Exercise Sheet: Bayesian Statistics

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Part I

Matrix & vector calculus

The exercises about Matrix & vector calculus are optional and can be skipped.

Exercise 1. ($\star$)Let A, B be $K \times K$ invertible matrices . Show that

$$(A+B)^{-1} = A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}$$

Exercise 2. ($\star\star$)[Woodbury matrix identity] Verify that

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}$$

if A and C are non-singular.

Exercise 3. ($\star\star$)[Sherman–Morrison formula] Let A be a $K \times K$ invertible matrix and u and v two $K \times 1$ column vectors. Verify that

$$\left(A+uv^\top\right)^{-1} = A^{-1} - \frac{1}{1+v^\top A^{-1}u}A^{-1}uv^\top A^{-1}$$

if $1+v^\top A^{-1}u \neq 0$, and if A is non-singular.

Exercise 4. ($\star\star\star$)[Block partition matrix inversion] Let A be $K \times K$ invertible matrix, and let $B = A^{-1}$ its inverse. Consider Partition

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix};\ B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

Namely, $B_{11} = \left[A^{-1}\right]_{11}$ is the upper corner of the A^{-1}, etc...
Show that

$$A_{11}^{-1} = B_{11} = B_{12}B_{22}^{-1}B_{21}$$
$$A_{11}^{-1}A_{12} = -B_{12}B_{22}^{-1}$$

Hint: Start by noticing that

$$AB = I \iff \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \iff \begin{cases} A_{11}B_{11} + A_{12}B_{21} & = I \\ A_{11}B_{12} + A_{12}B_{22} & = 0 \end{cases}$$

Part II

Random variables

Exercise 5. ($\star$)Let $y \in \mathcal{Y} \subseteq \mathbb{R}$ be a univariate random variable with CDF $F_y(\cdot)$. Consider a bijective function $h : \mathcal{Y} \to \mathcal{Z}$ with $z = h(y)$, and h^{-1} its inverse. The PDF of z is

$$F_z(z) = \begin{cases} F_Y(h^{-1}(z)) & \text{if } h \nearrow \\ \\ 1 - F_Y(h^{-1}(z)) & \text{if } h \searrow \end{cases}$$

Exercise 6. ($\star$)Let $y \in \mathcal{Y} \subseteq \mathbb{R}$ be a univariate random variable with PDF $f_y(\cdot)$. Consider a bijective function $h : \mathcal{Y} \to \mathcal{Z} \subseteq \mathbb{R}$ and let h^{-1} be the inverse function of h. Consider a univariate random variable such that $z = h(y)$. The PDF of z is

$$f_z(z) \ = f_y(y)|\det(\frac{\mathrm{d}y}{\mathrm{d}z})| \ = f_y(h^{-1}(z))|\det(\frac{\mathrm{d}}{\mathrm{d}z}h^{-1}(z))|$$

Exercise 7. ($\star$)Let $y \sim \text{Ex}(\lambda)$ r.v. with Exponential distribution with rate parameter $\lambda > 0$, and $f_{\text{Ex}(\lambda)}(y) = \lambda \exp(-\lambda y)1(y \geq 0)$. Let $z = 1 - \exp(-\lambda y)$. Calculate the PDF of z, and recognize its distribution.

Exercise 8. ($\star$)Prove the following properties

1. Let matrix $A \in \mathbb{R}^{q \times d}$, $c \in \mathbb{R}^q$, and $z = c + Ay$ then

$$\mathrm{E}(z) = \mathrm{E}(c + Ay) = c + A\mathrm{E}(y)$$

2. Let random variables $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$, and let functions ψ_1 and ψ_2 defined on $\mathcal{Z}$ and $\mathcal{Y}$, then

$$\mathrm{E}(\psi_1(z) + \psi_2(y)) = \mathrm{E}(\psi_1(z)) + \mathrm{E}(\psi_2(y))$$

3. If random variables $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$ are independent then

$$\mathrm{E}(\psi_1(z)\psi_2(y)) = \mathrm{E}(\psi_1(z))\mathrm{E}(\psi_2(y))$$

for any functions ψ_1 and ψ_2 defined on $\mathcal{Z}$ and $\mathcal{Y}$.

Exercise 9. ($\star$)Prove the following properties of the covariance matrix

1. $\mathrm{Cov}(z, y) = \mathrm{E}(zy^\top) - \mathrm{E}(z)\left(\mathrm{E}(y)\right)^\top$
2. $\mathrm{Cov}(z, y) = \left(\mathrm{Cov}(y, z)\right)^\top$
3. $\mathrm{Cov}_\pi(c_1 + A_1 z, c_2 + A_2 y) = A_1 \mathrm{Cov}_\pi(x, y) A_2^\top$, for fixed matrices A_1, A_2, and vectors c_1, c_2 with suitable dimensions.

4. If z and y are independent random vectors then $\text{Cov}(z, y) = 0$

Exercise 10. ($\star$)Prove that the (i, j)-th element of the covariance matrix between vector z and y is the covariance between their elements z_i and y_j:

$$[\text{Cov}(z, y)]_{i,j} = \text{Cov}(z_i, y_j)$$

Exercise 11. ($\star$)Prove the following properties of $\text{Var}(Y)$ for a random vector $y \in \mathcal{Y} \subseteq \mathbb{R}^d$

1. $\text{Var}(y) = \text{E}(yy^\top) - \text{E}(y)\,(\text{E}(y))^\top$
2. $\text{Var}(c + Ay) = A\text{Var}(y)A^\top$, for fixed matrix A, and vectors c with suitable dimensions.
3. $\text{Var}(y) \geq 0$; (semi-positive definite)

Exercise 12. ($\star$)Prove the following properties of characteristic functions

1. $\varphi_{A+Bx}(t) = e^{it^TA}\varphi_x(B^Tt)$ if $A \in \mathbb{R}^d$ and $B \in \mathbb{R}^{k\times d}$ are constants
2. $\varphi_{x+y}(t) = \varphi_x(t)\varphi_y(t)$ if and only if x and y are independent
3. if $M_x(t) = \text{E}(e^{t^Tx})$ is the moment generating function, then $M_x(t) = \varphi_x(-it)$

Exercise 13. ($\star$)Show that if $X \sim \text{Ex}(\lambda)$ then $\varphi_X(t) = \frac{\lambda}{\lambda - it}$.

Exercise 14. ($\star$)

1. Find $\varphi_X(t)$ if $X \sim \text{Br}(p)$.
2. Find $\varphi_Y(t)$ if $Y \sim \text{Bin}(n, p)$

Exercise 15. ($\star\star\star$)Prove the following statement related to the Bayesian theorem:
Assume a probability space $(\Omega, \mathscr{F}, P)$. Let a random variable $y : \Omega \to \mathcal{Y}$ with distribution $F(\cdot)$. Consider a partition $y = (x, \theta)$ with $x \in \mathcal{X}$ and $\theta \in \Theta$. Then the probability density function (PDF), or the probability mass function (PMF) of $\theta|x$ is

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)\text{d}F(\theta)} \tag{1}$$

Hint Consider cases where x is discrete and continuous. In the later case use the mean value theorem :

$$\int_A f(x)g(x)\text{d}x = f(\xi)\int_A g(x)\text{d}x$$

where $\xi \in A$ if A is connected, and $g(x) \geq 0$ for $x \in A$.

Exercise 16. ($\star$)Prove that:

1. if $Z \sim \text{N}(0, I)$ then $\varphi_Z(t) = \exp(-\frac{1}{2}t^Tt)$, where $Z \in \mathbb{R}^d$

2. if $X \sim \mathrm{N}(\mu, \Sigma)$ then $\varphi_X(t) = \exp(it^T\mu - \frac{1}{2}t^T\Sigma t)$, where $X \in \mathbb{R}^d$

Hint: Assume as known that if $Z \sim \mathrm{N}(0,1)$ then $\varphi_Z(t) = \exp(-\frac{1}{2}t^2)$, where $Z \in \mathbb{R}$

Exercise 17. ($\star$)Show the following properties of the Characteristic Function

1. $\varphi_x(0) = 1$ and $|\varphi_x(t)| \leq 1$ for all $t \in \mathbb{R}^d$
2. $\varphi_{A+Bx}(t) = e^{it^TA}\varphi_x(B^Tt)$ if $A \in \mathbb{R}^d$ and $B \in \mathbb{R}^{k\times d}$ are constants
3. x and y are independent then $\varphi_{x+y}(t) = \varphi_x(t)\varphi_y(t)$ (we do not proov the other way around)
4. if $M_x(t) = \mathrm{E}(e^{t^Tx})$ is the moment generating function, then $M_x(t) = \varphi_x(-it)$

Part III

Probability calculus

Exercise 18. ($\star$)Let a random variable $x \sim \text{IG}(a, b)$, a fixed value $c > 0$, and $y = cx$ then $y \sim \text{IG}(a, cb)$.

Exercise 19. ($\star\star\star$)Consider that x given z is distributed according to $\text{Ga}(\frac{n}{2}, \frac{nz}{2})$, and that z is distributed according to $\text{Ga}(\frac{m}{2}, \frac{m}{2})$; i.e.

$$\begin{cases} x|z & \sim \text{Ga}(\frac{n}{2}, \frac{nz}{2}) \\ z & \sim \text{Ga}(\frac{m}{2}, \frac{m}{2}) \end{cases}$$

Here, $\text{Ga}(\alpha, \beta)$ is the Gamma distribution with shape and rate parameters α and β, and PDF

$$f_{\text{Ga}(\alpha,\beta)}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} 1(x > 0)$$

1. Show that the compound distribution of x is F $x \sim \text{F}(n, m)$, where $\text{F}(n, m)$ is F distribution with numerator and denumerator degrees of freedom n and m, and PDF

$$f_{\text{F}(n,m)}(x) = \frac{1}{x\,\text{B}(\frac{n}{2}, \frac{m}{2})} \sqrt{\frac{(n\,x)^n\ m^m}{(n\,x+m)^{n+m}}} 1(x > 0)$$

2. Show that

$$\text{E}_{\text{F}(n,m)}(x) = \frac{m}{m-2}$$

3. Show that

$$\text{Var}_{\text{F}(n,m)}(x) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$$

Hint: If $\xi \sim \text{IG}(a, b)$ then $\text{E}_{\xi\sim\text{IG}(a,b)}(\xi) = \frac{b}{a-1}$, and $\text{Var}_{\xi\sim\text{IG}(a,b)}(\xi) = \frac{b^2}{(a-1)^2(a-2)}$

Exercise 20. ($\star\star$)Prove the following statement:
Let $x \sim \text{N}_d(\mu, \Sigma)$, $x \in \mathbb{R}^d$, and $y = (x-\mu)^\top \Sigma^{-1}(x-\mu)$. Then

$$y \sim \chi^2_d$$

Exercise 21. ($\star\star$)Let

$$\begin{cases} x|\xi & \sim \text{N}_d(\mu, \Sigma\xi) \\ \xi & \sim \text{IG}(a, b) \end{cases}$$

with PDF

$$f_{\text{N}_d(\mu,\Sigma\xi)}(x|\xi) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

$$f_{\text{IG}(a,b)}(\xi) = \frac{b^a}{\Gamma(a)} \xi^{-a-1} \exp\left(-\frac{b}{\xi}\right) 1_{(0,\infty)}(\xi)$$

Show that the marginal PDF of x is

$$f(x) = \int f_{\mathrm{N}_d(\mu,\Sigma\xi)}(x|\xi) f_{\mathrm{IG}(a,b)}(\xi)\mathrm{d}\xi$$
$$= \frac{2a^{-\frac{d}{2}}}{\pi^{\frac{n}{2}}\sqrt{\det(\frac{b}{a}\Sigma)}} \frac{\Gamma\left(a+\frac{d}{2}\right)}{\Gamma(a)} \left[1 + \frac{1}{2a}(x-\mu)^\top \left(\frac{b}{a}\Sigma\right)^{-1}(x-\mu)\right]^{-\frac{(2a+d)}{2}} \tag{2}$$

FYI: For $a = b = \frac{v}{2}$, the marginal PDF is the PDF of the d-dimensional Student T distribution.

The Following exercise is part of Homework 1

Exercise 22. ($\star\star\star$)

Let $x \sim \mathrm{T}_d(\mu, \Sigma, \nu)$. Recall that $x \sim \mathrm{T}_d(\mu, \Sigma, \nu)$ is the marginal distribution $f_x(x) = \int f_{x|\xi}(x|\xi) f_\xi(\xi)\mathrm{d}\xi$ of (x, ξ) where

$$x|\xi \sim \mathrm{N}_d(\mu, \Sigma\xi v)$$
$$\xi \sim \mathrm{IG}(\frac{v}{2}, \frac{1}{2})$$

Consider partition such that

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \qquad \Sigma = \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix},$$

where $x_1 \in \mathbb{R}^{d_1}$ and $x_2 \in \mathbb{R}^{d_2}$.

