UNIVERSITY OF CAMBRIDGE

MAMA/219, NST3AS/219, MAAS/219, PCAYM3/219

MAT3

MATHEMATICAL TRIPOS Part III

Thursday 11 June 2026 9:00am to 12:00pm

PAPER 219

ASTROSTATISTICS

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

Attempt no more than **THREE** questions.
There are **FOUR** 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 Consider an oversimplified model of the distribution of stars in our Galaxy. Suppose at distance $r > 0$ from Earth, the volumetric density of stars is

$$\rho(r) = \rho_0 \exp(-r/r_0),$$

where ρ_0 is the volumetric density (in units of stars per cubic parsec) at $r = 0$, and $r_0 > 0$ is a scale parameter. A satellite observes a large simple random sample of N stars, labelled $s = 1, \ldots, N$, and measures the parallax of each star s to determine its distance r_s. You have a catalog of the distance estimates $\boldsymbol{r} = \{r_s\}$ to these N observed stars. Assume that the errors on these distances are negligible. In each part below, show all steps.

(a) Derive the normalised probability density $P(r_s)$ of the distance $r_s > 0$ of a randomly selected star s in our Galaxy. Show that it is of the form

$$P(r_s) = C\, r_s^{\alpha} \exp(-\beta(r_s/r_0)),$$

and determine the values of the constants C, α, and β.

(b) Find the maximum likelihood estimate $\hat{r}_0$ for r_0, using the distance measurements $\boldsymbol{r}$ of the sample of N stars in your catalog. Evaluate the expected Fisher information. Compute the bias and variance of the estimator $\hat{r}_0$ and compare the latter against the Cramér-Rao lower bound.

(c) Henceforth, suppose that a knowledgeable astronomer informs you that, due to the limitations of the satellite, it could not observe stars with apparent fluxes (brightnesses at Earth) f_s fainter than a known flux limit $f_{\min}$, and therefore those stars (with $f_s < f_{\min}$) were not included in your catalog. However, any sampled stars with brighter fluxes ($f_s \geqslant f_{\min}$) are guaranteed to be included in your catalog. Let I_s be an indicator variable with value 1 if star s is observed, and 0 if not. Suppose that all stars have the same known intrinsic luminosity L_0. Derive and fully simplify the normalised probability density $P(r_s |\, I_s = 1)$ of the distance of a random observed star s included your catalog.

(d) Now suppose that each star has a different intrinsic luminosity L_s independently drawn from a Gaussian population distribution with known mean L_0 and known variance σ_L^2: $L_s \sim N(L_0, \sigma_L^2)$. Consider a random star s of unknown luminosity L_s at true distance r_s that would have been observed in the absence of the selection effect (i.e. if $f_{\min} = 0$). Derive $P(I_s = 1 |\, r_s)$, the probability of observing this star with the selection effect ($f_{\min} > 0$). You may express this in terms of the unit normal cumulative distribution function $\Phi(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-t^2/2)\, dt$. At what distance r_s is this selection probability equal to 0.5?

2 The physics of galaxy formation often produces complex galactic systems, each composed of a massive, central host galaxy surrounded by several satellite dwarf galaxies with different dynamical properties, such as the true angular momentum. We want to infer the mass of our Milky Way (MW) galaxy using measurements of the dynamical properties of its $N_{\rm sat}$ satellite dwarf galaxies. For each observed satellite s, we have a measurement d_s of (the absolute value of) its angular momentum x_s. We also have a catalog of K analogous galactic systems from a cosmological simulation, each comprising a massive host galaxy surrounded by $N_{\rm sat}$ associated satellite dwarf galaxies. Each row i in this catalog corresponds to system i and lists the log mass m_i of its massive host galaxy and the true angular momentum x_i^s of each associated satellite s ($s = 1, \ldots, N_{\rm sat}$, ordered in decreasing x_i^s). Each row can be thought of as a random draw $(\boldsymbol{x}_i, m_i)$ from a prior distribution $P(\boldsymbol{x}, m)$ encoding the correlations between the satellites' true angular momenta and the host galaxy's log mass within a system, induced by the physics of galaxy formation driving the simulation. Assume the MW galaxy and each of its satellites is a representative draw from this distribution.

