Bayesian Statistics III/IV (MATH3361/4071) Michaelmas term 2019

Problem class 2

Lecturer: Georgios Karagiannis georgios.karagiannis@durham.ac.uk

Exercise 1. ($\star\star$)Consider the Bayesian model

$$\begin{cases} x_i|\theta & \overset{\text{iid}}{\sim} \text{Ga}(\alpha,\beta),\ \forall i=1,...,n \\ (\alpha,\beta) & \sim \Pi(\alpha,\beta) \end{cases}$$

where Ga(a,β) is the Gamma distribution with expected value α/β. Specify a Jeffrey's prior for $\theta=(\alpha,\beta)$.

Hint-1: Gamma distr.: $x \sim \text{Ga}(a,b)$ has pdf $f(x)=\frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx)1_{(0,+\infty)}(x)$, and Expected value $\text{E}_{\text{Ga}}(x|a,b)=\frac{a}{b}$

Hint-2: You may also need that the second derivative of the logarithm of a Gamma function is the 'polygamma function of order 1'. Ie,

- $\digamma^{(0)}(\alpha)=\frac{\text{d}}{\text{d}\alpha}\log(\Gamma(a))$
- $\digamma^{(1)}(\alpha)=\frac{\text{d}^2}{\text{d}\alpha^2}\log(\Gamma(a))$

Hint-3: You may leave your answer in terms of function $\digamma^{(1)}(\alpha)$.

Hints:

- To calculate certain expectations, it may be useful to remember that the gamma family of distributions is an exponential family, with a certain canonical form (So please check again the corresponding Exercise from Homework 1).
- You may also need that the second derivative of the logarithm of a Gamma function is the 'polygamma function of order 1'. Ie,
  - $\digamma^{(0)}(\alpha)=\frac{\text{d}}{\text{d}\alpha}\log(\Gamma(a))$
  - $\digamma^{(1)}(\alpha)=\frac{\text{d}^2}{\text{d}\alpha^2}\log(\Gamma(a))$
- You may leave your answer in terms of function $\digamma^{(1)}(\alpha)$.

Solution. It is $\pi(\alpha,\beta)\propto\sqrt{\det(\mathscr{I}(\alpha,\beta))}\propto\sqrt{\det(\mathscr{I}_1(\alpha,\beta))}$ where

$$\mathscr{I}_1(\alpha,\beta)=-\text{E}_{\text{F}(x|\alpha,\beta)}\begin{bmatrix}\frac{\text{d}^2}{\text{d}\alpha^2}\log(f(x|\alpha,\beta)) & \frac{\text{d}^2}{\text{d}\alpha\text{d}\beta}\log(f(x|\alpha,\beta)) \\ \frac{\text{d}^2}{\text{d}\alpha\text{d}\beta}\log(f(x|\alpha,\beta)) & \frac{\text{d}^2}{\text{d}\beta^2}\log(f(x|\alpha,\beta))\end{bmatrix},\ \text{with}$$

So

$$f(x|\alpha,\beta)=\text{Ga}(x|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(a)}x^{\alpha-1}\exp(-\beta x)\Longrightarrow$$
$$\log(f(x|\alpha,\beta))=a\log(\beta)-\log(\Gamma(\alpha))-\beta x+(\alpha-1)\log(x)$$

So

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\alpha}\log(f(x|\alpha,\beta)) &= \log(\beta) - \frac{\mathrm{d}}{\mathrm{d}\alpha}\log(\Gamma(\alpha)) + \log(x) \\
\frac{\mathrm{d}}{\mathrm{d}\alpha^2}\log(f(x|\alpha,\beta)) &= -\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\log(\Gamma(\alpha)) = -\digamma^{(1)}(\alpha) \\
\frac{\mathrm{d}}{\mathrm{d}\beta}\log(f(x|\alpha,\beta)) &= \frac{\alpha}{\beta} - x \\
\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log(f(x|\alpha,\beta)) &= -\frac{\alpha}{\beta^2} \\
\frac{\mathrm{d}^2}{\mathrm{d}\alpha\mathrm{d}\beta}\log(f(x|\alpha,\beta)) &= \frac{1}{\beta}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}_{\mathrm{Ga}(a,b)}\left(\frac{\mathrm{d}}{\mathrm{d}\alpha^2}\log(f(x|\alpha,\beta))\right) &= -\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\log(\Gamma(\alpha)) = -\digamma^{(1)}(\alpha) \\
\mathrm{E}_{\mathrm{Ga}(a,b)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log(f(x|\alpha,\beta))\right) &= -\frac{\alpha}{\beta^2} \\
\mathrm{E}_{\mathrm{Ga}(a,b)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\alpha\mathrm{d}\beta}\log(f(x|\alpha,\beta))\right) &= \frac{1}{\beta}
\end{aligned}$$