Address the following:

1. Show that the marginal distribution of x_1 is such that

   $$x_1 \sim \mathrm{T}_{d_1}(\mu_1, \Sigma_1, \nu)$$

   Hint: Try to use the form $f_x(x) = \int f_{x|\xi}(x|\xi) f_\xi(\xi)\mathrm{d}\xi$.

2. Show that

   $$\xi|x_1 \sim \mathrm{IG}(\frac{1}{2}(d_1+v), \frac{1}{2}\frac{Q+v}{v})$$

   where $Q = (\mu_1 - x_1)^\top \Sigma_1^{-1}(\mu_1 - x_1)$.

   Hint: The PDF of $y \sim \mathrm{N}_d(\mu, \Sigma)$ is

   $$f(y) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right)$$

   Hint: The PDF of $y \sim \mathrm{IG}(a, b)$ is

   $$f_{\mathrm{IG}(a,b)}(y) = \frac{b^a}{\Gamma(a)} y^{-a-1} \exp(-\frac{b}{y}) 1_{(0,+\infty)}(y)$$

3. Let $\xi' = \xi\frac{v}{Q+v}$, with $Q = (\mu_1 - x_1)^\top \Sigma_1^{-1}(\mu_1 - x_1)$, show that

   $$\xi'|x_1 \sim \mathrm{IG}(\frac{v+d_1}{2}, \frac{1}{2})$$

4. Show that the conditional distribution of $x_2|x_1$ is such that

$$x_2|x_1 \sim \mathrm{T}_{d_2}(\mu_{2|1}, \dot{\Sigma}_{2|1}, \nu_{2|1})$$

where

$$\begin{aligned}
\mu_{2|1} =& \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \\
\dot{\Sigma}_{2|1} =& \frac{\nu + (\mu_1 - x_1)^\top \Sigma_1^{-1}(\mu_1 - x_1)}{\nu + d_1}\Sigma_{2|1} \\
\Sigma_{2|1} =& \Sigma_{22} - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top \\
\nu_{2|1} =& \nu + d_1
\end{aligned}$$

Hint: You can use the Example [Marginalization & conditioning] from the Lecture Handout

---

Exercise 23. ($\star\star\star$)Show that

1. If $x_i \sim \mathrm{N}_d(\mu_i, \Sigma_i)$ for $i = 1, .., n$ and $y = c + \sum_{i=1}^n B_i x_i$, then

$$y \sim \mathrm{N}_d(c + \sum_{i=1}^n \mu_i, \sum_{i=1}^n B_i \Sigma_i B_i^\top)$$

2. If $x_i \sim \mathrm{T}_d(\mu_i, \Sigma_i, v)$ for $i = 1, .., n$ and $z = c + \sum_{i=1}^n B_i x_i$, then

$$z \sim \mathrm{T}_d(c + \sum_{i=1}^n \mu_i, \sum_{i=1}^n B_i \Sigma_i B_i^\top, v)$$

Part IV

Bayesian paradigm and calculations

Exercise 24. ($\star$)Consider an i.i.d. sample $y_1, \ldots, y_n$ from the skew-logistic distribution with PDF

$$f(y_i|\theta) = \frac{\theta e^{-y_i}}{(1+e^{-y_i})^{\theta+1}}$$

with parameter $\theta \in (0, \infty)$. To account for the uncertainty about θ we assign a Gamma prior distribution with PDF

$$\pi(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}1(\theta \in (0,\infty)),$$

and fixed hyper parameters a, b specified by the researcher's prior info.

1. Derive the posterior distribution of θ.
2. Derive the predictive PDF for a future $z = y_{n+1}$.

The following exercise is about Bayesian treatment in problems with nuisance unknown parameters

- Assume observable quantities $y = (y_1, ..., y_n)$. Assume that the sampling distribution is $F(y|\theta)$ labeled by an unknown parameter $\theta \in \Theta$. Let $\theta = (\phi, \lambda)^\top$ with $\phi \in \Phi$ and $\lambda \in \Lambda$. Assume You are interested in learning parameter $\phi \in \Phi$, and You are not interested in learning the unknown parameter $\lambda \in \Lambda$; but both ϕ, λ are parts of the statistical model parameterisation. The unknown quantity $\lambda \in \Lambda$ is called nuisance parameter. We an call $\phi \in \Phi$ parameter of interest.
- In Bayesian Stats, learning (or quantifying uncertainty about) parameter of interest ϕ under the presence of a nuisance parameter $\lambda \in \Lambda$ is performed according to the Bayesian paradigm as usual: You specify a prior $\Pi(\phi, \lambda)$ with PDF/PMF $\pi(\phi, \lambda) = \pi(\phi|\lambda)\pi(\lambda)$ on the joint space of ALL Your unknown parameters $\theta = (\phi, \lambda)^\top$; you compute the joint posterior distribution $\mathrm{d}\Pi(\theta|y)$ of $\theta = (\phi, \lambda)^\top$ via the Bayesian theorem. Reasonably, Your posterior degree of believe about the parameter of interest ϕ given the data $y = (1_1, ..., y_n)$ is given through the marginal posterior distribution $\mathrm{d}\Pi(\phi|y)$.
- To summarize; Specify the Bayesian model as:

$$\begin{cases} y|\overbrace{\phi, \lambda}^{=\theta} \sim \mathrm{d}F(y|\overbrace{\phi, \lambda}^{=\theta}) & \text{, the statistical model} \\ (\underbrace{\phi, \lambda}_{=\theta}) \sim \mathrm{d}\Pi(\underbrace{\phi, \lambda}_{=\theta}) & \text{, the prior model} \end{cases}$$

  The joint posterior of θ given y is $\mathrm{d}\Pi(\theta|y) = \mathrm{d}\Pi(\lambda|y, \phi)\mathrm{d}\Pi(\phi|y)$ is with PDF/PMF

- $$\pi(\overbrace{\phi, \lambda}^{=\theta}|y) = \frac{f(y|\overbrace{\phi, \lambda}^{=\theta})\pi(\overbrace{\phi, \lambda}^{=\theta})}{f(y)} = \underbrace{\frac{f(y|\phi, \lambda)\pi(\lambda|\phi)}{f(y|\phi)}}_{=\pi(\lambda|y,\phi)} \underbrace{\frac{f(y|\phi)\pi(\phi)}{f(y)}}_{=\pi(\phi|y)} = \pi(\lambda|y, \phi)\pi(\phi|y)$$

  The (marginal) likelihood $f(y|\phi)$ of y given ϕ is

$$f(y|\phi) = \underbrace{\int_\Lambda f(y|\overbrace{\phi, \lambda}^{=\theta})\mathrm{d}\Pi(\lambda|\phi)}_{=\mathrm{E}_{\Pi(\lambda|\phi)}(f(y|\phi,\lambda)|\phi)} = \begin{cases} \int_\Lambda f(y|\phi, \lambda)\pi(\lambda|\phi)\mathrm{d}\lambda & \text{, if } \lambda \text{ cont} \\ \sum_{\forall\lambda\in\Lambda} f(y|\phi, \lambda)\pi(\lambda|\phi) & \text{, if } \lambda \text{ discr} \end{cases}$$

  The PDF/PMF $\pi(\phi|y)$ of marginal posterior $\mathrm{d}\Pi(\phi|y)$ of ϕ is

$$\pi(\phi|y) = \underbrace{\int_\Lambda \pi(\overbrace{\phi, \lambda}^{=\theta}|y)\mathrm{d}\lambda}_{=\mathrm{E}_{\Pi(\lambda|y)}(\pi(\phi|y,\lambda))} \qquad \text{or equivalently} \qquad \pi(\phi|y) = \frac{f(y|\phi)\pi(\phi)}{f(y)}$$

  The predictive distribution $G(z|y)$ of the next outcome $z = (y_{n+1}, ...y_{n+m})$ given y has pdf/pmf

$$g(z|y) = \int f(y|\overbrace{\phi, \lambda}^{=\theta})\mathrm{d}\Pi(\overbrace{\phi, \lambda}^{=\theta}|y)$$

  and the marginal likelihood $f(y)$ is

$$f(y) = \int f(y|\overbrace{\phi, \lambda}^{=\theta})\pi(\overbrace{\phi, \lambda}^{=\theta})\mathrm{d}\phi\mathrm{d}\lambda$$

Exercise 25. (⋆⋆⋆)Assume observable quantities $y = (y_1, ..., y_n)$ forming the available data set of size n. Assume that the observations are drawn i.i.d. from a sampling distribution which is judged to be in the Normal parametric <-story

family of distributions $\mathrm{N}(\mu, \sigma^2)$ with unknown mean μ and variance σ^2. We are interested in learning μ and the next outcome $z = y_{n+1}$. We do not care about σ^2.

Assume You specify a Bayesian model <-set-up

$$\begin{cases} y_i|\mu,\sigma^2 \sim \mathrm{N}(\mu,\sigma^2), \text{ for all } i=1,...,n & \text{, Statistical model} \\ \mu|\sigma^2 \sim \mathrm{N}(\mu_0, \sigma^2 \frac{1}{\tau_0}) & \text{, prior} \\ \sigma^2 \sim \mathrm{IG}(a_0, k_0) & \text{, prior} \end{cases}$$

1. Show that

$$\sum_{i=1}^{n}(y_i-\theta)^2 = n(\bar{y}-\theta)^2 + ns^2,$$

where $s^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

2. Show that the joint posterior distribution $\Pi(\mu,\sigma^2|y)$ is such as

$$\mu|y,\sigma^2 \sim \mathrm{N}(\mu_n, \sigma^2\frac{1}{\tau_n})$$
$$\sigma^2|y \sim \mathrm{IG}(a_n, k_n)$$

with

$$\mu_n = \frac{n\bar{y}+\tau_0\mu_0}{n+\tau_0}; \qquad \tau_n = n+\tau_0; \qquad a_n = a_0+n$$

$$k_n = k_0 + \frac{1}{2}ns_n^2 + \frac{1}{2}\frac{\tau_0 n(\mu_0-\bar{y})^2}{n+\tau_0}$$

Hint: It is

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

where

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

3. Show that the marginal posterior distribution $\Pi(\mu|y)$ is such as

$$\mu|y \sim \mathrm{T}_1\left(\mu_n, \frac{k_n}{a_n}\frac{1}{\tau_n}, 2a_n\right)$$

Hint-1: If $x \sim \mathrm{IG}(a,b)$, $y = cx$, then $y \sim \mathrm{IG}(a, cb)$.

Hint-2: The definition of Student T is considered as known

4. Show that the predictive distribution $\Pi(z|y)$ is Student T such as

$$z|y \sim \mathrm{T}_1\left(\mu_n, \frac{k_n}{a_n}(\frac{1}{\tau_n}+1), 2a_n\right)$$

Hint-1: Consider that

$$\mathrm{N}(x|\mu_1,\sigma_1^2)\,\mathrm{N}(x|\mu_2,\sigma_2^2) \;=\; \mathrm{N}(x|m,v^2)\,\mathrm{N}(\mu_1|\mu_2,\sigma_1^2+\sigma_2^2)$$

where

$$v^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}; \quad m = v^2\left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right)$$

Hint-2: The definition of Student T is considered as known

The following is about the Normal linear model of regression.

Exercise 26. ($\star\star\star$)(Normal linear regression model with unknown error variance) <-story

Consider we are interested in recovering the mapping

$$x \overset{\eta(x)}{\longmapsto} y$$

in the sense that y is the response (output quantity) that depends on x which is the independent variable (input quantity) in a procedure; E.g.:,

- y: precipitation in log scale
- $x = (\text{longitude}, \text{latitude})$: geographical coordinates.

It is believed that the mapping $\eta(x)$ can be represented as an expansion of d known polynomial functions $\{\phi_j(x)\}_{j=0}^{d-1}$ such as

$$\eta(x) = \sum_{j=0}^{d-1} \phi_j(x)\beta_j = \Phi(x)^\top \beta; \quad \text{with } \Phi(x) = (\phi_0(x), ..., \phi_{d-1}(x))^\top$$

where $\beta \in \mathbb{R}^d$ is unknown.

Assume observable quantities (data) in pairs (x_i, y_i) for $i = 1, ..., n$; (E.g. from the i-th station at location x_i I got the reading y_i). Assume that the response observations $y = (y_1, ..., y_n)$ may be contaminated by noise with unknown variance; such that

$$y_i = \eta(x_i) + \epsilon_i$$

where $\epsilon_i \sim \text{N}(0, \sigma^2)$ with unknown σ^2.

You are interested in learning β, but you do not care about σ^2. Also you want to learn the value of y_f at an untried x_f (i.e. the precipitation at any other location).

Consider the Bayesian model <-set-up

$$\begin{aligned} y|\beta, \sigma^2 &\sim \text{N}(\Phi\beta, I\sigma^2); \text{ the sampling distr} \\ \beta|\sigma^2 &\sim \text{N}(\mu_0, V_0\sigma^2); \text{prior distr} \\ \sigma^2 &\sim \text{IG}(a_0, k_0) \text{prior distr} \end{aligned}$$

where Φ is the design matrix $[\Phi]_{i,j} = \Phi_j(x_i)$.

