UNIVERSITY OF CAMBRIDGE

MAMA/216, EGT6/216, NST3AS/216, MAAS/216

MAT3

MATHEMATICAL TRIPOS Part III

Friday, 9 June, 2023 1:30 pm to 4:30 pm

PAPER 216

BAYESIAN MODELLING AND COMPUTATION

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

Attempt **ALL** questions.
There are **FOUR** questions in total.
Questions 1 and 3 are worth 20 marks each. Questions 2 and 4 are worth 15 marks each.

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 In a normal linear model, the observations $Y \in \mathbb{R}^n$ have a $N(X\beta, \Sigma_\varepsilon)$ distribution, where $X \in \mathbb{R}^{n\times p}$ is a fixed design matrix, and $\beta \in \mathbb{R}^p$ is a parameter of interest. We assign a prior distribution $N(0, \Sigma)$ to the parameter β where Σ has full rank.

(a) Find the posterior distribution of β.

(b) Suppose that $\Sigma_\varepsilon = \sigma^2 \alpha I_n$ and $\Sigma = \sigma^2 I_p$ for some constant $\alpha > 0$. We wish to estimate the hyperparameter $\hat{\sigma}^2$ by Empirical Bayes. Find its maximum marginal likelihood estimator.

(c) Let $B = \{x \in \mathbb{R}^p : |x| \leqslant 1\}$ be the unit ball in $\mathbb{R}^p$. Consider a formal decision problem with decision space B and utility function defined by $U^*(x, \beta) = (x^T\beta)^2$ for any $x \in B$. What is the Bayes decision rule?

(d) Now suppose that $\Sigma_\varepsilon = I_n$ and Σ is an arbitrary positive definite matrix. A statistical method requires drawing one sample from the posterior distribution of β given the first i observations $(Y_1, \ldots, Y_i)$, for each integer $i = 1, \ldots, n$. Describe an algorithm to do this with a computational complexity $\mathcal{O}(p^3 + np^2 + n^2p)$.

(e) Let $G = (V, E)$ be a phylogenetic tree, where each vertex represents a genetic variant of a virus. The edges E are directed from a root vertex $v_0 \in V$ to the leaves of the tree, where each edge $(v, v') \in E$ points from an older variant v to a more recent variant v'. The function $a : V \setminus \{v_0\} \to V$ maps a variant to its most recent ancestor, i.e., $a(v)$ is the unique vertex with $(a(v), v) \in E$.

For each variant $i \in V$, the observation Y_i measures the variant's reproductive rate in an animal model. We model the observations as i.i.d. with distribution

$$\log Y_i \mid \beta_i, \tau \sim N(\beta_i, 1/\tau).$$

We place a prior on the parameters $\beta = (\beta_i : i \in V)$ and τ with the following density with respect to the Lebesgue measure:

$$\pi(\beta, \tau) \propto \exp\left(-\tau + \sum_{i\in V, i\neq v_0} -J_i(\beta_i - \beta_{a(i)})^2 - c\|\beta\|^2\right),$$

where $c > 0$ and $J_i > 0$ for $i \in V$ are constant.

Consider a systematic scan Gibbs sampler targeting the posterior distribution of (β, τ). Find the complete conditional distributions $\pi(\beta \mid Y, \tau)$ and $\pi(\tau \mid Y, \beta)$. How does the computational complexity of each iteration scale with the number of variants $n = |V|$?

[*Hint: Let $G = (\{1, \ldots, n\}, E)$ be a tree, and suppose that a positive definite matrix $M \in \mathbb{R}^{n\times n}$ has $M_{i,j} \neq 0$ if and only if $\{i, j\} \in E$. A Cholesky decomposition $M = LL^T$ may be computed in $\mathcal{O}(n)$ FLOPs. In addition, L has $\mathcal{O}(n)$ non-zero entries.*]

2 (a) Let π be a probability measure on $(\mathbb{R}^d, \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-algebra. Let K be a π-reversible Markov kernel. Define what it means for a Markov chain with kernel K to be *geometrically ergodic*.

Define what it means for a set $A \in \mathcal{B}$ to be α-*small* in the Markov chain. Give sufficient drift–minorisation conditions for the Markov chain to be geometrically ergodic.

(b) Consider an exponential family likelihood

$$f(x \mid \theta) = \exp(T(x)^T \theta - Z(\theta))$$

with parameter $\theta \in \mathbb{R}^d$ and sufficient statistic T. We assign the natural conjugate prior to θ:

$$\pi(\theta) = \exp(\theta^T \lambda_1 - Z(\theta)\lambda_2 - \tilde{Z}(\lambda))$$

where $\lambda_1 \in \mathbb{R}^d$ and $\lambda_2 \geqslant 0$ and $\lambda = (\lambda_1^T, \lambda_2)^T$.

What is the posterior distribution $\pi(\theta \mid x)$? What is the Bayes estimator, $\hat{\theta}$, for the parameter θ under the quadratic loss? Express your answer in terms of $\tilde{Z}$.

