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

Exercise Sheet Handout: Bayesian Statistics

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

Part I

Random variables

Exercise 1. ($\star$)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

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

Solution. It is $z = h(y) \Leftrightarrow y = h^{-1}(z)$

For if $h \nearrow$ it is

$$F_z(z) = \mathrm{P}(Z \leq z) = \mathrm{P}(h^{-1}(Z) \leq h^{-1}(z)) = \mathrm{P}(Y \leq h^{-1}(z)) = F_Y(h^{-1}(z))$$

For if $h \searrow$ it is

$$F_z(z) = \mathrm{P}(Z \leq z) = \mathrm{P}(h^{-1}(Z) \geq h^{-1}(z)) = \mathrm{P}(Y \geq h^{-1}(z)) = 1 - F_Y(h^{-1}(z))$$

Exercise 2. ($\star$)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

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

Solution. It is $z = h(y) \Leftrightarrow y = h^{-1}(z)$

For if $h \nearrow$ it is

$$F_z(z) = \mathrm{P}(Z \leq z) = \mathrm{P}(h^{-1}(Z) \leq h^{-1}(z)) = \mathrm{P}(Y \leq h^{-1}(z)) = F_Y(h^{-1}(z))$$

and

$$f_z(z) = \frac{\mathrm{d}}{\mathrm{d}z}F_z(z) = \frac{\mathrm{d}}{\mathrm{d}z}F_Y(h^{-1}(z)) = \frac{\mathrm{d}}{\mathrm{d}h^{-1}}F_Y(h^{-1})\det(\frac{\mathrm{d}}{\mathrm{d}z}h^{-1}(z))$$

For if $h \searrow$ it is

$$F_z(z) = \mathrm{P}(Z \leq z) = \mathrm{P}(h^{-1}(Z) \geq h^{-1}(z)) = \mathrm{P}(Y \geq h^{-1}(z)) = 1 - F_Y(h^{-1}(z))$$

and

$$f_z(z) = \frac{\mathrm{d}}{\mathrm{d}z} F_z(z) = \frac{\mathrm{d}}{\mathrm{d}z}\left[1 - F_Y(h^{-1}(z))\right] = -\frac{\mathrm{d}}{\mathrm{d}h^{-1}} F_Y(h^{-1}) \det(\frac{\mathrm{d}}{\mathrm{d}z} h^{-1}(z))$$

but $\det(\frac{\mathrm{d}}{\mathrm{d}z} h^{-1}(z)) < 0$ because $h \searrow$. So in both cases:

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

Exercise 3. ($\star$)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 $\mathcal{Y}$.

Solution.

1. It is

$$\mathrm{E}(z) = \mathrm{E}(c + Ay) = \int (c + Ay)\, \mathrm{d}F(y) = c + A \int y \mathrm{d}F(y) = c + A\mathrm{E}(y)$$

2. It is

$$\begin{aligned}\mathrm{E}(\psi_1(z) + \psi_2(y)) &= \int (\psi_1(z) + \psi_2(y))\, \mathrm{d}F((z,y)) = \int \psi_1(z) \mathrm{d}F((z,y)) + \int \psi_1(z) \mathrm{d}F((z,y)) \\ &= \int \psi_1(z) \mathrm{d}F(z) + \int \psi_1(z) \mathrm{d}F(z) = \mathrm{E}(\psi_1(z)) + \mathrm{E}(\psi_2(y))\end{aligned}$$

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

$$\mathrm{d}F(z,y) = \mathrm{d}F(z)\mathrm{d}F(y)$$

It is

$$\mathrm{E}(\psi_1(z)\psi_2(y)) = \int (\psi_1(z)\psi_2(y))\, \mathrm{d}F((z,y)) = \left(\int \psi_1(z) \mathrm{d}F(z)\right)\left(\int \psi_2(y) \mathrm{d}F(y)\right)$$

Exercise 4. ($\star$)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 $\mathrm{Cov}(z,y) = 0$

Solution.

1. It is

$$\begin{aligned}\text{Cov}(z,y) &= \text{E}\left((z-\text{E}(z))(y-\text{E}(y))^\top\right)\\ &\text{E}\left(zy^\top - z\text{E}(y)^\top - \text{E}(z)y^\top + \text{E}(z)\text{E}(y)^\top\right)\\ &= \text{E}(zy^\top) - \text{E}(z)\left(\text{E}(y)\right)^\top\end{aligned}$$

2. It is

$$\begin{aligned}\left(\text{Cov}(y,z)\right)^\top &= \left(\text{E}\left((z-\text{E}(z))(y-\text{E}(y))^\top\right)\right)^\top = \text{E}\left(\left((z-\text{E}(z))(y-\text{E}(y))^\top\right)\right)^\top\\ &= \text{E}\left((y-\text{E}(y))(z-\text{E}(z))^\top\right) = \text{Cov}(y,z)\end{aligned}$$

3. It is

$$\begin{aligned}\text{Cov}(c_1+A_1z, c_2+A_2y) &= \text{E}\left((c_1+A_1z)(c_2+A_2y)^\top\right) - \text{E}(c_1+A_1z)\left(\text{E}(c_2+A_2y)\right)^\top\\ &= ... = A_1\left(\text{E}(zy^\top) - \text{E}(z)\left(\text{E}(y)\right)^\top\right)A_2^\top = A_1\text{Cov}(z,y)A_2^\top\end{aligned}$$

4. Obviously since

$$\text{Cov}(z,y) = 0 \iff \text{Cov}(z_i,y_j) = \begin{cases} i=j \\ i\neq j \end{cases}$$

Exercise 5. ($\star$)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:

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

Solution.

It is

$$\begin{aligned}\left[\text{Cov}(z,y)\right]_{i,j} &= \left[\text{E}(zy^\top) - \text{E}(z)\left(\text{E}(y)\right)^\top\right]_{i,j} =\\ &= \left[\text{E}(zy^\top)\right]_{i,j} - \left[\text{E}(z)\left(\text{E}(y)\right)^\top\right]_{i,j}\\ &= \text{E}(z_iy_j^\top) - \text{E}(z_i)\left(\text{E}(y_j)\right)^\top = \text{Cov}(z_i,y_j)\end{aligned}$$

Exercise 6. ($\star$)Prove the following properties of $\text{Var}(Y)$ for a random vector $y\in\mathcal{Y}\subseteq\mathbb{R}^d$

1. $\text{Var}(y) = \text{E}(yy^\top) - \text{E}(y)\left(\text{E}(y)\right)^\top$
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)

Solution.

1. $\text{Var}(y) = \text{Cov}(y,y) = \text{E}(yy^\top) - \text{E}(y)\left(\text{E}(y)\right)^\top$
2. $\text{Var}(c+Ay) = \text{Cov}(c+Ay, c+Ay) = A\text{Cov}(y,y)A^\top = A\text{Var}(y)A^\top$

3. For any vector $x \in \mathbb{R}^q$

$$\begin{aligned} t^\top \text{Var}(y) t &= t^\top \text{E}\left((y - \text{E}(y))(y - \text{E}(y))^\top\right) t \\ &= \text{E}\left(\left(t^\top (y - \text{E}(y))\right)\left(t^\top (y - \text{E}(y))\right)^\top\right) \\ &= \text{E}\left(zz^\top\right) = \text{E}\left(\sum_{j=1}^{d} z_j^2\right) \geq 0 \end{aligned}$$

for $z = t^\top (y - \text{E}(y))$.

Exercise 7. ($\star$)Prove the following properties of characteristic functions

1. $\varphi_{A+Bx}(t) = e^{it^T A}\varphi_x(B^T t)$ 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 $M_x(t) = \text{E}(e^{t^T x})$ is the moment generating function, then $M_x(t) = \varphi_x(-it)$

Solution.

1. It is

$$\varphi_{A+Bx}(t) = \text{E}(e^{it^T(A+Bx)}) = \text{E}(e^{A+it^T Bx}) = \text{E}(e^{it^T A}e^{iB^T tx}) = e^{it^T A}\text{E}(e^{i(B^T t)x}) = e^{it^T A}\varphi_x(B^T t)$$

2. straightforward
3. straightforward

Exercise 8. ($\star$)Show that if $X \sim \text{Ex}(\lambda)$ then $\varphi_X(t) = \frac{\lambda}{\lambda - it}$.

Solution. It is

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itX} \underbrace{\lambda e^{-\lambda x} 1(X > 0)}_{= f_{\text{Ex}}(x|\lambda)} \text{d}x = \lambda \int_{-\infty}^{\infty} e^{-x(\lambda - itX)} \text{d}x = \frac{\lambda}{\lambda - it}$$

Exercise 9. ($\star$)

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

Solution.

1. It is

$$\varphi_X(t) = \sum_{x=0,1} e^{itX} P(X = x) = e^{it0}(1 - p) + e^{it1}p = (1 - p) + pe^{it}$$

2. Because Binomial r.v. results as a summation of n IID Bernoulli r.v., it is $Y = \sum_{i=1}^{n} X_i$, where $X_i \sim \text{Br}(p)$ $i = 1, ..., n$ and IID. Then

$$\varphi_Y(t) = \varphi_{\sum X_i}(t) = \prod_{i=1}^{n} \varphi_{X_i}(t) = ((1 - p) + pe^{it})^n$$

Exercise 10. ($\star\star\star$)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

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

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

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

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

Solution. We consider separately two cases.

x **is discrete:**

Let $\Theta_0 \subseteq \Theta$ be any sub-set of Θ; I need to show that

$$P(\theta \in \Theta_0|x) = \frac{\int_{\Theta_0} f(x|\theta)\mathrm{d}F(\theta)}{\int_\Theta f(x|\theta)\mathrm{d}F(\theta)} = \begin{cases} \int_{\Theta_0} \frac{f(x|\theta)f(\theta)}{\int_\Theta f(x|\theta)\mathrm{d}F(\theta)}\mathrm{d}\theta & , \theta \text{ cont.} \\ \sum_{\theta\in\Theta_0} \frac{f(x|\theta)f(\theta)}{\int_\Theta f(x|\theta)\mathrm{d}F(\theta)} & , \theta \text{ discr.} \end{cases}$$

By Bayes theorem it is

$$P(\theta \in \Theta_0|x) = \frac{P(\Theta_0, x)}{P(x)}$$

where $P(x) = \int_\Theta f(x|\theta)\mathrm{d}F(\theta)$ and $P(\Theta_0, x) = \int_{\Theta_0} f(x|\theta)\mathrm{d}F(\theta)$.

x **is continuous:**

Let $\Theta_0 \subseteq \Theta$ be any sub-set of Θ; because the probability $P(x) = 0$, I need to show that

$$\lim_{r\to 0} P(\theta \in \Theta_0|B_r(x)) = \frac{\int_{\Theta_0} f(x|\theta)\mathrm{d}F(\theta)}{\int_\Theta f(x|\theta)\mathrm{d}F(\theta)} = \begin{cases} \int_{\Theta_0} \frac{f(x|\theta)f(\theta)}{\int_\Theta f(x|\theta)\mathrm{d}F(\theta)}\mathrm{d}\theta & , \theta \text{ cont.} \\ \sum_{\theta\in\Theta_0} \frac{f(x|\theta)f(\theta)}{\int_\Theta f(x|\theta)\mathrm{d}F(\theta)} & , \theta \text{ discr.} \end{cases}$$

for an open ball $B_r(x) = \{x' \in \mathcal{X} : |x' - x| < r\}$. By Bayes theorem

$$P(\theta \in \Theta_0|B_r(x)) = \frac{P(\Theta_0, B_r(x))}{P(B_r(x))}$$

where

$$P(\Theta_0, B_r(x)) = \int_{\Theta_0}\left[\int_{B_r(x)} f(\zeta|\theta)\mathrm{d}\zeta\right]\mathrm{d}F(\theta)$$

$$P(B_r(x)) = \int_{\Theta}\left[\int_{B_r(x)} f(\zeta|\theta)\mathrm{d}\zeta\right]\mathrm{d}F(\theta)$$

By mean value theorem[1] there exists $\zeta' \in B_r(y)$ such as

$$\int_{B_r(x)} f(\zeta|\theta)\mathrm{d}\zeta = f(\zeta'|\theta)\int_{B_r(x)} \mathrm{d}\zeta = f(\zeta'|\theta)\ \|B_r(x)\|$$

Then

$$P(\theta \in \Theta_0|B_r(x))\ = \frac{\int_{\Theta_0}\left[f(\zeta'|\theta)\ \|B_r(x)\|\right]\mathrm{d}F(\theta)}{\int_{\Theta}\left[f(\zeta'|\theta)\ \|B_r(x)\|\right]\mathrm{d}F(\theta)} \quad \xrightarrow{r\to 0} \quad \frac{\int_{\Theta_0} f(\zeta|\theta)\mathrm{d}F(\theta)}{\int_{\Theta} f(\zeta|\theta)\mathrm{d}F(\theta)}$$

Exercise 11. ($\star$)Prove that:

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

Solution.

1. It is

$$\varphi_Z(t) = \mathrm{E}(\exp(it^T Z)) = \mathrm{E}(\exp(i\sum_{j=1}^{d}(t_j Z_j))) = \mathrm{E}(\prod_{j=1}^{d}\exp(it_j Z_j)) = \prod_{j=1}^{d}\mathrm{E}(\exp(it_j Z_j))$$

$$= \prod_{j=1}^{d}\varphi_{Z_j}(t) = \prod_{j=1}^{d}\exp(-\frac{1}{2}t_j^2) = \exp(-\frac{1}{2}\sum_{j=1}^{d}t_j^2) = \exp(-\frac{1}{2}t^T t)$$

2. Assume a matrix L such as $\Sigma = LL^T$. It is $X = \mu + LZ$. Then

$$\varphi_X(t) = \varphi_{\mu+LZ}(t) = e^{it^T\mu}\varphi_Z(L^T t) = e^{it^T\mu}\exp(-\frac{1}{2}(L^T t)^T L^T t)$$

$$= e^{it^T\mu}\exp(-\frac{1}{2}t^T LL^T t) = \exp(it^T\mu - \frac{1}{2}t^T\Sigma t)$$

Exercise 12. ($\star$)Show the following properties of the Characteristic Function

1. $\varphi_x(0) = 1$ and $|\varphi_x(t)| \le 1$ for all $t \in \mathbb{R}^d$
2. $\varphi_{A+Bx}(t) = e^{it^T A}\varphi_x(B^T t)$ 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 $M_x(t) = \mathrm{E}(e^{t^T x})$ is the moment generating function, then $M_x(t) = \varphi_x(-it)$

Solution.

