UNIVERSITY OF CAMBRIDGE

MAMA/219, NST3AS/219, MAAS/219

MAT3

MATHEMATICAL TRIPOS Part III

Wednesday 11 June 2025 9:00 am to 12:00 pm

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 a simplified three-rung local distance ladder for the determination of the Hubble Constant. This involves using parallax and two types of standard candles, Cepheid stars and supernovae, as distance indicators. The absolute magnitudes of Cepheid stars (after correcting for their period-luminosity relation), are independent draws from a Gaussian population distribution, i.e. $M_c^* \sim N(M_0^*, \sigma_*^2)$, with unknown population mean M_0^* and known population variance σ_*^2. The absolute magnitudes of the supernovae are independently drawn from a Gaussian population distribution, $M_s \sim N(M_0^{\rm SN}, \sigma_{\rm SN}^2)$ with unknown population mean $M_0^{\rm SN}$ and known population variance $\sigma_{\rm SN}^2$.

In the first rung, we measure the parallax of stars within a nearby galaxy, the Large Magellanic Cloud (LMC), to determine an unbiased estimate $\hat{\mu}_{\rm LMC}$ of the LMC distance modulus with a Gaussian measurement error with variance $\sigma_{\rm LMC}^2$. (The distance modulus for a distance d is defined as $\mu = 25 + 5\log_{10}(d\ {\rm Mpc}^{-1})$, where Mpc is a megaparsec.) Within the LMC, we also observe $N_{\rm LMC}$ Cepheid stars. We measure the apparent magnitude $m_c^{\rm LMC}$ for each Cepheid star (corrected for the period-luminosity relation) labelled $c = 1, \ldots, N_{\rm LMC}$, with negligible measurement error.

In the second rung, we observe Cepheid stars in the same galaxies in which we observe Type Ia supernovae. In each of G calibrator galaxies, labelled $g = 1, \ldots, G$, we observe $N_{\rm Ceph}$ Cepheid stars and measure each one's apparent magnitudes m_c^g for $c = 1, \ldots, N_{\rm Ceph}$, with negligible measurement error. In each calibrator galaxy g, we also measure the apparent magnitude $m_{\rm SN}^g$ of a supernova, with negligible measurement error.

In the third, most distant, rung, we only observe Type Ia supernovae. They are distant enough that they participate in the smooth, overall expansion of the Universe. We measure the apparent magnitude $m_{\rm SN}^i$ of each supernova (SN), labelled $i = 1, \ldots, N_{\rm SN}$, with negligible measurement error. Each SN i in this set follows the Hubble law, the linear relation between their recession velocities $v_i = cz_i$ and their distances d_i: $d_i = c\,z_i/H_0$, where c is the speed of light and z_i is the redshift. Assume the redshift is measured exactly for each supernova in this set. In this set, only one supernova is observed in each galaxy, and every supernova is independent. The units of the Hubble constant H_0 are km s^{-1}Mpc^{-1}. Define $h = H_0/(100\ {\rm km\ s^{-1}\ Mpc^{-1}})$, $\theta = 5\log_{10} h$, and $\alpha = 5/\ln 10$. In each part below, show all steps.

(a) In the first rung, define the statistic $\hat{S}_1 \equiv \bar{m}_{\rm LMC} - \hat{\mu}_{\rm LMC}$, where

$$\bar{m}_{\rm LMC} = \frac{1}{N_{\rm LMC}} \sum_{c=1}^{N_{\rm LMC}} m_c^{\rm LMC}$$

is the sample mean of the apparent magnitudes of Cepheids observed in the LMC. Derive the sampling distribution of $\hat{S}_1$ in terms of the parameters. What is its expectation value and variance?

[QUESTION CONTINUES ON THE NEXT PAGE]

(b) In the second rung, define the statistics $\hat{S}_2^g \equiv m_{\rm SN}^g - \bar{m}^g$, where

$$\bar{m}^g = \frac{1}{N_{\rm Ceph}} \sum_{c=1}^{N_{\rm Ceph}} m_c^g$$

is the sample mean of the apparent magnitudes of Cepheids observed in galaxy g. Derive the sampling distribution of $\hat{S}_2^g$ in terms of the parameters. What is its expectation value and variance?

(c) In the third rung, derive the sampling distribution of the statistics $\hat{S}_3^i = m_{\rm SN}^i - f(z_i)$ in terms of the parameters, where $f(z) \equiv 25 + 5\log_{10}(cz/(100\text{ km s}^{-1}))$.