1. Show that the joint posterior distribution $\text{d}\Pi(\beta, \sigma^2|y)$ is such as

$$\beta|y, \sigma^2 \sim \text{N}(\mu_n, V_n\sigma^2); \qquad \sigma^2|y \sim \text{IG}(a_n, k_n)$$

with

$$V_n^{-1} = \Phi^\top\Phi + V_0^{-1}; \qquad \mu_n = V_n\left((\Phi^\top\Phi)^{-1}\Phi y + V_0^{-1}\mu_0\right); \qquad a_n = \frac{n}{2} + a_0$$

$$k_n = \frac{1}{2}(y-\Phi\hat{\beta}_n)^\top(y-\Phi\hat{\beta}_n) - \frac{1}{2}\mu_n^\top V_n^{-1}\mu_n + \frac{1}{2}\left(\mu_0^\top V_0^{-1}\mu_0 + y^\top\Phi^\top(\Phi^\top\Phi)^{-1}\Phi y\right) + k_0$$

Hint-1:

$$(y-\Phi\beta)^\top(y-\Phi\beta) = (\beta-\hat{\beta}_n)^\top\left[\Phi^\top\Phi\right](\beta-\hat{\beta}_n) + S_n; \quad S_n = (y-\Phi\hat{\beta}_n)^\top(y-\Phi\hat{\beta}_n); \quad \hat{\beta}_n = (\Phi^\top\Phi)^{-1}\Phi y$$

Hint-2: If $\Sigma_1 > 0$ and $\Sigma_2 > 0$ symmetric

$$-\frac{1}{2}(x-\mu_1)\Sigma_1^{-1}(x-\mu_1)^\top - \frac{1}{2}(x-\mu_2)\Sigma_2^{-1}(x-\mu_2)^\top = -\frac{1}{2}(x-m)V^{-1}(x-m)^\top + C$$

where

$$V^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1}; \quad m = V\left(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2\right); \quad C = \frac{1}{2}m^\top V^{-1}m - \frac{1}{2}\left(\mu_1^\top\Sigma_1^{-1}\mu_1 + \mu_2^\top\Sigma_2^{-1}\mu_2\right)$$

2. Show that the marginal posterior of β given y is

$$\beta|y \sim \mathrm{T}_d(\mu_n, V_n\frac{k_n}{a_n}, 2a_n)$$

3. Show that the predictive distribution of an outcome $y_f = \Phi_f\beta + \epsilon$ with $\Phi_f = (\phi_0(x_f), ..., \phi_{d-1}(x_f))$ and$\epsilon \sim \mathrm{N}(0, \sigma^2)$ at untried location x_f is

$$y_f|y \sim \mathrm{T}_d(\mu_n, [\Phi^\top\Phi + 1]\frac{k_n}{a_n}, 2a_n)$$

Consider that

$$\mathrm{N}(x|\mu_1, \sigma_1^2)\,\mathrm{N}(x|\mu_2, \sigma_2^2) = \mathrm{N}(x|m, v^2)\,\mathrm{N}(\mu_1|\mu_2, \sigma_1^2 + \sigma_2^2)$$

where

$$v^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}; \quad m = v^2\left(\frac{\mu_1}{\sigma_1^2} + \frac{\mu_2}{\sigma_2^2}\right)$$

Hint-2: The definition of Student T is considered as known

Exercise 27. ($\star\star$)Let $y = (y_1, ..., y_n)$ be observables drawn iid from sampling distribution $y_i|\theta \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \theta^2)$ for all $i = 1, ..., n$, where $\theta \in \mathbb{R}$ is unknown. Specify a conjugate prior density for θ up to an unknown normalizing constant.

Exercise 28. ($\star\star$)If the sampling distribution $F(\cdot|\theta)$ is discrete and the prior $\Pi(\theta)$ is proper, then the posterior $\Pi(\theta|y)$ is always proper.

Exercise 29. ($\star\star$)If the sampling distribution $F(\cdot|\theta)$ is continuous and the prior $\Pi(\theta)$ is proper, then the posterior $\Pi(\theta|y)$ is almost always proper.

The Limit Comparison Theorem for Improper Integrals

General: Let integrable functions $f(x)$, and $g(x)$ for $x \geq a$.

Let

$$0 \leq f(x) \leq g(x), \quad \text{for } x \geq a$$

Then

$$\int_a^\infty g(x)\mathrm{d}x < \infty \implies \int_a^\infty f(x)\mathrm{d}x < \infty$$
$$\int_a^\infty f(x)\mathrm{d}x = \infty \implies \int_a^\infty g(x)\mathrm{d}x = \infty$$

Type I: Let integrable functions $f(x)$, and $g(x)$ for $x \geq a$, and let $g(x)$ be positive.

Let

$$\lim_{n\to\infty} \frac{f(x)}{g(x)} = c$$

Then

- If $c \in (0, \infty)$:
$$\int_a^\infty g(x)\mathrm{d}x < \infty \iff \int_a^\infty f(x)\mathrm{d}x < \infty$$
- If $c = 0$:
$$\int_a^\infty g(x)\mathrm{d}x < \infty \implies \int_a^\infty f(x)\mathrm{d}x < \infty$$
- If $c = \infty$:
$$\int_a^\infty f(x)\mathrm{d}x = \infty \implies \int_a^\infty g(x)\mathrm{d}x = \infty$$

Type II: Let integrable functions $f(x)$, and $g(x)$ for $a < x \leq b$, and let $g(x)$ be positive.

Let

$$\lim_{n\to a^+} \frac{f(x)}{g(x)} = c$$

Then

- If $c \in (0, \infty)$:
$$\int_a^\infty g(x)\mathrm{d}x < \infty \iff \int_a^\infty f(x)\mathrm{d}x < \infty$$
- If $c = 0$:
$$\int_a^\infty g(x)\mathrm{d}x < \infty \implies \int_a^\infty f(x)\mathrm{d}x < \infty$$
- If $c = \infty$:
$$\int_a^\infty f(x)\mathrm{d}x = \infty \implies \int_a^\infty g(x)\mathrm{d}x = \infty$$

Note: A useful test function is

$$\int_0^\infty \left(\frac{1}{x}\right)^p \mathrm{d}x \quad \begin{cases} < \infty & , \text{ when } p > 1 \\ = \infty & , \text{ when } p \leq 1 \end{cases}$$

Exercise 30. ($\star\star$)Consider the Bayesian model

$$\begin{cases} x|\sigma & \sim \mathrm{N}(0, \sigma^2) \\ \sigma & \sim \mathrm{Ex}(\lambda) \end{cases}$$

where $\mathrm{Ex}(\lambda)$ is the exponential distribution with mean $1/\lambda$. Show that the posterior distribution is not defined always.

- HINT: Precisely, show that the posterior is not defined in the case that you collect only one observation $x = 0$.

Exercise 31. ($\star\star$)Consider the Bayesian model

$$\begin{cases} x|\sigma & \sim \mathrm{N}(0, \sigma^2) \\ \sigma & \sim \Pi(\sigma) \end{cases}$$

where $\Pi(\sigma)$ is an improper prior distribution with density such as $\pi(\sigma) \propto \sigma^{-1} \exp(-a\sigma^{-2})$ for $a > 0$. Show that we can use this prior on Bayesian inference.

The Following exercise is part of Homework 1

Exercise 32. ($\star\star$)Let x be an observation. Consider the Bayesian model

$$\begin{cases} x|\theta & \sim \mathrm{Pn}(\theta) \\ \theta & \sim \Pi(\theta) \end{cases}$$

where $\mathrm{Pn}(\theta)$ is the Poisson distribution with expected value θ. Consider a prior $\Pi(\theta)$ with density such as $\pi(\theta) \propto \frac{1}{\theta}$. Show that the posterior distribution is not always defined.

Hint-1: It suffices to show that the posterior is not defined in the case that you collect only one observation $x = 0$.

Hint-2: Poisson distribution: $x \sim \mathrm{Pn}(\theta)$ has PMF

$$\mathrm{Pn}(x|\theta) = \frac{\theta^x \exp(-\theta)}{x!} 1(x \in \mathbb{N})$$

The next exercise is about the Sequential processing of data via Bayes theorem

Exercise 33. ($\star\star$)Assume that observable quantities $x_1, x_2, ...$ are generated i.i.d by a process that can be modeled as a sampling distribution $\mathrm{N}(\mu, \sigma^2)$with known σ^2 and unknown μ.

1. Assume that you have collected an observation x_1. Specify a prior $\Pi(\mu)$ on μ as $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$ where μ_0, σ_0^2 are known.

   - Derive the posterior $\Pi(\theta|x_1)$.

   Next assume that you additionally aobserve an additional observation x_2 after collecting x_1. Consider the posterior $\Pi(\mu|x_1)$ as the current state of your knowledge about θ.

   - Derive the posterior $\Pi(\mu|x_1, x_2)$ in the light of the new additional observation x_2.

2. Assume that you have collected two observations (x_1, x_2). Specify a prior $\Pi(\mu)$ on μ as $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$ where μ_0, σ_0^2 are known.

   - Derive the posterior $\Pi(\theta|x_1, x_2)$ in the light of the observations (x_1, x_2).

3. What do you observe:

Hint: We considered the identity

$$-\frac{1}{2}\sum_{i=1}^{n}\frac{(y-\mu_i)^2}{\sigma_i^2} = -\frac{1}{2}\frac{(y-\hat{\mu})^2}{\hat{\sigma}^2} + c(\hat{\mu},\hat{\sigma}^2),$$

$$c(\hat{\mu},\hat{\sigma}^2) = -\frac{1}{2}\sum_{i=1}^{n}\frac{\mu_i^2}{\sigma_i^2} + \frac{1}{2}(\sum_{i=1}^{n}\frac{\mu_i}{\sigma_i^2})^2(\sum_{i=1}^{n}\frac{1}{\sigma_i^2})^{-1}; \quad \hat{\sigma}^2 = (\sum_{i=1}^{n}\frac{1}{\sigma_i^2})^{-1}; \quad \hat{\mu} = \hat{\sigma}^2(\sum_{i=1}^{n}\frac{\mu_i}{\sigma_i^2})$$

where $c(\hat{\mu},\hat{\sigma}^2)$ is constant w.r.t. y.

Part V

Exchangeability

We work on the proofs of the following theorems:

- Marginal distributions of finite exchangeable sequences $y_1, y_2, \ldots, y_k$ are invariant under permutations; i.e.:
$$\mathrm{d}F(y_{\mathfrak{p}(1)}, y_{\mathfrak{p}(2)}, \ldots, y_{\mathfrak{p}(k)}) = \mathrm{d}F(y_1, y_2, \ldots, y_k) \text{ for all } \mathfrak{p} \in \mathfrak{P}_n. \tag{3}$$
In particular, for $k = 1$, it follows that all y_i are identically distributed (but not necessarily independently, as stated in the Lecture notes)
- (Marginal) Expectations of finite exchangeable sequences $y_1, y_2, \ldots, y_k$ are all identical:
$$\mathrm{E}(g(y_i)) = \mathrm{E}(g(y_1)) \text{ for all } i = 1, ..., k \text{ and all functions } g\colon \mathcal{Y} \to \mathbb{R} \tag{4}$$
- (Marginal) Variances of finite exchangeable sequences $y_1, y_2, \ldots, y_k$ are all identical:
$$\mathrm{Var}(y_i) = \mathrm{Var}(y_1). \tag{5}$$
- Covariances between elements of finite exchangeable sequences $y_1, y_2, \ldots, y_k$ are all identical:
$$\mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(y_1, y_2) \text{ whenever } i \neq j. \tag{6}$$

Just for your information The properties above are implied by the following general theorem. However, you should not use this theorem, directly, to solve the exercises below...

Theorem. *Consider an exchangeable sequence $y_1, \ldots, y_n$. Let $g\colon \mathcal{Y}^k \to \mathbb{R}$ be any function of k of these, where $k \leq n$. Then, for any permutation $\pi \in \Pi_n$,*

$$E(g(Y_{\mathfrak{p}(1)}, Y_{\mathfrak{p}(2)}, \ldots, Y_{\mathfrak{p}(k)})) = E(g(Y_1, Y_2, \ldots, Y_k)) \tag{7}$$

This is not an exercise to solve. Feel free to read the solution of this exercise, as it may help you understand the the Interpretation of the 'representation Theorem with $0-1$ quantities'.