1. It is $\varphi_x(0) = \mathrm{E}(e^{i0^T x}) = \mathrm{E}(1) = 1$. Also

$$|\varphi_x(t)| = \left|\mathrm{E}(e^{it^T x})\right| = \left|\int\left(\cos(t^T x) + i\sin(t^T x)\right)\mathrm{d}F(x)\right| \le \int\left|\cos(t^T x) + i\sin(t^T x)\right|\mathrm{d}F(x) \le \int 1\mathrm{d}F(x) = 1$$

2. It is

$$\varphi_{A+Bx}(t) = \mathrm{E}(e^{it^T(A+Bx)}) = \mathrm{E}(e^{it^T A + Bit^T x}) = \mathrm{E}(e^{Ai}e^{i(B^T t)^\top x}) = e^{it^T A}\varphi_x(B^T t)$$

[1] $\int_A f(x)g(x)\mathrm{d}x = f(\xi)\int_A g(x)\mathrm{d}x$ where $\xi \in A$ if A is connected, and $g(x) \ge 0$ for $x \in A$.

3. It is

$$\varphi_{x+y}(t) = \mathrm{E}(e^{it^T(x+y)}) = \mathrm{E}(e^{it^T x} e^{it^T y}) = \mathrm{E}(e^{it^T x})\mathrm{E}(e^{it^T y}) = \varphi_x(t)\varphi_y(t)$$

Part II

Probability calculus

Exercise 13. ($\star$)Let a random variable $x \sim \text{IG}(a, b)$, a fixed value $c > 0$, and $y = cx$ then $y \sim \text{IG}(a, cb)$.

Solution. It is $y = cx$ and $x = \frac{1}{c}y$

$$f(y) = f_{\text{IG}(a,b)}(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \quad \propto (\frac{1}{c}y)^{-a-1}\exp(-\frac{b}{\frac{1}{c}y})1_{(0,+\infty)}(\frac{1}{c}y)\frac{1}{c}$$

$$\propto y^{-a-1}\exp(-\frac{cb}{y})1_{(0,+\infty)}(y) = f_{\text{IG}(a,cb)}(y)$$

Exercise 14. ($\star\star\star$)

What is in this box is just for your information, you are not required to learn, and can be skipped.

Definition. [a] Consider a random variable x distributed according to $F(x|z)$ with a PDF/PMF $f(x|z)$ labeled by an unknown parameter z. Consider that z is again distributed according to some other distribution $\Pi(z)$ with PDF/PMF $\pi(z)$. Then:

1. This dependency is denoted as
$$x|z \sim F(x|z)$$
$$z \sim \Pi(z)$$
2. Distribution $F(x|z)$ is called conditional or parametrized distribution
3. Distribution $\Pi(z)$ is called mixing or latent distribution; z is called mixing or latent variable.
4. Distribution $G(x)$ that results by marginalizing (integrating) the conditional distribution $F(x|z)$ with respect to $\Pi(z)$ as
$$G(x) = \int F(x|z)\mathrm{d}\Pi(z)$$
is called the compound (mixture) distribution of x. The compound distribution $G(x)$ has PDF/PMF
$$g(x) = \int f(x|z)\mathrm{d}\Pi(z) = \begin{cases} \int g(x|z)\pi(z)\mathrm{d}z & , z \text{ cont.} \\ \\ \sum_{\forall z} g(x|z)\pi(z) & , z \text{ discr.} \end{cases}$$

$G(x)$ is also called continuous mixture distribution when the mixing/latent variable z is continuous. $G(x)$ is also called finite mixture when z is discrete.

Proposition. *From the Lecture notes, we get*

$$E_G(x) = E_\Pi\left(E_F(x|z)\right)$$
$$Var_G(x) = E_\Pi\left(Var(x|z)\right) + Var_\Pi\left(E_F(x|z)\right)$$

[a] From WIKIPEDIA

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

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

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

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

1. Show that the compound distribution of x is F $x \sim \mathrm{F}(n, m)$, where $\mathrm{F}(n, m)$ is F distribution with numerator and denumerator degrees of freedom n and m, and PDF
$$f_{\mathrm{F}(n,m)}(x) = \frac{1}{x\,\mathrm{B}\left(\frac{n}{2}, \frac{m}{2}\right)} \sqrt{\frac{(n\,x)^n\ m^m}{(n\,x+m)^{n+m}}} 1(x > 0)$$
2. Show that
$$\mathrm{E}_{\mathrm{F}(n,m)}(x) = \frac{m}{m-2}$$

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 $\text{Var}_{\xi\sim\text{IG}(a,b)}(\xi) = \frac{b^2}{(a-1)^2(a-2)}$

Solution.

1. It is

$$f_{\text{Ga}(\frac{n}{2},\frac{nz}{2})}(x|z) = \frac{(\frac{nz}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{nz}{2}x} 1(x>0) \; ; \qquad f_{\text{Ga}(\frac{m}{2},\frac{m}{2})}(z) = \frac{(\frac{m}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})} z^{\frac{m}{2}-1} e^{-\frac{m}{2}z} 1(z>0)$$

So:

$$\begin{aligned}
f(x) &= \int f_{\text{Ga}(\frac{n}{2},\frac{nz}{2})}(x|z) f_{\text{Ga}(\frac{m}{2},\frac{m}{2})}(z) \text{d}z \\
&= \int \overbrace{\frac{(\frac{nz}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{nz}{2}x} 1(x>0)}^{=f_{\text{Ga}(\frac{n}{2},\frac{nz}{2})}(x|z)} \overbrace{\frac{(\frac{m}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})} z^{\frac{m}{2}-1} e^{-\frac{m}{2}z} 1(z>0)}^{=f_{\text{Ga}(\frac{m}{2},\frac{m}{2})}(z)} \text{d}z \\
&= \frac{(\frac{n}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \frac{(\frac{m}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})} 1(x>0) x^{\frac{n}{2}-1} \int_0^\infty z^{\frac{n}{2}} e^{-\frac{nx}{2}z} z^{\frac{m}{2}-1} e^{-\frac{m}{2}z} \text{d}z \\
&= \frac{(\frac{n}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \frac{(\frac{m}{2})^{\frac{m}{2}}}{\Gamma(\frac{m}{2})} 1(x>0) x^{\frac{n}{2}-1} \int_0^\infty z^{\frac{n}{2}+\frac{m}{2}-1} e^{-(\frac{m}{2}+\frac{nx}{2})z} \text{d}z \\
&= \frac{(\frac{n}{2})^{\frac{n}{2}} (\frac{m}{2})^{\frac{m}{2}}}{\text{B}(\frac{n}{2},\frac{m}{2})} 1(x>0) x^{\frac{n}{2}-1} \left(\frac{m}{2}+\frac{nx}{2}\right)^{-(\frac{n}{2}+\frac{m}{2})} \\
&= \frac{(n)^{\frac{n}{2}}(m)^{\frac{m}{2}}}{\text{B}(\frac{n}{2},\frac{m}{2})} \frac{1}{x} \sqrt{\frac{x^n}{(m+nx)^{n+m}}} 1(x>0) \\
&= \frac{1}{x\,\text{B}(\frac{n}{2},\frac{m}{2})} \sqrt{\frac{(n\,x)^n\; m^m}{(n\,x+m)^{n+m}}} 1(x>0)
\end{aligned}$$

2. It is

$$\begin{aligned}
\text{E}(x) \quad &= \text{E}_{\text{Ga}(\frac{m}{2},\frac{m}{2})}\left(\text{E}_{\text{Ga}(\frac{n}{2},\frac{nz}{2})}(x|z)\right) \quad = \text{E}_{z\sim\text{Ga}(\frac{m}{2},\frac{m}{2})}\left(\frac{1}{z}\right) \\
&= \text{E}_{\xi\sim\text{IG}(\frac{m}{2},\frac{m}{2})}(\xi) \quad = \frac{\frac{m}{2}}{\frac{m}{2}-1} = \frac{m}{m-2}
\end{aligned}$$

3. It is

$$\begin{aligned}
\text{Var}(x) =& \text{E}_{\text{Ga}(\frac{m}{2},\frac{m}{2})}\left(\text{Var}_{\text{Ga}(\frac{n}{2},\frac{nz}{2})}(x|z)\right) + \text{Var}_{\text{Ga}(\frac{m}{2},\frac{m}{2})}\left(\text{E}_{\text{Ga}(\frac{n}{2},\frac{nz}{2})}(x|z)\right) \\
=& \text{E}_{\text{Ga}(\frac{m}{2},\frac{m}{2})}\left(\frac{2}{nz^2}\right) + \text{Var}_{\text{Ga}(\frac{m}{2},\frac{m}{2})}\left(\frac{1}{z}\right) && = \frac{2}{n}\text{E}_{\text{Ga}(\frac{m}{2},\frac{m}{2})}\left(\frac{1}{z^2}\right) + \text{Var}_{\text{Ga}(\frac{m}{2},\frac{m}{2})}\left(\frac{1}{z}\right) \\
=& \frac{2}{n}\text{E}_{\xi\sim\text{IG}(\frac{m}{2},\frac{m}{2})}\left(\xi^2\right) + \text{Var}_{\xi\sim\text{IG}(\frac{m}{2},\frac{m}{2})}(\xi) \\
=& \frac{2}{n}\left(\frac{(\frac{m}{2})^2}{(\frac{m}{2}-1)(\frac{m}{2}-2)}\right) + \left(\frac{\frac{m}{2}}{\frac{m}{2}-1}\right) && = \ldots = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}
\end{aligned}$$

Exercise 15. ($\star\star$)Prove the following statement:
Let $x \sim \mathrm{N}_d(\mu, \Sigma)$, $x \in \mathbb{R}^d$, and $y = (x-\mu)^\top \Sigma^{-1}(x-\mu)$. Then

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

Solution. It is

$$y = (x-\mu)^\top \Sigma^{-1}(x-\mu) = \left(\Sigma^{-1/2}(x-\mu)\right)^\top \left(\Sigma^{-1/2}(x-\mu)\right) = z^\top z = \sum_{i=1}^{d} z_i^2$$

where $z = \Sigma^{-1/2}(x-\mu)$, and $z \sim \mathrm{N}_d(0, I)$. Because $z_i \sim \mathrm{N}(0,1)$, it is $\sum_{i=1}^d z_i^2 \sim \chi^2_d$ (from stats concepts 2).

Exercise 16. ($\star\star$)Let

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

with PDF

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

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

Show that the marginal PDF of x is

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

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

Solution. It is

$$\int f_{\mathrm{N}_d(\mu,\Sigma\xi)}(x|\xi) f_{\mathrm{IG}(a,b)}(\xi)\mathrm{d}\xi =$$

$$= \int \underbrace{\left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sqrt{\det(\Sigma\xi)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \frac{\Sigma^{-1}}{\xi}(x-\mu)\right)}_{=\mathrm{N}_d(x|\mu,\Sigma\xi)} \underbrace{\frac{b^a}{\Gamma(a)} \xi^{-a-1} \exp\left(-\frac{b}{\xi}\right) 1_{(0,\infty)}(\xi)}_{=\mathrm{IG}(\xi|a,b)} \mathrm{d}\xi$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sqrt{\det(\Sigma)}} \frac{b^a}{\Gamma(a)} \int \xi^{-a-1-\frac{d}{2}} \exp\left(-\frac{1}{\xi}\left[\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + b\right]\right) \mathrm{d}\xi$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sqrt{\det(\Sigma)}} \frac{b^a}{\Gamma(a)} \Gamma\left(a+\frac{d}{2}\right) \left[\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + b\right]^{-\left(a+\frac{d}{2}\right)}$$

$$= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sqrt{\det(\Sigma)}} \frac{b^{-\frac{d}{2}}}{\Gamma(a)} \Gamma\left(a+\frac{d}{2}\right) \left[\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\frac{1}{b} + 1\right]^{-\frac{(2a+d)}{2}}$$

$$= \frac{2a^{-\frac{d}{2}}}{\pi^{\frac{n}{2}} \sqrt{\det(\frac{b}{a}\Sigma)}} \frac{\Gamma\left(a+\frac{d}{2}\right)}{\Gamma(a)} \left[1 + \frac{1}{2a}(x-\mu)^\top \left(\frac{b}{a}\Sigma\right)^{-1}(x-\mu)\right]^{-\frac{(2a+d)}{2}}$$

The Following one will be given as Homework

Exercise 17. (⋆⋆⋆)

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

Solution.

1. From what is given, it is $x|\xi \sim \mathrm{N}_d(\mu, \Sigma\xi v)$ and $\xi \sim \mathrm{IG}(\frac{v}{2}, \frac{1}{2})$ namely,

$$f_x(x) = \int f_{x_1,x_2|\xi}(x_1, x_2|\xi) f_\xi(\xi)\mathrm{d}\xi = \int f_{x_2|\xi,x_1}(x_2|\xi, x_1) f_{x_1|\xi}(x_1|\xi) f_\xi(\xi)\mathrm{d}\xi$$

It is

$$\begin{aligned}f_{x_1}(x_1) &= \int\int f_{x_1,x_2|\xi}(x_1, x_2|\xi) f_\xi(\xi)\mathrm{d}\xi\mathrm{d}x_2 = \int\int f_{x_2|\xi,x_1}(x_2|\xi, x_1) f_{x_1|\xi}(x_1|\xi) f_\xi(\xi)\mathrm{d}\xi\mathrm{d}x_2\\ &= \int \left(\int f_{x_2|\xi,x_1}(x_2|\xi, x_1)\mathrm{d}x_2\right) f_{x_1|\xi}(x_1|\xi) f_\xi(\xi)\mathrm{d}\xi = \int f_{x_1|\xi}(x_1|\xi) f_\xi(\xi)\mathrm{d}\xi\end{aligned}$$

Because $x_1|\xi \sim \mathrm{N}_{d_1}(\mu_1, \Sigma_1\xi v)$, and $\xi \sim \mathrm{IG}(\frac{v}{2}, \frac{1}{2})$, it is $x_1 \sim \mathrm{T}_{d_1}(\mu_1, \Sigma_1, \nu)$ from the statement of the question.