(c) Let μ be the density in $\mathbb{R}^d$

$$\mu(\theta) = \exp(\theta^T \lambda_1 - F(\theta)\lambda_2 - \tilde{F}(\lambda))$$

where $F : \mathbb{R}^d \to \mathbb{R}$ satisfies $|F(\theta) - Z(\theta)| < C < \infty$ for all $\theta \in \mathbb{R}^d$. Consider a Metropolis–Hastings algorithm targeting the prior distribution π with proposal kernel $q(\theta, \cdot) = \mu(\cdot)$ for all $\theta \in \mathbb{R}^d$. Prove that this Markov chain is geometrically ergodic.

3 Let K be a Markov kernel on $(\mathcal{X}, \mathcal{B})$ with stationary distribution π.

(a) Define the *asymptotic variance* of the Markov chain Monte Carlo estimator for $\pi(\psi)$ for some function $\psi \in L^2(\pi)$. State an upper bound on the asymptotic variance in terms of the Markov chain's spectral gap.

(b) Show that the operator $K : L^2(\pi) \to L^2(\pi)$ defined by $Kf(x) = \int f(y)K(x, dy)$ is self-adjoint if the Markov kernel is reversible.

(c) The *discrete-time Dirichlet energy* associated to K is defined by $\mathcal{E}_K(f) = \langle f, (I - K)f\rangle_\pi$. Prove that

$$\mathcal{E}_K(f) = \frac{1}{2}\mathbb{E}\big((f(X_2) - f(X_1))^2\big),$$

where (X_1, X_2) is a Markov chain with kernel K and $X_1 \sim \pi$.

(d) Suppose that K is π-reversible. We say that K satisfies a *discrete-time Poincaré inequality* with constant $C \in (0, \infty)$ if, for all $f \in L^2(\pi)$,

$$\mathrm{Var}_\pi(f) \leqslant C\mathcal{E}_{K^2}(f).$$

Prove that this inequality holds if and only if, for all integers $t > 0$, and all $f \in L^2(\pi)$,

$$\mathrm{Var}_\pi(K^t f) \leqslant \left(1 - \frac{1}{C}\right)^t \mathrm{Var}_\pi(f).$$

(e) Let K and Q be two different, π-reversible Markov kernels. Suppose that for any $x \in \mathcal{X}$ and any measurable set $A \in \mathcal{B}$ with $x \notin A$, we have $K(x, A) \geqslant Q(x, A)$. Using part (c), prove that $\mathcal{E}_K(f) \geqslant \mathcal{E}_Q(f)$ for all $f \in L^2(\pi)$.

Suppose further that K and Q have a positive spectrum, so there exist projection valued measures S_K and S_Q such that $K = \int_0^1 \lambda S_K(d\lambda)$ and $Q = \int_0^1 \lambda S_Q(d\lambda)$.

By considering the operators $\int_0^1 \sqrt{\lambda} S_K(d\lambda)$ and $\int_0^1 \sqrt{\lambda} S_Q(d\lambda)$, or otherwise, prove an inequality relating the spectral gaps of K and Q. [You may cite any result from the course.]

4 (a) Define the *Hamiltonian Monte Carlo* algorithm targeting a probability distribution π in $\mathbb{R}^d$. [You may reference the Leapfrog iteration without definition.]

(b) Let $(X_t)_{t\geqslant 0}$ be an Itô diffusion process solving the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t.$$

Fix $D(x) = \frac{\sigma(x)\sigma(x)^T}{2}$, and let

$$b(x) = -[D(x) + Q(x)]\nabla H(x) + \Gamma(x),$$
$$\Gamma_i(x) = \sum_{j=1}^{d} \frac{\partial}{\partial x_j}(D_{ij}(x) + Q_{ij}(x)),$$

where $b_i + \sum_{j=1}^d \frac{\partial}{\partial x_j} D_{ij}$ is π-integrable for each i.

State necessary and sufficient conditions such that the diffusion process $(X_t)_{t\geqslant 0}$ has stationary distribution π.

(c) The *Underdamped Langevin Dynamics* is a diffusion process $(X_t, P_t)_{t\geqslant 0}$ which solves the stochastic differential equation:

$$\begin{aligned} dP_t &= -\gamma P_t dt - \eta \nabla U(X_t)dt + (\sqrt{2\gamma\eta})dB_t \\ dX_t &= P_t dt, \end{aligned}$$

where X_t and P_t take values in $\mathbb{R}^d$, $(B_t)_{t\geqslant 0}$ is a d-dimensional Brownian motion, and $\gamma, \eta > 0$ are constants.

Find the stationary distribution of $(X_t, P_t)_{t\geqslant 0}$. [*Hint: Use part (b) with $Q(x)$ not depending x.*]

Define the Euler–Maruyama discretisation of this process with step size δ.

END OF PAPER