IEOR E4703 Monte-Carlo Simulation
Martin Haugh Due: 12.55pm Thursday 27 April 2017

Assignment 8: Answer Q2, Q4 and one other question

1. **(Estimating Genotype Frequencies (from Robert & Casella))**
The data in the table below displays observed genotype frequencies from blood-type data. The effect of a dominant *allele* creates a missing data problem. For example, if someone has genotype AA or AO then only A will be observed. Similarly, a person with genotype BB or BO will result in an observation of B. The goal of the study is to estimate the unknown parameters p_A, p_B and $p_O = 1 - p_A - p_B$.

Genotype	Probability	Observed	Probability	Frequency
AA	p_A^2	A	$p_A^2 + 2p_Ap_O$	$n_A = 186$
AO	$2p_Ap_O$			
BB	p_B^2	B	$p_B^2 + 2p_Bp_O$	$n_B = 38$
BO	$2p_Bp_O$			
AB	$2p_Ap_B$	AB	$2p_Ap_B$	$n_{AB} = 13$
OO	p_O^2	AO	p_O^2	$n_O = 284$

We will assume a Dirichlet $D(\alpha_1, \alpha_2, \alpha_3)$ prior so that

$$\pi(p_A, p_B) \propto p_A^{\alpha_1 - 1} p_B^{\alpha_2 - 1} (1 - p_A - p_B)^{\alpha_3 - 1}. \tag{1}$$

(a) Write out the likelihood of the data and the posterior distribution. How would you estimate p_A, p_B and p_O using an MCMC algorithm?

(b) We can also view this model as a *missing data* problem. Describe the likelihood in this missing data formulation.

Hint: Let z_{AA}, z_{AO}, z_{BB} and z_{BO} denote the number of AA, AO, BB and BO genotypes, respectively, in the data-set. They are unobserved whereas $n_A := z_{AA} + z_{AO}$ and $n_B := z_{BB} + z_{BO}$ are observed.

(c) Derive a Gibbs sampler to sample from your posterior. In particular, what are the various conditional distributions?

(d) Implement your Gibbs sampler and after discarding a suitable number of burn-in samples, plot histograms of the marginal distributions of p_A, p_B and p_O. (One should generally perform some convergence diagnostics as well so feel free to do this. It should be easy to recycle your code from the last assignment.)

2. **(Gibbs Sampling in a DAG After Observing Some Nodes)**
In the case of a general acyclic directed graph (DAG) the lecture slides claim that

$$p(x_i \mid \mathbf{x}_{-i}) = \frac{1}{Z}\, p(x_i \mid pa(x_i)) \prod_{j \in ch(i)} p(x_j \mid pa(x_j)) \tag{2}$$

where $pa(x_i)$ and $ch(i)$ are the *parent* and *children* nodes, respectively, of x_i, and Z is the normalization constant

$$Z = \sum_{x_i} p(x_i \mid pa(x_i)) \prod_{j \in ch(i)} p(x_j \mid pa(x_j)).$$

(a) Prove (2). *Hint:* Use the representation of the joint distribution of a DAG that we gave in the lecture notes.

(b) Suppose now a subset of the nodes $E_{Obs} \subset \{x_1, \ldots, x_K\}$ have been observed. Let $E_{UnObs} := \{x_1, \ldots, x_K\} \setminus E_{Obs}$ denote the unobserved nodes. Explain clearly how you could use Gibbs sampling to simulate from $p(E_{UnObs} \mid E_{Obs})$.

