UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday, 9 June, 2017 9:00 am to 12:00 pm

PAPER 216

BAYESIAN MODELLING AND COMPUTATION

*Attempt no more than **FIVE** 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

A biologist models the partition of N animals into 4 classes in terms of a single parameter $\theta \in [0,1]$ as follows

$$(Y_1, Y_2, Y_3, Y_4) \sim \text{Multinomial}\left(N\,;\, \frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right).$$

We put a uniform prior distribution on θ, and observe a partition $y = (y_1, y_2, y_3, y_4)$.

(a) Construct a Gibbs sampler for the posterior distribution of θ with a data augmentation $Z \mid y, \theta \sim \text{Binomial}(y_1\,;\, \theta/(2+\theta))$, explaining how to sample each step.

(b) Define *adaptive rejection sampling*, and justify why this algorithm can be applied to draw i.i.d. samples from the posterior distribution of θ.

(c) Define an algorithm which takes as input a Uniform$(0,1)$ random variable U, and i.i.d. Gamma$(1,1)$ random variables $G_1, \ldots, G_N$, which are independent of U, and outputs an exact sample of the posterior distribution of θ.

2

We are given a Phylogenetic tree, a binary tree in which the leaves represent n animal species and the internal nodes represent ancestors of those species. Every node v is associated to a random variable X_v taking values in $\{\mathrm{A}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}^k$, where each entry represents a DNA base present in a specific site of the genome. We define an evolutionary model parametrised by a Markov kernel K in the space $\{\mathrm{A}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}$. The distribution of $\{X_v; v \text{ a node in the tree}\}$ satisfies the global Markov property on the tree. The conditional distribution of X_v given its parent $X_{p(v)}$ on the tree is

$$\mu(X_v \mid X_{p(v)}) = \prod_{i=1}^{k} K(X_{p(v)}(i), X_v(i)).$$

The distribution of X_{v_0} for the root node v_0 is uniform on $\{\mathrm{A}, \mathrm{G}, \mathrm{T}, \mathrm{C}\}^k$.

Suppose we observe the value of X_v for every leaf v of the tree. Derive the Expectation Maximisation update for the maximum likelihood estimate of K. If any quantity cannot be derived analytically, specify an algorithm to compute it, and justify your choice.

3

Let Y_i be the number of failures observed in water pumps at nuclear plant i during a time period of length t_i. Consider the hierarchical model

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Poisson}(t_i\theta_i) \quad \text{independent for } i = 1, \ldots, n, \\ \theta_i \mid b &\sim \text{Gamma}(a, b) \quad \text{independent for } i = 1, \ldots, n, \\ b &\sim \text{Gamma}(c, 1). \end{aligned}$$

(a) Derive a Gibbs sampler for the posterior distribution of $(\theta_1, \ldots, \theta_n, b)$.

(b) Prove that the Markov chain on the parameter $\theta = (\theta_1, \ldots, \theta_n)$ defined by the Gibbs sampler satisfies the drift condition for geometric ergodicity with the Lyapunov function $V(\theta) = 1 + (\sum_{i=1}^n \theta_i)^2$.

[*Hint:* A Gamma(a, b) distribution has probability density function

$$f(x) = \frac{x^{a-1} \exp(-bx) b^a}{\Gamma(a)}$$

for $x \in [0, \infty)$, mean a/b, and variance a/b^2. A Poisson(λ) distribution has probability mass function

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for $x \in \{0, 1, 2, \ldots\}$.]

4

(a) Define the mean-field variational inference problem for the posterior distribution $\mu(\cdot \mid y)$ of a vector of parameters $X = (X_1, \ldots, X_p)$ with observables y. Provide an expression for the optimal mean-field marginal distribution of X_1 as a function of fixed marginal distributions for $X_2, \ldots, X_k$, and prove that it is optimal.

(b) Let $(X^{(t)})_{t \geqslant 0}$ be a $\mu(\cdot \mid y)$-reversible Markov chain with kernel K, and let $K^t\nu$ be the law of $X^{(t)}$ if $X^{(0)} \sim \nu$. Define a variational approximation

$$q^{*t} = \arg\min_{q \in \mathcal{Q}_t} \mathrm{KL}(q \,\|\, \mu(\cdot \mid y)), \tag{1}$$

in the family $\mathcal{Q}_t = \{q\,;\, q = K^t\nu, \nu(x) = \prod_{i=1}^p \nu_i(x_i), \nu_i \in \tilde{\mathcal{Q}}\}$, where $\mathrm{KL}(\cdot \,\|\, \cdot)$ denotes the Kullback–Leibler divergence and $\tilde{\mathcal{Q}}$ is some parametric family of distributions. Prove that q^{*t} is not necessarily equal to $K^t q^{*0}$ by constructing a counterexample.