Exercise 34. ($\star\star\star\star\star$)(Representation Theorem with $0-1$ quantities). If $y_1, y_2, ...$ is an infinitely exchangeable sequence of $0-1$ random quantities with probability measure P, there exists a distribution function Π such that the joint mass function $p(y_1, ..., y_n)$ for $y_1, ..., y_n$ has the form

$$p(x_1, ..., x_n) = \int_0^1 \prod_{i=1}^n \underbrace{\theta^{y_i}(1-\theta)^{1-y_i}}_{f_{\mathrm{Br}(\theta)}(y_i|\theta)} \mathrm{d}\Pi(\theta)$$

where

$$\Pi(t) = \lim_{n\to\infty} \Pr(\frac{1}{n}\sum_{i=1}^n y_i \leq t) \quad \text{and} \quad \theta \overset{\text{as}}{=} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n y_i$$

aka θ is the limiting relative frequency of 1s, by SLLN

Hint: (Helly's theorem [modified]) Given a sequence of distribution functions $\{F_1, F_2, ...\}$ that satisfy the tightness condition; [for each $\epsilon > 0$ there is a such that for all sufficiency large i it is $F_i(a) - F_i(-a) > 1 - \epsilon$], there exists a distribution F and a sub-sequence $\{F_{i_1}, F_{i_2}, ...\}$ such that $F_{i_j} \to F$.

Exercise 35. ($\star\star$)Clearly a set of independent and identically distributed random variables form an exchangeable sequence. Thus sampling with replacement generates an exchangeable sequence. What about sampling without replacement? Prove that sampling n items from N distinct objects without replacement (where $n \leq N$) is exchangeable.

Exercise 36. ($\star\star$)Let $Y_1, \ldots, Y_n$ be an exchangeable sequence, and let g be any function on $\mathcal{Y}$. Show, directly from the definition of exchangeability in the summary notes) that $\mathrm{E}(g(Y_i))$ does not depend on i:

$$\mathrm{E}(g(Y_i)) = \mathrm{E}(g(Y_1)) \text{ for all } i \in \{2, \ldots, n\} \tag{8}$$

For ease of exposition, you may restrict your proof to the case $i = 2$.

Exercise 37. ($\star\star$)Let $Y_1, \ldots, Y_n$ be an exchangeable sequence. Use

$$\mathrm{E}(g(Y_i)) = \mathrm{E}(g(Y_1)) \text{ for all } i \in \{2, \ldots, n\} \tag{9}$$

to show that $\mathrm{Var}(Y_i)$ does not depend on i:

$$\mathrm{Var}(Y_i) = \mathrm{Var}(Y_1) \text{ for all } i \in \{2, \ldots, n\} \tag{10}$$

Exercise 38. ($\star\star$)Let $Y_1, \ldots, Y_n$ be an exchangeable sequence. By expanding $var(\sum_{k=1}^{n} Y_k)$, show that when $i \neq j$,

$$cov(Y_i, Y_j) \geq -\frac{var(Y_1)}{n-1} \tag{11}$$

Exercise 39. ($\star$)What does

$$cov(Y_i, Y_j) \geq -\frac{var(Y_1)}{n-1}$$

imply about the correlation of infinite exchangeable sequences?

Part VI

Sufficiency

Exercise 40. ($\star\star$)Consider the Bayesian model

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

Hint-1: The PDF of $x \sim \text{G}(a, b)$ is $\text{Ga}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) 1_{(0,+\infty)}(x)$

Hint-2: The PDF of $x \sim \text{Ex}(\theta)$ is $\text{Ex}(x|\theta) = \text{Ga}(x|1, \theta)$

1. Show that the parametric model is member of the Exponential family, and the sufficient statistic for a sample of observables $x = (x_1, ..., x_n)$.
2. Show that the posterior distribution θ given x is Gamma and compute its parameters.
3. Show that the predictive distribution $G(z|x)$ of a future z given $x = (x_1, ..., x_n)$, has PDF

$$g(z|x) = \frac{a^*(b^*)^{a^*}}{(z + b^*)^{a^*+1}} 1(x \geq 0)$$

Exercise 41. ($\star\star\star$)Consider the Bayesian model

$$\begin{cases} x_i|\theta & \overset{\text{IID}}{\sim} \text{Mu}_k(\theta) \\ \theta & \sim \text{Di}_k(a) \end{cases}$$

where $\theta \in \Theta$, with $\Theta = \{\theta \in (0, 1)^k | \sum_{j=1}^k \theta_j = 1\}$ and $\mathcal{X}_k = \{x \in \{0, ..., n\}^k | \sum_{j=1}^k x_j = 1\}$.

Hint-1: Mu_k denotes the Multinomial probability distribution with PMF

$$\text{Mu}_k(x|\theta) = \begin{cases} \prod_{j=1}^k \theta_j^{x_j} & \text{, if } x \in \mathcal{X}_k \\ 0 & \text{, otherwise} \end{cases} \tag{12}$$

Hint-2: $\text{Di}_k(a)$ denotes the Dirichlet distribution with PDF

$$\text{Di}_k(\theta|a) = \begin{cases} \frac{\Gamma(\sum_{j=1}^k a_j)}{\prod_{j=1}^k \Gamma(a_j)} \prod_{j=1}^k \theta_j^{a_j - 1} & \text{, if } \theta \in \Theta \\ 0 & \text{, otherwise} \end{cases}$$

1. Show that the parametric model (27) is a member of the $k - 1$ exponential family.
2. Compute the likelihood $f(x_{1:n}|\theta)$, and find the sufficient statistic $t_n := t_n(x_{1:n})$.
3. Compute the posterior distribution. State the name of the distribution, and expresses its parameters with respect to the observations and the hyper-parameters of the prior. Justify your answer.
4. Compute the probability mass function of the predictive distribution for a future observation $y = x_{n+1}$ in closed form.

   Hint $\Gamma(x) = (x - 1)\Gamma(x - 1)$.

Exercise 42. ($\star\star$)Suppose that the vector $\mathbf{x} = (x, y, z)$ has a trinomial distribution depending on the index n and the parameter $\varpi = (\pi, \rho, \sigma)$ where $\pi + \rho + \sigma = 1$, that is

$$p(\mathbf{x}|\varpi) = \frac{n!}{x!\, y!\, z!}\pi^x \rho^y \sigma^z \qquad (x + y + z = n).$$

Show that this distribution is in the two-parameter exponential family.

Exercise 43. ($\star$)Establish the formula

$$(n_0^{-1} + n^{-1})^{-1}(\bar{x} - \theta_0)^2 = n\bar{x}^2 + n_0\theta_0^2 - n_1\theta_1^2$$

where $n_1 = n_0 + n$ and $\theta_1 = (n_0\theta_0 + n\bar{x})/n_1$.

This formula is often used to 'complete the square' in quadratiforms.

The following is a proof of a theorem

Exercise 44. ($\star\star\star$)Let $y_1, y_2...$ be an infinitely exchangeable sequence of random quantities. Let $t = t(y_1, ..., y_n)$ be a statistic for a finite $n \geq 1$. Then t is predictive sufficient if, and only if, it is parametric sufficient.

The following is a proof of a theorem

Exercise 45. ($\star\star\star$)(This is a theorem in Handout 5) Prove the following statement.

Let $t : \mathcal{Y} \to \mathcal{T}$ be a statistic. Then t is a parametric sufficient statistic for θ in the Bayesian sense if and only if the likelihood function $L(\cdot|\cdot)$ on $\mathcal{Y} \times \Theta$ can be factorized as the product of a kernel function k on $\mathcal{Y} \times \Theta$ and a residue function ρ on Θ as

$$L(y; \theta) = k(t(y)|\theta)\rho(y). \tag{13}$$

Part VII

Priors

Exercise 46. ($\star\star$)Let $y = (y_1, ..., y_n)$ be observable quantities, generated from an exponential family of distributions as

$$y_i|\theta \overset{\text{iid}}{\sim} \text{Ef}(u, g, h, c, \phi, \theta, c),\ i = 1, ..., n$$

with density

$$\text{Ef}(y_i|u, g, h, c, \phi, \theta, c) = u(y_i)g(\theta)^n \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(y_i))$$

and assume a conjugate prior $\Pi(\theta)$ with pdf/pmf

$$\pi(\theta) = \tilde{\pi}(\theta|\tau)\ \propto g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j)$$

1. Show that the posterior $\Pi(\theta|y)$ of θ has pdf/pmf $\pi(\theta|y) = \tilde{\pi}(\theta|\tau^*)$ with $\tau^* = (\tau_0^*, \tau_1^*, ..., \tau_k^*)$, $\tau_0^* = \tau_0 + n$, and $\tau_j^* = \sum_{i=1}^{n} h_j(x_i) + \tau_j$ for $j = 1, ..., k$, and pdf/pmf

$$\pi(\theta|y) = \pi(\theta|\tau^*)\ \propto g(\theta)^{\tau^*} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j^*) \tag{14}$$

   The operation $*$ here is addition $\tau * t(y) \longmapsto \tau + t(y) = \tau^*$

2. Show that the predictive distribution $G(z|y)$ for a new outcome $z = (y_{n+1}, ..., y_{n+m})$ has pdf/ pmf

$$g(z|y) = \prod_{i=1}^{m} u(z_i)\frac{K(\tau + t(y) + t(z))}{K(\tau + t_n(y))} \tag{15}$$

   where $t(z) = (m, \sum_{i=1}^{m} h_1(z_i), ..., \sum_{i=1}^{m} h_k(z_i))$.

Exercise 47. ($\star\star$)

1. Show that the skew-logistic family of distributions, with

$$f(x|\theta) = \frac{\theta e^{-x}}{(1 + e^{-x})^{\theta+1}} \tag{16}$$

   for $x \in \mathbb{R}$, labeled by $\theta > 0$, is a member of the exponential family and identify the factors u, g, h, ϕ, θ, c.

2. Show that the Gamma distribution

$$f(\theta|\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\theta^{\alpha_0-1}e^{-\beta_0\theta} \tag{17}$$

   with hyperparameters $w_0 \coloneqq (\alpha_0, \beta_0)$ (where $\alpha_0 > 1$ and $\beta_0 > 0$) is conjugate for i.i.d. sampling from for the skew-logistic distribution. Relate the hyperparameters (τ_0, τ_1) to the standard parameters (α_0, β_0) of the gamma distribution.

3. Given an i.i.d. sample $x_1, \ldots, x_n$ from the skew-logistic distribution, and assuming that the prior is Gamma(α_0, β_0), derive the posterior distribution of θ.

4. Given an i.i.d. sample $x_1, \ldots, x_n$ from the skew-logistic distribution, and assuming that the prior is Gamma(α_0, β_0), derive the predictive PDF for a future $y = x_{n+1}$ up to a normalising constant.

5. Give a minimal sufficient statistic for θ under i.i.d. sampling from the skew-logistic distribution. Would this statistic still be sufficient if we had chosen a prior for θ which was not a Gamma distribution?

The exercise below is theoretical, and very challenging.

Exercise 48. ($\star\star\star$)Prove the following statement.

Consider a PDF/PMF $f(x|\theta)$ with $x \in \mathcal{X}$ and $\theta \in \Theta$ where $\mathcal{X}$ does not depend on θ. If there exist a parametrised conjugate prior family $\mathcal{F} = (\pi(\theta|\tau), \tau \in T)$ with $\dim(\Lambda) < \infty$, then $f(x|\tau)$ is number of the exponential family.

Hint: Use the Pitman-Koopman Lemma:

> **Lemma.** *(Pitman-Koopman Lemma) If a family of distributions is such that for a large enough sample size there exist a sufficient statistic of constant dimension, then the family is an exponential family if the support does not depend on θ.*

PS: Please think about the importance of this result (!!!)

Exercise 49. ($\star\star$)Suppose that you have a prior distribution for the probability θ of success in a certain kind of gambling game which has mean 0.4, and that you regard your prior information as equivalent to 12 trials. You then play the game 25 times and win 12 times. What is your posterior distribution for θ?

Exercise 50. ($\star\star$)Find a (two-dimensional) sufficient statistic for (α, β) given an n-sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ from the two-parameter gamma distribution

$$p(x|\alpha, \beta) = \{\beta^\alpha \Gamma(\alpha)\}^{-1} x^{\alpha-1} \exp(-x/\beta) \qquad (0 < x < \infty)$$

where the parameters α and β can take any values in $0 < \alpha < \infty, 0 < \beta < \infty$.

Exercise 51. ($\star\star$)Suppose that your prior for θ with PDF

$$\pi(\theta) = \frac{2}{3}\mathrm{N}(\theta|0, 1) + \frac{1}{3}\mathrm{N}(\theta|1, 1)$$

that a single observation $x \sim \mathrm{N}(\theta, 1)$ turns out to equal 2. What is your posterior probability that $\theta > 1$?

Hint

We should use the following identity, discussed in the Lecture notes, :

$$-\frac{1}{2}\sum_{i=1}^{n}\frac{(y-\mu_i)^2}{\sigma_i^2} = -\frac{1}{2}\frac{(y-\hat{\mu})^2}{\hat{\sigma}^2} + c(\hat{\mu}, \hat{\sigma}^2),$$

$$c(\hat{\mu}, \hat{\sigma}^2) = -\frac{1}{2}\sum_{i=1}^{n}\frac{\mu_i^2}{\sigma_i^2} + \frac{1}{2}(\sum_{i=1}^{n}\frac{\mu_i}{\sigma_i^2})^2(\sum_{i=1}^{n}\frac{1}{\sigma_i^2})^{-1};$$

$$\hat{\sigma}^2 = (\sum_{i=1}^{n}\frac{1}{\sigma_i^2})^{-1} \quad ; \quad \hat{\mu} = \hat{\sigma}^2(\sum_{i=1}^{n}\frac{\mu_i}{\sigma_i^2})$$

where $c(\hat{\mu}, \hat{\sigma}^2)$ is constant w.r.t. y.

