UNIVERSITY OF CAMBRIDGE

MAT3, MAMA

MATHEMATICAL TRIPOS Part III

Monday, 3 June, 2019 9:00 am to 12:00 pm

PAPER 216

BAYESIAN MODELLING AND COMPUTATION

*Attempt no more than **FOUR** questions.*

*There are **SIX** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper
Rough paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 A computer network has vertices V and edges E. Each vertex represents a computer and there is an edge between two computers if there is communication between them. The network has no cycles, and each vertex is adjacent to at most 4 others. Some of the computers are infected by malware, and an engineer investigates a subset $V_1 \subset V$. If a computer is infected, the engineer has a probability $1/4$ of detecting the malware, and the outcome of the engineer's investigation on different computers in V_1 is independent, given the state — infected or not — of each computer. The engineer detected malware in a subset $V_2 \subset V_1$ of the investigated computers.

The engineer wants to estimate the number of infected computers in the network. Let X_i be a random variable taking values in $\{0, 1\}$, such that $X_i = 1$ indicates an infection in computer i. The engineer puts a prior distribution p on the state $X_V = (X_i)_{i \in V}$ of all the computers, defined by

$$p(x_V) = Z \exp\left(\sum_{\{i,j\} \in E} (2x_i - 1)(2x_j - 1) - \frac{1}{10} \sum_{i \in V} x_i \right) \quad \text{for } x_V \in \{0, 1\}^V,$$

where Z is a normalising constant.

Provide a plausible justification for this prior distribution and write down the posterior distribution of X_V given the outcome of the engineer's investigations.

Describe efficient algorithms to perform each of the following tasks. You may cite any algorithm defined in the course, but must justify your choice and explain how to use its output.

(a) Finding the most likely configuration $x_V^* \in \{0, 1\}^V$ for the variable X_V under the posterior distribution.

(b) Computing the exact posterior mean of the number of infected computers.

(c) Estimating the posterior probability, q, that the number of infected computers is more than twice the number of vertices in V_2. The variance of your estimator should be at most $q(1 - q)/1000$.

In each case, how does the computational cost of the algorithm depend on the number of vertices $|V|$? Explain briefly.

2 Let π be a probability measure on $(\mathbb{R}^d, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra, and suppose $X = (X_1, \ldots, X_d) \sim \pi$.

(a) Define the random scan Gibbs sampler with target distribution π.

(b) Let K^t be the t-step transition kernel of the Gibbs sampler in part (a). Now, let ν be the law of the random vector $F(X) = (f(X_1), f(X_2), \ldots, f(X_d))$, where $f : \mathbb{R} \to \mathbb{R}$ is a continuous bijection. And let Q^t be the t-step transition kernel for the random scan Gibbs sampler with target distribution ν. Prove that if $Y \sim K^t(x, \cdot)$, then $F(Y) \sim Q^t(F(x), \cdot)$.

(c) Using the result of part (b), prove that if the kernel K satisfies

$$\|K^t(x, \cdot) - \pi\|_{TV} \leqslant 2^{-t} \quad \forall x \in \mathbb{R}^d,$$

then, the kernel Q satisfies

$$\|Q^t(x, \cdot) - \nu\|_{TV} \leqslant 2^{-t} \quad \forall x \in \mathbb{R}^d.$$

3 Let ν and μ be probability measures on $\mathbb{R}^d$. For a random vector X taking values in $\mathbb{R}^d$, let X_i be the ith coordinate of X, and X_{-i} be a vector containing all coordinates except i. Denote $\nu(x_i \mid x_{-i})$ the complete conditional density of the ith coordinate under ν with respect to the Lebesgue measure on $\mathbb{R}$. We shall assume that all complete conditionals of ν and μ have densities with respect to Lebesgue measure. Define

$$p(x, i) = \frac{\mu(x_i \mid x_{-i})}{\nu(x_i \mid x_{-i})} \quad \text{for } x \in \mathbb{R}^d, i \in \{1, 2, \ldots, d\},$$

and suppose there exist c_1, c_2, such that $0 < c_1 < p(x, i) < c_2$ for all $x \in \mathbb{R}^d$ and $i \in \{1, \ldots, d\}$.

A Tempered Gibbs Sampling Markov chain $(X(t))_{t \geqslant 1}$ is defined by the following iteration. Given the state at time t, $X(t)$, we sample the next state by first sampling a coordinate i in $\{1, 2, \ldots, d\}$ with probability proportional to $p(X(t), i)$, setting $X_{-i}(t+1) = X_{-i}(t)$, and sampling $X_i(t+1)$ from the complete conditional $\mu(x_i \mid x_{-i} = X_{-i}(t))$.

(a) Prove that $(X(t))_{t \geqslant 1}$ is π-reversible for some probability measure π, and find π.

(b) Suppose that $(X(t))_{t \geqslant 1}$ is stationary and geometrically ergodic and let $h : \mathbb{R}^d \to [0, 1]$. Prove that the estimator

$$\hat{H}_n = \frac{d}{n} \sum_{i=1}^{n} \frac{1}{\sum_{i=1}^{d} p(X(t), i)} h(X(t))$$

converges in probability to $H = \int h(x)\nu(dx)$, i.e. for all $\epsilon > 0$, $\Pr(|\hat{H}_n - H| > \epsilon) \to 0$ as $n \to \infty$.