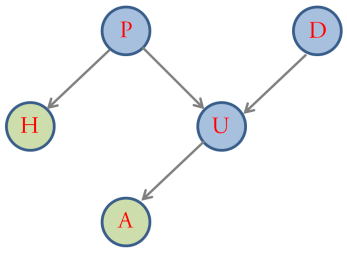
3. **(Exercise 3.1 from Barber: the Party Animal)**
Consider the following list of binary random variables:

- $P \in \{0, 1\}$: worker went to a party
- $D \in \{0, 1\}$: worker is not motivated at work
- $H \in \{0, 1\}$: worker has headache
- $U \in \{0, 1\}$: worker under-performing at work
- $A \in \{0, 1\}$: boss angry

Suppose the joint distribution of these random variables decomposes as follows.

$$p(A, D, H, P, U) = p(P)p(D)p(H \mid P)p(U \mid D, P)p(A \mid U)$$

This distribution corresponds to the following Bayes Network



To complete the above specification, assume that

- $p(P = 1) = 0.2$
- $p(D = 1) = 0.4$
- $p(A = 1 \mid U = 1) = 0.95$, $p(A = 1 \mid U = 0) = 0.5$
- $p(H = 1 \mid P = 0) = 0.2$, $p(H = 1 \mid P = 1) = 0.9$
- $p(U = 1 \mid P = 1, D = 1) = 0.999$, $p(U = 1 \mid P = 1, D = 0) = 0.9$, $p(U = 1 \mid P = 0, D = 1) = 0.9$, $p(U = 1 \mid P = 0, D = 0) = 0.01$.

It can be shown (via tedious calculations) that $p(P = 1 \mid H = 1, A = 1) = \mathbf{0.6097}$. You will confirm this via MCMC in parts (a) and (b) below! (In these parts there is no need to run convergence diagnostics since we know the answer in advance but note that in general, convergence diagnostics should always be performed.)

(a) Write a Metropolis-Hastings based MCMC code that generates samples from the conditional distribution

$$\mathbb{P}(P, D, U \mid H = 1, A = 1) \propto \mathbb{P}(P, D, U, H = 1, A = 1)$$

where your proposal distribution is one that flips the bit value of a randomly chosen element from the set $\{P, D, U\}$. For example, if D is the randomly chosen element and its current value if 0 then its value in the next sample (if accepted) will be 1 with the values of P and U unchanged.

Compute the average of the samples for P, and check that it converges to the value that you computed in part (a). (The more diligent among you might want to implement the Gelman-Rubin approach to diagnose when convergence to stationarity has occurred. In general you should do this!!)

(b) Write a Gibbs sampler MCMC code that generates samples from the conditional distribution

$$\mathbb{P}(P, D, U \mid H = 1, A = 1) \propto \mathbb{P}(P, D, U, H = 1, A = 1).$$

Cycle through the variables $\{P, D, U\}$ in a round robin fashion. Note that conditioned on all other variables, the transition probability of a binary variable X is easy to compute. In particular,

$$\mathbb{P}(X = 1 \mid X^c = x^c) \propto \mathbb{P}(X = 1, X^c = x^c),$$

where X^c denotes all the other variables other than X, and $X^c = x^c$ denotes that all these other binary variables are set to the values in x^c.

Compute the time average of the samples for P, and see if it converges to the value that you computed in part (a).

4. **(Empirical Bayes)**
Assume $X \mid \theta$ is exponential with density $f(x \mid \theta) = e^{-x/\theta}/\theta$ and corresponding CDF, $F(x \mid \theta)$. Let $\pi(\theta)$ be the prior on θ. Then $m_\pi(x) := \int_\Theta f(x \mid \theta)\pi(\theta)\, d\theta$ is the marginal density of X and $M_\pi(x) := \int_0^x m_\pi(x)\, dx$ is the corresponding CDF.

   (a) Show that $\theta = (1 - F(x \mid \theta))/f(x \mid \theta)$.

   (b) Show that Bayes estimator of θ with respect to π is

$$\delta(x) := \frac{1 - M_\pi(x)}{m_\pi(x)}.$$

   (c) Suppose you observe $X_i \mid \theta_i$ for $i = 1, \ldots, n+1$. Explain how you would estimate θ_{n+1} in the empirical Bayes fashion, using the result in (b).