The Exercise below has been set as Homework 2

Exercise 52. ($\star\star$)Let $x = (x_1, ..., x_n)$ be observables. Consider a Bayesian model such as

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

Hint-1 Poisson distribution $x \sim \text{Pn}(\lambda)$ has PMF: $\text{Pn}(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)1_{\mathbb{N}}(x)$, where $\mathbb{N} = \{0, 1, 2, ...\}$ and $\lambda > 0$.

Hint-2 Gamma distribution $x \sim \text{Ga}(a, b)$ has PDF: $\text{Ga}(x|a, b) = \frac{b^a}{\Gamma(a)}x^{a-1} \exp(-bx)1_{(0,\infty)}(x)$, with $a > 0$ and $b > 0$.

Hint-2 Negative Binomial distribution $x \sim \text{Nb}(r, \theta)$ has PMF: $\text{Nb}(x|r, \theta) = \binom{r+x-1}{r-1}\theta^r(1-\theta)^x 1_{\mathbb{N}}(x)$ with $\theta \in (0, 1)$, $r \in \mathbb{N} - \{0\}$, and $\mathbb{N} = \{0, 1, 2, ...\}$.

1. Compute the likelihood in the aforesaid Bayesian model.
2. Show that the sampling distribution is a member of the exponential family.
3. Specify the PDF of the conjugate prior distribution $\Pi(\lambda)$ of λ, and identify the parametric family of distributions as $\lambda \sim \text{Ga}(a, b)$, with $a > 0$, and $b > 0$. While you are deriving the conjugate prior distribution of λ, discuss which of the prior hyper-parameters can be considered as the 'strength of the prior information and which can be considered as summarizing the prior information.
4. Compute the PDF of the posterior distribution of λ, identify the posterior distribution as a Gamma distribution $\text{Ga}(\tilde{a}, \tilde{b})$, and compute the posterior hyper-parameters $\tilde{a}$, and $\tilde{b}$.
5. Compute the PMF of the predictive distribution of a future outcome $y = x_{n+1}$, identify the name of the resulting predictive distribution, and compute its parameters.

Exercise 53. ($\star\star$)Consider the Bayesian model

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

where $\theta \in \Theta$, with $\Theta = \{\theta \in (0, 1)^k | \sum_{j=1}^k \theta_j = 1\}$ and $\mathcal{X}_k = \{x \in \{0, 1\}^k | \sum_{j=1}^k x_j = 1\}$.

Hint-1: Mu_k denotes the Multinomial probability distribution with PMF

$$\text{Mu}_k(x|\theta) = \begin{cases} \prod_{j=1}^k \theta_j^{x_j} & \text{, if } x \in \mathcal{X}_k \\ 0 & \text{, otherwise} \end{cases}$$

Hint-2: $\text{Di}_k(a)$ denotes the Dirichlet distribution with PDF

$$\text{Di}_k(\theta|a) = \begin{cases} \frac{\Gamma(\sum_{j=1}^k a_j)}{\prod_{j=1}^k \Gamma(a_j)} \prod_{j=1}^k \theta_j^{a_j - 1} & \text{, if } \theta \in \Theta \\ 0 & \text{, otherwise} \end{cases}$$

1. Derive the conjugate prior distribution for θ, and recognize that it is a Dirichlet distribution family of distributions.

2. Verify that the prior distribution you derived above is indeed conjugate by using the definition.

Exercise 54. ($\star\star$)Consider the Bayesian model

$$\begin{cases} x|\theta & \sim \text{Ga}(a,\theta) \\ \theta & \sim \Pi(\theta) \end{cases}$$

where $\text{Ga}(a,\theta)$ is the Gamma distribution with expected value a/θ and density function

$$f(x|a,\theta) = \frac{\theta^a}{\Gamma(a)} x^{a-1} \exp(-\theta x) 1\left(x \geq 0\right).$$

Specify a Jeffrey's prior density for θ.

The Exercise below has been set as Homework 2

Exercise 55. ($\star\star$)Assume observation x sampled from a Maxwell distribution with density

$$f(x|\theta) = \sqrt{\frac{2}{\pi}} \theta^{3/2} x^2 \exp(-\frac{1}{2}\theta x^2).$$

Find the Jeffreys prior density for the parameter θ.

Exercise 56. ($\star\star$)Consider the trinomial distribution

$$\begin{aligned} p(x,y|\pi,\rho) &= \frac{n!}{x!\,y!\,z!} \pi^x \rho^y \sigma^z, \qquad (x+y+z=n) \\ &\propto \pi^x \rho^y (1-\pi-\rho)^{n-x-y}. \end{aligned}$$

Specify a Jeffreys' prior for (π,ρ).

HINT: It is $\text{E}(x) = n\pi$, $\text{E}(y) = n\rho$.

Exercise 57. ($\star\star$)Suppose that x has a Pareto distribution $\text{Pa}(\xi,\gamma)$ where ξ is known but γ is unknown, that is,

$$p(x|\gamma) = \gamma \xi^\gamma x^{-\gamma-1} I_{(\xi,\infty)}(x).$$

Use Jeffreys' rule to find a suitable reference prior for γ.

Exercise 58. ($\star\star$)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 $\text{Ga}(a,\beta)$ 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,

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

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

Exercise 59. ($\star\star\star$)Consider the the model of Normal linear regression where the observables are pairs (ϕ_i, y_i) for $i = 1, ..., n$, assumed to be modeled according to the sampling distribution

$$y_i|\beta, \sigma^2 \sim \mathrm{N}(\phi_i^\top \beta, \sigma^2)$$

for $i = 1, ..., n$ with unknown $(\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$. Namely,

$$y|\beta, \sigma^2 \sim \mathrm{N}_n(\Phi\beta, I\sigma^2)$$

where $y = (y_1, ...y_n)$, and Φ is the design matrix. Here β is d-dimensional. Find the Jeffreys' priors for (β, σ^2).

Hint: Recall your AMV: $\frac{\mathrm{d}}{\mathrm{d}x}x^\top Ax = 2Ax$, $\frac{\mathrm{d}}{\mathrm{d}x}(c + Ax) = A$, and $\frac{\mathrm{d}}{\mathrm{d}x}(A(x))^\top = \left(\frac{\mathrm{d}}{\mathrm{d}x}A(x)\right)^\top$.

Hint: If $y|\beta, \sigma^2 \sim \mathrm{N}_n(\Phi\beta, I\sigma^2)$, then $\mathrm{E}_{y|\beta,\sigma^2\sim\mathrm{N}_n(\Phi\beta, I\sigma^2)}\left((y - \Phi\beta)^\top (y - \Phi\beta)\right) = n\sigma^2$.

Exercise 60. ($\star\star$)Consider the Bayesian model

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

where $\mathrm{Pn}(\theta)$ is the Poisson distribution with expected value θ. Specify a Jeffreys' prior for θ.

Hint: Poisson distribution: $x \sim \mathrm{Pn}(\theta)$ has PMF

$$\mathrm{Pn}(x|\theta) = \frac{\theta^x \exp(-\theta)}{x!}1(x \in \mathbb{N})$$

Exercise 61. ($\star\star$)Consider the Bayesian model

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

where $\mathrm{Pn}(\theta)$ is the Poisson distribution with expected value θ. Specify a Maximum entropy prior under the constrain $\mathrm{E}(\theta) = 2$ and reference measure such as $\pi_0(\theta) = \frac{1}{\sqrt{\theta}}$. In particular, you also have to state the name of the derived Maximum entropy prior distribution and report the values of its parameters.

Hint-1: Poisson distribution: $x \sim \mathrm{Pn}(\theta)$ has PMF

$$\mathrm{Pn}(x|\theta) = \frac{\theta^x \exp(-\theta)}{x!}1(x \in \mathbb{N})$$

Hint-2: Gamma distribution: $x \sim \mathrm{Ga}(a, b)$ has PDF

$$\mathrm{Ga}(x|a, b) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-\beta x)1(x > 0)$$

Part VIII

Decision theory

Exercise 62. ($\star\star\star$) Show that, under the squared error loss, if two unbiased and independent real estimators δ_1 and δ_2 are distinct and satisfy

$$R(\theta, \delta_1) = \mathrm{E}_F(\theta - \delta_1(x))^2 = R(\theta, \delta_2) = \mathrm{E}_F(\theta - \delta_2(x))^2,$$

the estimator δ_1 is not admissible.

Hint: Consider $\delta_3 = (\delta_1 + \delta_2)/2$ or $\delta_4 = \delta_1^a \delta_2^{1-a}$.

Hint: Jensen's inequality: Let $g(\cdot)$ be a convex function and X be a random variable follwoing a distribution F then

$$g(\mathrm{E}_F(X)) \leq \mathrm{E}_F(g(X))$$

Extend this result to all strictly convex losses and construct a counter-example when the loss function is not convex.

Part IX

Point estimation

Exercise 63. ($\star\star$)Consider a Bayesian model

$$\begin{cases} x_i|\sigma^2 & \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2), \qquad i = 1, ..., n \\ \sigma^2 & \sim \text{IG}(a, b) \end{cases}$$

where, $\mu \in \mathbb{R}$ is known and $a > 0$, $b > 0$. We denote the observables as $x = (x_1, ..., x_n)$.

Find the Bayesian parametric point estimator of σ^2, for the squared, and zero-one loss functions.

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

Hint: The posterior distribution of σ^2 is $\text{IG}(\text{d}\sigma^2|a + \frac{1}{2}n, b + \frac{n}{2}s^2)$, where s^2 is the sample variance.

Exercise 64. ($\star\star$)Consider a Bayesian model

$$\begin{cases} x_i|\theta & \overset{\text{iid}}{\sim} \text{Ex}(\theta), \qquad i = 1, ..., n \\ \theta & \sim \text{Ga}(a, b) \end{cases}$$

where $a > 0$, $b > 0$. We denote the observables as $x = (x_1, ..., x_n)$.

1. Find the Bayesian parametric point estimator of θ, for the squared, and zero-one loss functions.

   Hint: The posterior distribution of θ is $\text{Ga}(a + n, b + n\bar{x})$

2. Find the Bayesian predictive point estimator of unobserved y, for the squared, and zero-one loss functions.

   Hint The predictive pmf of y is

$$g(y|x) = \frac{(b + n\bar{x})^{a+n}}{\Gamma(a + n)} \frac{\Gamma(a + n + 1)}{\Gamma(1)} (b + n\bar{x} + y)^{-(a+n)-1}$$

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

Hint: Exponential distr.: $x \sim \text{Ex}(b)$ has pdf $f(x) = b \exp(-bx) 1_{(0,+\infty)}(x)$,

---

Exercise 65. ($\star\star$)Consider observables $x = (x_1, ..., x_n)$. Consider the Bayesian model

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

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 $B \sim \text{LN}(\mu_A, \sigma_A^2)$ with parameters μ_A, σ_A^2 if $B = \exp(A)$ where $A \sim \text{N}(\mu_A, \sigma_A^2)$.

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

Hint-3: It is

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

where

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

Exercise 66. ($\star\star$)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, $p(\vec{x}|y) = p(x_1, \ldots, x_n|y)$ based on some data y. We decide to use the following loss function:

$$\ell(\hat{\vec{x}}, \vec{x}) = \sum_{i=1}^{n} (1 - \delta(\hat{x}_i, x_i)) \tag{18}$$

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.

The following exercise is given as a Homework 3. Borrowed from Dr I. H. Jermyn @ Durham U.

Exercise 67. ($\star\star$) [1]This exercise is based on a problem that arises in image processing. Look at the first row of Fig 1. If we were to observe the sunflower field from above, the sunflowers would be spread uniformly over it. Viewed from an angle, the sunflowers cluster at the top of the picture due to the effect of perspective. We would like to be able to tell from this clustering at what angle the camera was pointing and its height above the ground. We will not solve this problem here (it is rather difficult in general), but instead look at an idealized and simplified version of it.

Consider the left hand image in the second row of the figure. It shows 200 points sampled at random uniformly from the unit square. On the right, is a transformation of these points similar to that undergone by the sunflower image, except that here only the 'y-coordinate', the vertical position in the image, has been affected.

