UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4021/9021 — Applied Bayesian Analysis

Day of examination: Thursday, 30 November 2023

Examination hours: 15.00 – 19.00

This problem set consists of 5 pages.

Appendices: None

Permitted aids: Approved calculator, and one sheet of paper with the candidate's own personal notes.

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

Answer whether each of the claims listed below is true or false, and give a brief justification of your answer.

(a) To find the Laplace approximation (lazy Bayes) of a posterior distribution, we do not need to compute the marginal likelihood.

(b) For any exchangeable sequence $y_1, \dots, y_n$ of binary observations, there exists a random variable θ and a prior $\pi(\theta)$ such that the observations are conditionally independent and identically distributed, given θ.

(c) The Metropolis-Hastings algorithm provides a technique of estimating the marginal likelihood of a model.

Problem 2

We say that a random variable y follows the half-normal distribution with parameter $\theta > 0$, written $y \sim \mathrm{HN}(\theta)$, if its density is

$$f(y, \theta) = \sqrt{\frac{2}{\pi}}\theta \exp\left(-\frac{1}{2}\theta^2 y^2\right), \quad \text{for } y > 0.$$

(a) Verify that $f(y, \theta)$ is indeed a valid probability density.

Hint: You may use the fact that $\int_0^\infty \exp(-x^2)\,\mathrm{d}x = \sqrt{\pi}/2$.

(Continued on page 2.)

(b) Assume now that $y_1, \dots, y_n$ are independently sampled from the $\mathrm{HN}(\theta)$ distribution. Write down the log-likelihood function $\ell(\theta)$, and find the maximum likelihood estimator $\widehat{\theta}$, expressed in terms of $w_n = n^{-1}\sum_{i=1}^n y_i^2$. Also exhibit a normal approximation for the distribution of $\widehat{\theta}$.

Hint: You may use that

$$\begin{aligned} \mathbb{E}[y] &= \frac{1}{\theta}\sqrt{\frac{2}{\pi}} \\ \mathrm{Var}(y) &= \frac{1}{\theta^2}\left(1 - \frac{2}{\pi}\right). \end{aligned}$$

(c) Find the Jeffreys prior. Is this proper or improper?

(d) We will now introduce a conjugate prior for θ. Suppose that θ follows the Nakagami distribution in the prior, which has density

$$g(\theta) = \frac{2}{\Gamma(\alpha)}\left(\frac{\alpha}{\beta}\right)^{\alpha}\theta^{2\alpha-1}\exp\left\{-\frac{\alpha}{\beta}\theta^2\right\}, \quad \text{for } \theta > 0,$$

where $\alpha, \beta > 0$ are parameters. Here,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,\mathrm{d}t$$

is the Gamma function. Show that the posterior distribution for $\theta \mid y_1, \dots, y_n$ is also a Nakagami distribution, and find the posterior parameters α_n, β_n.

(e) Find the Laplace approximation (lazy Bayes) of the posterior distribution of θ given $y_1, \dots, y_n$. Explain briefly how you would use this to construct a 95% credibility interval for θ.

(f) Show that the Bayes estimate of θ under squared error loss equals

$$\theta_B = \frac{\Gamma(\alpha_n + 1/2)}{\Gamma(\alpha_n)}\sqrt{\frac{\beta_n}{\alpha_n}}.$$

(g) Suppose that $x \sim \mathrm{N}(0, 1/\theta^2)$ is normally distributed with mean 0 and variance $1/\theta^2$. That is, it has the density

$$h(x, \theta) = \frac{\theta}{\sqrt{2\pi}}\exp\left\{-\frac{\theta^2}{2}x^2\right\}.$$

Show that if $y = |x|$ then $y \sim \mathrm{HN}(\theta)$. Use this result to explain how you would sample from the half-normal distribution.

(Continued on page 3.)

(h) We say that a random variable y is distributed according to the Student's t-distribution with parameters $\mu, \lambda, \nu > 0$ if it emits the density

$$\mathrm{St}(y; \mu, \lambda, \nu) = \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left[1 + \frac{\lambda(y-\mu)^2}{\nu}\right]^{-\nu/2-1/2}, \quad \text{for } y \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\lambda, \nu > 0$. For the $\mathrm{Naka}(\alpha, \beta)$ prior, let $\bar{f}(y)$ denote the predictive distribution of the next data point y_{n+1} given the observed data $y_1, \ldots, y_n$. Show that

$$\bar{f}(y) = 2\,\mathrm{St}(y; 0, \lambda, \nu),$$

for $y > 0$, where you need to identify the parameters λ, ν. Where is the extra factor of 2 coming from?

(i) Assume that we are able to sample from the normal, the Nagakami and the Student's t-distributions. Give two separate recipes for sampling from the predictive distribution of the next data point y_{n+1} given the data $y_1, \ldots, y_n$.

Problem 3

Suppose we have observed a sequence of pairs of data points $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ for $i = 1, \ldots, n$. We map each input x_i to a vector of p features,

$$\boldsymbol{\phi}(x_i) = (\phi_0(x_i), \phi_1(x_i), \ldots, \phi_{p-1}(x_i))^\top,$$

where by convention $\phi_0(x) = 1$. The vector of outcomes $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is modelled using linear regression on the the features,

$$\pi(\mathbf{y} \mid \mathbf{w}, \beta) = \mathrm{N}(\mathbf{y}; \Phi\mathbf{w}, \beta^{-1} I_n), \tag{1}$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_{p-1})^\top$ is the vector of coefficients, $\beta > 0$ is the precision (inverse variance) of the model, I_n is the $n \times n$ identity matrix and

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{p-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_{p-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \cdots & \phi_{p-1}(x_n) \end{pmatrix}$$

is the $n \times p$ design matrix. Here,

$$\mathrm{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}$$

is the density of the n-dimensional Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and (symmetric and positive definite) covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, defined for $\mathbf{x} \in \mathbb{R}^n$.

