MIDTERM EXAM

Answer any three of the four questions. You can take up to 2 hours for the exam. It must be handed in by 6:30PM today, March 12, either by email to sims@princeton.edu or on paper at my office, 209 JRRB.

(1) We wish to use MCMC to generate a sample from a posterior distribution proportional to $f(\theta)$, where θ is one-dimensional and we have an efficient program for computing $f(\theta)$ for any θ. we are considering possible "jump distributions" for a Metropolis MCMC algorithm. The notation is that θ_j is the draw at the j'th iteration, $\hat{\theta}_{j+1}$ is the proposal for θ_{j+1}, before the accept/reject rule has been applied, and κ is a normalizing constant that makes the density integrate to one. In describing the jump pdf we use θ' for $\hat{\theta}_{j+1}$ and θ for θ_j. Here are four possible jump pdf's:

i) $\dfrac{\kappa}{e^{\theta'-\theta}+e^{2(\theta-\theta')}}$

ii) $\dfrac{\kappa}{e^{\theta'-\theta}+e^{\theta-\theta'}}$

iii) $e^{\theta-\theta'}$ for $\theta' > \theta$, 0 for $\theta' < \theta$

iv) uniform on $(\theta - .3, \theta + .3)$

(a) Which (if any) of these jump distributions would work to make the Metropolis algorithm have the fixed point property: that if θ_j is a draw from the target density, θ_{j+1} will be also.
The jump distribution density has to be invariant to switches of its two arguments for Metropolis to work. (i) and (iii) fail this requirement. Many students thought (iv) was not symmetric. But its pdf is $\mathbf{1}[|\theta - \theta'| < .3]$, which is certainly symmetric.

(b) For the ones that do satisfy the fixed point property (if any do) are there forms of f for which the jump distribution would clearly not lead to convergence to the target?
(iv) is symmetric, but if the target f has $f(\theta) = 0$ over any interval of length greater than or equal to .3, the M-H draws will not cross the gap. Nonetheless, such draws do still formally have the fixed point property — if our initial draw is from f, the next will also have f as marginal distribution. It's just that the fixed point property by itself does not guarantee convergence of the algorithm.

(c) Suppose that instead of the Metropolis algorithm we use Metropolis-Hastings, but with the same set of possible jump distributions. For each jump distribution, specify how your answer would change (or that it would remain the same).
M-H does not require symmetry of the jump distribution. It uses in its accept/reject decision the ratio of $f(\theta')/p(\theta',\theta)$ to $f(\theta)/p(\theta,\theta')$, where $p(\theta',\theta)$ is the jump density from which θ' is drawn. In the Metropolis algorithm, the same decision rule for

Date: May 10, 2020.

accept/reject is based simply on $f(\theta')/f(\theta)$. Since for (ii) and (iv) the jump distribution is symmetric, the answer is the same. With symmetry of the jump distribution, Metropolis is identical to M-H. (i) is asymmetric, but positive over the whole real line. It is therefore has the fixed-point property in M-H and would be likely to converge. (iii) is very asymmetric, and is zero over half the real line. Its jumps can only be in one direction, so it obviously can't lead to convergence. Nonetheless, it has the fixed point property. It will always have $p(\theta', \theta)$ positive, and will always have $p(\theta, \theta') = 0$. Thus $f(\theta)/p(\theta, \theta') = \infty$, meaning that the new draw is never accepted. The M-H algorithm with this jump density repeats the initial draw forever. But if the initial draw was from a marginal distribution matching f, all the other (identical) draws also have that marginal distribution, so the fixed-point property is satisfied.

(2) We have a sample of unemployment spells of length $\{t_i, i = 1, \ldots, N\}$ and for each observation we have a single covariate x_i. The x_i values are all positive. Our model for the length of the unemployment spells is that they are i.i.d. across i and each has an exponential distribution with pdf

$$x_i \beta e^{-x_i \beta t_i} .$$

However, some of the spells are top-coded, meaning that the actual t_i is larger than some cutoff value T, and that what we have for that observation is only T, indicating that it has been top-coded, not the actual t_i, together with x_i. Describe a Gibbs sampling algorithm that treats both the t_i values for top-coded observations and β as part of the unknown parameter vector and delivers, upon convergence, a sample from the joint posterior on β and the topcoded t_i values. It may be helpful to recall that a Gamma(n, a) pdf is proportional to $a^n z^{n-1} e^{-az}$, and is easy to sample from.

The distribution of a single observation is a continuous density over $(0, T)$ equal to $x\beta \exp(-x\beta t)$, plus a weight of $\int_T^\infty x\beta e^{-x\beta t}\, dt = e^{-x\beta T}$ on the point $t = T$. Thus the likelihood of the full sample using N for the full sample size and n for the number of top-coded observations, is

$$\Big(\prod_{t_i < T} x_i\Big) \beta^{N-n} e^{-\beta \sum x_i t_i} ,$$

where the $\sum x_i t_i$ term is a sum over all observations, with $t_i = T$ for the topcoded ones.