[1]This exercise is a modified version of the one from Dr Jermyn's lecture notes in Bayesian statistics 2015-2016

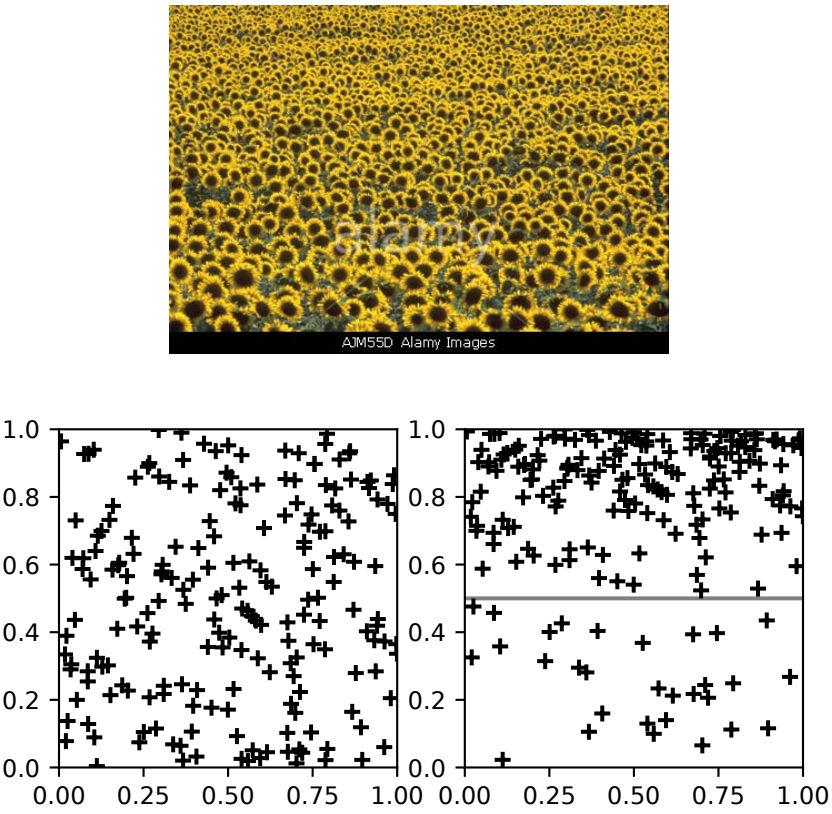

Figure 1: Two sampled point patterns. (Sunflower image © Soren Breiting / Alamy.)

If we were asked whether the right hand image had been sampled from a uniform distribution on the unit square, I am sure we would all say 'no'. The question is how we can justify this response. The first part of the exercise is an alternative to the examples given in lectures, showing the need to use Bayesian and not ad hoc methods to obtain sensible answers in many inference problems.

The second part of the exercise is about inferring the camera angle given the data points using Bayesian methods, and tests various technical issues.

1. **Classical treatment.**

   A classical statistical technique to address this problem might go like this. Let's define a 'statistic', a quantity to be calculated from the data, and whose properties we will study to create a test. For example, in a coin tossing experiment, we will get some sequence of heads and tails. A statistic might be the number of heads.

   In this case, one possibility is to divide the unit square into two halves, top and bottom, and to count the number r of points in the bottom half.

   (a) If we assume that both point patterns were sampled from a uniform distribution on the unit square, which of the two sets of points is more probable? Does this help to justify the inference that the right hand image was not sampled from a uniform distribution?

   (b) If the total number of points is n, what is the probability distribution for r (the 'sampling distribution') under the assumption of sampling from the uniform distribution on the unit square?

   (c) What are the mean and variance of this distribution?

(d) For the right hand image in the bottom row of Fig 1, by visual inspection, extract (approximately) the value of the statistic r.

(e) Using a normal approximation to the sampling distribution, perform a significance test under the null hypothesis H_0 that the sampling was uniform. What is your conclusion?

(f) How do we reconcile the answer to Sub-question 1a with the result of the hypothesis test? When we calculate the probability of observing r points in the bottom half of the unit square, what are we actually calculating?

(g) What would be the result of the hypothesis test if we defined r as the number of points in the left half of the unit square?

(h) Is this a reasonable thing to do? Why?

What is being introduced in the last question is a possible alternative hypothesis, specifying the nature of the non-uniformity. The problem is that this alternative hypothesis has no place in the classical statistical testing methodology: all we have is H_0.

As a result of this deficiency of standard hypothesis testing, other methods have been developed. Likelihood methods in classical statistics take alternatives into account by using two (or more) hypotheses and comparing the probabilities of the data under each of them.

In our example, we could take a non-uniform model, and then compute the probabilities of the data under both uniform and non-uniform models. This would work in one sense: the probability of the data would be higher under some non-uniform models than under the uniform model. Unfortunately, there are many non-uniform models. Some of them, those with probability densities concentrated around the data points, assign very high probability to the observed data, and yet we do not accept them as valid explanations. This is a usuall phenomenon in Frequentist statistics: there are many hypotheses that predict the observed data with near certainty, and maximum likelihood is powerless to discount them.

2. **Bayesian treatment.**

To disallow these extreme possibilities (if indeed they are unreasonable), we have to assign probabilities to the possible non-uniformities. One way to do this is via a choice of a parameterized family of non-uniformities. Any parameterized model implicitly assigns probability zero to any non-family member, and hence is Bayesian by default.

In the case of the image processing problem, we know a lot about the types of distortion that arise (essentially perspective), and we can construct a reasonable family quite easily. Without going into details, and making several approximations, the coordinates $(x, y) \in [0, 1]^2$ of a point in the image distorted by camera viewing angle are related to the coordinates $(u, v) \in [0, 1]^2$ of the same point in the undistorted image that would be taken by a camera looking vertically downwards, by the following equations:

$$x = u \tag{19}$$

$$y = \frac{v(1+t)}{1+tv} \tag{20}$$

where $t = \tan(\alpha)$ is the tangent of the angle α between the camera viewing direction and vertically downwards—in fact, this was the exact transformation used to convert the left hand image in Fig 1 to the right hand image.

(a) Suppose we knew t. Derive the probability density $f(u, v|t)$ for a point (x, y) in the distorted image, given t, and given that the sampling density was uniform on the unit square in the undistorted image, i.e.

$$f(u, v|t) = 1 \tag{21}$$

(b) Write down the corresponding probability density, given t, of a set of points (x_1, y_1), …, (x_n, y_n) sampled independently from the non-uniform density.

(c) Suppose we have no reason to favour any particular value of $\alpha \in [0, \pi/2]$ before we see the data. Write down the prior probability density for α.

(d) From the answer to Sub-question 2c, derive the prior probability density of t. [Hint: $\frac{\mathrm{d}}{\mathrm{d}t}(\tan^{-1} t) = \frac{1}{1+t^2}$.]

(e) Hence write down the posterior probability density for t up to an overall normalization factor, given the data points (x_1, y_1), …, (x_n, y_n).

(f) From this result, derive the equation satisfied by the MAP estimate of t. (Taking logarithms makes things easier.)

(g) By expanding the log posterior probability density about 0 to second order in t, find the MAP estimate for t when t is small.

(h) Find the MAP estimate of α when α is small.

Solution.

1. **Classical treatment.**

(a) The probability that a single point falls in the infinitesimal element $\mathrm{d}u\mathrm{d}v$ at point (u, v) is:

$$\mathrm{d}F(u, v) = \mathsf{P}_F(u \in \mathrm{d}u, v \in \mathrm{d}v) = \mathrm{d}u\mathrm{d}v \tag{22}$$

The probability that n points sampled independently fall in the infinitesimal elements $\mathrm{d}u_1\mathrm{d}v_1$, …, $\mathrm{d}u_n\mathrm{d}v_n$ at points (u_1, v_1), …, (u_n, v_n) is therefore

$$\mathrm{d}F(u_1, v_1, \ldots, u_n, v_n) = \prod_{i=1}^{n} \mathrm{d}F(u_i, v_i) = \prod_{i=1}^{n} \mathrm{d}u_i\mathrm{d}v_i \tag{23}$$

This does not depend on the data points (u_i, v_i) and so is the same for both patterns.[2]

If we want to justify the idea that the set of points in the right hand image did not come from a uniform distribution, this obviously does not help, since both cases are the same.

(b) The probability of one point landing in the bottom half of the square is $\frac{1}{2}$ for a uniform distribution. The probability of any *given* set of r points lying in the bottom half (and thus $n - r$ lying in the top half) is then $(\frac{1}{2})^r(\frac{1}{2})^{n-r}$. The probability of some set of r points lying in the bottom half is thus given by the binomial distribution:

$$\mathsf{P}(r|n, \text{uniform}) = \binom{n}{r}\frac{1}{2^n} \tag{24}$$

(c) The mean of a binomial distribution is np and the variance is $np(1-p)$, meaning, in this case, that the mean is $\frac{n}{2}$ and the variance is $\frac{n}{4}$. For the given example, these are mean 100 and standard deviation $\sqrt{50} \simeq 7$.

(d) There are about $r = 30$ points in the bottom half of the square. (The exact number is not so relevant here.)

(e) The standardized value of the statistic r is $z = (r - 100)/7$, so for $r = 30$ we get $z = -10$, i.e. 10 standard deviations below the mean. The probability of finding this or a more extreme value in the bottom half of the square is then $2(1 - \Phi(10))$, a number that is vanishingly small. The null hypothesis is thus convincingly rejected.

[2] There is a subtlety here. The above probability is that the points fall in the given elements with the given labelling. Because there are $n!$ ways to label n points, strictly speaking, the probability of a configuration is $n!$ times the above, with the understanding that now the distribution is defined on sets of unlabelled points.

(f) Superficially there seems to be a contradiction between the fact that the probabilities of the two sets of data are the same, but the hypothesis test so strongly rejects the null hypothesis. In fact, of course, the probability of the data and the probability computed in the hypothesis test are completely different. The former is the probability of a particular set of positions for points. The latter is a different in two ways. Remember that the probability of an individual configuration is constant under the uniform hypothesis. Then first, the probability of a particular value of r is given by the integral of this constant over all possible positions of the points that keeps the same number in the bottom half of the square. Second, the probability in the hypothesis test is then the sum of these probabilities for all values of r 'more extreme', i.e. further from the mean, than the value we observe. For the hypothesis test, then, we calculate the probability of a large, indeed in this case infinite, set of conceivable data sets, none of which we have actually observed.

Naturally, with such a difference in the probabilities, the conclusions to be drawn are different.

(g) The test does not justify rejecting the null hypothesis.

(h) It seems unreasonable because we know or suspect that the non-uniformity is in the vertical direction, and the statistic should be some measure of this non-uniformity. However, this alternative hypothesis has no place in the classical statistical testing methodology: all we have is H_0.

2. **Bayesian treatment.**

Part X

Credible regions

Exercise 68. (⋆⋆⋆)Consider the Bayesian model

$$\begin{cases} x_i|\sigma^2 & \overset{\text{IID}}{\sim} \text{N}(\mu, \sigma^2), \qquad i = 1, ..., n \\ \sigma^2 & \sim \text{IG}(a, b) \end{cases} \tag{25}$$

where μ is known, and the prior hyper-parameters a, b are known.

1. Calculate the posterior distribution of σ^2 of Bayesian model (25). Justify your calculations with brief comments.

   Hint: It is

   $$\sigma^2|x_1, ..., x_n \sim \text{IG}(a^*, b^*)$$

   with $a^* = \frac{n}{2} + a$ and $b^* = b + \frac{1}{2}ns^2$.

2. Prove that the predictive distribution for a future $y = x_{n+1} \in \mathbb{R}$, given the Bayesian model (25), is Student T such as

   $$\text{T}(\mu, \frac{b + \frac{1}{2}ns^2}{a + \frac{1}{2}n}, 2a + n)$$

   Justify your calculations with brief comments.

3. Find the C_γ (posterior) parametric HPD credible interval for σ^2. In particular, find the system of equations whose roots are the boundaries of the HPD credible interval. Justify your calculations with brief comments.

4. Find the C_γ predictive HPD credible interval for a future $y = x_{n+1} \in \mathbb{R}$. Justify your calculations with brief comments.

5. Consider a sample size $n = 40$, sample variance $s^2 = 2$, fixed hyper-parameters $a = 10$, $b = 5$, and known $\mu = 1$. Consider that the 97.5% quantile of the standard Student t distribution with 14 degrees of freedom is equal to $t_{60,\,0.975} \approx 2$. Compute the values of the boundaries of the 95% predictive HPD credible interval $C_{\gamma=0.05}$ for $y = x_{n+1} \in \mathbb{R}$. Justify your calculations with brief comments.

Hint: $x \sim \text{IG}(a, b)$, then the PDF of x is

$$f_{\text{IG}(a,b)}(x) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp(-\frac{b}{x}) 1_{(0,+\infty)}(x)$$

Hint: Student distribution: $x \sim \text{T}(\mu, v, a)$ iff $\sqrt{1/v}(x - \mu) \sim \text{T}(0, 1, a)$; $\text{T}(0, 1, a)$ is the standard Student t_a distribution.

Hint: Student distribution: $x \sim \text{T}(\mu, v, a)$, then $E(x) = \mu$ and $Var(x) = \frac{a}{a-2}v$.

