MA 589 — Computational Statistics

Project 5

(Due: Monday, 11/26/18)

1. A genome can have a number of alterations to its DNA sequence. A common type of alteration is a *copy number variation*: some regions might have been *amplified* or *deleted*. To measure possible changes in copy number, we can collect data from a comparative genomic hybridization (CGH) array. Roughly, for each probe i in the array that corresponds to a position in a genomic sequence we measure Y_i, the log of the ratio between the copy number of an individual at position i and a reference for a normal copy number.

   Suppose you have a simple CGH assay with only $n = 200$ probes. You decide to model the states of each probe as "deleted" (state 1), "normal" (state 2), and "duplicated" (state 3), and set a Markov chain for the transitions between states with probabilities

$$P = \begin{bmatrix} 0.50 & 0.50 & 0 \\ 0.05 & 0.90 & 0.05 \\ 0 & 0.5 & 0.5 \end{bmatrix}$$

   The chain starts at "normal", i.e., state 2. For the emission Y_i at state X_i of the chain, you assume that

$$Y_i \,|\, X_i = s \overset{\text{iid}}{\sim} N(\mu_s, \sigma_s^2), \quad i = 1, \ldots, n,$$

   with

State s	1	2	3
μ_s	-1	0	1
σ_s	0.7	0.5	0.7

   Your task is to infer, given the observations Y in the file `cgh.csv`, which positions are duplicated and deleted.

   (a) Using the forward algorithm, compute $\log \mathbb{P}(Y)$.

   (b) Using the Viterbi algorithm, obtain the MAP estimate $\widehat{X}$ for X. What is $\log \mathbb{P}(\widehat{X} \,|\, Y)$?

   (c) Plot Y as points and, for each probe i, the mean μ at $\widehat{X}_i$. It's better to connect the means using a solid, thicker line. Comment on the plot; for instance, does $\widehat{X}$ seem to provide a reasonable fit? Are there regions that you would believe to differ from $\widehat{X}$, say, as being amplified instead of normal according to $\widehat{X}$?

   (d) What is the probability that the *last* probe has a normal copy number given Y? How much more likely is it, again given Y, for the last probe to be in a deleted region instead of a duplicated region?

2. The human genome has genetic markers based on "single nucleotide polymorphisms" (SNPs) that can be used to assess the genetic risk of diseases. For instance, consider the $p = 11$ markers in file `hla_snps.csv`, where `dbsnp` is an identifier and `position` is the genomic position; these markers belong to the human leukocyte antigen (HLA) gene complex region, located in the short arm of chromosome 6. This complex encodes proteins that regulate the immune system and thus are connected to autoimmune disorders such as type I diabetes and celiac disease. File `hla_study.csv` lists the results from a (hypothetical) study with $n = 100$ individuals: `status` indicates disease status and the other columns count the number of allele copies of each HLA marker[1].

   Let us assume that the disease status Y_i of the i-th individual can be modeled using logistic regression with genotypes $\mathbf{x}_i = (x_{ij})_{j=1,\dots,p}$, $x_{ij} \in \{0, 1, 2\}$:

$$Y_i \,|\, \beta_0, \beta \overset{\text{ind}}{\sim} \mathsf{Bern}\Big[\text{logit}^{-1}\big(\beta_0 + \mathbf{x}_i^\top \beta\big)\Big], \quad i = 1, \dots, n.$$