Hence

$$\mathscr{I}_1(\alpha,\beta) = -\mathrm{E}_{\mathrm{Ga}(a,b)}\begin{bmatrix} -\digamma^{(1)}(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & -\frac{\alpha}{\beta^2} \end{bmatrix} = \begin{bmatrix} \digamma^{(1)}(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{bmatrix}$$

Therefore

$$\pi(\alpha,\beta) \propto \sqrt{\det(\mathscr{I}(\alpha,\beta))} \propto \sqrt{\det(\mathscr{I}_1(\alpha,\beta))} = \sqrt{\digamma^{(1)}(\alpha)\frac{\alpha}{\beta^2} + \frac{1}{\beta^2}} = \frac{1}{\beta}\sqrt{\digamma^{(1)}(\alpha)\alpha + 1}$$

Exercise 2. ($\star\star$)Consider observables $x = (x_1, ..., x_n)$. Consider the Bayesian model

$$\begin{cases} x_i|\theta & \overset{\text{IID}}{\sim} \mathrm{N}(\theta, 1), \quad i = 1, ..., n \\ \theta & \sim \mathrm{d}\Pi(\theta) \end{cases}$$

where $\pi(\theta) \propto 1$ and that we have only one observable. Consider the LINEX loss function

$$\ell(\theta,\delta) = \exp(c(\theta-\delta)) - c(\theta-\delta) - 1$$

1. Show that $\ell(\theta,\delta) \geq 0$
2. Find the Bayes estimator $\hat{\delta}$ under LINEX loss function and under the given Bayesian model.

   Hint-1: Random variable B follows a log-normal distribution $B \sim \mathrm{LN}(\mu_A, \sigma_A^2)$ with parameters μ_A, σ_A^2 if $B = \exp(A)$ where $A \sim \mathrm{N}(\mu_A, \sigma_A^2)$.

   Hint-2: If $B \sim \mathrm{LN}(\mu_A, \sigma_A^2)$ then $\mathrm{E}_{\mathrm{LN}(\mu_A,\sigma_A^2)}(B) = \exp(\mu_A + \frac{\sigma_A^2}{2})$.

   Hint-3: It is

   $$-\frac{1}{2}\frac{(\mu-\mu_1)^2}{v_1^2} - \frac{1}{2}\frac{(\mu-\mu_2)^2}{v_2^2}... - \frac{1}{2}\frac{(\mu-\mu_n)^2}{v_n^2} = -\frac{1}{2}\frac{(\mu-\hat{\mu})^2}{\hat{v}^2} + C$$

   where

   $$\hat{v}^2 = \left(\sum_{i=1}^{n}\frac{1}{v_i^2}\right)^{-1}; \quad \hat{\mu} = \hat{v}^2\left(\sum_{i=1}^{n}\frac{\mu_i}{v_i^2}\right); \quad C = \frac{1}{2}\frac{\hat{\mu}^2}{\hat{v}^2} - \frac{1}{2}\sum_{i=1}^{n}\frac{\mu_i^2}{v_i^2}$$

Solution. So

1. Let $g(x) = \exp(cx) - cx - 1$ with $x = \theta - \delta$. I observe that g is differential with $g'(x) = c(\exp(cx) - 1)$. Also $g(\cdot)$ has a minimum at $x = 0$, as $g'(0) = 0$, $g''(0) > 0$, and in general

$$g'(x) : \begin{cases} < 0, & \text{for } x < 0 \\ = 0, & \text{for } x = 0 \\ > 0, & \text{for } x > 0 \end{cases}$$

Moreover, it is $\lim_{x\to-\infty} g(x) = \lim_{x\to+\infty} g(x) = +\infty$ and $g(0) = 0$. Hence $g(x) > 0$, $\forall x > 0$ In other words, $\ell(\theta, \delta) \geq 0$, $\forall \theta > 0$.