Exercise 69. (⋆⋆)Consider the Bayesian model

$$\begin{aligned} y_i|\theta &\overset{\text{iid}}{\sim} \text{N}\left(\theta, \sigma^2\right), \; i = 1, ..., n \\ \theta &\sim \text{N}\left(\mu_0, \sigma_0^2\right) \end{aligned}$$

where μ_0, σ_0^2 are fixed hyper-parameters, and θ unknown.

1. Derive the $1-\frac{a}{2}$ HPD credible posterior interval for θ.

   Hint-1: It is

   $$\sum_{i=1}^{n}\frac{(x-\mu_i)^2}{\sigma_i^2}=\frac{(x-\hat{\mu})^2}{\hat{\sigma}^2}+\text{const ind of } x$$

   where $\hat{\sigma}^2=\left(\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\right)^{-1}$ and $\hat{\mu}=\hat{\sigma}^2\left(\sum_{i=1}^{n}\frac{\mu_i}{\sigma_i^2}\right)$.

   Hint-2: The 97.5% quantile of the standard Normal distribution is 1.959964.

2. What size your dataset need to have in order to satisfy a 0.95% HPD credible posterior interval for θ which has length of 1 unit? Consider that $\sigma^2=4$ and $\sigma_0^2=9$.

Exercise 70. ($\star\star\star\star$)A random sample, $\boldsymbol{x}=(x_1,x_2,...,x_n)$, of size n is taken from N(θ,σ^2), where σ^2 is known. Suppose there exist positive constants; a, ϵ, M and c (small values of a and ϵ are of interest), such that in the interval I_a, defined by

$$\bar{x}-\lambda_a\sqrt{\sigma^2/n}\leq\theta\leq\bar{x}+\lambda_a\sqrt{\sigma^2/n}$$

where $2\Phi(-\lambda_a)=a$, the prior density of θ lies between $c(1-\epsilon)$ and $c(1+\epsilon)$: and outside I_a it is bounded by Mc. Then the posterior density $\pi(\theta|x)$ satisfies the inequalities

$$\frac{1-\epsilon}{(1+\epsilon)(1-a)+Ma}\sqrt{\frac{1}{2\pi\sigma^2/n}}\exp(-\frac{1}{2}\frac{(\bar{x}-\theta)^2}{\sigma^2/n})\quad\leq\pi(\theta|x)\leq\quad\frac{1+\epsilon}{(1-\epsilon)(1-a)+Ma}\sqrt{\frac{1}{2\pi\sigma^2/n}}\exp(-\frac{1}{2}\frac{(\bar{x}-\theta)^2}{\sigma^2/n})$$

inside I_a, and

$$0\quad\leq\pi(\theta|x)\leq\quad\frac{M}{(1-\epsilon)(1-a)}\sqrt{\frac{1}{2\pi\sigma^2/n}}\exp(-\frac{1}{2}\lambda_a^2)$$

outside I_a.

Comment This is a nice result that shows how the posterior PDF of the mean of the (what we call) Normal model with known variance behaves.

Part XI

Hypothesis tests

Exercise 71. ($\star\star$)Consider a Bayesian model

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

Hint-1 Poisson distribution has PMF: $\text{Pn}(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda) 1_{\mathbb{N}}(x)$

Hint-2 Gamma distribution has PDF: $\text{Ga}(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) 1_{(0,\infty)}(x)$, with $E(x) = a/b$, $Var(x) = a/b^2$.

Hint-3 Negative Binomial distribution has PMF: $\text{Nb}(x|r, \theta) = \binom{r+x-1}{r-1} \theta^r (1-\theta)^x 1_{\mathbb{N}}(x)$. with $\theta \in (0, 1)$, $r \in \mathbb{N}$.

1. Show that the sampling distribution (param. model) is a member of the exponential family
2. Compute the likelihood
3. Specify the PDF of the conjugate prior distribution of λ, and identify the distribution.
4. Compute the PDF of the posterior distribution of λ, and identify the distribution.
5. Compute the PMF of the predictive distribution of $y = x_{n+1}$,and identify the distribution.
6. Consider that we are interested in testing the hypothesis whether $\lambda = \lambda_0$, (where λ_0 is a fixed known number), or not.
   (a) Design the test of hypotheses in Bayesian framework: Namely, set pair of hypotheses, specify priors, and compute the associated Bayes Factor.
   (b) Compute the posterior probability that $\lambda = \lambda_0$.
   (c) Perform the hypothesis test to test if $\lambda = 2$ or not based on the Jeffrey's scaling rule, by considering that
   - we have collected two observations $x_1 = 2$, $x_2 = 3$,
   - a priori the probability that $\{\lambda = 2\}$ is 0.5,
   - given $\{\lambda \neq 2\}$, the prior distr. of λ is a conjugate one with $E(\lambda) = 2$, and $Var(\lambda) = 1$.

Exercise 72. ($\star\star\star$) Let $y = (y_1, ..., y_n)$ observables and consider the Bayesian hypothesis test

$$\text{H}_0 : y_i|\theta_0 \overset{\text{IID}}{\sim} \text{N}(\theta_0, \sigma^2),\ i = 1, ..., n \qquad \text{vs} \qquad \text{H}_1 : \begin{cases} y_i|\theta & \overset{\text{IID}}{\sim} \text{N}(\theta, \sigma^2),\ i = 1, ..., n \\ \theta & \sim \text{N}(\mu_0, \sigma_0^2) \end{cases}$$

with $\pi_j = \text{P}_\Pi(\theta \in \Theta_j)$ for $j = 0, 1$. Let $\theta_0, \sigma^2, \mu_0, \sigma_0^2$ be fixed values.

1. Let $\text{B}_{01}(y)$ denote the Bayes factor $\text{B}_{01}(y)$ defined as the ratio of the posterior probabilities of H_0 and H_1 over the ratio of the prior probabilities of H_0 and H_1. Calculate

$$\text{B}_{01}(y) = \frac{\left(\frac{\sigma^2}{n}\right)^{-\frac{1}{2}}}{\left(\frac{\sigma^2}{n} + \sigma_0^2\right)^{-\frac{1}{2}}} \frac{\exp\left(-\frac{1}{2}\frac{(\bar{y}-\theta_0)^2}{\frac{\sigma^2}{n}}\right)}{\exp\left(-\frac{1}{2}\frac{(\bar{y}-\mu_0)^2}{\frac{\sigma^2}{n}+\sigma_0^2}\right)};$$

2. Let $\mathsf{P}_\Pi(\mathrm{H}_0|y)$denote the posterior probability of the null hypothesis H_0. Calculate:

$$\mathsf{P}_\Pi(\mathrm{H}_0|y) = \left(1 + \frac{1-\pi_0}{\pi_0} \frac{\left(\frac{\sigma^2}{n} + \sigma_0^2\right)^{-\frac{1}{2}}}{\left(\frac{\sigma^2}{n}\right)^{-\frac{1}{2}}} \frac{\exp\left(-\frac{1}{2}\frac{(\bar{y}-\mu_0)^2}{\frac{\sigma^2}{n}+\sigma_0^2}\right)}{\exp\left(-\frac{1}{2}\frac{(\bar{y}-\theta_0)^2}{\frac{\sigma^2}{n}}\right)}\right)^{-1}$$

Hint-1: It is

$$-\frac{1}{2}\sum_{i=1}^{n}\frac{(x-\mu_i)^2}{\sigma_i^2} = -\frac{1}{2}\frac{(x-\hat{\mu})^2}{\hat{\sigma}^2} + C(\hat{\mu},\hat{\sigma}^2)$$

$$\hat{\sigma}^2 = (\sum_{i=1}^{n}\frac{1}{\sigma_i^2})^{-1}; \qquad \hat{\mu} = \hat{\sigma}^2(\sum_{i=1}^{n}\frac{\mu_i}{\sigma_i^2}); \qquad C(\hat{\mu},\hat{\sigma}^2) = \underbrace{\frac{1}{2}\frac{(\sum_{i=1}^{n}\frac{\mu_i}{\sigma_i^2})^2}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}} - \frac{1}{2}\sum_{i=1}^{n}\frac{\mu_i^2}{\sigma_i^2}}_{=\text{independent of } x}$$

Hint-2: It is $\sum_{i=1}^{n}(x_i-\theta)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta)^2$

Exercise 73. ($\star\star\star$)Recall the linear regression model mapping in a linear form the dependent variable y with a set of regressors $\{\Phi_j\}_{j\in\mathscr{M}}$ where $\mathscr{M}$ is the set of size d that includes the labels of the available regressors; e.g.

$$y_i|\beta,\sigma^2 \sim \mathrm{N}\left(\sum_{j\in\mathscr{M}}\Phi_{i,j}\beta_j, I\sigma^2\right), \quad \text{for } i=1,...,n$$

where the regression coefficients $\{\beta_j\}_{j\in\mathscr{M}}$ and the noise variance σ^2 are unknown.

Let $\mathscr{M}_0$ and $\mathscr{M}_1$ denote two sets of regressors (nested or not) with $\dim(\mathscr{M}_j) = d_j$. We are interested in testing whether the linear model with $\mathscr{M}_0$ set of regressors or that with $\mathscr{M}_0$ set of regressors models the data generating processes 'better'. I.e. we test

$$\mathrm{H}_0: \begin{cases} y|\beta_{\mathscr{M}_0},\sigma^2 & \sim \mathrm{N}(\Phi_{\mathscr{M}_0}\beta_{\mathscr{M}_0}, I\sigma^2) \\ \beta_{\mathscr{M}_0}|\sigma^2 & \sim \mathrm{N}(\mu_{\mathscr{M}_0}, V_{\mathscr{M}_0}\sigma^2) \\ \sigma^2 & \sigma^2 \sim \mathrm{IG}(a,k) \end{cases} \quad \text{v.s.} \quad \mathrm{H}_1: \begin{cases} y|\beta_{\mathscr{M}_1},\sigma^2 & \sim \mathrm{N}(\Phi_{\mathscr{M}_1}\beta_{\mathscr{M}_1}, I\sigma^2) \\ \beta_{\mathscr{M}_1}|\sigma^2 & \sim \mathrm{N}(\mu_{\mathscr{M}_1}, V_{\mathscr{M}_1}\sigma^2) \\ \sigma^2 & \sigma^2 \sim \mathrm{IG}(a,k) \end{cases}$$

1. Calculate the Bayes factor $B_{01}(y)$ as

$$B_{01}(y); = \sqrt{\frac{|V_1|}{|V_0|}}\sqrt{\frac{|V_0^*|}{|V_1^*|}}\left(\frac{k_0^*}{k_1^*}\right)^{-\frac{n}{2}-a};$$

2. Calculate the posterior marginal probability $\mathsf{P}_\Pi(\mathrm{H}_0|y)$ as

$$\mathsf{P}_\Pi(\mathrm{H}_0|y) = \left(1 + \frac{1-\pi_0}{\pi_0}\sqrt{\frac{|V_0|}{|V_1|}}\sqrt{\frac{|V_1^*|}{|V_0^*|}}\left(\frac{k_1^*}{k_0^*}\right)^{-\frac{n}{2}-a}\right)^{-1}$$

where for $j=0,1$

$$k_j^* = k + \frac{1}{2}\mu_j^\top V_j^{-1}\mu_j - \frac{1}{2}\left(\mu_j^*\right)^\top\left(V_j^*\right)^{-1}\mu_j^* + \frac{1}{2}y^\top y$$

$$V_j^* = \left(V_j^{-1} + \Phi_j^\top\Phi_j\right)^{-1}; \qquad \mu_j^* = V_j^*\left(V_j^{-1}\mu_j + \Phi_j^\top y\right)$$

Hint: You may use the following identity:

$$(y-\Phi\beta)^\top(y-\Phi\beta)+(\beta-\mu)^\top V^{-1}(\beta-\mu)=(\beta-\mu^*)^\top\left(V^*\right)^{-1}(\beta-\mu^*)+S^*;$$
$$S^*=\mu^\top V^{-1}\mu-(\mu^*)^\top\left(V^*\right)^{-1}(\mu^*)+y^\top y;\qquad V^*=\left(V^{-1}+\Phi^\top\Phi\right)^{-1};\qquad \mu^*=V^*\left(V^{-1}\mu+\Phi^\top y\right)$$

The Exercise below has been set as Homework 4

Exercise 74. ($\star\star$)

1. Consider the Single vs. General alternative Bayesian hypothesis test

$$\mathrm{H}_0:\theta=\theta_0\qquad\text{vs}\qquad \mathrm{H}_1:\theta\neq\theta_0.$$

or more formally

$$\mathrm{H}_0:\begin{cases}y|\theta_0 & \sim F(y|\theta_0)\\ \theta_0 \text{ is fixed}\end{cases}\qquad\text{vs}\qquad \mathrm{H}_1:\begin{cases}y|\theta & \sim F(y|\theta)\\ \theta & \sim \Pi_1(\theta),\ \theta\in\Theta_1\end{cases}\tag{26}$$

Show that the Bayes factor B_{01} of H_0 and H_1 can be computed as

$$\mathrm{B}_{01}=\frac{\pi_1\left(\theta_0|y\right)}{\pi_1\left(\theta_0\right)}$$

2. In the above hypothesis test, assume that $y|\theta\sim\mathrm{Bin}(n,\theta)$ has Binomial sampling distribution with unknown parameter θ.
   (a) Find an approximation of B_{01}, when n is large.
   (b) For large n, show that $\mathrm{B}_{01}>1$ when $z_0=\frac{p-\theta_0}{\sqrt{u}}$, with $p=\frac{y}{n}$ and $u=\frac{p(1-p)}{n}$, satisfies $|z|<\max\left(\sqrt{k},0\right)$ for some k that depends of n, p, and $\pi_1\left(\theta_0\right)$.
   (c) Let the conditional prior $\Pi_1(\theta)$ be a Uniform distribution with positive mass above the interval $[0,1]$. Show that this choice of the conditional prior $\Pi_1(\theta)$ can create a "paradox" when compared with fixed size tests.