2. From what is given, it is $x|\xi \sim \mathrm{N}_d(\mu, \Sigma\xi v)$, and hence $x_1|\xi \sim \mathrm{N}_d(\mu_1, \Sigma_1\xi v)$ as marginal of a Normal distribution. From the Bayes Theorem, it is

$$\begin{aligned}f_{\xi|x_1}(\xi|x_1) &\propto f_{x_1|\xi}(x_1|\xi) f(\xi)\\ &\propto \xi^{-\frac{d_1}{2}} \exp\left(-\frac{1}{2}(x_1 - \mu_1)^\top (\Sigma_1\xi v)^{-1}(x_1 - \mu_1)\right) \times \xi^{-\frac{d_1+v}{2}-1} \exp\left(-\frac{1}{\xi}\frac{1}{2}\right)\\ &\propto \xi^{-\frac{d_1+v}{2}-1} \exp\left(-\frac{1}{\xi}\frac{1}{2}\left[(x_1 - \mu_1)^\top \Sigma_1^{-1}(x_1 - \mu_1)\frac{1}{v} + 1\right]\right)\\ &\propto \xi^{-\frac{d_1+v}{2}-1} \exp\left(-\frac{1}{\xi}\frac{1}{2}\frac{Q+v}{v}\right)\end{aligned}$$

This is the kernel of the Inverse Gamma distribution, and hence I can recognize that

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

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

$$\begin{aligned}f(\xi'|x_1) = f_{\mathrm{IG}(\frac{1}{2}(d_1+v), \frac{1}{2}\frac{Q+v}{v})}(\xi|x_1)\left|\frac{\mathrm{d}\xi}{\mathrm{d}\xi'}\right| &\propto (Q\xi')^{-\frac{d_1+v}{2}-1} \exp(-\frac{1}{2}\frac{Q+v}{v}\frac{1}{\frac{Q+v}{v}\xi'})1_{(0,+\infty)}(\frac{Q+v}{v}\xi')\frac{Q+v}{v}\\ &\propto (\xi')^{-\frac{d_1+v}{2}-1} \exp(-\frac{1}{2}\frac{1}{\xi'})1_{(0,+\infty)}(\xi') = f_{\mathrm{IG}(\frac{v+d_1}{2}, \frac{1}{2})}(\xi')\end{aligned}$$

So

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

4. I will try to show that

$$x_2|\xi', x_1 \sim \mathrm{N}_{d_2}\left(\mu_{2|1}, (v+d_1)\dot{\Sigma}_{2|1}\xi'\right)$$
$$\xi'|x_1 \sim \mathrm{IG}(\frac{v+d_1}{2}, \frac{1}{2})$$

which leads to

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

since because

$$f_{x_2|x_1}(x_2|x_1) = \int f_{x_2|\xi,x_1}(x_2|\xi, x_1) f_{\xi}(\xi|x_1)\mathrm{d}\xi$$

- I have calculated that

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

where $\xi' = \xi\frac{v}{Q+v}$ with $Q = (\mu_1 - x_1)^\top \Sigma_1^{-1}(\mu_1 - x_1)$.

- It is (from multivariate Normal properties of the Example in the Hint)

$$x_2|\xi, x_1 \sim \mathrm{N}_{d_2}\left(\mu_{2|1}, (\underbrace{\Sigma_{22} - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top}_{=\Sigma_{2|1}})\xi v\right) \equiv \mathrm{N}_{d_2}\left(\mu_{2|1}, \Sigma_{2|1} v\xi\right)$$

where $\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$. If I rearrange the parameters in order to appear $\xi' = \xi\frac{v}{Q+v}$ in the covariance I get

$$x_2|\xi, x_1 \sim \mathrm{N}_{d_2}\left(\mu_{2|1}, \Sigma_{2|1} v\xi' \frac{v+Q}{v}\frac{v+d_1}{v+d_1}\right)$$

By setting

$$\dot{\Sigma}_{2|1} = \Sigma_{2|1}\frac{v+Q}{v+d_1}$$

I get

$$x_2|\xi', x_1 \sim \mathrm{N}_{d_2}\left(\mu_{2|1}, (v+d_1)\dot{\Sigma}_{2|1}\xi'\right)$$

So I have

$$x_2|\xi', x_1 \sim \mathrm{N}_{d_2}\left(\mu_{2|1}, (v+d_1)\dot{\Sigma}_{2|1}\xi'\right)$$
$$\xi'|x_1 \sim \mathrm{IG}(\frac{v+d_1}{2}, \frac{1}{2})$$

which gives that $x_2|x_1 \sim \mathrm{T}_{d_2}(\mu_{2|1}, \dot{\Sigma}_{2|1}, \nu_{2|1})$ with $\nu_{2|1} = v + d_1$. So the distribution of $x_2|x_1$ is $x_2|x_1 \sim \mathrm{T}_{d_2}(\mu_{2|1}, \dot{\Sigma}_{2|1}, \nu_{2|1})$.

Note: Alternatively, one could prove sub-questions (2) and (4) by performing several pages of Matrix calculations to show that

$$\begin{aligned} f_X(x|\mu,\Sigma) =& \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}\det(\Sigma)^{\frac{1}{2}}}\left(1+\frac{1}{v}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)^{-\frac{\nu+d}{2}} \\ =& ... \\ =& \frac{\Gamma(\frac{\nu+d_1}{2})}{\Gamma(\frac{\nu}{2})\nu^{\frac{d_1}{2}}\pi^{\frac{d_1}{2}}\det(\Sigma_1)^{\frac{1}{2}}}\left(1+\frac{1}{v}(x_1-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(x_1-\mu_1)\right)^{-\frac{\nu+d_1}{2}} \\ & \times \frac{\Gamma(\frac{\nu_{2|1}+d_2}{2})}{\Gamma(\frac{\nu_{2|1}}{2})\nu_{2|1}^{\frac{d_2}{2}}\pi^{\frac{d_2}{2}}\det(\dot{\Sigma}_{2|1})^{\frac{1}{2}}}\left(1+\frac{1}{v_{2|1}}(x_2-\mu_{2|1})^{\mathrm{T}}\dot{\Sigma}_{2|1}^{-1}(x_2-\mu_{2|1})\right)^{-\frac{\nu_{2|1}+d_2}{2}} \end{aligned}$$

see Raiffa, H., & Schlaifer, R. (1961; Section 8.3). Applied statistical decision theory. This requires a lot of vector and matrix calculus.

Exercise 18. ($\star\star\star$)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
$$z \sim \mathrm{T}_d(c + \sum_{i=1}^n \mu_i, \sum_{i=1}^n B_i \Sigma_i B_i^\top, v)$$

Solution.

1. For any $a \in \mathbb{R}^d$
$$a^\top y = a^\top \left(c + \sum_{i=1}^n B_i x_i\right) = a^\top c + \sum_{i=1}^n a^\top B_i x_i = a^\top c + \sum_{i=1}^n \left(B_i^\top a\right)^\top x_i$$
follows a univariate Normal distribution. So y follows a d-dimensional Normal by definition. Also
$$\mathrm{E}(y) = \mathrm{E}(c + \sum_{i=1}^n B_i x_i) = c + \sum_{i=1}^n \mu_i$$
and
$$\mathrm{Var}(y) = \mathrm{Var}(c + \sum_{i=1}^n B_i x_i) = \sum_{i=1}^n B_i \mathrm{Var}(x_i) B_i^\top = \sum_{i=1}^n B_i \Sigma_i B_i^\top$$
So by definition $y \sim \mathrm{N}_d(c + \sum_{i=1}^n \mu_i, \sum_{i=1}^n B_i \Sigma_i B_i^\top)$.
2. It is
$$z = c + \sum_{i=1}^n B_i x_i = c + \sum_{i=1}^n B_i \left(\mu_i + y_i \sqrt{v\xi}\right) = \left(c + \sum_{i=1}^n B_i \mu_i\right) + \left(\sum_{i=1}^n B_i y_i\right) \sqrt{v\xi}$$

for $y_i \sim \mathrm{N}_d(0, \Sigma_i)$ and $\xi \sim \mathrm{IG}(\frac{v}{2}, \frac{1}{2})$, and hence

$$z = \left(c + \sum_{i=1}^{n} B_i \mu_i \right) + \tilde{y}\sqrt{v\xi}$$

where $\tilde{y} \sim \mathrm{N}_d(0, \sum_{i=1}^{n} B_i \Sigma_i B_i^\top)$. Hence, $z \sim \mathrm{T}_d(c + \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} B_i \Sigma_i B_i^\top, v)$ by definition.

Part III

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 19. ($\star\star\star\star\star$)(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

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

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 $F_{i_j} \to F$.

Solution. Let the sum of random quantities be $S_n = \sum_{i=1}^{n} y_i$, and assume that the sum S_n is equal to value s_n; i.e. $S_n = t_n$. By exchangeability, for $0 \leq t_n < n$, it is

$$p(S_n = t_n) = \binom{n}{t_n} p(y_{\mathfrak{p}(1)}, ..., y_{\mathfrak{p}(n)})$$

for any permutation operator $\mathfrak{p}$. For finite N, let $N \geq n \geq t_n \geq 0$,

$$\begin{aligned}
p(S_n = t_n) &= \sum_{t_N=0}^{N} p(S_n = t_n|S_N = t_N)p(S_N = t_N) \\
&= \underbrace{\sum_{t_N=0}^{t_n-1} p(S_n = t_n|S_N = t_N)p(S_N = t_N)}_{=0} \qquad (8) \\
&\qquad + \sum_{y_N=y_n}^{N-(n-y_n)} p(S_n = t_n|S_N = t_N)p(S_N = t_N) \\
&\qquad\quad + \underbrace{\sum_{t_N=N-(n-t_n)+1}^{N} p(S_n = t_n|S_N = t_N)p(S_N = t_N)}_{=0} \qquad (9) \\
&= \sum_{y_N=y_n}^{N-(n-y_n)} p(S_n = t_n|S_N = t_N)p(S_N = t_N)
\end{aligned}$$

The terms in (8, 9) are zero because $p(S_n = t_n|S_N = t_N) = 0$ for $t_N < t_n$ and $t_N > N - (n - t_n)$ because we contrition on $S_N = t_N$.

We work out on $p(S_n = t_n|S_N = t_N)$ which is the conditional probability for S_n given $S_N = t_N$. We observe that the random variable $S_n|S_N = t_N$ follows a Hypergeometric distribution $S_n|S_N = t_N \sim \text{Hy}(t_N, N - t_N, n)$. This is because it describes a Hypergeometric experiment[2]. i.e., $S_n = t_n$ is the number of successes (random draws for which the object drawn has a specified feature) in n random draws without replacement, from a finite population of size N that contains exactly $S_N = t_N$ objects of that feature, wherein each draw is either a success or a failure (aka $x_i = 0$ or 1). Hence, $p(S_n = t_n|S_N = t_N)$ is a Hypergeometric PMF, namely

$$p(S_n = t_n|S_N = t_N) = \text{Hy}(S_n = t_n|t_N, N - t_N, n) = \frac{\binom{t_N}{t_n}\binom{N-t_N}{n-t_n}}{\binom{N}{n}}, \ 0 \leq t_n \leq n$$

[2] https://en.wikipedia.org/wiki/Hypergeometric_distribution

Rewriting the binomial coefficients by rearranging the terms in the product, we get

$$p(S_n = t_n) = \sum \binom{N}{n}^{-1} \binom{t_N}{t_n} \binom{N - t_N}{n - t_n} p(S_N = t_N)$$
$$= \binom{n}{t_n} \sum \frac{(t_N)_{t_n} (N - t_N)_{n-t_n}}{(N)_n} p(S_N = t_N)$$

where $(y)_r = y(y-1)...(y-r+1)$.

Now, define a function $\Pi_N(\theta)$ on $\mathbb{R}$ as the step function which is zero for $\theta < 0$, and has steps of size $p(S_N = t_N)$ at $\theta = t_N/N$ for $t_N = 0, 1, 2, ..., N$. Then, by changing variable we get,

$$p(S_n = t_n) = \binom{n}{t_n} \int_0^1 \frac{(\theta N)((1-\theta)N)_{n-t_n}}{(N)_n} \mathrm{d}\Pi_N(\theta).$$

This result holds for any finite N. Now we need to consider $N \to \infty$. In the limit, we get

$$\lim_{N\to\infty} \frac{(\theta N)((1-\theta)N)_{n-t_n}}{(N)_n} = \theta^{t_n}(1-\theta)^{n-t_n} = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \tag{10}$$

Note that function $\Pi_N(t)$ is a step function, starting at zero and ending at one with N steps of varying sizes at particular values of t. By Helly's theorem, there exists a subsequence $\{\Pi_{N_1}, \Pi_{N_2}, ...\}$ such that

$$\lim_{N_j\to\infty} \Pi_{N_j} = \Pi$$

where Π is a distribution function.

Exercise 20. ($\star\star$)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.

Solution. Sampling without replacement is clearly not iid. However, it is exchangeable. Assume that we sample n items from N distinct objects without replacement, we have that:

$$f(y_1, ..., y_n) = \frac{1}{N^{\underline{n}}} = \frac{(N-n)!}{N!} \tag{11}$$

Clearly, the probability mass function does not depend on the ordering of the sequence. Therefore the sequence is exchangeable.