   The main goal here is to infer the effect of each marker in the disease status. However, as it is common in these studies, we do not expect that every marker is actually associated with the status. To account for this, we perform *variable selection* and set

$$\beta_j \,|\, \theta_j \overset{\text{ind}}{\sim} N\Big(0, (1 - \theta_j)\tau_0^2 + \theta_j \tau^2\Big), \quad j = 1, \dots, p,$$

   where $\theta_j \in \{0, 1\}$ indicates if the j-th marker is associated with the status, $\tau_0^2 = 10^{-4}$ and $\tau^2 = 1$. Since we have no prior knowledge about the intercept β_0, we can set $\mathbb{P}(\beta_0) \propto 1$. The variable inclusion indicators θ are not independent, but have dependencies coded according to a graph G where two markers u and v are connected, $(u, v) \in G$, if and only if $|\text{pos}_u - \text{pos}_v| < 50{,}000$:

$$\mathbb{P}(\theta) \propto \exp\left\{ \frac{h}{2} \sum_{u=1}^{p} (2\theta_u - 1) + \frac{1}{T} \sum_{(u,v) \in G} (2\theta_u - 1)(2\theta_v - 1) \right\},$$

   with $h = -100$ and $T = 100$. To perform inference, we devise a Gibbs sampler to estimate $\mathbb{P}(\beta 0, \beta, \theta \,|\, Y)$.

   (a) Show that, with $\sigma^2(\theta_j) := (1 - \theta_j)\tau_0^2 + \theta_j \tau^2$,

$$\mathbb{P}(\theta_j \,|\, \beta_0, \beta, \theta_{[-j]}, Y) \propto \big(\sigma^2(\theta_j)\big)^{-\frac{1}{2}} \exp\left\{ -\frac{\beta_j^2}{2\sigma^2(\theta_j)} + \theta_j \left(h + \frac{2}{T} \sum_{v \in N_j} (2\theta_v - 1) \right) \right\}.$$

   Remember to use logs when computing $\mathbb{P}(\theta_j = 1 \,|\, \beta_j, \theta_{N_j})$.

   Sampling from $\beta_0 \,|\, \beta, \theta, Y$ and $\beta_j \,|\, \beta_0, \beta_{[-j]}, \theta, Y$ is trickier because there is no closed form expression for the distributions, but we can use Metropolis-Hastings steps[2].

[1] Real-world genome-wide association studies (GWAS) consider millions of markers from tens of thousands of individuals.

[2] This is known as *Metropolis-within-Gibbs*, a hybrid MCMC strategy.

For the proposal, we exploit a Laplace approximation to the conditional distribution; recall from lecture that we derived a routine `step-laplace-mh`$(\beta^{(t)}; u, z, \phi, \beta_0, \omega)$ that draws the next move from $\beta^{(t)}$ in a Metropolis-Hastings sampler for

$$u_i \,|\, \beta \overset{\text{ind}}{\sim} \mathsf{Bern}\Big[\text{logit}^{-1}\big(z_i\beta + \phi_i\big)\Big] \quad \text{and} \quad \beta \sim N(\beta_0, \omega^{-1}).$$

(b) Suppose the chain is at the t-step and you've already sampled $\theta^{(t+1)}$ above. If X is the design matrix with rows $\mathbf{x}_i$ and j-th column denoted by X_j, show that the remainder of the Gibbs cycle is

$$\beta_0^{(t+1)} = \texttt{step-laplace-mh}\big(\beta_0^{(t)}; Y, \mathbf{1}_n, X\beta^{(t)}, 0, 0\big),$$

with $\mathbf{1}_n$ a vector of ones of length n, and, for $j = 1, \ldots, p$,

$$\beta_j^{(t+1)} = \texttt{step-laplace-mh}\big(\beta_j^{(t)}; Y, X_j, \\ \mathbf{1}_n\beta_0^{(t+1)} + X_{1:(j-1)}\beta_{1:(j-1)}^{(t+1)} + X_{(j+1):p}\beta_{(j+1):p}^{(t)}, 0, (1-\theta_j^{(t+1)})/\tau_0^2 + \theta_j^{(t+1)}/\tau^2\big).$$

(c) Implement the hybrid Gibbs sampler using the HLA study: draw inferences for which markers should be selected and their effects (coefficients) and use histogram and/or density plots to summarize their distribution; assess convergence using trace and autocorrelation plots for $\log \mathbb{P}(\beta_0, \beta, \theta \,|\, Y)$ and selected coefficients, and compute scale reduction factors and effective sample size ratios.