Part XII

Inference under model uncertainty

Exercise 75. ($\star\star$)Given a finite collection of models $\{\mathscr{M}_0, \mathscr{M}_1, \mathscr{M}_2, ..., \mathscr{M}_K\}$ the marginal model posterior probability of $\mathscr{M}_k$ can be calculated from Bayes factors as

$$\pi\left(\mathscr{M}_k|y\right) = \frac{\pi(\mathscr{M}_k)}{\pi(\mathscr{M}_0)} B_{k,0}(y) \Bigg/ \sum_{k'=0}^{K} \frac{\pi(\mathscr{M}_{k'})}{\pi(\mathscr{M}_0)} B_{k',0}(y) \ , \text{ for } k = 1, ..., K$$

Exercise 76. ($\star\star$)Let $B_{k,j}(y)$ be the Bayes factor of model $\mathscr{M}_k$ against model $\mathscr{M}_j$, for all $\forall k, i, j \in \mathcal{K}$. . Show that $B_{k,j}(y) = B_{k,i}(y)B_{i,j}(y)$, for all $\forall k, i, j \in \mathcal{K}$.

Exercise 77. ($\star\star$)Consider the Bayesian model

$$\begin{cases} y|\theta_k, \mathscr{M}_k \sim F(y|\theta_k, k) \\ \theta_k|\mathscr{M}_k \sim \Pi(\theta_k|k) \\ \mathscr{M}_k \sim \Pi(k) \end{cases} \tag{27}$$

Let z be a vector of future outcomes. Let the conditional predictive distribution for z given the observables y and for model $\mathscr{M}_k$ be $\mathrm{G}(z|y)$. Let the marginal predictive distribution for z given the observables y be $\mathrm{G}(z|y)$; This is the predictive distribution produced by BMA.

Show that

1. the predictive expectation of the future outcome produced by BMA is

$$\mathrm{E}_G\left(z|y\right) = \mathrm{E}_\Pi\left(\mathrm{E}_G\left(z|y,k\right)|y\right) = \sum_{k\in\mathcal{K}} \mathrm{E}_G\left(z|y,k\right)\pi(k|y), \forall k \in \mathcal{K}$$

2. the predictive variance of the future outcome produced by BMA is

$$\mathrm{Var}_G\left(z|y\right) = \sum_{k\in\mathcal{K}} \left(\mathrm{Var}_G\left(z|y,k\right) + \left(\mathrm{E}_G\left(z|y,k\right)\right)^2\right)\pi(k|y) - \left(\mathrm{E}_G\left(z\,|y\right)\right)^2, \forall k \in \mathcal{K}$$

Part XIII

Hierarchical Bayes

Relevant Vector Machine (Needs to be proofread)

Exercise 78. ($\star\star$)[Relevance Vector Machine]

Regarding the statistical model: Long story short (supplementary material)

Consider that we are interested in recovering the mapping

$$x \overset{\eta}{\longmapsto} \eta(x)$$

in the sense that $y \in \mathbb{R}$ is the response (output quantity) that depends on $x = (x_1, ..., x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$ which is the independent variable (input quantity) in a procedure; E.g.:,

- y: precipitation in log scale
- $x = (\text{longitude}, \text{latitude})$: geographical coordinates.

Consider a set of observed data $\{(y_i, x_i)\}_{i=1}^n$, which may be contaminated by additive noise of unknown variance; i.e.

$$y_i = \eta(x_i) + \epsilon_i,$$

where $\epsilon_i \overset{\text{IID}}{\sim} \text{N}\left(0, \sigma^2\right)$ and $\sigma^2 > 0$ is unknown. We wish to recover $\eta(x)$ by using the Tikhonov regularization on the functional space $\mathcal{H}$ such that

$$\eta = \arg\min_{\forall \tilde{\eta} \in \mathcal{H}} \left\{ \sum_{i=1}^{n} L\left(y_i - \tilde{\eta}\left(x_i\right)\right) + \lambda \left\|\tilde{\eta}\right\|_{\mathcal{H}}^2 \right\} \tag{28}$$

By assuming that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS), the solution to (29) is such that

$$\eta\left(x\right) = \sum_{j=1}^{n} k(x, x_j)\beta_j = k(x)^\top \beta$$

where $k(x) = \left(k\left(x, x_1\right), ..., k\left(x, x_n\right)\right)^\top$, $k\left(x, x_j\right)$ is the reproducing kernel (such as $k_\phi\left(x, x_j\right) = \exp\left(-\phi \left\|x - x_j\right\|^2\right)$ for some known parameter $\phi > 0$), and $\beta \in \mathbb{R}^{n+1}$ is an unknown vector.

Consider the following Bayesian model

$$\begin{cases} y|\beta, \sigma^2 & \sim \text{N}\left(K\beta, I\sigma^2\right) \\ \beta|\lambda & \sim \text{N}\left(0, D^{-1}\right), \quad D = (\lambda_0, \lambda_1, ..., \lambda_n) \\ \lambda_i & \overset{\text{iid}}{\sim} \text{d}\Pi\left(\lambda_i\right) \propto \lambda_i^{a-1} \exp\left(-b\lambda_i\right) \text{d}\lambda_i, \quad \forall i = 1, ..., n \\ \sigma^2 & \sim \text{d}\Pi\left(\sigma^2\right) \propto \left(\sigma^2\right)^{c-1} \exp\left(-\frac{1}{\sigma^2} d\right) \text{d}\sigma^2 \\ \beta, \sigma^2 & \text{a priori independent} \end{cases}$$

where $[K]_{i,j} = k(x_i, x_j)$ is a known matrix with size $n \times n$. The quantities $a > 0$, $b > 0$, $c > 0$, $d > 0$, and $\phi > 0$ are considered as fixed.

1. When $b = 0$, show that a necessary condition for a valid posterior inference is $a \in (-1/2, 0)$ for any choice of prior for τ (i.e. any choice of (c, d)).

2. Let $P = K\left(K^\top K\right)^{-1} K^\top$. Show that (2a) and (2b) are sufficient conditions for the Bayesian model to lead to a valid posterior inference

   (a) if $a > 0$ and $b > 0$, or

   (b) if $y^\top (I - P)\, y + 2d > 0$ and $c > -\frac{n}{2}$

3. Does the the improper Uniform prior on the joint $\log(\lambda_i)$ and $\log\left(\sigma^2\right)$, i.e. $\pi\left(\log(\lambda_i), \log\left(\sigma^2\right)\right) \propto 1$, lead to a valid inference?

4. Does the Jeffreys' prior $\pi(\lambda_i) \propto 1/\lambda_i$ lead to a valid inference?

Hint-1:

$$(y - K\beta)^\top (y - K\beta) + (\beta - \mu)^\top V^{-1} (\beta - \mu) = (\beta - \mu^*)^\top (V^*)^{-1} (\beta - \mu^*) + S^*;$$
$$S^* = \mu^\top V^{-1} \mu - (\mu^*)^\top (V^*)^{-1} (\mu^*) + y^\top y; \qquad V^* = \left(V^{-1} + K^\top K\right)^{-1}; \qquad \mu^* = V^* \left(V^{-1}\mu + K^\top y\right)$$

Hint-2: Sherman-Morrison-Woodbury formula:

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U \left(C^{-1} + VA^{-1}U\right)^{-1} VA^{-1}$$

Hint-3:

$$-\frac{y^\top y}{2\sigma^2} \leq -\frac{y^\top \left(I\sigma^2 + KD^{-1}K^\top\right)^{-1} y}{2} \leq -\frac{1}{2\sigma^2} y^\top (I - P)\, y$$

where $P = K\left(K^\top K\right)^{-1} K$.

Hint-4: It is given that $\int_{(0,\infty)} \frac{t^{-(a+1)}}{(\xi+t)^{1/2}} \mathrm{d}t < \infty$ if and only if $a \in (-1/2, 0)$.

Part XIV

Empirical Bayes

Exercise 79. ($\star\star$)Consider the following Bayesian model:

$$y_i|\theta_i \sim \text{Ex}(\theta_i);$$
$$\theta_i \sim \text{Ga}(a, b)$$

where $i = 1, ..., n$, $a > 2$, and $b > 2$.

1. Compute the posterior empirical Bayes estimate for θ_n under quadratic loss by learning estimates $\hat{a}$ and $\hat{b}$ of a and b from the method of moments.
2. Formulate the system of non-linear equations whose solution lead to the estimated of $\hat{a}$ and $\hat{b}$ of a and b from the ML-II method.
3. Compute the estimate $\hat{y}_{n+1}$ of y_{n+1} as the average of the marginal predictive $g(y_{n+1})$. Compute the EB posterior distribution of θ_{n+1}, as a function of the estimates $\hat{a}$ and $\hat{b}$ are known from part 2 and the estimate $\hat{y}_{n+1}$.

Part XV

Asymptotic posterior

Exercise 80. ($\star\star$)Consider the following Bayesian model:

$$x_{1:n} \sim \prod_{i=1}^{n} \mathrm{N}_2(x_i|\theta, I_2); \qquad \text{likelihood}$$

$$\pi(\theta) \propto 1, \qquad \text{prior}$$

where $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$.

1. Find the posterior distribution of $\theta|x_{1;n}$, by using Bayes theorem
2. Find the asymptotic posterior distribution of $\psi|x_{1:n}$, where $\psi = \theta_1/\theta_2$ as $n \to \infty$

Exercise 81. ($\star\star\star$)(Lemma **??**) If conditions (e1) and (e2), then

$$\lim_{n\to\infty} \pi(m_n|x_{1:n})|\Sigma_n|^{1/2} \leq (2\pi)^{-k/2}.$$

The equality holds when condition (e3) is satisfied.

Hint: Regularity conditions

e1 (Steepness) $\bar{\sigma}_n^2 \to 0$ as $n \to \infty$ where $\bar{\sigma}_n^2$ is the largest eigenvalue of Σ_n

e2 (Smoothness) For any $\epsilon > 0$ there exists N and $\delta > 0$ such that, for any $n > N$ and $\theta \in B_\delta(m_n)$, $\ddot{U}_n(\theta)$ exists and satisfies

$$I - A(\epsilon) \leq \ddot{U}_n(\theta)\,(\ddot{U}_n(m_n))^{-1} \leq I + A(\epsilon),$$

where I is the $k \times k$ identity matrix, and $A(\epsilon)$ is a $k \times k$ symmetric positive-semidefinete matrix whose largest eigenvalue tends to zero as $\epsilon \to 0$

e3 (Concentration) For any $\delta > 0$, as $n \to \infty$.

$$Q_n := \int_{B_\delta(m_n)} \pi_n(\theta)\mathrm{d}\theta \to 1. \tag{29}$$

Exercise 82. ($\star\star\star$)(Asymptotic normality under conjugacy) Assume that a collection of data $x_{1:n} = (x_1, ..., x_n)$ from a likelihood in canonical exponential family form

$$f(x_{1:n}|\psi) = a(x_{1:n}) \exp(x^\top \psi - b(\psi))$$

with canonical conjugate prior density

$$\pi(\psi; n_0, y_0) = \mathrm{Cef}(\psi|n_0, y_0),$$

where

$$\mathrm{Cef}(\psi; n, y) := c(n, y) \exp(ny^\top \psi - nb(\psi)).$$

The posterior density is

$$\pi(\psi|x_{1:n}) = \mathrm{Cef}(\psi|n_0 + n, ny_0 + n\bar{x}_n),$$

with $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$. Then

$$\Sigma_n^{-1/2}(\psi_n - b'(m_n)) \underset{D}{\longrightarrow} N(0, 1),$$

where

$$b'(m_n) = \nabla b(\psi)|_{\psi=m_n} = \frac{n_0 x_0 + n\bar{x}_n}{n_0 + n}$$

$$[b''(m_n)]_{i,j} = (\frac{\partial^2 b(\psi)}{\partial\psi_i \partial\psi_j})|_{\psi=m_n} = (n_0 + n)[\Sigma]^{-1}_{i,j}.$$