2.

- It is

$$\begin{aligned} \rho(\pi, \delta|x) &= \mathrm{E}_\Pi(\ell(\theta, \delta)|x) = \mathrm{E}_\Pi(\exp(c(\theta - \delta)) - c(\theta - \delta) - 1|x) \\ &= \exp(-cd)\mathrm{E}_\Pi(\exp(c\theta)|x) - c(\mathrm{E}_\Pi(\theta|x) - \delta) - 1) \end{aligned}$$

- I will minimize the posterior expected risk to find the Bayes estimator (rule). So, it is

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\delta}\rho(\pi, \delta|x) &= -c\exp(-c\delta)\mathrm{E}_\Pi(\exp(c\theta)|x) + c \\ 0 &= \left.\frac{\mathrm{d}}{\mathrm{d}\delta}\rho(\pi, \delta|x)\right|_{\delta=\delta^\pi} \\ 0 &= -c\exp(-c\delta^\pi)\mathrm{E}_\Pi(\exp(c\theta)|x) + c \\ \delta^\pi &= \frac{1}{c}\log(\mathrm{E}^\pi(\exp(c\theta)|x)) \end{aligned}$$

- By using the Bayes theorem,

$$\begin{aligned} \pi(\theta|x) &\propto \prod_{i=1}^n \mathrm{N}(x_i|\theta, 1)\pi(\theta) \propto \prod_{i=1}^n \mathrm{N}(x_i|\theta, 1) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^d (x_i - \theta)^2\right) \propto \exp\left(-\frac{1}{2}\frac{(\theta - \hat{\theta})^2}{\hat{v}^2} + \text{const...}\right) \end{aligned}$$

with

$$\hat{v}^2 = \left(\sum_{i=1}^n \frac{1}{v_i^2}\right)^{-1} = \left(\sum_{i=1}^n \frac{1}{1}\right)^{-1} = \frac{1}{n}; \qquad \hat{\theta} = \hat{v}^2\left(\sum_{i=1}^n \frac{\mu_i}{v_i^2}\right) = \frac{1}{n}\left(\sum_{i=1}^n \frac{x_i}{1}\right) = \bar{x}$$

So the posterior distribution is

$$\theta|x \sim \mathrm{N}(\bar{x}, 1/n)$$

- Now lets assume that $\mathrm{E}_\Pi(\exp(c\theta)|x) < \infty$.
- Then $\mathrm{E}_\Pi(\exp(c\theta)|x) = \mathrm{E}_{\mathrm{N}(\bar{x},1/n)}(\underbrace{\exp(c\theta)}_{=\tilde{\theta}}|x) = \mathrm{E}_{\mathrm{LN}(c\bar{x},c^2/n)}(\tilde{\theta}|x) = \exp(c\bar{x} + c^2/2n)$ (as an expected value of LN distribution). Hence,

$$\delta^\pi(x) = c\bar{x} + c^2/2n$$

.

Exercise 3. ($\star\star$)Suppose we wish to estimate the values of a collection of discrete random variables $\vec{X} = X_1, \ldots, X_n$. We have a posterior joint probability mass function for these variables, $p\left(\vec{x}|y\right) = p\left(x_1, \ldots, x_n|y\right)$ based on some data y. We decide to use the following loss function:

$$\ell(\hat{\vec{x}}, \vec{x}) = \sum_{i=1}^{n}(1 - \delta(\hat{x}_i, x_i)) \tag{1}$$

where $\delta(a, b) = 1$ if $a = b$ and zero otherwise.