(d) Let $\Delta M \equiv M_0^{\rm SN} - M_0^*$ and $\mathcal{M} \equiv M_0^{\rm SN} - \theta$. Derive the likelihood function of the unknown parameters $(M_0^*, \Delta M, \mathcal{M})$, up to a multiplicative constant, using the statistics of the observed data over all three rungs simultaneously, as defined in parts (a), (b), and (c) above.

(e) Derive the maximum likelihood estimators $(\hat{M}_0^*, \Delta\hat{M}, \hat{\mathcal{M}})$, checking 1st and 2nd-order conditions. Compare the variances of these estimators to the Cramér-Rao bound. What is the MLE $\hat{\theta}$ for θ? Derive the bias and variance σ_θ^2 of $\hat{\theta}$.

(f) What is the maximum likelihood estimator $\hat{h}$ for h? Approximate the fractional variance $\mathrm{Var}[\hat{h}/h]$ to lowest order in σ_θ^2. Suppose that with continued observations, the number of Hubble flow SNe, $N_{\rm SN}$, the number of calibrator galaxies, G, and the number of Cepheids observed in the LMC, $N_{\rm LMC}$, become arbitrarily large. What is the dominant remaining source of uncertainty in the estimate of the Hubble Constant?

2 Consider the linear regression of the quasar X-ray spectral index vs. bolometric luminosity in the presence of measurement error in both quantities and intrinsic dispersion. Consider the following probabilistic generative model,

$$\xi_i \sim N(\mu, \tau^2)$$

$$\eta_i | \, \xi_i \sim N(\alpha + \beta\xi_i, \sigma^2)$$

$$x_i | \, \xi_i \sim N(\xi_i, \sigma_x^2)$$

$$y_i | \, \eta_i \sim N(\eta_i, \sigma_y^2),$$

where all random draws are independent. The astronomer measures values $\mathcal{D} = \{x_i, y_i\}$, which are noisy measurements of the true luminosity ξ_i and the true spectral index η_i of each quasar. The measurement errors are independent with known variances (σ_x^2, σ_y^2), for N independent quasars, labelled $i = 1, \ldots, N$.

(a) Write down the joint distribution $P(y_i, x_i, \eta_i, \xi_i | \, \alpha, \beta, \sigma^2, \mu, \tau^2)$ for a single quasar.

(b) Derive the observed data likelihood function for all the quasars:

$$L(\alpha, \beta, \sigma^2, \mu, \tau^2) = \prod_{i=1}^{N} P(y_i, x_i | \, \alpha, \beta, \sigma^2, \mu, \tau^2).$$

Show all steps and maximally simplify.

(c) Adopt non-informative priors and write down the posterior probability density $P(\alpha, \beta, \sigma^2, \mu, \tau^2 | \, \mathcal{D})$ up to a constant. Describe an MCMC algorithm that will generate samples from the posterior density. Show that this algorithm respects detailed balance with the posterior density as the stationary distribution.

(d) Consider the following factorisation of the individual-quasar observed-data likelihood:

$$P(y_i, x_i | \, \alpha, \beta, \sigma^2, \mu, \tau^2) = P(y_i | \, x_i; \alpha, \beta, \sigma^2, \mu, \tau^2) \times P(x_i | \, \alpha, \beta, \sigma^2, \mu, \tau^2).$$

Derive explicitly the two densities on the right-hand side. Suppose the population distribution of the latent (true) independent variables $\{\xi_i\}$ is much wider than their individual measurement uncertainties σ_x. If $\tau \gg \sigma_x$, show that the full-sample likelihood factors as

$$L(\alpha, \beta, \sigma^2, \mu, \tau^2) \approx L_1(\alpha, \beta, \sigma^2) \times L_2(\mu, \tau^2),$$

so that the estimation of the regression parameters $(\alpha, \beta, \sigma^2)$ decouples from the estimation of the latent distribution of the independent variables. Find $L_1(\alpha, \beta, \sigma^2)$ and $L_2(\mu, \tau^2)$. What are the maximum likelihood estimators for μ, τ^2 ?

3 Consider a quasar whose stochastic brightness over time $y(t)$ can be modelled as a realisation of a Gaussian process,

$$y(t) \sim \mathcal{GP}(\mu, k(t, t')),$$

with prior mean level μ and a symmetric, stationary covariance kernel

$$k(t, t') = \exp(-|t - t'|/\tau)$$

where $\tau > 0$ is a characteristic timescale. An astronomer has measured the brightness of the quasar at known times t_1 and $t_2 = t_1 + \Delta t$, where $\Delta t > 0$ is the known observational cadence, yielding $y_1 \equiv y(t_1)$ and $y_2 \equiv y(t_2)$ with negligible measurement error. Denote the kernel values for indexed times as $R_{ij} \equiv k(t_i, t_j)$ and let $R \equiv k(0, 0)$ and $x \equiv \Delta t/\tau$. In all parts below, show all steps.

