UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday, 18 June, 2021 12:00 pm to 3:00 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 an oversimplified model of the distribution of stars in our Galaxy. Suppose at distance $r > 0$ (in units of parsecs) from Earth, the spatial density of stars is

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

where ρ_0 is the spatial 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 for a year and measures the parallax angle α_s (in units of arcsec) of each star, labelled $s = 1, \dots, N$. You have a catalog of the parallax measurements $\boldsymbol{\alpha} = \{\alpha_s\}$ to these N observed stars. Assume that the measurement errors of the parallaxes are negligible. In each part below, show all steps.

(a) Derive the probability density $P(r_s)$ of the distance $r_s > 0$ of a randomly selected star s in our Galaxy.

(b) Derive the probability density $P(\alpha_s)$ of the star's stellar parallax $\alpha_s > 0$. Show that

$$P(\alpha_s) = C \left(\frac{\alpha_s}{\alpha_0} \right)^{\gamma} e^{-\alpha_0/\alpha_s}.$$

Define C, γ, and α_0 in terms of previous quantities.

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

(d) 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.

(e) Now suppose that each star has a different intrinsic luminosity 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 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 Many astronomical time-domain phenomena exhibit periodicity. Consider a Gaussian process (GP) on the plane $\boldsymbol{x} \in \mathbb{R}^2$ with mean level μ and a squared exponential kernel: $g(\boldsymbol{x}) \sim \mathcal{GP}(\mu, k_{\text{SE}}(\boldsymbol{x}, \boldsymbol{x}'))$, where

$$k_{\text{SE}}(\boldsymbol{x}, \boldsymbol{x}') = A^2 \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{x'}|^2}{2l^2}\right).$$

Now consider the process $f(t) = g(\boldsymbol{u}(t))$ restricted to the unit circle:

$$\boldsymbol{u}(t) = (\sin \omega t, \cos \omega t)\,.$$

(a) Derive and fully simplify the covariance $k_{\boldsymbol{\theta}}(t, t') = \text{Cov}[f(t), f(t')]$ between $f(t)$ and $f(t')$, where $\boldsymbol{\theta} = (A, \omega, l)$. Is this covariance kernel stationary? What is the period T of functions drawn from a GP with this kernel? Justify your answer. You may find the following identities useful:

$$\sin\alpha - \sin\beta = 2\sin\left(\frac{\alpha - \beta}{2}\right)\cos\left(\frac{\alpha + \beta}{2}\right)$$

$$\cos\alpha - \cos\beta = -2\sin\left(\frac{\alpha + \beta}{2}\right)\sin\left(\frac{\alpha - \beta}{2}\right).$$

(b) Henceforth, consider a variable star whose brightness $f(t) \sim \mathcal{GP}(\mu, k_{\boldsymbol{\theta}}(t, t'))$ varies in time as a realisation of a GP with prior mean μ and covariance function $k_{\boldsymbol{\theta}}(t, t')$. An astronomer regularly observes the star every Δt days. Let $\boldsymbol{t} = (t_1, \ldots, t_N)^T$ be the grid of the first N observation times t_i, for $i = 1, \ldots, N$. At each time t_i, the astronomer records y_i, an unbiased measurement of the latent brightness $f_i = f(t_i)$, with Gaussian measurement error with variance σ_i^2. Let $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ be the vector of observed brightnesses, and assume all errors are independent. Derive the likelihood $P(\boldsymbol{y}|\,\boldsymbol{t}; \mu, \boldsymbol{\theta})$. [If you have not solved part (a), you may assume you can numerically evaluate the kernel $k_{\boldsymbol{\theta}}(t, t')$].

(c) Given a proper prior $P(\mu, \boldsymbol{\theta})$, write down an expression for the posterior $P(\mu, \boldsymbol{\theta}|\,\boldsymbol{y}, \boldsymbol{t})$, up to the normalisation constant. Briefly describe an MCMC algorithm that in principle will enable you to generate samples from this posterior. Prove that this algorithm respects detailed balance with the posterior as the stationary distribution. Briefly describe how you would in practice implement and evaluate the MCMC to estimate the posterior mean of the period of the variable star.

(d) Suppose you have successfully collected M independent samples $\{\mu_m, \boldsymbol{\theta}_m\}$, $m = 1, \ldots, M$, from the posterior. Denote the future brightness at the next observation time $t^* \equiv t_{N+1} = t_N + \Delta t$ as $f^* = f(t^*)$. How would you compute the posterior predictive mean and variance of the future f^* given the observed past data?