[*You can cite any result from the lecture notes.*]

4 A scientist wishes to apply Bayesian linear regression to a data set with responses $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n \times p}$. The model with parameter $\beta \in \mathbb{R}^p$ assumes

$$Y \mid X, \beta \sim N(X\beta, \sigma^2 I).$$

We only observe a subset $O \subset \{1, \ldots, n\} \times \{1, \ldots, p\}$ of the entries in the design matrix, so that the data consist of Y, and $X_O = (X_{i,j}; (i,j) \in O)$. We consider the entries observed, O, to be fixed and independent of X and Y. Furthermore, we assume that $X_{i,j} \sim \text{Bernoulli}(\pi_j)$ independently for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

The prior distribution makes all the parameters independent, with $\beta \sim N(0, \sigma_\beta^2 I)$ and $\pi_j \sim \text{Uniform}([0,1])$ for each $j = 1, \ldots, p$.

(a) Define the two steps of the Expectation-Maximisation algorithm for finding the maximum a posteriori (MAP) estimator of the parameters (β, π), with the unobserved covariates $X_U = (X_{i,j}; (i,j) \notin O)$ as latent variables.

(b) Write down the conditional distribution of X_U given Y, X_O, β, π.

(c) The E-step in the algorithm of part (a) yields a function $Q(\beta, \pi \mid \beta^{(t)}, \pi^{(t)})$, where $(\beta^{(t)}, \pi^{(t)})$ is the value of the parameters at iteration t. Prove that this function is of the form

$$-(\beta - d)^T A(\beta - d) + \sum_{j=1}^{p} r_j \log \pi_j + q_j \log(1 - \pi_j) + \text{constants}$$

for some coefficients $d \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$, $r \in \mathbb{R}^p$ and $q \in \mathbb{R}^p$. Write an expression for A as an expectation and prove that it is positive definite. Explain how the computational cost of finding the coefficients depends on the number of unobserved covariates in each row of X.

(d) Given the output of the E-step, solve the M-step. [You may express your answer in terms of the coefficients d, (r_j), and (q_j), defined in part (c).]

5 Let $Y \in \mathbb{R}^{n \times p}$ be a contingency table. A zero-inflated Poisson model makes the entries of Y independent, and the distribution of $Y_{i,j}$ a mixture of a point mass at 0 with weight π_j, and a Poisson distribution with mean $\mu_{i,j} = \alpha_i \beta_j$ with weight $1 - \pi_j$.

Let the prior distribution of the parameters be $\alpha_i \sim \text{Gamma}(1,1)$ for $i = 1 \ldots, n$, and $\beta_j \sim \text{Gamma}(1,1)$, $\pi_j \sim \text{Beta}(1,1)$ for $j = 1, \ldots, p$, with all parameters independent a priori.

We would like to sample the posterior distribution of the parameters $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\beta = (\beta_1, \ldots, \beta_p)$, and $\pi = (\pi_1, \ldots, \pi_p)$ given observations Y. Introduce auxiliary variables and define a Gibbs sampler such that the distributions required at each step may be sampled easily, and specify these distributions.

6 The Hamiltonian Monte Carlo Markov chain $(X_t, P_t)_{t\geqslant 1}$, where X_t and P_t are $\mathbb{R}^d$-valued random variables, is defined by the following iteration. Given the state at time t, (X_t, P_t), define $x(0) = X_t$, $p(0) = P_t$, and the Leapfrog recursion with step size ε:

$$\begin{aligned} p(t+\varepsilon/2) &= p(t) - \frac{\varepsilon}{2}\nabla U(x(t)) \\ x(t+\varepsilon) &= x(t) + \varepsilon p(t+\varepsilon/2) \\ p(t+\varepsilon) &= p(t+\varepsilon/2) - \frac{\varepsilon}{2}\nabla U(x(t+\varepsilon)). \end{aligned}$$

Apply this recursion L times to obtain $(x(L\varepsilon), p(L\varepsilon))$. Then set $X_{t+1} = x(\varepsilon L)$ with probability $\alpha(X_t, P_t, x(\varepsilon L), p(\varepsilon L))$, and otherwise set $X_{t+1} = X_t$. Finally, sample P_{t+1} from a $N(0, I)$ distribution.

(a) Suppose that the stationary distribution of the process has density $\pi(x, p) = Z^{-1}\exp(-U(x) - p^T p/2)$ with respect to the Lebesgue measure, where Z is a normalising constant. Write down the acceptance probability $\alpha(X_t, P_t, x(\varepsilon L), p(\varepsilon L))$.

Now suppose that $U(x) = x^T x/2$.

(b) Prove that $(x(L\varepsilon), p(L\varepsilon))$ is a linear transformation of $(x(0), p(0))$.

(c) Prove that if $L > 0$ is fixed, there exist C and ε_0, such that for all $\varepsilon < \varepsilon_0$,

$$1 - C\varepsilon^3 \leqslant \alpha(X_t, P_t, x(\varepsilon L), p(\varepsilon L)) \leqslant 1.$$

[*Hint: For any $M \in \mathbb{R}^{d\times d}$, $x^T M^T M x \leqslant \rho x^T x$ for all $x \in \mathbb{R}^d$, where ρ is the largest eigenvalue of $M^T M$. Furthermore, if $M^T M$ has the following block structure:*

$$M^T M = \begin{bmatrix} (1+O(\varepsilon^4))I & O(\varepsilon^3)I \\ O(\varepsilon^3)I & (1+O(\varepsilon^4))I \end{bmatrix}$$

then, $\rho = 1+O(\varepsilon^3)$. We say $f(\varepsilon) = O(g(\varepsilon))$ if $|f(\varepsilon)| \leqslant Kg(\varepsilon)$ for all $\varepsilon \leqslant \varepsilon_0$, for some constants K, ε_0.]

END OF PAPER