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

$$\mathrm{E}(g(Y_i)) = \mathrm{E}(g(Y_1)) \text{ for all } i \in \{2, \ldots, n\} \tag{12}$$

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

Solution. For ease of exposition, we show that $\mathrm{E}(g(Y_1)) = \mathrm{E}(g(Y_2))$. The general case follows similarly.

$$\mathrm{E}(g(Y_1)) = \sum_{(y_1,y_2,y_3,\ldots,y_n)\in\mathcal{Y}^n} g(y_1) f(y_1, y_2, y_3, \ldots, y_n) \tag{13}$$

and by exchangeability, we can swap the indices 1 and 2 in the probability mass function, so

$$= \sum_{(y_1,y_2,y_3,\ldots,y_n)\in\mathcal{Y}^n} g(y_1)f(y_2,y_1,y_3,\ldots,y_n) \tag{14}$$

and swapping y_1 and y_2 (we can always do this, exchangeability is not used here),

$$= \sum_{(y_2,y_1,y_3,\ldots,y_n)\in\mathcal{Y}^n} g(y_2)f(y_1,y_2,y_3,\ldots,y_n) = \mathrm{E}(g(Y_2)) \tag{15}$$

Exercise 22. ($\star\star$)Let $Y_1,\ldots,Y_n$ be an exchangeable sequence. Use

$$\mathrm{E}(g(Y_i)) = \mathrm{E}(g(Y_1)) \text{ for all } i \in \{2,\ldots,n\} \tag{16}$$

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

$$\mathrm{Var}(Y_i) = \mathrm{Var}(Y_1) \text{ for all } i \in \{2,\ldots,n\} \tag{17}$$

Solution. By the usual properties of variance,

$$\mathrm{Var}(Y_i) = \mathrm{E}(Y_i^2) - \mathrm{E}(Y_i)^2 \tag{18}$$

and now applying16 twice

$$\mathrm{Var}(Y_i) = \mathrm{E}(Y_1^2) - \mathrm{E}(Y_1)^2 = \mathrm{Var}(Y_1)$$

Exercise 23. ($\star\star$)Let $Y_1,\ldots,Y_n$ be an exchangeable sequence. By expanding $var(\sum_{k=1}^n Y_k)$, show that when $i \neq j$,

$$cov(Y_i,Y_j) \geq -\frac{var(Y_1)}{n-1} \tag{19}$$

Solution. It is

$$0 \leq var\left(\sum_{k=1}^n Y_k\right) = \sum_{k=1}^n var(Y_k) + 2\sum_{k=1}^{n-1}\sum_{\ell=k+1}^{n} cov\left(Y_k,Y_\ell\right) \tag{20}$$

and because, by exchangeability, $\mathrm{var}\left(Y_k\right) = var\left(Y_1\right)$ and $\mathrm{cov}\left(Y_k,Y_\ell\right) = \mathrm{cov}\left(Y_i,Y_j\right)$ for all $k \neq \ell$,

$$= n\,\mathrm{var}\left(Y_1\right) + (n^2-n)\mathrm{cov}\left(Y_i,Y_j\right) \tag{21}$$

where the n^2-n factor can be derived as follows: note that the pairs of indices (k,ℓ) appearing in the sum can be put into a matrix—the sum does not include the diagonal of this matrix (n pairs), but otherwise covers precisely half of it, and the full matrix has n^2 pairs, so there are $(n^2-n)/2$ terms in the sum.

Consequently,

$$\mathrm{Cov}\left(Y_i,Y_j\right) \geq -\frac{n\mathrm{var}\left(Y_1\right)}{n^2-n} = -\frac{\mathrm{var}\left(Y_1\right)}{n-1} \tag{22}$$

Exercise 24. ($\star$)What does

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

imply about the correlation of infinite exchangeable sequences?

Solution. The correlation must be non-negative: because, as $n \to \infty$, $cov(Y_i, Y_j) \geq 0$ for all $i \neq j$.

Part IV

Bayesian Calculations

Exercise 25. ($\star\star\star$)(Nuisance parameters are involved) <-story

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 family of distributions $\text{N}(\mu, \sigma^2)$ 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

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

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

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

with

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

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

Hint: 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

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

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

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

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

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

3. 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

Solution.

1. I use the Bayes theorem

$$\begin{aligned}
\pi(\mu,\sigma^2|y) \propto & f(y|\mu,\sigma^2)\pi(\mu,\sigma^2) = \prod_{i=1}^{n}\mathrm{N}(y_i|\mu,\sigma^2)\mathrm{N}(\mu|\mu_0,\sigma^2\frac{1}{\tau_0})\mathrm{IG}(\sigma^2|a_0,k_0) \\
& \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma^2/\tau_0}\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{a_0}{2}+1}\exp\left(-\frac{1}{\sigma^2}k_0\right) \\
& \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{1}{2}+\frac{a_0}{2}+1}\exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{1}-\frac{1}{2}\frac{(\mu-\mu_0)^2}{1/\tau_0}\right]-\frac{1}{\sigma^2}k_0\right)
\end{aligned}$$

It is

$$-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{1}-\frac{1}{2}\frac{(\mu-\mu_0)^2}{1/\tau_0} = -\frac{1}{2}\frac{(\mu-\mu_n)^2}{\underbrace{v_n^2}_{=1/\tau_n}}+C$$

where

$$\begin{aligned}
v_n =& \left(\sum_{i=1}^{n}\frac{1}{1}+\frac{1}{1/\tau_0}\right)^{-1} = \frac{1}{n+\tau_0} \implies \tau_n = n+\tau_0 \\
\mu_n =& v_n\left(\sum_{i=1}^{n}\frac{y_i}{1}+\frac{\mu_0}{1/\tau_0}\right) \implies \mu_n = \frac{n\bar{y}+\tau_0\mu_0}{n+\tau_0} \\
C_n =& \frac{1}{2}\frac{\mu_n^2}{v_n^2}-\frac{1}{2}\left(n\bar{y}^2+\tau_0\mu_0^2\right) = \text{...calc...} = \frac{1}{2}ns_n^2+\frac{1}{2}\frac{\tau_0 n(\mu_0-\bar{y})^2}{n+\tau_0}
\end{aligned}$$

So

$$\begin{aligned}
\pi(\mu,\sigma^2|y) \propto & \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}+\frac{n}{2}+\frac{a_0}{2}+1}\exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}\frac{(\mu-\mu_n)^2}{1/\tau_n}+C_n\right]-\frac{1}{\sigma^2}k_0\right) \\
& \propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(\mu-\mu_n)^2}{\sigma^2/\tau_n}\right)}_{\propto \mathrm{N}(\mu|\mu_n,\sigma^2/\tau_n)} \times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\overbrace{\frac{n}{2}+\frac{a_0}{2}}^{=a_n}+1}\exp\left(-\frac{1}{\sigma^2}\overbrace{(k_0+C_n)}^{=k_n}\right)}_{\propto \mathrm{IG}(\sigma^2|a_n,k_n)} \\
& \propto \mathrm{N}(\mu|\mu_n,\sigma^2/\tau_n)\mathrm{IG}(\sigma^2|a_n,k_n)
\end{aligned}$$

2. It is

$$\pi(\mu|y) = \int \pi(\mu, \sigma^2|y)\mathrm{d}\sigma^2 = \int \mathrm{N}(\mu|\mu_n, \sigma^2/\tau_n)\mathrm{IG}(\sigma^2|a_n, k_n)\mathrm{d}\sigma^2$$

by change of variable $\xi = \sigma^2 \frac{1}{2k_n}$, it is

$$\begin{aligned}\pi(\mu|y) &= \int \mathrm{N}(\mu|\mu_n, \xi 2k_n \frac{1}{\tau_n}\frac{2a_n}{2a_n})\mathrm{IG}(\xi|\frac{2a_n}{2}, \frac{1}{2})\mathrm{d}\xi = \int \mathrm{N}(\mu|\mu_n, \xi \frac{1}{\tau_n}\frac{k_n}{a_n}2a_n)\mathrm{IG}(\xi|\frac{2a_n}{2}, \frac{1}{2})\mathrm{d}\xi \\ &= \mathrm{T}_1(\mu|\mu_n, \frac{k_n}{a_n}\frac{1}{\tau_n}, 2a_n)\end{aligned}$$

3. It is

$$\begin{aligned}g(z|y) &= \int f(z|\mu, \sigma^2)\pi(\mu, \sigma^2|y)\mathrm{d}\mu\mathrm{d}\sigma^2 = \int \mathrm{N}(z|\mu, \sigma^2)\mathrm{N}(\mu|\mu_n, \sigma^2/\tau_n)\mathrm{IG}(\sigma^2|a_n, k_n)\mathrm{d}\mu\mathrm{d}\sigma^2 \\ &= \int \left[\int \mathrm{N}(z|\mu, \sigma^2)\mathrm{N}(\mu|\mu_n, \sigma^2/\tau_n)\mathrm{d}\mu\right] \mathrm{IG}(\sigma^2|a_n, k_n)\mathrm{d}\sigma^2\end{aligned}$$

Normal density is symmetric $\mathrm{N}(z|\mu, \sigma^2)\mathrm{N}(\mu|\mu_n, \sigma^2/\tau_n) = \mathrm{N}(\mu|z, \sigma^2)\mathrm{N}(\mu|\mu_n, \sigma^2/\tau_n)$, and by using the Hint

$$\int \mathrm{N}(\mu|z, \sigma^2)\mathrm{N}(\mu|\mu_n, \sigma^2/\tau_n)\mathrm{d}\mu = \int \mathrm{N}(\mu|\text{const.}, \text{const.})\mathrm{N}\left(z|\mu_n, \sigma^2\left[\frac{1}{\tau_n}+1\right]\right)\mathrm{d}\mu = \mathrm{N}\left(z|\mu_n, \sigma^2\left[\frac{1}{\tau_n}+1\right]\right)$$

So

$$g(z|y) = \int \mathrm{N}\left(z|\mu_n, \sigma^2\left[\frac{1}{\tau_n}+1\right]\right)\mathrm{IG}(\sigma^2|a_n, k_n)\mathrm{d}\sigma^2$$

by change the variable $\xi = \sigma^2 \frac{1}{2k_n}$, it is

$$g(z|y) = \int \mathrm{N}\left(z|\mu_n, \xi\left[\frac{1}{\tau_n}+1\right]\frac{k_n}{a_n}2a_n\right)\mathrm{IG}(\xi|\frac{2a_n}{2}, \frac{1}{2})\mathrm{d}\xi = \mathrm{T}_1\left(z|\mu_n, \left[\frac{1}{\tau_n}+1\right]\frac{k_n}{a_n}, 2a_n\right)$$

The following is about the Normal linear model of regression. ~~The calculations are too challenging.; (not anymore...)~~

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

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

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$\epsilon \sim \text{N}(0, \sigma^2)$ at untried location x_f is

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

Consider that

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

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

Solution.

1. I use the Bayes theorem

$$\begin{aligned}\pi(\mu,\sigma^2|y) \propto & f(y|\mu,\sigma^2)\pi(\mu,\sigma^2) = \mathrm{N}(y|\Phi\beta, I\sigma^2)\mathrm{N}(\beta|\mu_0,\sigma^2V_0)\mathrm{IG}(\sigma^2|a_0,k_0)\\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2}(y-\Phi\beta)^\top(I\sigma^2)^{-1}(y-\Phi\beta)\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2}(\beta-\mu_0)^\top(V_0\sigma^2)^{-1}(\beta-\mu_0)\right)\\ &\quad\times\left(\frac{1}{\sigma^2}\right)^{\frac{a_0}{2}+1}\exp\left(-\frac{1}{\sigma^2}k_0\right)\end{aligned}$$

but

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

so

$$\begin{aligned}\pi(\mu,\sigma^2|y) \propto & \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2}\frac{1}{\sigma^2}(\beta-\hat\beta_n)^\top\left[\Phi^\top\Phi\right](\beta-\hat\beta_n) - \frac{1}{2}\frac{1}{\sigma^2}S_n\right)\\ &\quad\times\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2}(\beta-\mu_0)^\top(V_0\sigma^2)^{-1}(\beta-\mu_0)\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{a_0}{2}+1}\exp\left(-\frac{1}{\sigma^2}k_0\right)\\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2}\frac{1}{\sigma^2}(\beta-\hat\beta_n)^\top\left[\Phi^\top\Phi\right](\beta-\hat\beta_n) - \frac{1}{2}\frac{1}{\sigma^2}(\beta-\mu_0)^\top V_0^{-1}(\beta-\mu_0)\right)\\ &\quad\times\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{a_0}{2}+1}\exp\left(-\frac{1}{2}\frac{1}{\sigma^2}S_n - \frac{1}{\sigma^2}k_0\right)\end{aligned}$$

but

$$-\frac{1}{2}(\beta-\hat\beta_n)^\top\left[\Phi^\top\Phi\right](\beta-\hat\beta_n) - \frac{1}{2}(\beta-\mu_0)^\top V_0^{-1}(\beta-\mu_0) = -\frac{1}{2}(\beta-\mu_n)^\top V_n^{-1}(\beta-\mu_n) + \frac{1}{2}C_n$$

$$V_n^{-1} = \Phi^\top\Phi + V_0^{-1}; \qquad \mu_n = V_n\left(\Phi^\top\Phi\hat\beta_n + V_0^{-1}\mu_0\right) = V_n\left((\Phi^\top\Phi)^{-1}\Phi y + V_0^{-1}\mu_0\right)$$

$$C_n = \frac{1}{2}\mu_n^\top V_n^{-1}\mu_n - \frac{1}{2}\left(\mu_0^\top V_0^{-1}\mu_0 + \hat\beta_n^\top\left[\Phi^\top\Phi\right]\hat\beta_n\right) = \frac{1}{2}\mu_n^\top V_n^{-1}\mu_n - \frac{1}{2}\left(\mu_0^\top V_0^{-1}\mu_0 + y^\top\Phi^\top(\Phi^\top\Phi)^{-1}\Phi y\right)$$

So

$$\pi(\mu,\sigma^2|y) \propto \underbrace{\left(\frac{1}{|V_n\sigma^2|}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}(\beta-\mu_n)^\top\left[V_n\sigma^2\right]^{-1}(\beta-\mu_n)\right)}_{\propto \mathrm{N}_d(\beta|\mu_n,V_n\sigma^2)} \times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\overbrace{\frac{n}{2}+a_0}^{a_n}+1}\exp\left(-\frac{1}{\sigma^2}\overbrace{\left[\frac{1}{2}S_n - C_n + k_0\right]}^{=k_n}\right)}_{\propto \mathrm{IG}(\sigma^2|a_n,k_n)}$$

So

$$\begin{cases}\mu|\sigma^2 \sim \mathrm{N}(\mu_n, \sigma^2V_n)\\ \sigma^2 \sim \mathrm{IG}(a_n,k_n)\end{cases}$$