(Continued on page 4.)

Recall the density of a Gamma distribution, given by

$$\text{Gamma}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\}, \quad \text{for } x > 0,$$

where $a, b > 0$ are parameters.

In the regression model (1), use the following joint conjugate prior for $\mathbf{w}$ and β,

$$\pi(\mathbf{w}, \beta) = \text{N}(\mathbf{w}; \mathbf{m}, \beta^{-1} S)\text{Gamma}(\beta; a, b),$$

where $\mathbf{m} \in \mathbb{R}^p$ is a prespecified vector and $S \in \mathbb{R}^{p \times p}$ is a prespecified symmetric positive definite matrix.

(a) Show that the log prior density can be written as

$$\log \pi(\mathbf{w}, \beta) = -\frac{\beta}{2}(\mathbf{w} - \mathbf{m})^\top S^{-1}(\mathbf{w} - \mathbf{m}) + \left(a + \frac{p}{2} - 1\right) \log \beta - b\beta + \text{constant},$$

where the constant does not depend on $\mathbf{w}$ or β.

Hint: Recall that if A is an $n \times n$ matrix and $\lambda > 0$ is a positive scalar, then $|\lambda A| = \lambda^n |A|$.

(b) Verify that

$$\frac{1}{2}(\mathbf{w} - \mathbf{m})^\top S^{-1}(\mathbf{w} - \mathbf{m}) + \frac{1}{2}(\mathbf{y} - \Phi\mathbf{w})^\top(\mathbf{y} - \Phi\mathbf{w})$$
$$= \frac{1}{2}(\mathbf{w} - S_n\mathbf{z})^\top S_n^{-1}(\mathbf{w} - S_n\mathbf{z}) - \frac{1}{2}\mathbf{z}^\top S_n \mathbf{z} + \frac{1}{2}\mathbf{m}^\top S^{-1}\mathbf{m} + \frac{1}{2}\mathbf{y}^\top\mathbf{y},$$

where

$$S_n^{-1} = S^{-1} + \Phi^\top \Phi,$$
$$\mathbf{z} = S^{-1}\mathbf{m} + \Phi^\top \mathbf{y}.$$

(c) Show that the posterior distribution $\pi(\mathbf{w}, \beta \,|\, \mathbf{y})$ takes the same functional form as the prior,

$$\pi(\mathbf{w}, \beta \,|\, \mathbf{y}) = \text{N}(\mathbf{w}; \mathbf{m}_n, \beta^{-1} S_n)\text{Gamma}(\beta; a_n, b_n),$$

and specify the remaining posterior parameters $\mathbf{m}_n, a_n$ and b_n.

(d) Now suppose we are given a new input $x' \in \mathbb{R}^d$. It can be shown that

$$\int_{\mathbb{R}^p} \exp\left\{-\frac{\beta}{2}(\mathbf{w} - \mathbf{m}_n)^\top S_n^{-1}(\mathbf{w} - \mathbf{m}_n) - \frac{\beta}{2}(y' - \phi(x')^\top\mathbf{w})^2\right\} \mathrm{d}\mathbf{w}$$
$$= \beta^{-p/2}|T|^{-1}(2\pi)^{p/2} \exp\left\{-\frac{\beta}{2}\mathbf{m}_n^\top S_n^{-1}\mathbf{m}_n - \frac{\beta}{2}(y')^2 + \frac{\beta}{2}\mathbf{v}^\top T\mathbf{v}\right\},$$

(Continued on page 5.)

where

$$\mathbf{v} = S_n^{-1}\mathbf{m}_n + y'\boldsymbol{\phi}(x'),$$
$$T^{-1} = S_n^{-1} + \boldsymbol{\phi}(x')\boldsymbol{\phi}(x')^\top.$$

Use this result to show that the predictive distribution of the corresponding outcome y' is a Student's t-distribution (see problem 2), and identify its parameter ν (you do not have to identity the other parameters μ and λ.)

(e) Show that the marginal likelihood $\pi(\mathbf{y})$ can be written as

$$\pi(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}} \frac{b^a}{b_n^{a_n}} \frac{\Gamma(a_n)}{\Gamma(a)} \frac{|S_n|^{1/2}}{|S|^{1/2}}.$$

Problem 4

Suppose we want to sample from a target distribution on the positive reals $\mathbb{R}_{>0}$, with density $\pi(x)$, which we are only able to evaluate up to a constant. We therefore use the Metropolis-Hastings algorithm. Since the density vanishes for all negative inputs, we do not want to waste time on negative proposed values. Therefore, we use a log-scale proposal. More precisely, if the chain is in position x, we first sample $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, for some prespecified $\sigma > 0$. We then let the proposed value be

$$x' = \exp\{\log x + \varepsilon\},$$

which is guaranteed to be positive.

(a) Use transformation of variables to find an expression for the proposal density $q(x' \,|\, x)$ of x' given x.

(b) Use the expression obtained in **(a)** to show that the acceptance probability for the algorithm can be written as

$$\alpha(x' \,|\, x) = \min\left\{1, \frac{\pi(x')x'}{\pi(x)x}\right\}.$$

(c) Suppose we run the chain to obtain a sample $X_1, \dots, X_S$ for some large number S. Explain briefly how you would assess the quality of the output of the chain.

(d) How would you use the output $X_1, \dots, X_S$ to estimate the variance associated to the distribution π?

THE END - GOOD LUCK!