UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday, 8 June, 2018 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

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 Let $(X_i)_{i\geqslant 1}$ be a μ-reversible Markov chain on $\mathcal{X}$.

(a) Define geometric ergodicity for $(X_i)_{i\geqslant 1}$.

(b) Let $X_1 = x$ with probability 1, and define $m_n = \lfloor n^{1/3} \rfloor$. Prove that if $(X_i)_{i\geqslant 1}$ is geometrically ergodic and $Y \sim \mu$, then for any measurable function $f : \mathcal{X} \to \mathbb{R}$, with $\sup_{x\in\mathcal{X}} |f(x)| < \infty$,

$$\limsup_{n\to\infty} \frac{1}{n - m_n} \mathrm{Var}\left(\sum_{i=m_n+1}^{n} f(X_i) \right) \leqslant \gamma \mathrm{Var}(f(Y))$$

for some $\gamma < \infty$ which does not depend on f.

(c) Using the result of part (b) prove that under the same assumptions,

$$\limsup_{n\to\infty} \frac{1}{n} \mathrm{Var}\left(\sum_{i=1}^{n} f(X_i) \right) \leqslant \gamma \mathrm{Var}(f(Y))$$

for some $\gamma < \infty$ which does not depend on f.

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

2 Let $\mathcal{C}$ be a compact, convex subset of $\mathbb{R}^2$.

(a) Define the Hit-and-Run algorithm which produces approximate samples from the uniform distribution on $\mathcal{C}$.

(b) Prove that the Markov kernel $K(x, dy)$ in the Hit-and-Run algorithm admits a density, $p(x, y)$, with respect to the Lebesgue measure and that this density satisfies $\inf_{x,y\in\mathcal{C}} p(x, y) > 0$.

(c) State the drift condition for geometric ergodicity.

(d) Using part (b), prove that the algorithm is geometrically ergodic.

3 Let Y_i be the number of trains departing more than 10 minutes late from London King's Cross, out of a total of n_i trains, on the ith day of the year. For each day, we have a vector $x_i \in \mathbb{R}^p$ of independent variables. For example, x_i may contain an indicator for the event of snow on day i, among other variables. The relationship between x_i and Y_i is modelled as follows,

$$
\begin{aligned}
&Y_i \mid \theta_i \sim \text{Binomial}\left(n_i, \frac{e^{\theta_i}}{1+e^{\theta_i}}\right) && \\
&\theta_i \sim N(x_i^\top \beta + \sigma^2 Z_i, \sigma_0^2) && \text{for } i = 1, \dots, 365, \text{ independent,} \\
&Z_i = \sqrt{\rho} Z_{i-1} + \xi_i && \text{for } i = 2, \dots, 365, \\
&Z_1 \sim N(0,1), \quad \xi_i \sim N(0, 1-\rho) && \text{for } i = 2, \dots, 365, \text{ independent.}
\end{aligned}
$$

The parameters in the model are $\beta \in \mathbb{R}^p$, $\sigma^2 > 0$, $\sigma_0^2 > 0$, $\rho \in (0,1)$. We put an improper prior distribution $p(\beta, \sigma^2, \sigma^2, \rho) = 1/(\sigma^2 \sigma_0^2)$ on the parameters.

(a) How would you interpret a coefficient β_j for $j \in \{1, \dots, p\}$? Why might it be desirable to make θ_i random, as opposed to making it equal to its expected value $x_i^\top \beta$? Discuss the role of the parameters $\sigma^2 + \sigma_0^2$ and ρ in this model.

(b) Consider a Gibbs sampler targeting the posterior distribution of the variables $\beta, \sigma^2, \sigma_0^2, \rho, \theta, Z$ conditional on x and Y. Propose algorithms to draw <u>exact</u> samples from the following conditional distributions and justify your choice.

i) $p(Z \mid \beta, \sigma^2, \sigma_0^2, \rho, \theta, x, Y)$,

ii) $p(\theta \mid \beta, \sigma^2, \sigma_0^2, \rho, Z, x, Y)$.

4 A factor analysis model for observations $(Y_1, \dots, Y_n)$ with $Y_i \in \mathbb{R}^p$ for $i = 1, \dots, n$, assumes that each vector is independent and

$$Y_i = \Lambda Z_i + \xi_i$$

where $Z_i \sim N(0, I_k)$, $\xi_i \sim N(0, \sigma^2 I_p)$ are independent, and the matrix $\Lambda \in \mathbb{R}^{p \times k}$ is a parameter. You may assume σ^2 is fixed.

(a) What is the marginal distribution of Y_i?

(b) In the case $k = 1$, derive an explicit formula for the parameter update in the EM algorithm for finding the maximum likelihood estimator of Λ.

(c) Consider now a general model with parameters θ, latent variables Z, and observables Y. Prove that an iteration of the EM algorithm for finding the maximum likelihood estimator of θ cannot decrease the likelihood function.

5 Let $(Y_i)_{i\geqslant 0}$ be a Markov chain with state space $\mathbb{R}^d$, and π a probability density function on the same space. A step of the Markov chain may be simulated as follows. Given Y_i, propose a state $Z = g(Y_i, V_i)$, where V_i is a random variable in $\mathbb{R}^d$ with distribution ν and independent of Y_i. The function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ satisfies $g(g(y,v),-v) = y$. Then, with probability $\min\{1, \pi(Z)/\pi(Y_i)\}$ set $Y_{i+1} = Z$, and otherwise set $Y_{i+1} = Y_i$.

Show that Hamiltonian Monte Carlo is a Markov chain of this form by specifying g and ν for that algorithm. In general, which conditions on g and ν make $(Y_i)_{i\geqslant 0}$ π-reversible? Is it necessary that ν be normal?

6 You are given a collection of n bank notes, some of which are counterfeits. Let Y_i be 1 if bank note i is genuine, and 0 if it is a counterfeit. Let $x_i \in \mathbb{R}^p$ be a vector of features of bank note i, such as the weight and size. We apply a Probit regression model, which assumes

$$Y_i \sim \text{Bernoulli}(\mu_i), \quad \mu_i = \Phi(x_i^\top \beta),$$

independent for $i = 1, \dots, n$, where Φ is the standard normal cumulative distribution function. We put a prior $N(0, \sigma^2 I)$ on the parameter β. For a bank note which is not in the training set, with features x_{test}, you are asked to estimate the posterior mean of $\Phi(x_{\text{test}}^\top \beta)$, the probability that it is genuine.

(a) Given i.i.d. samples $\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(n)}$ from the posterior distribution $p(\beta \mid Y)$, write down the Monte Carlo estimator for the desired posterior mean.

(b) Derive the gradient of the log-posterior $g(\beta) = \nabla_\beta \log p(\beta \mid Y)$, and explain why this can be used as a control variate.

(c) Suppose that the covariance matrix of the vector $(\Phi(x_{\text{test}}^\top \beta^{(1)}), g(\beta^{(1)})^\top)$ is known. Derive the control variates estimator with minimal variance, and prove that it has smaller variance than the Monte Carlo estimator of part (a).

END OF PAPER