2. It is

$$\pi(\beta|y) = \int \pi(\beta, \sigma^2|y)\mathrm{d}\sigma^2 = \int \mathrm{N}(\beta|\mu_n, V_n\sigma^2)\mathrm{IG}(\sigma^2|a_n, k_n)\mathrm{d}\sigma^2$$

by change the variable $\xi = \sigma^2 \frac{1}{2k_n}$, it is

$$\pi(\beta|y) = \int \mathrm{N}(\beta|\mu_n, \xi 2k_n V_n \frac{2a_n}{2a_n})\mathrm{IG}(\xi|\frac{2a_n}{2}, \frac{1}{2})\mathrm{d}\xi = \int \mathrm{N}(\beta|\mu_n, \xi V_n \frac{k_n}{a_n} 2a_n)\mathrm{IG}(\xi|\frac{2a_n}{2}, \frac{1}{2})\mathrm{d}\xi$$
$$= \mathrm{T}_d(\beta|\mu_n, \frac{k_n}{a_n}V_n, 2a_n)$$

3. It is

$$g(y_f|y) = \int f(y_f|\Phi_f\beta, \sigma^2)\pi(\beta, \sigma^2|y)\mathrm{d}\beta\mathrm{d}\sigma^2 = \int \mathrm{N}(y_f|\Phi_f\beta, \sigma^2)\mathrm{N}(\beta|\mu_n, V_n\sigma^2)\mathrm{IG}(\sigma^2|a_n, k_n)\mathrm{d}\beta\mathrm{d}\sigma^2$$
$$= \int \underbrace{\left[\int \mathrm{N}(y_f|\Phi_f\beta, \sigma^2)\mathrm{N}(\beta|\mu_n, V_n\sigma^2)\mathrm{d}\beta\right]}_{=A} \mathrm{IG}(\sigma^2|a_n, k_n)\mathrm{d}\sigma^2$$

by change of variable for $\xi' = \Phi_f\beta \sim \mathrm{N}(\Phi_f\mu_n, \Phi_f^\top V_n\Phi_f\sigma^2)$

$$A = \int \mathrm{N}(y_f|\xi', \sigma^2)\mathrm{N}(\xi'|\Phi_f\mu_n, \Phi_f^\top V_n\Phi_f\sigma^2)\mathrm{d}\xi'$$

because Normal is symmetric around the mean

$$A = \int \mathrm{N}(\xi'|y_f, \sigma^2)\mathrm{N}(\xi'|\Phi_f\mu_n, \Phi_f^\top V_n\Phi_f\sigma^2)\mathrm{d}\xi'$$

by using the Hint

$$A = \int \mathrm{N}(\xi'|\text{const.}, \text{const.})\mathrm{N}\left(y_f|\Phi_f\mu_n, \sigma^2\left[\Phi_f^\top V_n\Phi_f + 1\right]\right)\mathrm{d}\xi = \mathrm{N}\left(y_f|\Phi_f\mu_n, \sigma^2\left[\Phi_f^\top V_n\Phi_f + 1\right]\right)$$

So

$$g(y_f|y) = \int \mathrm{N}\left(y_f|\Phi_f\mu_n, \sigma^2\left[\Phi_f^\top V_n\Phi_f + 1\right]\right)\mathrm{IG}(\sigma^2|a_n, k_n)\mathrm{d}\sigma^2$$

by change the variable $\xi = \sigma^2 \frac{1}{2k_n}$, it is

$$g(y_f|y) = \int \mathrm{N}\left(y_f|\Phi_f\mu_n, \xi\left[\Phi_f^\top V_n\Phi_f + 1\right]\frac{k_n}{a_n}2a_n\right)\mathrm{IG}(\xi|\frac{2a_n}{2}, \frac{1}{2})\mathrm{d}\xi = \mathrm{T}_1\left(y_f|\Phi_f\mu_n, \left[\Phi_f^\top V_n\Phi_f + 1\right]\frac{k_n}{a_n}, 2a_n\right)$$

So

$$y_f|y \sim \mathrm{T}_1\left(\Phi_f\mu_n, \left[\Phi_f^\top V_n\Phi_f + 1\right]\frac{k_n}{a_n}, 2a_n\right)$$

, or equiv.

$$y(x_f)|y \sim \mathrm{T}_1\left(\phi^\top(x_f)\mu_n, \left[\Phi_f^\top V_n\Phi_f + 1\right]\frac{k_n}{a_n}, 2a_n\right)$$

Part V

Sufficiency

Exercise 27. ($\star\star$)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

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

Solution.

1. The parametric model is

$$\text{Ex}(x|\theta) = \theta \exp(-\theta x) 1(x \geq 0)$$

It is member of the exponential family

$$\text{Ef}_1(x|u, g, h, c, \phi, \theta, c) = u(x) g(\theta) \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) h_j(x))$$

with $u(x_{1:n}) = 1$, $g(\theta) = \theta$, $c_1 = -1$, $\phi_1(\theta) = \theta$, $h_1(x) = x$. The sufficient statistic is $t_n = (n, \sum_{i=1}^n x_i)$.

2. I can get the posterior by using the Bayes theorem

$$\begin{aligned} \pi(\theta|x) \propto f(x|\theta)\pi(\theta|a, b) \qquad & \propto \prod_{i=1}^{n} \text{Ex}(x_i|\theta) \text{Ga}(\theta|a, b) \\ & \propto \theta^n \exp(-\theta \sum_{i=1}^{n} x_i) \theta^{a-1} \exp(-\theta b) \ \propto \theta^{a+n-1} \exp(-\theta(\sum_{i=1}^{n} x_i + b)) \\ & \propto \text{Ga}(\theta|\underbrace{a + n}_{=a^*}, \underbrace{b + \sum_{i=1}^{n} x_i}_{=b^*}) \end{aligned}$$

3. By using the definition of the predictive distribution, it is ...

$$g(z|x) = \int_{\mathbb{R}_+} f(z|\theta)\pi(\theta|x)\mathrm{d}\theta \overset{z\geq 0}{=} \int_{\mathbb{R}_+} \theta \exp(-\theta z)\, \frac{(b^*)^{a^*}}{\Gamma(a^*)}\theta^{a^*-1}\exp(-\theta b^*)\mathrm{d}\theta$$
$$= \frac{(b^*)^{a^*}}{\Gamma(a^*)}\int_{\mathbb{R}_+} \theta^{a^*+1-1}\exp(-\theta(z+b^*))\mathrm{d}\theta = \frac{(b^*)^{a^*}}{\Gamma(a^*)}\frac{\Gamma(a^*+1)}{(z+b^*)^{a^*+1}} = \frac{a^*(b^*)^{a^*}}{(z+b^*)^{a^*+1}}$$
$$= \frac{a^*(b^*)^{a^*}}{(z+b^*)^{a^*+1}}$$

Exercise 28. ($\star\star\star$)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

$$\text{Mu}_k(x|\theta) = \begin{cases} \prod_{j=1}^k \theta_j^{x_j} & \text{, if } x \in \mathcal{X}_k \\ 0 & \text{, otherwise} \end{cases} \tag{23}$$

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

Solution.

1. There are $k-1$ independent parameters in $\text{Mu}_k(\theta)$ because $\sum_{j=1}^k \theta_j = 1$. I consider as parameters $(\theta_1, ..., \theta_{k-1})$ and the last one is a function of them as $\theta_k = 1 - \sum_{j=1}^{k-1}\theta_j$.

   It is

$$\text{Mu}_k(x|\theta) = \prod_{j=1}^k \theta_j^{x_j} = \prod_{j=1}^{k-1}\theta_j^{x_j}(1-\sum_{j=1}^{k-1}\theta_j)^{1-\sum_{j=1}^{k-1}x_j} = (1-\sum_{j=1}^{k-1}\theta_j)\exp(\sum_{j=1}^{k-1}x_j\log(\frac{\theta_j}{1-\sum_{j=1}^{k-1}\theta_j}))$$

This is the $k-1$ exponential family PDF with

$$u(x)=1; \qquad g(\theta)=(1-\sum_{j=1}^{k-1}\theta_j); \qquad c=(1,...,1)$$

$$h(x)=(x_1,...x_{k-1}); \qquad \phi(\theta)=(\log(\frac{\theta_1}{1-\sum_{j=1}^{k-1}\theta_j}),...,\log(\frac{\theta_{k-1}}{1-\sum_{j=1}^{k-1}\theta_j})),$$

2. The likelihood is

$$f(x_{1:n}|\theta)=\prod_{i=1}^{n}\text{Mu}_k(x_i|\theta)=\prod_{j=1}^{k}\theta_j^{\sum_{i=1}^{n}x_{i,j}}=\prod_{j=1}^{k}\theta_j^{x_{*,j}}=(1-\sum_{j=1}^{k-1}\theta_j)^n\exp\left(\sum_{j=1}^{k-1}x_{*,j}\log(\frac{\theta_j}{1-\sum_{j=1}^{k-1}\theta_j})\right)$$

and the sufficient statistic is

$$t_n=(n,x_{*,1},...,x_{*,k-1})$$

3. It is

$$\pi(\theta|x_{1:n})=\prod_{i=1}^{n}\text{Mu}_k(x_i|\theta)\text{Di}_k(\theta|a)\propto\prod_{j=1}^{k}\theta_j^{x_{*,j}}\prod_{j=1}^{k}\theta_j^{a_j-1}=\prod_{j=1}^{k}\theta_j^{x_j+a_{*,j}-1}\propto\text{Di}_k(\theta|\tilde{a})$$

where $\tilde{a}=(\tilde{a}_1,...,\tilde{a}_k)$, with $\tilde{a}_j=a_j+x_{*,j}$ for $j=1,...,k$. So the posterior is $\theta|x_{1:n}\sim\text{Di}_k(\tilde{a})$.

4. It is

$$\begin{aligned}
p(y|x_{1:n}) &= \int \text{Mu}_k(y|\theta)\text{Di}_k(\theta|\tilde{a})\text{d}\theta=\int\prod_{j=1}^{k}\theta_j^{y_j}\frac{\Gamma(\sum_{j=1}^{k}\tilde{a}_j)}{\prod_{j=1}^{k}\Gamma(\tilde{a}_j)}\prod_{j=1}^{k}\theta_j^{\tilde{a_j}-1}\text{d}\theta\\
&= \frac{\Gamma(\sum_{j=1}^{k}\tilde{a}_j)}{\prod_{j=1}^{k}\Gamma(\tilde{a}_j)}\int\prod_{j=1}^{k}\theta_j^{y_j+\tilde{a_j}-1}\text{d}\theta=\frac{\Gamma(\sum_{j=1}^{k}\tilde{a}_j)}{\prod_{j=1}^{k}\Gamma(\tilde{a}_j)}\frac{\prod_{j=1}^{k}\Gamma(y_j+\tilde{a}_j)}{\Gamma(\sum_{j=1}^{k}(y_j+\tilde{a}_j))}\\
&= \frac{\Gamma(\sum_{j=1}^{k}\tilde{a}_j)}{\prod_{j=1}^{k}\Gamma(\tilde{a}_j)}\frac{\prod_{j=1}^{k}\Gamma(y_j+a_j+x_{*,j})}{\Gamma(2n+a_j))}\\
&= \frac{\Gamma(a_*+x_{*,*})}{\prod_{j=1}^{k}\Gamma(a_j+x_{*,j})}\frac{\prod_{j=1}^{k}\Gamma(y_j+a_j+x_{*,j})}{\Gamma(2n+a_j))}\\
&= \frac{\Gamma(n+a_*)}{\Gamma(2n+a_j))}\prod_{j=1}^{k}\frac{\Gamma(y_j+a_j+x_{*,j})}{\Gamma(a_j+x_{*,j})}=\frac{\prod_{j=1}^{k}\prod_{\ell=0}^{y_j-1}(a_j+x_{*,j}+\ell)}{\prod_{\ell=0}^{n-1}(a_*+n+\ell)}
\end{aligned}$$

Exercise 29. ($\star\star$)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.

Solution. It is

$$
\begin{aligned}
p(\mathbf{x}|\varpi) &= \frac{n!}{x!\,y!\,z!}\pi^x\rho^y\sigma^z \\
&= \underbrace{\left(\frac{n!}{x!\,y!\,(n-x-y)!}\right)}_{=u(x,y)}\underbrace{\exp\{n\log(1-\pi-\rho)\}}_{=g(\pi,\rho)}\exp[\underbrace{1}_{=c_1}\underbrace{x}_{=h_1(x,y)}\underbrace{\log\{\pi/(1-\pi-\rho\}}_{=\phi_1(\pi,\rho)}+\underbrace{1}_{=c_2}\underbrace{y}_{=h_2(x,y)}\underbrace{\log\{\rho/(1-\pi-\rho)\}}_{=\phi_2(\pi,\rho)}] \\
&= u(x,y)\times g(\pi,\rho)\times\exp[c_1h_1(x,y)\phi_1(\pi,\rho)+c_2h_2(x,y)\phi_2(\pi,\rho)]
\end{aligned}
$$

Exercise 30. ($\star$)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.

Solution. Elementary manipulation gives

$$
\begin{aligned}
&n\bar{x}^2+n_0\theta_0^2-(n+n_0)\left(\frac{n\bar{x}+n_0\theta_0}{n+n_0}\right)^2 \\
&= \frac{1}{n+n_0}[\{n(n+n_0)-n^2\}\bar{x}^2+\{n_0(n+n_0)-n_o^2\}\theta_0^2-2(nn_0)\bar{x}\theta] \\
&= \frac{nn_0}{n+n_0}[\bar{x}^2+\theta_0^2-2\bar{x}\theta_0] = (n_0^{-1}+n^{-1})^{-1}(\bar{x}-\theta)^2.
\end{aligned}
$$

The following is a proof of a theorem

Exercise 31. ($\star\star\star$)(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

$$L(y;\theta)=k(t(y)|\theta)\rho(y). \tag{24}$$

Solution.

($\Longleftarrow$) I need to show that for any set $\Theta'\subseteq\Theta$ it is $\Pi(\Theta'|y)=\Pi(\Theta'|t)$, where $t=t(y)$.

For any Θ' and Y', it is

$$\Pi(\Theta'|Y') = \int\Pi(\Theta'|y)\mathrm{d}F(y|Y') = \mathrm{E}_{y|Y'}(\Pi(\Theta'|y)\,|\,Y')$$

Let's take $\mathcal{Y}'=\mathcal{Y}(t)=\{y:t(y)=t'\}$ for some t' .

