has 3 states: AA (homozygous dominant), Aa (heterozygous), and aa (homozygous recessant). We see that G_1 and G_2 are the parents of G_3 and G_4. We assume the prior on the root nodes, $p(G_1)$ and $p(G_2)$, is equal to

$$p(G_1) = [p^2, 2p(1-p), (1-p)^2] \tag{9}$$

for known p. Furthermore, the CPDs $p(G_3|G_1, G_2)$ and $p(G_4|G_1, G_2)$ are given by Mendel's laws. Thus there are no unknown parameters in the CPDs for G_3 or G_4.

O_i is a noisy measurement of G_i and also has 3 states. For simplicity, we assume a tabular CPD

$$p(O_i = o|G_i = g, \boldsymbol{\nu}) = \nu_{g,o} \tag{10}$$

Note that measurements are not available for person 1 (who was just introduced into the model to simplify the likelihood).
In addition to genotype information, we have phenotype information. Let $Y_i \in \mathbb{R}$ be some response variable we are studying (e.g., blood pressure), and let $\mathbf{x}_i \in \mathbb{R}^d$ be some vector of covariates (e.g, age, weight, etc). The response also depends on the genotype. We assume the following linear model,

$$p(y_i|\mathbf{x}_i, G_i, \lambda, \mathbf{w}) \quad = \quad \mathcal{N}(y_i|\mathbf{w}^T[I(G_i = 1)\ I(G_i = 2)\ I(G_i = 3)\ \mathbf{x}_i], \lambda^{-1}) \tag{11}$$

(Note that we could replace one of the $I(G_i = k)$ terms with a constant of 1, due to the sum-to-one constraint.) Thus $\mathbf{w}$ is $d + 3$ dimensional.
The overall model is

$$p(\mathbf{y}, \mathbf{O}, \mathbf{G}|\boldsymbol{\theta}) \quad = \quad \prod_{i=1}^{2} p(G_i) \prod_{i=3}^{4} p(G_i|G_1, G_2) \prod_{i=2}^{n} p(y_i|\mathbf{x}_i, G_i, \lambda, \mathbf{w}) p(o_i|G_i, \boldsymbol{\nu}) \tag{12}$$

where $\boldsymbol{\theta} = (\mathbf{w}, \lambda, \boldsymbol{\nu})$ are all the parameters.

1. Write down the expected complete data log likelihood, where expectations are wrt the hidden G variables. Hint: the $p(\mathbf{G})$ term is independent of $\boldsymbol{\theta}$ and can be dropped. Also, we can rewrite $p(o_i|G_i, \boldsymbol{\nu})$ as follows:

$$p(o_i|G_i, \boldsymbol{\nu}) = \prod_{g=1}^{3} \prod_{o=1}^{3} \boldsymbol{\nu}_{g,o}^{I(G_i=g, o_i=o)} \tag{13}$$

2. What expected sufficient statistics need to be computed in the E step?

3. What update equations should you use in the M step? (For simplicity, just compute the MLE rather than a MAP estimate.)

6 Mean field for image denoising

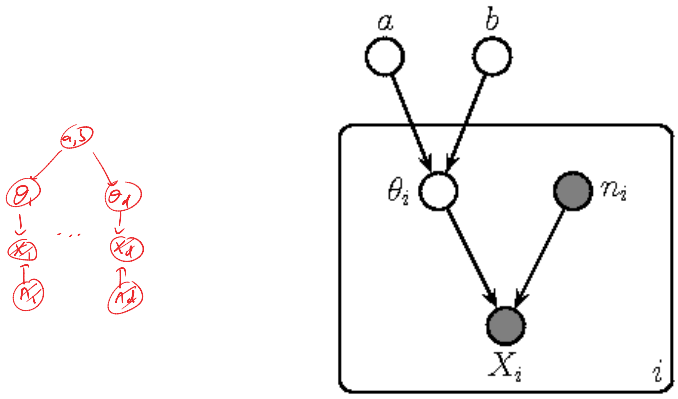
Modify the PMTK function `gibbsIsingImageDenoiseDemo` to use mean field inference instead of Gibbs sampling. Plot the posterior mean estimate of the latent image for different numbers of iterations of mean field updates.