(a) First, consider the case of measuring a single satellite ($N_{\rm sat} = 1$) of the MW. The measurement d is an unbiased estimate of x, with a Gaussian error of known variance σ^2, and is conditionally independent from m given x. Write down the likelihood function $P(d|\, x)$. Write down an expression for the normalised posterior probability density of the MW log mass, $P(m|\, d)$, and the posterior mean $\mathbb{E}[m|\, d]$, conditional on this single satellite's measurement d.

(b) Derive an estimate $\bar{m}_{\boldsymbol{w}} = \sum_{i=1}^{K} m_i\, w_i$ of the posterior mean of the Milky Way log mass, $\mathbb{E}[m|\, d]$, that uses the catalog of simulated systems. Specify the importance weights w_i.

(c) Because the importance weights are unequal, not all K prior samples contribute equally to the posterior estimate. By considering the variance of your estimator, derive an expression for the effective sample size $K_{\rm eff} \leqslant K$.

(d) Now consider the case of Gaussian measurements $\boldsymbol{d} = \{d_s\}$ of $N_{\rm sat} > 1$ dwarf satellites of the MW, labelled $s = 1, \ldots, N_{\rm sat}$. Each measurement has a known error variance σ_s^2. Derive an estimate for the posterior mean MW log mass, simultaneously conditioning on the independent measurements of all of these satellites. You may assume that, given its true angular momentum, each satellite's measurement is conditionally independent from the log mass m of the central galaxy, and the angular momenta of the other satellites and their measurements.

(e) Suppose that we can compute $P(m|\, \boldsymbol{d})$ (with proper normalisation) but cannot directly sample from it, and we want to compute the posterior mean using importance sampling. Prove that the optimal importance function from which to sample to approximate the posterior mean of m, in the sense of minimum variance, is

$$Q^*(m) = \frac{|m|\, P(m|\, \boldsymbol{d})}{\int |m|\, P(m|\, \boldsymbol{d})\, dm}$$

(You may use Jensen's Inequality: $\mathbb{E}[g(X)] \geqslant g(\mathbb{E}[X])$ where g is a convex function and X is a generic random variable.) Explain why this is not likely to be a useful importance function.

3 A quasar's light curve $f(t)$ is its brightness as a function of time and is often modelled as a realisation of an Orstein-Uhlenbeck (O-U) process, also called a damped random walk. The O-U process is a Gaussian process (GP):

$$f(t) \sim \mathcal{GP}(m(t), k(t, t'))$$

with prior mean function $m(t) = c$ and covariance function or kernel

$$\mathrm{Cov}[f(t), f(t')] = k(t, t') = A^2 \exp(-|t - t'|/\tau)$$

with characteristic amplitude $A^2 = \tau\sigma^2/2$. The short-term variability of the process is controlled by σ^2 and the characteristic timescale for the quasar brightness to revert to the mean c is τ. In a doubly-lensed quasar system, a galaxy, along the line of sight between the quasar and Earth, acts as a strong gravitational lens and produces two observed images of the same quasar. However, the brightness time series of image 2 will have a constant time delay (horizontal shift) Δt and constant magnification (vertical shift) Δm relative to image 1 due to strong lensing effects. An astronomer measures the brightness time series $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ of two images ($i \in \{1, 2\}$) of a lensed quasar at observation times $\boldsymbol{t}$, so that

$$y_{1,j} = f(t_j) + \epsilon_{1,j}$$

$$y_{2,j} = f(t_j - \Delta t) + \Delta m + \epsilon_{2,j}$$

for N times t_j ($j = 1, \ldots, N$). The heteroskedastic measurement errors $\epsilon_{i,j}$ are mutually independent, zero-mean Gaussian random variables with known variances $\sigma^2_{i,j}$.

Given a value of the Hubble constant H_0, a lens model M_l produces a probabilistic prediction of the time delay Δt. Therefore, an estimate of the time delay from the analysis of the light curve data can be used with a lens model M_l to estimate the Hubble constant.

(a) Derive the marginal likelihood function $P(\boldsymbol{y}_1, \boldsymbol{y}_2 | \, \boldsymbol{t}, \boldsymbol{\theta})$ of the time delay, magnification, and O-U process parameters $\boldsymbol{\theta} = (\Delta t, \Delta m, c, A, \tau)$, with the two time series $\boldsymbol{y}_1, \boldsymbol{y}_2$ considered jointly.