The problem did not state what prior should be used for β. An exponential or Gamma prior would be easy to handle, but even simpler is a flat prior, which we'll use in this answer.

Note that if we're just interested in inference about β, there's no need for MCMC. The likelihood is, as a function of β, proportional to a $\Gamma(N - n + 1, \sum_i x_i t_i)$ density. That is, it has the same form as if there were no top-coding, except that the shape parameter, in the Gamma is reduced by n. So the posterior mean for β is $(N - n + 1)/\sum x_i t_i$ rather than $(N + 1)/\sum x_i t_i$.

This analytic form for the posterior on β means we could do better than Gibbs sampling. We could make i.i.d. draws directly from the Gamma posterior on β, then for each drawn β_j make draws for all the top-coded observation t_i's from their distributions conditional on

β and $t_i > T$. For observation i this conditional distribution is exponential on (T, ∞), with rate parameter $x_i\beta_j$.

But the question asked about Gibbs sampling. For Gibbs sampling, one would draw from the conditional distributions of the top-coded t_i's conditional on a draw β_j, then, rather than drawing β_{j+1} from its marginal, drawing it from the $\text{Gamma}(N+1, \sum x_i t_i)$ that is its conditional density given all the t_i's (now using the drawn t_i's for the top-coded observatins rather than T). This will give us an MCMC sample that converges to the true posterior density for β and for $\{t_i \mid t_i > T\}$. Because its draws for β_i will, like most MCMC draws, be serially correlated, the draws will converge more slowly than if we drew each time directly from the marginal posterior on β.

(3) We have data on a sample of women in age groups 20-24, 25-30, and 30-35. For each observation we have the age group, marital status (1 for married, 0 otherwise), and the state of residence, which is either New Jersey or Pennsylvania. We calculate the fraction married in the sample for each of the six age and state groups, but the sample is of modest size, and we wonder whether it is really necessary to distinguish New Jersey and Pennsylvania.

(a) Explain how to do a likelihood ratio test of the hypothesis that the probabilities of marriage don't depend on state of residence.

We have three age groups and two states, for a total of six groups. In each group we observe n_{ij} the number of married women in age group i and state j, and N_{ij}, the total number of women in the group. Clearly the probability of marriage is likely to depend on age, and we have made no assumption that the age distribution is the same across states. so it makes sense to test the hypothesis that within each age group the probabilities of marriage are the same across states. So the unknown parameters are p_{ij}, $i = 1, 2, 3$, $j = 1, 2$ and the likelihood is

$$\prod_{i,j} p_{ij}^{n_{ij}} (1 - p_{ij})^{N_{ij} - n_{ij}} .$$

The MLE is then $p_{ij} = n_{ij}/N_{ij}$, all i, j. The null hypothesis is that $p_{i1} = p_{i2}$ for all i. Under that restriction there are just three parameters, p_i, $i = 1, 2, 3$, and for each age group the two state subgroups are combinded, so the MLE has $p_i = (n_{i1} + n_{i2})/(N_{i1} + N_{i2})$. The likelihood ratio test then compares twice the log likelihoods at their MLE's to a χ^2 distribution with 3 degrees of freedom (6 parameters, reduced to 3 under the null).

To form a likelihood, one has to have a model. The model above is a "fully saturated" model, plus a restricted version of it. It was a common choice by students on the exam. However, other models were possible. Here are hree other interpretations of the question:

(i) Linear probability model, with two dummy variables, one for state and one for age group on the right, and a married dummy on the left. To use this to form a likelihood ratio, you needed to use ML on it, which is not the same as OLS, and most (not all) who used this model were vague or silent about how to maximize the likelihood. The LR just restricted the coefficient of the state dummy to zero.

This model could not be used in the last part, evaluating the probability that the probability of marriage is higher in PA in all age categories.

(ii) A simple model that ignores the age group data. This could be an LPM with only one right-hand-side variable, or simply using the $p_j^n(1-p_j^n)$ as the likelihood for each cell j, but with j indexing only states, not state x age. This is OK in part (a), but again is not usable for (c).

(iii) Modeling each of the 12 married x state x age cells as having its own p_{msa} probability, with those probabilities summing to one. Then the likelihood, and posterior on the probabilities with a flat prior, has the form of the Dirichlet. This is an elegant approach, works on all parts of the problem, and validates the hint which was unhelpful for the other models. The drawback of this approach is that specifying how the draws from the 12-parameter posterior should be used to answer the question was less transparent. And the sole person who used this modeling approach abandoned it in part (c).

(b) Explain how to apply the BIC (i.e. Schwarz) criterion to the same hypothesis.
The BIC compares twice the difference in log likelihoods not to a tail quantile of the $\chi^2(3)$ distribution, but to $3\log T$, where T is sample size.