(c) Let $(Z^{(t)})_{t \geqslant 1}$ be i.i.d. Uniform$(0, 1)$ random variables, and given a random variable $X^{(0)}$, define a Markov chain $(X^{(t)})_{t \geqslant 0}$ using the recursion $X^{(t)} = f(X^{(t-1)}, Z^{(t)})$, for $t \geqslant 1$, where f is a deterministic function. Let K be the kernel of this Markov chain, and consider the variational problem in Eq. 1, letting $\tilde{\mathcal{Q}}$ be the set of univariate normal distributions. Assume that $f(x, z)$ and the logarithmic posterior density $\log \mu(x \mid y)$ are everywhere differentiable with respect to the parameter x, for every z and every y, and the gradients can be computed easily. Suggest an algorithm to solve this variational problem and justify your choice.

5

A Gaussian process classification model with binary outcomes $(Y_i)_{1\leqslant i\leqslant n}$ is defined by $\Pr(Y_i = 1) = \Phi(f(x_i))$, where Φ is the CDF of a $N(0,1)$ distribution and the function $f : \mathbb{R}^m \to \mathbb{R}$ has a Gaussian process prior distribution with mean 0 and covariance function,

$$K(z_1, z_2) = \sigma^2 \exp\left[-\frac{1}{2}(z_1 - z_2)^\top A(z_1 - z_2)\right], \quad \text{for } z_1, z_2 \in \mathbb{R}^m.$$

The parameter σ^2 is the variance of the marginal prior distribution of $f(x)$ at any value of x. The parameter A is a diagonal matrix with $A_{ii} = \tau_i^{-1}$, and defines the covariance between values of f at different points. The prior distribution makes $\sigma^{-2}, \tau_1, \ldots, \tau_m$ i.i.d. Gamma$(1, 1)$.

(a) Suppose you implement a Gibbs sampler for the posterior distribution which alternates sampling the full conditionals of 3 blocks of variables: $(f(x_1), \ldots, f(x_n))$, σ^{-2}, and $(\tau_1, \ldots, \tau_m)$, and you observe that it takes a long time to converge to the stationary distribution. Provide a plausible explanation for this.

(b) A Metropolis–Hastings algorithm targeting the marginal posterior of $\sigma^2, \tau_1, \ldots, \tau_m$ given $y_1, \ldots, y_n$ might be more efficient. However, it is not possible to compute the marginal likelihood $\mu(y \mid \sigma^2, \tau)$. Instead, you decide to implement a pseudo-marginal Metropolis–Hastings algorithm. Define this algorithm with a given proposal kernel q and explain how to implement it using importance sampling.

6

A probability density function $\mu : \mathbb{R}^d \to \mathbb{R}^+$ is not everywhere differentiable, so it is not possible to simulate it through Hamiltonian Monte Carlo. We define a smoothed version, ν, through

$$\log \nu(x) = C + \min_y \left[\log \mu(y) + \lambda \|x - y\|_2^2\right]$$

where C is a constant not depending on x, and $\lambda > 0$. The density ν is differentiable everywhere and there is an efficient algorithm to compute its gradient. Consider a Hamiltonian dynamics with positions x, momenta p, and Hamiltonian

$$H(x, p) = -\log \nu(x) + \frac{p^\top p}{2}.$$

Let $T_{\varepsilon,L}$ be the function that maps an initial condition to the output of L steps of leapfrog integration for this Hamiltonian dynamics with step size ε. Now, consider the Markov chain which iterates the following steps for $n = 1, 3, 5, \dots$: Given (X_n, P_n), first draw $P_{n+1} \sim N(0, I)$, and set $X_{n+1} = X_n$. Then, define $(X', -P') = T_{\varepsilon,L}(X_{n+1}, P_{n+1})$ and set $X_{n+2} = X'$ and $P_{n+2} = P'$ with probability $\alpha(X_{n+1}, P_{n+1}, X', P')$. Otherwise, set $X_{n+2} = X_{n+1}$ and $P_{n+2} = P_{n+1}$.

Define an acceptance probability $\alpha(X_{n+1}, P_{n+1}, X', P')$ which ensures that this Markov chain has stationary distribution μ, and prove that μ is the stationary distribution. You may cite the fact that leapfrog integration is reversible and volume preserving.

END OF PAPER