(a) Suppose μ and τ are known. We wish to predict the quasar's brightness $y_3 = y(t_3)$ at a future third time $t_3 = t_2 + \Delta t$. Derive and fully simplify the posterior predictive distribution $P(y_3 | \, y_2, y_1)$, and derive the posterior predictive mean and variance. Is y_3 conditionally independent from y_1 given y_2? Justify your answer.

(b) Derive and fully simplify the limiting values of the posterior predictive mean and variance of y_3 given the observed data as $\Delta t/\tau \to \infty$.

(c) An astronomer now additionally observes $y_3 = y(t_3)$ without measurement error. Derive and fully simplify an expression for the likelihood function $P(\boldsymbol{y} | \, \boldsymbol{t}, \mu, \tau)$, where $\boldsymbol{y} = (y_1, y_2, y_3)^T$ and $\boldsymbol{t} = (t_1, t_2, t_3)^T$, in the form of a product of three univariate probability densities. Assuming τ is known and μ is unknown, derive and fully simplify the maximum likelihood estimator $\hat{\mu}$ for μ. Is $\hat{\mu}$ unbiased? Compute the variance of this estimator in the cases of $\Delta t/\tau \to 0$ and $\Delta t/\tau \to \infty$. Justify your answers.

4 Consider two different, independent experiments, A and B, for estimating a scalar cosmological parameter θ. Experiment A has a vector of nuisance parameters $\boldsymbol{\alpha}$ and yields dataset $\boldsymbol{D}_{\rm A}$. Its likelihood function is denoted $L_A(\theta, \boldsymbol{\alpha}) \equiv P(\boldsymbol{D}_{\rm A}|\, \theta, \boldsymbol{\alpha})$. Experiment B has a vector of nuisance parameters $\boldsymbol{\beta}$ and yields an independent dataset $\boldsymbol{D}_{\rm B}$. Its likelihood function is denoted $L_B(\theta, \boldsymbol{\beta}) \equiv P(\boldsymbol{D}_{\rm B}|\, \theta, \boldsymbol{\beta})$. Assume the proper priors on the cosmological parameter and nuisance parameters are separable, i.e. $\pi(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \pi(\theta)\, \pi(\boldsymbol{\alpha})\, \pi(\boldsymbol{\beta})$, and that each prior factor can both be numerically evaluated and sampled from.

(a) Considering Experiment A by itself, describe an algorithm to generate (possibly weighted) samples from the posterior $\mathcal{P}_{\rm A}(\theta, \boldsymbol{\alpha}) \equiv P(\theta, \boldsymbol{\alpha}|\, \boldsymbol{D}_{\rm A})$ and to compute the evidence $\mathcal{Z}_{\rm A}$. Describe a method to estimate the marginal posterior, $\mathcal{P}_{\rm A}(\theta) = \int \mathcal{P}_{\rm A}(\theta, \boldsymbol{\alpha})\, {\rm d}\boldsymbol{\alpha}$.

(b) Considering Experiments A and B jointly, write down expressions for the joint posterior $\mathcal{P}_{\rm AB}(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta}) \equiv P(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta}|\, \boldsymbol{D}_{\rm A}, \boldsymbol{D}_{\rm B})$, the evidence $\mathcal{Z}_{\rm AB}$, and the marginal posterior $\mathcal{P}_{\rm AB}(\theta)$.

(c) Suppose we ultimately only care about the marginal posterior $\mathcal{P}_{\rm AB}(\theta)$, and the high dimensionalities of the nuisance parameters of the individual experiments, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, make sampling the joint posterior $\mathcal{P}_{\rm AB}(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta})$ computationally impossible. Define functions $f_{\rm A}(\theta)$ and $f_{\rm B}(\theta)$ that act as effective likelihood functions so that Bayes' Theorem works on the marginal space of θ, i.e.

$$f_{\rm A}(\theta)\, f_{\rm B}(\theta)\, \pi(\theta) = \mathcal{P}_{\rm AB}(\theta)\, \mathcal{Z}_{\rm AB}.$$

(d) Devise a method to estimate the functions $f_{\rm A}(\theta)$ and $f_{\rm B}(\theta)$ using the outputs of the methods described for the analysis of individual experiments in part (a), thereby circumventing the sampling of the joint posterior $\mathcal{P}_{\rm AB}(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

(e) Describe a method to compute the evidence $\mathcal{Z}_{\rm AB}$ without sampling of the joint space of $(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

END OF PAPER