1. Derive an expression for the estimated values, found by minimizing the expectation of the loss function. [Hint: use linearity of expectation.]
2. When the probability distribution is a posterior distribution in some problem, this type of estimate is sometimes called 'maximum posterior marginal' (MPM) estimate. Explain why this name is appropriate.
3. Explain in words what the loss function is measuring. Compare with the loss function for MAP estimation.

Solution.

1. We have that

$$E\left(\ell(\hat{\vec{x}}, \vec{X})|y\right) = E\left(1 - \sum_{i=1}^{n}\delta(\hat{x}_i, X_i)|y\right) \tag{2}$$

$$= n - \sum_{i=1}^{n} E\left(\delta(\hat{x}_i, X_i)|y\right) = n - \sum_{i=1}^{n}\sum_{x_i\in\mathcal{X}_i}\delta(\hat{x}_i, x_i)p\left(x_i|y\right) = n - \sum_{i=1}^{n} p\left(\hat{x}_i|y\right) \tag{3}$$

To minimize this, it suffices to minimize each term of the sum separately, and so we have for each $i \in \{1, \ldots, n\}$ separately:

$$\hat{x}_i^* = \arg\max_{x_i\in\mathcal{X}_i} p\left(x_i|y\right) \tag{4}$$

[The x_i in (4) corresponds to the $\hat{x}_i$ in ; we simply drop the hat to keep notation as simple as possible.]

Contrast this with the MAP estimate, which requires optimisation over the full joint pmf:

$$\hat{\vec{x}}^* = \arg\max_{\vec{x}\in\vec{\mathcal{X}}} p\left(\hat{x}_i|y\right) \tag{5}$$

2. The individual terms of the sum in (3) are the posterior marginal distributions for each x_i found from the joint distribution $p\left(\hat{x}_i|y\right)$. The estimate for each x_i is found by maximizing its own posterior marginal distribution, hence the name.

3. MPM estimation has attractive properties. Like MAP estimation, it can be defined on any set (albeit with the same caveats as for MAP with continuous variables). On the other hand, it does not insist that all variables 'match' to get the minimum loss. Rather, it counts the number that match by summing over the individual zero-one losses. The loss function for MAP estimation, on the other hand, imposes a different penalization since it is

$$\ell(\hat{\vec{x}}, \vec{x}) = 1 - \delta(\hat{\vec{x}}, \vec{x}) = 1 - \prod_{i=1}^{n}\delta(\hat{x}_i, x_i) \tag{6}$$

where by taking the product, the loss will be one unless *all* variables match.

Exercise 4. (Example from the Lecture's handout) Consider a Bayesian model

$$\begin{cases} y_i|\mu & \overset{\text{iid}}{\sim} \mathrm{N}_d(\mu, \Sigma), \qquad i = 1, ..., n \\ \mu & \sim \mathrm{N}_d(\mu_0, \Sigma_0) \end{cases}$$

where uncertain $\mu \in \mathbb{R}^d$, $d \geq 1$, and known Σ, μ_0, Σ_0. Find the C_a parametric HPD credible set for μ.

Hint-1: If $z = (z_1, ..., z_d)^\top$ such as $z_j \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$ for $j = 1, ..., d$, and $\xi = z^\top z = \sum_{j=1}^{d} z_j^2$, then $\xi \sim \chi^2_d$

Hint-2: It is

$$-\frac{1}{2}\sum_{i=1}^{n}(x-\mu_i)^\top\Sigma_i^{-1}(x-\mu_i)) = -\frac{1}{2}(x-\hat{\mu})^\top\hat{\Sigma}^{-1}(x-\hat{\mu})) + C(\hat{\mu}, \hat{\Sigma}) \quad ;$$

$$\hat{\Sigma} = (\sum_{i=1}^{n}\Sigma_i^{-1})^{-1}; \quad \hat{\mu} = \hat{\Sigma}(\sum_{i=1}^{n}\Sigma_i^{-1}\mu_i);$$

$$C(\hat{\mu}, \hat{\Sigma}) = \underbrace{\frac{1}{2}(\sum_{i=1}^{n}\Sigma_i^{-1}\mu_i)^\top(\sum_{i=1}^{n}\Sigma_i^{-1})^{-1}(\sum_{i=1}^{n}\Sigma_i^{-1}\mu_i) - \frac{1}{2}\sum_{i=1}^{n}\mu_i^\top\Sigma_i^{-1}\mu_i}_{=\text{independent of } x}$$

Solution.