(c) Explain how to use MCMC (or direct calculation, if it looks feasible) to evaluate the posterior probability, under a flat prior, that the probabilities of marriage are higher in Pennsylvania than New Jersey in all age categories. For this part, it might be helpful to recall that if $\{z_1, \ldots, z_n\}$ are in the unit simplex (i.e. all positive and summing to one), then

$$p(z \mid \alpha) = \prod_{i=1}^{n} z_i^{(\alpha_i - 1)}$$

is proportional to the Dirichlet pdf, which is easy to sample from.
The hint to this part was an unintentional red herring if you took the most common modeling approaches. If, as most did, you directly model the conditional probabilities of marriage given state and age, or just given state, the full set of p_{ij}'s does not sum to one, so they are not jointly Dirichlet in the posterior. Instead, each p_{ij} is independent in the posterior, with p_{ij} distributed as a $\text{Beta}(n_{ij}+1, N_{ij}-n_{ij}+1)$. (Of course a Beta distribution is a special case of a Dirichlet, but I should have made the hint refer to the Beta.) MCMC is not needed here: One could just draw repeatedly from the six independent Beta pdf's and count the fraction of such draws in which all three age categories have $p_{i2} > p_{i1}$, assuming Pennsylvania is state 2. With the 12-cell Dirichlet approach, one would count the proportion of draws in which

$$\frac{p_{1,2,j}}{p_{1,2,j}+p_{0,2,j}} > \frac{p_{1,1,j}}{p_{1,1,j}+p_{0,1,j}}\,.$$

for all $j = 1, 2, 3$, where p_{ikj} is the probability of the cell with marriage dummy i, state dummy k, and age indicator j, and assuming $k = 2$ is PA.

(4) Suppose we have an i.i.d. sample $\{Y_i, X_i, i = 1, \ldots, N\}$ and we know that in fact $Y_i = X_i^3$ for every i. That is, we know there is an exact, non-stochastic relation connecting

Y to X for every observation. We could nonetheless estimate a linear regression of Y on X from the sample, and use the usual sandwich estimator to form a covariance matrix for the estimated coefficients in the linear regression.

(a) Show that we may find a positive definite sandwich covariance matrix for the coefficients β in the regression. Since the relation between Y and X is non-stochastic, is it misleading to present non-zero standard errors on the regression coefficients? If not, where is the randomness coming from?

We'll assume the linear regression includes a constant, though the question didn't say whether it did or not. The sandwich estimator of the variance of $\hat{\beta}_{OLS}$ is

$$(\sum Z_i' Z_i)^{-1} \sum Z_i Z_i' \hat{u}_i^2 (\sum Z_i' Z_i)^{-1} ,$$

where we're using $Z_i = (X_i, 1)$ and $\hat{u}_i = Y_i - Z_i \hat{\beta}_{OLS}$. $Z'Z = \sum Z_i' Z_i$ will in large enough samples be positive definite if $E[Z_i' Z_i]$ is positive definite. If the distribution of X_i has enough points of support, the linear regression will not fit perfectly and $\sum Z_i' Z_i \hat{u}_i^2$ will be non-singular. So we certainly may find a positive definite covariance matrix for the OLS estimates, according to the sandwich estimator.

This is the *pre*-sample standard deviation of the OLS parameter estimates. If the person interpreting the results understands this, it is not misleading. It describes uncertainty only about what variation we can expect in the estimated coefficient if we re-estimate with a new sample and the same joint distribution for (X_i, Y_i). On the other hand, if the person interpreting the results wants to understand uncertainty about a prediction of Y^* given an observation X^*, using $\hat{Y} = (X^*, 1)\hat{\beta}_{OLS}$ as the predictor, the sandwich variance estimator is useless. Once we know what the value of X^* is, we know exactly what $Y^* = (X^*)^3$ is, and thus we know exactly what the prediction error $Y^* - \hat{Y}$ is. The repeated-sample uncertainty, which arises entirely from uncertainty about what sample values of X will be drawn, is irrelevant. This is true even if X^* has been drawn from the same distribution as the original sample. Once we have seen the value of X^*, the distribution it was drawn from does not matter: the prediction error is not uncertain at all, conditional on X^*.

(b) Describe what happens to the sandwich estimator in this setup if the distribution from which the X_i values are drawn has finitely many points of support, possibly even fewer than the length of the X vector.

This question was too vaguely worded. I hoped you would recognize that when the X_i distribution has many fewer points of support than we have observations, uncertainty about what X_i values will appear in the sample eventually completely disappears (every possible value of X_i will have been observed), so that the repeated-sample uncertainty about $\hat{\beta}_{OLS}$ is entirely uncertainty about how many times each possible X_i value occurs in the sample. If there are k possible values of X_i, the true joint distribution of (Y_i, X_i) can be parameterized by the probabilities of those k points: $\{p_j, j = 1, \ldots, k\}$ and $\{n_j, j = 1, \ldots, k\}$, the number of times value j has appeared in the sample, is a sufficient statistic. The OLS estimator is a function of the $\{n_j\}$, and their distribution is known (it is multinomial, with $\{p_j\}$ as parameters.) Thus likelihood-based inference is possible without any distributional assumptions beyond

the finiteness of the support. However, the question did not ask about this. A good answer would simply have observed that the uncertainty in this case arises entirely out of the n_j's.