It is

$$
\begin{aligned}
\Pi(\Theta'|y=y') =& \Pi(\Theta'|\mathcal{Y}') = \mathrm{E}_{y|Y'}(\Pi(\theta|y) = \mathrm{E}_{y|\mathcal{Y}'}\left(\int_{\Theta'}\mathrm{d}\Pi(\theta|y)\right) = \mathrm{E}_{y|\mathcal{Y}'}\left(\frac{\int_{\Theta'}L(y;\theta)\mathrm{d}\Pi(\theta)}{\int_{\Theta}L(y;\theta)\mathrm{d}\Pi(\theta)}\,|\,\mathcal{Y}'\right) \\
&= \mathrm{E}_{y|\mathcal{Y}'}\left(\frac{\int_{\Theta'}k(t(y)|\theta)\rho(y)\mathrm{d}\Pi(\theta)}{\int_{\Theta}k(t(y)|\theta)\rho(y)\mathrm{d}\Pi(\theta)}\,|\,\mathcal{Y}'\right) = \mathrm{E}_{y|\mathcal{Y}'}\left(\frac{\int_{\Theta'}k(t(y)|\theta)\mathrm{d}\Pi(\theta)}{\int_{\Theta}k(t(y)|\theta)\mathrm{d}\Pi(\theta)}\,|\,Y'\right)
\end{aligned}
$$

Because the fraction inside the expectation is the same for all values $y \in \mathcal{Y}'$, we may suppress it. So

$$\Pi(\Theta'|y) = \frac{\int_{\Theta'} k(t(y)|\theta)\mathrm{d}\Pi(\theta)}{\int_{\Theta} k(t(y)|\theta)\mathrm{d}\Pi(\theta)} = \Pi(\Theta'|t)$$

($\Longleftarrow$) I ll construct a 'kernel' $\kappa(y|\theta)$ invariant for each $y \in \mathcal{Y}(t)$ where $\mathcal{Y}(t) = \{y : t(y) = t\}$. I set

$$\kappa(y|\theta) = \frac{f(y|\theta)}{f(y)}\,; \qquad\qquad \rho(y) = f(y)\,; \tag{25}$$

so by Bayes theorem

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{f(y)} = \pi(\theta)\frac{f(y|\theta)}{f(y)} = \pi(\theta)\kappa(y|\theta)$$

Because t is parametric sufficient for any $y, y' \in \mathcal{Y}(t) = \{y : t(y) = t\}$ it is

$$\pi(\theta|y) = \pi(\theta|t) = \pi(\theta|y')$$

so

$$\pi(\theta|y) = \pi(\theta|y') \implies \pi(\theta)\kappa(y|\theta) = \pi(\theta)\kappa(y'|\theta) \overset{\pi(\theta)\neq 0}{\Longrightarrow} \kappa(y|\theta) = \kappa(y'|\theta)$$

From every $t \in \mathcal{T}$ I choose one $y \in \mathcal{Y}(t)$, I set $k(t|\theta) = \kappa(y|\theta)$, and I substitute it in (25), so I get

$$\kappa(y|\theta) = \frac{f(y|\theta)}{f(y)} \implies f(y|\theta) = \kappa(y|\theta)f(y) \implies f(y|\theta) = k(t(y)|\theta)\rho(y)$$

The following is a proof of a theorem

Exercise 32. ($\star\star\star$)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.

Solution. Let $\mathcal{Y}(t) = \{y :\ t = t(y)\}$, and let $z = (y_{n+1}, ..., y_{n+m})$ then

$$p(z|t) = \frac{p(z,t)}{p(t)} = \frac{\int_{\mathcal{Y}(t)} p(y_{n+1:n+m}, y_{1:n})\mathrm{d}y_{1:n}}{p(t)}$$

$$(repr.theor.) = \frac{1}{p(t)}\int_{\mathcal{Y}(t)}[\int_{\Theta}\prod_{i=1}^{n+m} f(y_i|\theta)\mathrm{d}\Pi(\theta)]\mathrm{d}y_{1:n} = \frac{1}{p(t)}\int_{\Theta}[\int_{\mathcal{Y}(t)}\prod_{i=n+1}^{n+m} f(y_i|\theta)\prod_{i=1}^{n} f(y_i|\theta)\mathrm{d}y_1\cdots\mathrm{d}y_n]\mathrm{d}\Pi(\theta)$$

$$= \frac{1}{p(t)}\int_{\Theta}\prod_{i=n+1}^{n+m} f(y_i|\theta)f(t|\theta)\mathrm{d}\Pi(\theta)$$

$$= \int_{\Theta}\prod_{i=n+1}^{n+m} f(y_i|\theta)\underbrace{\frac{f(t|\theta)\mathrm{d}\Pi(\theta)}{p(t)}}_{=\mathrm{d}\Pi(\theta|t)} = \int_{\Theta}\prod_{i=n+1}^{n+m} f(y_i|\theta)\mathrm{d}\Pi(\theta|t).$$

So

$$p(z|t) = \int_{\Theta}\prod_{i=n+1}^{n+m} f(y_i|\theta)\mathrm{d}\Pi(\theta|t); \qquad\qquad p(z|y) = \int_{\Theta}\prod_{i=n+1}^{n+m} f(y_i|\theta)\mathrm{d}\Pi(\theta|y)$$

Therefore, we have parametric sufficiency $p(z|t) = p(z|y)$ if and only if $\mathrm{d}\Pi(\theta|t) = \mathrm{d}\Pi(\theta|y)$ for all $\mathrm{d}\Pi(\theta)$.

Part VI

Priors

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.

Solution.

1. the posterior distribution $\Pi(\mu|x_1)$ has PDF

$$\begin{aligned}\pi(\mu|x_1) &\propto \overbrace{\mathrm{N}(x_1|\mu, \sigma^2)}^{\text{likelihood}}\overbrace{\mathrm{N}(\mu|\mu_0, \sigma_0^2)}^{\text{prior}} \\ &\propto \exp(-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2})\exp(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}) \\ &\propto \exp(-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2} - \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}) \\ &\propto \exp(-\frac{1}{2}\frac{(\mu-\hat{\mu}_1)^2}{\hat{\sigma}_1^2}) \propto \mathrm{N}(\mu|\hat{\mu}_1, \hat{\sigma}_1^2) \qquad (26)\end{aligned}$$

   where $\hat{\sigma}_1^2 = (\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2})^{-1}$, and $\hat{\mu}_1 = \hat{\sigma}_1^2(\frac{x_1}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})$. In (26), we recognized the kernel of the Normal PDF. Hence, $\mu|x_1 \sim \mathrm{N}(\hat{\mu}_1, \hat{\sigma}_1^2)$

Then the posterior distribution $\Pi(\mu|x_1, x_2)$ has PDF

$$\begin{aligned}\pi(\mu|x_1,x_2) &\propto \overbrace{(x_2|\mu,\sigma^2)}^{\text{likelihood}}\overbrace{\mathrm{N}(\mu|\hat{\mu}_1,\hat{\sigma}_1^2)}^{\text{prior}}\\ &\propto \exp(-\frac{1}{2}\frac{(x_2-\mu)^2}{\sigma^2})\exp(-\frac{1}{2}\frac{(\mu-\hat{\mu}_1)^2}{\hat{\sigma}_1^2})\\ &\propto \exp(-\frac{1}{2}\frac{(x_2-\mu)^2}{\sigma^2}-\frac{1}{2}\frac{(\mu-\hat{\mu}_1)^2}{\hat{\sigma}_1^2})\\ &\propto \exp(-\frac{1}{2}\frac{(\mu-\hat{\mu}_2)^2}{\hat{\sigma}_2^2}) \propto \mathrm{N}(\mu|\hat{\mu}_2,\hat{\sigma}_2^2) \end{aligned} \tag{27}$$

where $\hat{\sigma}_2^2 = (\frac{1}{\sigma^2}+\frac{1}{\hat{\sigma}_1^2})^{-1} = (\frac{1}{\sigma^2}+\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2})^{-1}$, and $\hat{\mu}_2 = \hat{\sigma}_1^2(\frac{x_2}{\sigma^2}+\frac{\hat{\mu}_1}{\hat{\sigma}_1^2}) = \hat{\sigma}_2^2(\frac{x_1}{\sigma^2}+\frac{x_2}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})$. In (26), we recognized the kernel of the Normal PDF. Hence, $\mu|x_1, x_2 \sim \mathrm{N}(\hat{\mu}_2, \hat{\sigma}_2^2)$.

2. The posterior distribution $\Pi(\mu|x_1, x_2)$ has PDF

$$\begin{aligned}\pi(\mu|x_1,x_2) &\propto \overbrace{\mathrm{N}(x_1|\mu,\sigma^2)\mathrm{N}(x_2|\mu,\sigma^2)}^{\text{likelihood}}\overbrace{\mathrm{N}(\mu|\mu_0,\sigma_0^2)}^{\text{prior}}\\ &\propto \exp(-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2})\exp(-\frac{1}{2}\frac{(x_2-\mu)^2}{\sigma^2})\exp(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2})\\ &\propto \exp(-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2}-\frac{1}{2}\frac{(x_2-\mu)^2}{\sigma^2}-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2})\\ &\propto \exp(-\frac{1}{2}\frac{(\mu-\hat{\mu})^2}{\hat{\sigma}^2}) \propto \mathrm{N}(\mu|\hat{\mu},\hat{\sigma}^2) \end{aligned} \tag{28}$$

where $\hat{\sigma}^2 = (\frac{1}{\sigma^2}+\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2})^{-1}$, and $\hat{\mu} = \hat{\sigma}^2(\frac{x_1}{\sigma^2}+\frac{x_2}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})$. In (28), we recognized the kernel of the Normal PDF. Hence, $\mu|x_1, x_2 \sim \mathrm{N}(\hat{\mu}, \hat{\sigma}^2)$

3. It is easy to see that $\hat{\mu}_2 = \hat{\mu}$, and $\hat{\sigma}_2^2 = \hat{\sigma}^2$, from (1) and (2). We observe the two Learning Scenarios are equivalent in the sense that they lead to the same posterior $\mathrm{d}\Pi(\mu|x_1, x_2)$ at the end posterior $\mathrm{d}\Pi(\mu|x_1, x_2)$ in a single application of Bayes theorem with the full data $x = (x_1, x_2)$.

Exercise 34. ($\star\star$)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.

Solution. It is

$$f(y) \leq \sum_{\forall y} f(y) = \sum_{\forall y} \overbrace{\int f(y|\theta)\mathrm{d}\Pi(\theta)}^{f(y)=} \overset{\text{Fubini}}{=} \int \sum_{\forall y} f(y|\theta)\mathrm{d}\Pi(\theta) = \int \mathrm{d}\Pi(\theta) = 1$$

Exercise 35. ($\star\star$)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.

Solution. It is

$$\int f(y)\mathrm{d}y = \int_{\forall y} \overbrace{\int_{\forall\theta} f(y|\theta)\mathrm{d}\Pi(\theta)}^{f(y)=}\mathrm{d}y \overset{\text{Fubini}}{=} \int_{\forall\theta}\int_{\forall y} f(y|\theta)\mathrm{d}y\mathrm{d}\Pi(\theta) = \int \mathrm{d}\Pi(\theta) = 1$$

So it is $f(y) < \infty$ for every set of y (possibly) apart from a finite number of y's with 'probability' zero.

Exercise 36. ($\star\star$)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

$$\pi(\theta|y) = \pi(\theta|\tau^*)\ \propto g(\theta)^{\tau^*} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j^*) \tag{29}$$

   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

$$g(z|y) = \prod_{i=1}^{m} u(z_i)\frac{K(\tau + t(y) + t(z))}{K(\tau + t_n(y))} \tag{30}$$

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

Solution.

1. According to the Bayes theorem, where

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta) \propto g(\theta)^n \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{i=1}^{n} h_j(y_i)))\, g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j)$$

$$\propto g(\theta)^{n+\tau_0} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{i=1}^{n} h_j(y_i) + \tau_j)))\ \propto \tilde{\pi}(\theta|y, \tau + t(y)).$$

2. According to the predictive pdf/pmf equation, and assuming that θ is a continuous random quantity, it is

$$g(z|y) = \int_\Theta \prod_{l=1}^{m} f(z_l|\theta)\pi(\theta|y)\mathrm{d}\theta = \int_\Theta \prod_{l=1}^{m} u(z_l)g(\theta)^m \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{l=1}^{m} h_j(z_l))))$$

$$\times \frac{1}{K(\tau + t(y))} g(\theta)^{n+\tau_0} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{i=1}^{n} h_j(y_i) + \tau_j)))\mathrm{d}\theta$$

$$= \prod_{l=1}^{m} u(z_l)\frac{1}{K(\tau + t_n(y))} \underbrace{\int_\Theta g(\theta)^{n+m+\tau_0} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{l=1}^{m} h_j(z_l) + \sum_{i=1}^{n} h_j(y_i) + \tau_j)))\mathrm{d}\theta}_{=K(\tau+t(y)+t(z))}$$

   For the case where θ is a discrete random quantity, the proof is similar.

Exercise 37. ($\star\star$)

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

$$f(x|\theta) = \frac{\theta e^{-x}}{(1+e^{-x})^{\theta+1}} \tag{31}$$

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

$$f(\theta|\alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta^{\alpha_0 - 1} e^{-\beta_0 \theta} \tag{32}$$

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?

Solution.

1. It is

$$\begin{aligned} f(x|\theta) &= \frac{\theta e^{-x}}{(1+e^{-x})^{\theta+1}} \\ &= e^{-x}\theta \frac{1}{(1+e^{-x})} \exp\left(-\theta \log(1+e^{-x})\right) \\ &= e^{-x}\theta \exp\left(-\theta \log(1+e^{-x})\right) \\ &= \frac{e^{-x}}{(1+e^{-x})} \theta \exp\left(-\theta \log(1+e^{-x})\right) \end{aligned}$$

So it is a member of the exponential distribution family

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

with $k = 1$, and

$$u(x) = \frac{e^{-x}}{(1+e^{-x})}, \qquad g(\theta) = \theta, \qquad h(x) = \log(1+e^{-x}), \qquad \phi(\theta) = \theta, \qquad c = -1.$$

2. Following the corresponding theorem,

$$\begin{aligned} \pi(\theta|\tau) &\propto g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} c_j \phi_j(\theta) \tau_j) \\ &\propto \theta^{\tau_0} \exp(-1\theta\tau_1) \ \propto \theta^{(\tau_0+1)-1} \exp(-\theta\tau_1) \end{aligned}$$

where we recognize the Gamma PDF which identifies a Gamma distribution $\theta|\tau \sim \text{Ga}(\tau_0+1, \tau_1)$.