7 Gibbs sampling for the related cancer rates model

Consider the following example from (**?**, p24). Suppose we measure the number of people at risk of cancer, n_i, in various cities, and the number of people who died in these cities, x_i. We assume $x_i \sim \text{Bin}(n_i, \theta_i)$, and we want to estimate the θ_i. One approach is to estimate them all separately, but this will suffer from the sparse data problem (underestimation of the rate of cancer due to small n_i). Another approach is to assume all the θ_i are the same; this is called **parameter tying**. The resulting **pooled** MLE is just $\hat{\theta} = \frac{\sum_i x_i}{\sum_i n_i}$. A compromise approach is to assume that the θ_i are similar, but that there may be city-specific variations. This is modeled by assuming the θ_i are drawn from some common distribution, say $\theta_i \sim \text{Beta}(a, b)$. If a and b were fixed, then the θ_i would be conditionally independent, and there would be no borrowing of statistical strength. So instead we will estimate a and b, as well as the θ_i.
Our assumptions are illustrated in Figure 2. Thus the full joint distribution can be written as

$$p(\mathbf{x}, \mathbf{n}, \boldsymbol{\theta}, a, b) \quad = \quad \prod_{i=1}^{n} p(x_i|n_i, \theta_i) p(\theta_i|a, b) p(a, b) \tag{14}$$

$$\quad = \quad \prod_{i=1}^{n} \text{Bin}(x_i|n_i, \theta_i) \text{Beta}(\theta_i|a, b) p(a) p(b) \tag{15}$$



(a) (b)

Figure 2: (a) Hierarchical model of cancer rates represented as a directed graphical model. (b) Same as (a), but using plate notation.

Let us put vague Gamma(0.01,0.01) priors on both a and b. The data is shown below.

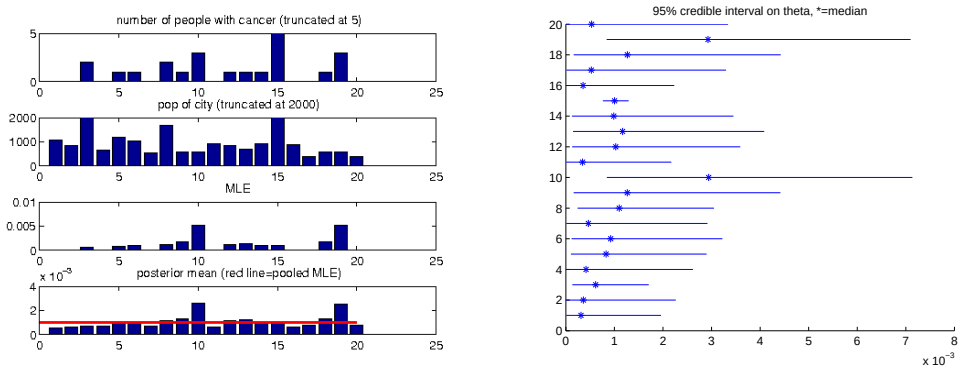
```
x = [0 0 2 0 1 1 0 2 1 3 0 1 1 1 54 0 0 1 3 0];
n = [1083 855 3461 657 1208 1025 527 1668 583 582 917 857 ...
    680 917 53637 874 395 581 588 383];
```

Perform Gibbs sampling in this model,[1] and estimate the posterior means $E[\theta_i|\mathcal{D}]$. Use these to create a plot similar to Figure 3(a).[2] Also, plot the posterior 95% credible intervals for $p(\theta_i|\mathcal{D})$, as in Figure 3(b).[3] Include your source code and results.

[1] Since the Gamma prior is not conjugate to the Beta likelihood, sampling from the full conditionals $p(a|b, \boldsymbol{\theta})$ and $p(b|a, \boldsymbol{\theta})$ is a bit tricky. You could use Metropolis within Gibbs. But the easiest way is to probably use BUGS, either directly, or calling it from R or Matlab. (This requires that you install BUGS on your computer, which is easiest if you are using Windows.) For an example of how to use BUGS to solve a very similar problem, for modeling related Gaussian random variables, see `http://www.cs.ubc.ca/~murphyk/Software/MATBUGS/schools_writeup/schools_writeup.html`.

[2] We see that the posterior mean is shrunk towards the pooled estimate more strongly for cities with small sample sizes n_i. For example, city 1 and city 20 both have a 0 observed cancer incidence rate, but city 20 has a smaller population, so its rate is shrunk more towards the population-level estimate (i.e., it is closer to the horizontal red line).

[3] We see that the cities with the highest posterior median rates of cancer are 10 and 19, which are also the ones with the highest MLE. However, we see that the posterior uncertainty is large. Conversely, for city 15, which has a very large number of measured people (53,637 people), the posterior uncertainty is very small. Consequently this city has the largest impact on the posterior estimate of a and b.



(a) (b)

Figure 3: (a) First row: Number of cancer incidents x_i in 20 cities in Missouri. Second row: population size n_i. The largest city (number 15) has a population of $n_{15} = 53637$ and $x_{15} = 54$ incidents, but we truncate the vertical axes of the first two rows so that the differences between the other cities are visible. Third row: MLE $\hat{\theta}_i$. Fourth row: posterior mean $E[\theta_i|\mathcal{D}]$. The red line is $E[a/(a+b)|\mathcal{D}]$, the population-level mean. (b) Posterior 95% credible intervals on the cancer rates. Figure created using `ebCancerExample`, which uses empirical Bayes. Visually identical results can be obtained using `mhMissouriCancer`, which uses MCMC.