(e) Suppose $\boldsymbol{\theta}$ are known, μ is unknown, and $l \ll \Delta t/T \ll 1/N$. Using an improper flat prior on μ, derive an approximate posterior density $P(\mu|\,\boldsymbol{y}, \boldsymbol{t}; \boldsymbol{\theta})$. What are the posterior mean and variance? Derive an approximate probability density of future f^* given past data, $P(f^*|\,t^*; \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{\theta})$, marginalising over the uncertainty in μ. What are the posterior predictive mean and variance of f^*? Simplify for the case of $\sigma_i = 0$.

3 Suppose we have a sample of $N > 3$ supernovae in the nearby universe, labelled $s = 1, \ldots, N$. For each supernova s, we obtain a measurement D_s of the supernova's latent absolute magnitude M_s. The measurement, however, is perturbed by zero-mean Gaussian error with known variance σ^2: $D_s | M_s \sim N(M_s, \sigma^2)$. Assume all measurement errors are independent. We assume that the latent absolute magnitudes are independently drawn from a Gaussian population distribution: $M_s \sim N(M_0, \tau^2)$, with population mean M_0 and population variance τ^2.

(a) Suppose you knew M_0 and τ^2. Using the population distribution as a prior, derive the posterior density $P(M_s | M_0, \tau^2, D_s)$ for an individual s. What is the posterior mean estimate $\tilde{M}_s$ of M_s? Express this using $b = \tau^2/(\tau^2 + \sigma^2)$. What is the posterior variance?

(b) Regard your posterior mean $\tilde{M}_s$ as a point estimate of the latent M_s. Compute and fully simplify the mean squared error $\mathbb{E}[(\tilde{M}_s - M_s)^2]$ of this estimate, where the total expectation is taken over both the measurement distribution and the population distribution. Compare this to the mean squared error $\mathbb{E}[(D_s - M_s)^2]$ of the individual estimate D_s. Which is smaller?

(c) Derive and fully simplify the probability density $P(D_s | M_0, \tau^2)$ of the estimate D_s given the population M_0 and τ^2. Next, suppose we know τ^2 but not M_0. Henceforth, adopt an improper flat prior on M_0. Using the full sample of objects, derive the posterior density $P(M_0 | \tau^2, \boldsymbol{D})$, where $\boldsymbol{D}$ is the vector of values $\{D_s\}$. What are the posterior mean and variance?

(d) Now suppose we know neither M_0 nor τ^2. Henceforth, adopt an improper power-law prior $P(\tau^2) \propto (\tau^2)^k$ on $\tau^2 \geqslant 0$ (and zero for $\tau^2 < 0$), where the exponent k is restricted to integer values. Derive an expression for the unnormalised posterior density $P(\tau^2 | \boldsymbol{D})$. For what values of k can the posterior be normalised? Comment on the suitability of the choice $k = 0$.

(e) For a general k consistent with the above constraints, write down the full joint posterior density $P(\boldsymbol{\theta} | \boldsymbol{D})$ of all parameters $\boldsymbol{\theta} = (\{M_s\}, M_0, \tau^2)$, up to a normalisation constant. Construct a Gibbs sampling algorithm that generates an MCMC to sample this posterior density by deriving a sequence of proposed moves that are always accepted. Specify the order of your sequence. Assume you have access to algorithms that generate random draws from the following probability densities:

   i) Gaussian: $N(x | a, b^2)$,
   ii) Inverse gamma: $\text{Inv-Gamma}(x | a, b) \propto x^{-(a+1)} \exp(-b/x)$, $x > 0$.

   Suppose at the start of cycle t, $\boldsymbol{\theta}^t \sim P(\boldsymbol{\theta} | \boldsymbol{D})$. Derive, up to a normalisation constant, the probability density of $\boldsymbol{\theta}^{t+1}$ after a full update of all parameters.

4 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 parts (a)-(e) 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. In each part below, show all steps.

(a) Derive the normalised posterior density $P(\theta|\,y)$. What are the posterior mean $\tilde{\theta}$ and the posterior variance σ_θ^2?

(b) Derive the evidence or marginal likelihood Z.

(c) Evaluate the posterior expectation of the log-likelihood function: $\mathbb{E}_{\theta|y}[\ln \mathcal{L}(\theta)]$, where the expectation is taken with respect to the posterior density $P(\theta|\,y)$. This provides a measure of model fit to the data.

(d) The Kullback-Leibler divergence, or relative entropy, between two distributions $P(\theta)$ and $Q(\theta)$,

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

provides a measure of discrepancy between the two distributions. In particular, the K-L divergence $D_{\mathrm{KL}}[P(\theta|y)\,||\,\pi(\theta)]$ measures the information compression moving from the prior to the posterior. Evaluate $D_{\mathrm{KL}}[P(\theta|y)\,||\,\pi(\theta)]$.

(e) In the specific Gaussian case, does the following equality hold?

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

Verify your answer explicitly using your results from parts (b), (c), and (d).

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

END OF PAPER