I will use the Definition of HPD credible interval.

- First, I compute the posterior of μ. It is

$$\begin{aligned} \pi(\mu|y) \propto f(y|\mu)\pi(\mu) &= \prod_{i=1}^{n}\mathrm{N}_d(y_i|\mu, \Sigma)\mathrm{N}_d(\mu|\mu_0, \Sigma_0) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^\top\Sigma^{-1}(y_i-\mu) - \frac{1}{2}(\mu-\mu_0)^\top\Sigma_0^{-1}(\mu-\mu_0)\right) \\ &\propto \exp\left(-\frac{1}{2}(\mu-\hat{\mu}_n)^\top\hat{\Sigma}_n^{-1}(\mu-\hat{\mu}_n)\right) \end{aligned}$$

  where

$$\hat{\Sigma}_n = (n\Sigma^{-1} + \Sigma_0^{-1})^{-1}; \qquad \hat{\mu}_n = \hat{\Sigma}_n(n\Sigma^{-1}\bar{y} + \Sigma_0^{-1}\mu_0)$$

  I recognize that $\pi(\mu|y) = \mathrm{N}_d(\mu|\hat{\mu}_n, \hat{\Sigma}_n)$, and hence $\mu|y \sim \mathrm{N}_d(\hat{\mu}_n, \hat{\Sigma}_n)$

- Now let's implement Definition of HPD credible interval. So,

$$\begin{aligned} C_a &= \left\{\mu \in \mathbb{R}^d : \pi(\mu|y) \geq k_a\right\} \\ &= \left\{\mu \in \mathbb{R}^d : \mathrm{N}_q(\mu|\hat{\mu}_n, \hat{\Sigma}_n) \geq k_a\right\} \\ &= \left\{\mu \in \mathbb{R}^d : (\mu-\hat{\mu}_n)^\top\hat{\Sigma}_n^{-1}(\mu-\hat{\mu}_n) \leq \underbrace{-\log(2\pi\det(\hat{\Sigma}_n)))k_a}_{=\tilde{k}_a}\right\} \end{aligned} \tag{7}$$

and I want the smallest constant $\tilde{k}_a$ (aka the largest constant k_a) such that

$$\Pr_{\Pi}(\mu \in C_a|y) \geq 1 - a \Longleftrightarrow$$

$$\Pr_{\Pi}\left(\underbrace{(\mu - \hat{\mu}_n)^\top \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n)}_{=\xi} \leq \tilde{k}_a\right) \geq 1 - a \tag{8}$$

- I need to find quantile $\tilde{k}_a$. This requires to find the distribution of ξ. I know that

  $$\xi = (\mu - \hat{\mu}_n)^\top \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n) \sim \chi^2_d \tag{9}$$

  because $\xi = z^\top z = \sum_{j=1}^n z_j$ with $z = L^{-1}(\mu - \hat{\mu}_n) \sim \mathrm{N}_d(0, I_d)$ where L is the lower matrix of the Cholesky decomposition of $\hat{\Sigma}_n = L^\top L$.

  Hence Eq. 8, (due to Eqs. 7, 9) becomes

  $$\Pr_{\chi^2_d}((\mu - \hat{\mu}_n)^\top \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n) \leq \tilde{k}_a) = 1 - a \tag{10}$$

  which means that, $\tilde{k}_a$ is the $1 - a$ quantile of the χ^2_d distribution, aka $\tilde{k}_a = \chi^2_{d,1-a}$

- Hence, the C_a parametric HPD credible set for μ is

  $$C_a = \{\mu \in \mathbb{R}^d : (\mu - \hat{\mu}_n)^\top \hat{\Sigma}_n^{-1}(\mu - \hat{\mu}_n) \leq \chi^2_{d,1-a}\}$$