Also

$$K(\tau) = \int_{\mathbb{R}_+} g(\theta)^{\tau_0} \exp(\sum_{j=1}^{k} \phi_j(\theta)\tau_j)\mathrm{d}\theta = \int_{\mathbb{R}_+} \theta^{\tau_0}\exp(-\theta\tau_1)\mathrm{d}\theta = \frac{\Gamma(\tau_0+1)}{\tau_1^{\tau_0+1}}$$

since $\int_{\mathbb{R}_+} \text{Ga}(\theta|\tau_0+1,\tau_1)\mathrm{d}\theta = 1$. To easy the notation with $\theta|\tau \sim \text{Ga}(a,b)$, we can set $(a,b) = (\tau_0+1,\tau_1)$.

3. By using the Bayes theorem

$$\begin{aligned}
\pi(\theta|x_{1:n}) \propto f(x_{1:n}|\theta)\pi(\theta|a,b) \qquad &\propto \prod_{i=1}^{n} f(x_i|\theta)\text{Ga}(\theta|a,b) \\
&\propto \theta^n \exp(-\theta\sum_{i=1}^{n}\log(1+e^{-x_i}))\theta^{a-1}\exp(-\theta b) \\
&\propto \theta^{a+n-1}\exp(-\theta(\sum_{i=1}^{n}\log(1+e^{-x_i})+b)) \\
&\propto \text{Ga}(\theta|\underbrace{a+n}_{=a^*}, \underbrace{b+\sum_{i=1}^{n}\log(1+e^{-x_i})}_{=b^*})
\end{aligned}$$

4. By using the definition for the predictive PDF, it is

$$\begin{aligned}
f(y|x_{1:n}) &= \int_{\mathbb{R}_+} f(y|\theta)\text{Ga}(\theta|a^*,b^*)\mathrm{d}\theta \\
&\propto \int_{\mathbb{R}_+} \frac{e^{-y}}{(1+e^{-y})}\theta\exp(-\theta\log(1+e^{-y}))\theta^{a^*-1}\exp(-\theta b^*)\mathrm{d}\theta \\
&\propto \frac{e^{-y}}{(1+e^{-y})}\int_{\mathbb{R}_+}\theta^{a^*+1-1}\exp(-\theta(b^*+\log(1+e^{-y})))\mathrm{d}\theta \\
&\propto \frac{e^{-y}}{(1+e^{-y})}\frac{\Gamma(a^*+1)}{(b^*+\log(1+e^{-y}))^{a^*+1}} \\
&\propto \frac{e^{-y}}{(1+e^{-y})}\frac{1}{(b^*+\log(1+e^{-y}))^{a^*+1}}
\end{aligned}$$

5. Because the parametric model is member of the exponential family, the minimal sufficient statistic is $(n, \sum_{i=1}^{n} h(x_i)) = (n, \sum_{i=1}^{n}\log(1+e^{-x_i}))$. By the Neyman factorisation theorem sufficiency only depends on the likelihood: it does not depend on our choice of prior.

The exercise below is theoretical, and very challenging.

Exercise 38. ($\star\star\star$)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 (!!!)

Solution. Assume $\exists \mathcal{F}$ s.t. $\mathcal{F} = (\pi(\theta|\tau), \tau \in T)$, with $\dim(T) < \infty$. Let's take $\pi(\theta|\tau)$ as a (conjugate) prior. Then the posterior is $\pi(\theta|x_{1:n}) \propto f(x_{1:n}|\theta)\pi(\theta|\tau)$. Due to the conjugacy $\exists \tau^*(x_{1:n})$ s.t.

$$\pi(\theta|\tau^*(x_{1:n})) = \frac{f(x_{1:n}|\theta)\pi(\theta|\tau)}{p(x_{1:n})} \iff$$

$$f(x_{1:n}|\theta) = \underbrace{\frac{\pi(\theta|\tau^*(x_{1:n}))}{\pi(\theta|\tau)}}_{=h(\tau^*(x_{1:n}),\theta)} \underbrace{p(x_{1:n})}_{=g(x_{1:n})}$$

Due to Newman factorisation criterion $\tau^*(x_{1:n})$ is a sufficient statistic, and its dimentionality does not depend on θ. Then because of the Pitman-Koopman Lemma, the $f(x_{1:n}|\theta)$ is a member of the exponential family.

- Before that, we knew that parametric models which are members of the exponential family have conjugate priors. The result above says that if there is a conjugate prior for a parametric model whose sample space does not depend on the unknown parameter, then the parametric model is member of the exponential family. Of course it does not apply to Uniform or Pareto parametric models whose sample space depends on the unknown parameter.

Exercise 39. ($\star\star$)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?

Solution. I will make Bayesian Statistical Inference. The parametric model is the Bernoulli distribution, because the experiment is a Bernoulli experiment. The limiting frequency of successes, aka $\theta \in (0, 1)$ is the uncertain parameter for which I want to perform inference. To account uncertainty about the unknown parameter θ, I will assign a prior distribution. For my computational convenience, I will try to see if there exist a conjugate prior. If it exists, I will assign a conjugate prior for θ. So...

Consider the Bayesian model

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

where $\theta \in \mathbb{R}$.

- I will try to find a conjugate prior in order to do it

  The likelihood is such that

  $$f(x_{1:n}|\theta) = \prod_{i=1}^{n} \text{Br}(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = (1-\theta)^n \exp(\log(\frac{\theta}{1-\theta}) \sum_{i=1}^{n} x_i)$$

  The parametric model $f(\cdot|\theta)$ is the Bernoulli distribution which belongs to the exponential family with $u(x) = 1$, $g(\theta) = (1-\theta)$, $c_1 = 1$, $\phi_1(\theta) = \log(\frac{\theta}{1-\theta})$, $h_1(x) = x$.

  The corresponding conjugate prior has pdf such as

  $$\pi(\theta|\tau) = K(\tau)(1-\theta)^{\tau_0} \exp(\log(\frac{\theta}{1-\theta})\tau_1) = K(\tau)\theta^{(\tau_1+1)-1}(1-\theta)^{(\tau_0-\tau_1+1)-1}$$

  where $K(\tau) = \int_0^1 \theta^{(\tau_1+1)-1}(1-\theta)^{(\tau_0-\tau_1+1)-1}\text{d}\theta = \frac{\Gamma(\tau_1+1)\Gamma(\tau_0-\tau_1+1)}{\Gamma(\tau_0+2)}$

  Since we recognize that the prior distribution is Beta, we perform a re-parametrization, as

  $$\pi(\theta|\tau) = \text{Be}(\theta|a, b)$$

where $a = \tau_1 + 1$, $b = \tau_0 - \tau_1 + 1$.

- According to my prior info:
  - my prior info worth 12 trials, so the effective number of observations that the prior distribution contributes is $\tau_0 = 12$,
  - the a priori mean of θ is 0.4, so $\mathrm{E}(\theta) = \frac{a}{a+b} = 0.4$. So
    - $a + b = (\tau_1 + 1) + (\tau_0 - \tau_1 + 1) = \tau_0 + 2 = 14$
    - $\frac{a}{a+b} = 0.4 \iff a = 5.6$ and $b = 8.4$
- By using the Bayes theorem, or the theorem about the conjugate posteriors of parametric models in Exponential family, I get a posterior

$$\theta|x_{1:n} \sim \mathrm{Be}(\underbrace{\sum_{i=1}^{n} x_i + a}_{=a^*}, \underbrace{n - \sum_{i=1}^{n} x_i + b}_{=b^*}) \equiv \mathrm{Be}(12 + 5.6, 13 + 8.4) \equiv \mathrm{Be}(17.6, 21.4)$$

Exercise 40. ($\star\star$)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 $0 < \alpha < \infty$, $0 < \beta < \infty$.

Solution. Because

$$\begin{aligned} p(\boldsymbol{x}|\alpha, \beta) &= \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta) \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \exp(-1x/\beta + 1(a-1)\log(x)) \end{aligned}$$

it belongs to the exponential family, the sufficient statistic is $(n, \sum_{i=1}^n x_i, \sum_{i=1}^n \log(x_i))$ or equivalently $(n, \sum_{i=1}^n x_i, \prod_{i=1}^n x_i)$.

Exercise 41. ($\star\star$)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.

Solution. The sampling distribution is

$$f(y_i|\theta) = \mathrm{N}(y_i|\theta, \theta^2) \ \propto (\theta^2)^{-1/2} \exp(-\frac{1}{2}\frac{(y_i - \theta)^2}{\theta^2}) \propto |\theta|^{-1} \exp(-\frac{1}{2} y_i^2 \frac{1}{\theta^2} + y_i \frac{1}{\theta})$$

and hence it belongs to the exponential family with $g(\theta) = |\theta|^{-1}$, $c_1 = -\frac{1}{2}$, $\phi_1(\theta) = \frac{1}{\theta^2}$, $h_1(y_i) = y_i^2$, $c_2 = 1$, $\phi_2(\theta) = \frac{1}{\theta}$, $h_2(y_i) = y_i$.

The corresponding conjugate prior has pdf such as

$$\pi(\theta) = \tilde{\pi}(\theta|\tau) \propto |\theta|^{-\tau_0} \exp(-\frac{1}{2}\frac{1}{\theta^2}\tau_1 + \frac{1}{\theta}\tau_2), \qquad \text{where } \tau = (\tau_0, \tau_1, \tau_2).$$

I actually cannot recognize it as a standard distribution in this case. The posterior distribution has pdf such as

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta) = \prod_{i=1}^{n} \mathrm{N}(y_i|\theta, \theta^2)\pi(\theta) \propto |\theta|^{-(\tau_0+n)} \exp(-\frac{1}{2}\frac{1}{\theta^2}(\tau_1 + \sum_{i=1}^n y_i^2) + \frac{1}{\theta}(\tau_2 + \sum_{i=1}^n y_i))$$

Namely, $\pi(\theta|y) = \tilde{\pi}(\theta|\tau^*)$, with $\tau* = (\tau_0 + n, \tau_1 + \sum_{i=1}^n y_i^2, \tau_2 + \sum_{i=1}^n y_i)$; so it is conjugate.

Exercise 42. ($\star\star$)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.

Solution. The prior is of the form

$$\pi(\theta) = \varpi_1 \mathrm{N}(\theta|\mu_1, 1) + \varpi_2 \mathrm{N}(\theta|\mu_2, 1)$$

where $\varpi_1 = 2/3$, $\varpi_2 = 1/3$, $\mu_1 = 0$, and $\mu_2 = 1$.

- The posterior pdf will be of the form

$$\pi(\theta|x=2) = \varpi_1^* \pi_1(\theta|x=2) + \varpi_2^* \pi_2(\theta|x=2)$$

  with each component computed as follows.

- For the components of the mixture: For the 1st component, it is

$$\begin{aligned}
\pi_1(\theta|x=2) &\propto f(x|\theta)\pi_1(\theta) \ \propto \ \mathrm{N}(x|\theta,1)\mathrm{N}(\theta|\mu_1,1) \\
&\propto \exp(-\frac{1}{2}\frac{(x-\theta)^2}{1})\exp(-\frac{1}{2}\frac{(\theta-\mu_1)^2}{1}) \\
&= \exp(-\frac{1}{2}\frac{(\theta-x)^2}{1} - \frac{1}{2}\frac{(\theta-\mu_1)^2}{1}) \\
&\overset{\text{Hint}}{\propto} \exp(-\frac{1}{2}\frac{(\theta-\hat{\mu}_1)^2}{\hat{\sigma}_1^2}) \ \propto \ \mathrm{N}(\theta|\hat{\mu}_1, \hat{\sigma}_1^2)
\end{aligned}$$

  where $\hat{\sigma}_1^2 = 1/2$ $\hat{\mu_1} = \frac{x+\mu_1}{2} = 1$, because of $x = 2$, and $\mu_1 = 0$, and Hint . So

$$\pi_1(\theta|x=2) = \mathrm{N}(\theta|1, 1/2).$$

For the 2nd component, it is

$$\begin{aligned}\pi_2(\theta|x=2) \propto f(x|\theta)\pi_2(\theta) \;&\propto\; \mathrm{N}(x|\theta,1)\mathrm{N}(\theta|\mu_2,1)\\ &\propto \exp(-\frac{1}{2}\frac{(x-\theta)^2}{1})\exp(-\frac{1}{2}\frac{(\theta-\mu_2)^2}{1})\\ &= \exp(-\frac{1}{2}\frac{(\theta-x)^2}{1}-\frac{1}{2}\frac{(\theta-\mu_2)^2}{1})\\ &\overset{\text{Hint}}{\propto} \exp(-\frac{1}{2}\frac{(\theta-\hat{\mu}_2)^2}{\hat{\sigma}_2^2}) \;\propto\; \mathrm{N}(\theta|\hat{\mu}_2,\hat{\sigma}_2^2)\end{aligned}$$

where $\hat{\sigma}_2^2 = 1/2$, $\hat{\mu_2} = \frac{x+\mu_2}{2} = \frac{3}{2}$, because of $x=2$, $\mu_2 = 1$, and Hint.