(b) Adopt an independent, wide, Gaussian prior on the time delay $\Delta t \sim N(0, \sigma^2_{\text{prior}})$, and assume a proper prior $P(\boldsymbol{\nu})$ is specified for the other parameters $\boldsymbol{\nu} = (\Delta m, c, A, \tau)$. Write down a posterior density $P(\boldsymbol{\theta} | \, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{t})$. Describe an MCMC algorithm to generate samples from this posterior density.

(c) Prove that your MCMC algorithm respects detailed balance with a stationary distribution equal to the posterior distribution.

(d) Describe a method to construct a numerical approximation to the marginal posterior of the time delay $P(\Delta t | \, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{t})$ that uses samples from your MCMC described in part (b).

(e) Assume an independent, uniform prior on $H_0 \sim U(H_0^{\min}, H_0^{\max})$. Using the probabilistic prediction from a lens model M_l, $P(\Delta t | \, H_0, M_l)$, derive an expression for the unnormalised posterior probability density $P(H_0 | \, \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{t}, M_l)$ that uses your approximation in part (d).

[QUESTION CONTINUES ON THE NEXT PAGE]

(f) Suppose there are $m > 1$ candidate lens models $\{M_l\}$, labelled $l = 1, \ldots, m$, each with equal prior probability $P(M_l)$. Using Bayesian model averaging, derive an expression for the posterior $P(H_0 | \boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{t})$ averaging over all candidate models.

4 The Kullback-Leibler (K-L) divergence, or relative entropy, between two distributions $P(\theta)$ and $Q(\theta)$ is defined as:

$$D_{\mathrm{KL}}[\,P(\theta)\,||\,Q(\theta)\,] = \int P(\theta) \ln \left(\frac{P(\theta)}{Q(\theta)} \right) \mathrm{d}\theta.$$

(a) Consider a Bayesian inference problem with observed data y, parameter θ, likelihood function $\mathcal{L}(\theta) = P(y|\,\theta)$, and a proper prior $\pi(\theta)$. For part (i) below, assume the specific case wherein y and θ are scalars and the likelihood and prior are chosen to be univariate Gaussian: $P(y|\,\theta) = N(y|\,\theta, \sigma^2)$ and $\pi(\theta) = N(\theta|\,0, \tau^2)$. The measurement variance σ^2 and the prior variance τ^2 are known.

(i) Derive the normalised posterior $P(\theta|\,y)$ and define the posterior mean $\tilde{\theta}$ and posterior variance σ^2_θ. Using the specific Gaussian assumptions above, determine whether or not the following equality holds by explicitly evaluating each individual term:

$$\ln Z = \mathbb{E}_{\theta|y}[\,\ln \mathcal{L}(\theta)\,] - D_{\mathrm{KL}}[\,P(\theta|\,y)\,||\,\pi(\theta)\,],$$

where Z is the marginal likelihood or evidence, $\mathbb{E}_{\theta|y}[\,\ln \mathcal{L}(\theta)\,]$ is the posterior expectation of the log likelihood, and $D_{\mathrm{KL}}[\,P(\theta|\,y)\,||\,\pi(\theta)\,]$ is the K-L divergence between the posterior and prior.

(ii) Does this equality hold in the general case for arbitrary, but valid, likelihoods $\mathcal{L}(\theta)$ and proper priors $\pi(\theta)$? Prove or disprove.

(b) Consider a family of distributions $Q_{\boldsymbol{\phi}}(\theta)$, parametrised by $\boldsymbol{\phi}$, used to approximate the posterior $P(\theta|\,y)$. We wish to find the best approximation $Q_{\boldsymbol{\phi}^*}(\theta)$ by minimising $D_{\mathrm{KL}}[\,Q_{\boldsymbol{\phi}}(\theta)\,||\,P(\theta|\,y)\,]$ with respect to $\boldsymbol{\phi}$.

(i) Derive a general expression for the evidence lower bound (ELBO), and show that minimising the K-L divergence is equivalent to maximising the ELBO.

(ii) Assume a Gaussian family of approximating distributions $Q_{\boldsymbol{\phi}}(\theta) = N(\theta|\,\mu_Q, \sigma^2_Q)$, where $\boldsymbol{\phi} = (\mu_Q, \sigma^2_Q)$. Evaluate the ELBO for the specific Gaussian case described in part (a) above, and maximise it to find the variational parameter values $\boldsymbol{\phi}^*$ of the best approximator $Q_{\boldsymbol{\phi}^*}(\theta)$.

END OF PAPER