So

$$\pi_2(\theta|x=2) = \mathrm{N}(\theta|3/2, 1/2)$$

- For the posterior weights, it is

$$\varpi_1^* = \frac{\varpi_1 f_1(x)}{\varpi_1 f_1(x) + \varpi_2 f_2(x)}; \qquad\qquad \varpi_2^* = \frac{\varpi_2 f_2(x)}{\varpi_1 f_1(x) + \varpi_2 f_2(x)};$$

So, I compute

$$\begin{aligned}f_1(x) = \int f(x|\theta)\pi_1(\theta)\mathrm{d}\theta \quad &= \int \mathrm{N}(x|\theta,1)\mathrm{N}(\theta|\mu_1,1)\mathrm{d}\theta\\ &= \int \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\frac{(\theta-x)^2}{1})\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\frac{(\theta-\mu_1)^2}{1})\mathrm{d}\theta\\ &= \frac{1}{2\pi}\int \exp(-\frac{1}{2}\frac{(\theta-x)^2}{1}-\frac{1}{2}\frac{(\theta-\mu_1)^2}{1})\mathrm{d}\theta\\ &\overset{\text{Hint}}{=} \frac{1}{2\pi}\int \exp(-\frac{1}{2}\frac{(\theta-\hat{\mu}_1)^2}{\hat{\sigma}_1^2}-\frac{1}{2}(\frac{x^2}{1}+\frac{\mu_1^2}{1})+\frac{1}{2}(x+\mu_1)^2(\frac{1}{1}+\frac{1}{1})^{-1})\mathrm{d}\theta\\ &= \frac{1}{2\pi}\underbrace{\int \exp(-\frac{1}{2}\frac{(\theta-\hat{\mu}_1)^2}{\hat{\sigma}_1^2})\mathrm{d}\theta}_{=\sqrt{2\pi\hat{\sigma}_1^2}}\exp(-\frac{1}{2}(x^2+\mu_1^2)+\frac{1}{2}\frac{(x+\mu_1)^2}{2})\\ &= \frac{1}{2\pi}\sqrt{2\pi\hat{\sigma}_1^2}\exp(-\frac{1}{2}(x^2+\mu_1^2)+\frac{1}{4}(x+\mu_1)^2)\\ &= \frac{1}{2\sqrt{\pi}}\exp(-\frac{1}{2}(2^2+1^2)+\frac{1}{4}(2+1)^2)\end{aligned}$$

because of Hint, $\hat{\mu}_1 = \frac{x+\mu_1}{2} = 1$, $\hat{\sigma}_1^2 = 1/2$, $x=2$, $\mu_1 = 0$. Hence,

$$f_1(x=2) = \frac{1}{2\sqrt{\pi}}\exp(-1) \approx 0.10.$$

Also, I compute

$$
\begin{aligned}
f_2(x) = \int f(x|\theta)\pi_2(\theta)\mathrm{d}\theta \qquad &= \int \mathrm{N}(x|\theta,1)\mathrm{N}(\theta|\mu_2,1)\mathrm{d}\theta \\
&= \int \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\frac{(\theta-x)^2}{1})\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\frac{(\theta-\mu_2)^2}{1})\mathrm{d}\theta \\
&= \frac{1}{2\pi}\int \exp(-\frac{1}{2}\frac{(\theta-x)^2}{1} - \frac{1}{2}\frac{(\theta-\mu_2)^2}{1})\mathrm{d}\theta \\
&\overset{\text{Hint}}{=} \frac{1}{2\pi}\int \exp(-\frac{1}{2}\frac{(\theta-\hat{\mu}_2)^2}{\hat{\sigma}_2^2} - \frac{1}{2}(\frac{x^2}{1}+\frac{\mu_2^2}{1}) + \frac{1}{2}(x+\mu_2)^2(\frac{1}{1}+\frac{1}{1})^{-1})\mathrm{d}\theta \\
&= \frac{1}{2\pi}\underbrace{\int \exp(-\frac{1}{2}\frac{(\theta-\hat{\mu}_2)^2}{\hat{\sigma}_2^2})\mathrm{d}\theta}_{=\sqrt{2\pi\hat{\sigma}_2^2}}\exp(-\frac{1}{2}(x^2+\mu_2^2) + \frac{1}{2}\frac{(x+\mu_2)^2}{2}) \\
&= \frac{1}{2\pi}\sqrt{2\pi\hat{\sigma}_2^2}\exp(-\frac{1}{2}(x^2+\mu_2^2) + \frac{1}{4}(x+\mu_2)^2) \\
&= \frac{1}{2\sqrt{\pi}}\exp(-\frac{1}{2}(2^2+(\frac{3}{2})^2) + \frac{1}{4}(2+\frac{3}{2})^2)
\end{aligned}
$$

because of Hint, and $\hat{\mu}_2 = \frac{x+\mu_2}{2} = \frac{3}{2}$, $\hat{\sigma}_2^2 = 1/2$, since $x = 2$, $\mu_2 = 1$. Hence,

$$f_2(x=2) = \frac{1}{2\sqrt{\pi}}\exp(-\frac{1}{4}) \approx 0.22$$

So

$$
\begin{aligned}
\varpi_1^* &= \frac{\varpi_1 f_1(x=2)}{\varpi_1 f_1(x=2) + \varpi_2 f_2(x=2)} = \frac{2\exp(-1)}{2\exp(-1)+\exp(-1/4)} \approx 0.48 \\
\varpi_2^* &= \frac{\varpi_2 f_2(x=2)}{\varpi_1 f_1(x=2) + \varpi_2 f_2(x=2)} = \frac{\exp(-1/4)}{2\exp(-1)+\exp(-1/4)} \approx 0.52
\end{aligned}
$$

- The posterior becomes

$$\pi(\theta|x=2) = \varpi_1^*\mathrm{N}(\theta|1,1/2) + \varpi_2^*\mathrm{N}(\theta|3/2,1/2)$$

- It is

$$
\begin{aligned}
\pi(\theta>1|x=2) = 1-\pi(\theta\leq 1|x=2) &= 1 - \int_{-\infty}^{1}\pi(\theta|x=2)\mathrm{d}\theta \\
&= 1 - \varpi_1^*\int_{-\infty}^{1}\pi_1(\theta|x=2)\mathrm{d}\theta - \varpi_2^*\int_{-\infty}^{1}\pi_2(\theta|x=2)\mathrm{d}\theta \\
&= 1 - \varpi_1^*\int_{-\infty}^{1}\mathrm{N}(\theta|1,\frac{1}{2})\mathrm{d}\theta - \varpi_2^*\int_{-\infty}^{1}\mathrm{N}(\theta|\frac{3}{2},\frac{1}{2})\mathrm{d}\theta \\
&= 1 - \varpi_1^*\Phi(\frac{0}{\sqrt{1/2}}) - \varpi_2^*\Phi(-\sqrt{1/2}) \\
&= 1 - \varpi_1^*\Phi(0) - \varpi_2^*\Phi(-\sqrt{1/2}) \\
&\approx 0.63
\end{aligned}
$$

The Exercise below has been set as Homework 2.

Exercise 43. ($\star\star$)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.

Solution.

1. The likelihood is

$$f(x|\lambda) = \prod_{i=1}^{n} \text{Pn}(x_i|\lambda) \ = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \lambda^{\sum_{i=1}^n x_i} \exp(-n\lambda) \tag{33}$$

2. The k parameter exponential family of distributions has the form

$$\text{Ef}_k(x|u, g, h, \phi, \theta, c) = u(x)g(\theta)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(x)); \qquad x \in \mathcal{X}$$

and if sampling space $\mathcal{X}$ does not depend on θ it is also called regular. So I just need to bring the sampling density distribution in this form. It is

$$\text{Pn}(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)1_{\mathbb{N}}(x) \ = \frac{1}{x!}\exp(-\lambda)\exp(x\log(\lambda))1_{\mathbb{N}}(x)$$

So $\text{Pn}(\lambda)$ is member of the regular 1-parameter exponential family with

$$u(x) = \frac{1}{x!}1_{\mathbb{N}-\{0\}}(x), \qquad g(\lambda) = \exp(-\lambda), \qquad h_1(x) = x, \qquad \phi_1(\lambda) = \log(\lambda), \qquad c_1 = 1.$$

The sampling space $\mathcal{X}$ does not depend on the uncertain parameter λ and hence it is a regular exponential family of distributions.

3. There are two ways to derive the conjugate prior. I will present both.

Way-1 (Theorem 20 from the Handout)

The sampling distribution is member of the 1- regular exponential distribution family, as the density of the sampling density distribution $\text{Pn}(x|\lambda)$ can be written in the form

$$\text{Pn}(x|\lambda) = u(x)g(\lambda)\exp(\sum_{j=1}^{k} c_j\phi_j(\lambda)h_j(x)); \qquad x \in \mathcal{X}$$

with

$$u(x) = \frac{1}{x!}1_{\mathbb{N}-\{0\}}(x), \qquad g(\lambda) = \exp(-\lambda), \qquad h_1(x) = x, \qquad \phi_1(\lambda) = \log(\lambda), \qquad c_1 = 1.$$

Since the sampling space $\mathcal{X}$ of the sampling distribution does not depend on the unknown parameter λ, (Theorem 20 from the Handout) the conjugate prior is

$$\begin{aligned} \pi(\lambda) &\propto g(\lambda)^{\tau_0}\exp(c_1\tau_1\phi_1(\lambda)) \\ &= \exp(-\lambda\tau_0)\exp(\tau_1\log(\lambda)) \\ &= \lambda^{\tau_1}\exp(-\lambda\tau_0) \\ &\propto \text{Ga}(\lambda|a,b), \text{ for } a = \tau_1 + 1,\ b = \tau_0 \end{aligned} \tag{34}$$

So the conjugate prior is $\lambda \sim \text{Ga}(\lambda|a,b)$ with $a > 0$ and $b > 0$.

Way-2 (Theorem 12 in the Handout)

The likelihood can be written as

$$f(x|\lambda) = \prod_{i=1}^{n}\text{Pn}(x_i|\lambda) \ = \underbrace{\lambda^{\sum_{i=1}^{n}x_i}\exp(-n\lambda)}_{=k(t(x)|\lambda)}\underbrace{\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)}_{=\rho(x)} \tag{35}$$

where a kernel of the likelihood is $k(t(x)|\lambda) = \lambda^{\sum_{i=1}^{n}x_i}\exp(-n\lambda)$, with sufficient statistics $t(x) = (n, \sum_{i=1}^{n}x_i)$, and $\rho(x) = \left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)$ is the residual term of it. The dimensionality of the sufficient statistic $t(x)$ does not depend on the sample size n, and the observables are iid. Hence, (Theorem 12 in the Handout) the conjugate prior results as the aforesaid likelihood kernel from (39) where the sufficient statistics are replaced by a priori hyper-parameters $\tau = (\tau_0, \tau_1)$, such as

$$\pi(\lambda) \propto k(\tau|\lambda) = \lambda^{\tau_1}\exp(-\tau_0\lambda) \propto \text{Ga}(\lambda|a,b), \text{ for } a = \tau_1 + 1,\ b = \tau_0 \tag{36}$$

where I recognize the kernel of the Gamma distribution. So the conjugate prior is $\lambda \sim \text{Ga}(a,b)$ with $a > 0$ and $b > 0$.

In (38) and (40) , as strength of the prior information can be considered the parameter τ_0 (and hence b) because it substitutes the sample size n in the likelihood (37). In (38) and (40) , as prior information summary can be considered the parameter τ_1 (and hence a) because it substitutes the summary $\sum_{i=1}^{n}x_i$ in the likelihood (37).

4. According to the definition, the posterior PDF can be computed via the Bayes theorem

$$\begin{aligned}\pi(\lambda|x) \propto f(x|\lambda)\pi(\lambda) \quad &\propto \prod_{i=1}^{n} \text{Pn}(x_i|\lambda)\text{Ga}(\lambda|a,b)\\ &\propto \left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)\lambda^{\sum_{i=1}^n x_i}\exp(-n\lambda)\times\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-\lambda b)\\ &\propto \lambda^{\sum_{i=1}^n x_i+a-1}\exp(-\lambda(n+b))\\ &\propto \text{Ga}(\lambda|\sum_{i=1}^{n}x_i+a, n+b)\end{aligned}$$

So the posterior distribution is $\lambda|x \sim \text{Ga}(\tilde{a},\tilde{b})$, $\tilde{a}=\sum_{i=1}^n x_i+a$, $\tilde{b}=n+b$.

- Alternatively, we could use the Theorem in the Lecture notes stating the properties of the Conjugate priors... I.e. $\lambda|x \sim \text{Ga}(\sum_{i=1}^n x_i+(\tau_1+1), n+(\tau_0))$ –It is up to you...

5. According to the definition, the predictive PMF is

$$\begin{aligned}g(y|x) &= \int_{(0,\infty)} f(y|\lambda)\pi(\lambda|x)\text{d}\lambda \quad = \int_{(0,\infty)}\text{Pn}(y|\lambda)\text{Ga}(\lambda|\tilde{a},\tilde{b})\text{d}\lambda\\ &= \int_{(0,\infty)}\frac{1}{y!}\lambda^y\exp(-\lambda)1_{\mathbb{N}-\{0\}}(y)\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\lambda^{\tilde{a}-1}\exp(-\lambda\tilde{b})\text{d}\lambda\\ &= \frac{1}{y!}\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}1_{\mathbb{N}-\{0\}}(y)\int_{(0,\infty)}\lambda^{y+\tilde{a}-1}\exp(-\lambda(\tilde{b}+1))\text{d}\lambda\\ &= \frac{1}{y!}\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\frac{\Gamma(y+\tilde{a})}{(\tilde{b}+1)^{y+\tilde{a}}}1_{\mathbb{N}-\{0\}}(y) \quad = \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{\Gamma(y+\tilde{a})}{\Gamma(\tilde{a})}1_{\mathbb{N}-\{0\}}(y)\\ &= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{(y+\tilde{a}-1)(y+\tilde{a}-2)\cdots(\tilde{a})\cancel{\Gamma(\tilde{a})}}{\cancel{\Gamma(\tilde{a})}}1_{\mathbb{N}-\{0\}}(y)\\ &= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y(y+\tilde{a}-1)(y+\tilde{a}-2)\cdots(\tilde{a})1_{\mathbb{N}-\{0\}}(y)\\ &= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{(y+\tilde{a}-1)!}{(\tilde{a}-1)!}1_{\mathbb{N}-\{0\}}(y) \quad = \frac{(y+\tilde{a}-1)!}{(\tilde{a}-1)!y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y1_{\mathbb{N}-\{0\}}(y)\\ &= \binom{y+\tilde{a}-1}{\tilde{a}-1}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(1-\frac{\tilde{b}}{\tilde{b}+1})^y1_{\mathbb{N}-\{0\}}(y) \quad = \text{Nb}(y|\tilde{a},\frac{\tilde{b}}{\tilde{b}+1})\end{aligned}$$

where $\tilde{a}=\sum_{i=1}^n x_i+a$, $\tilde{b}=n+b$.

The Exercise below has been set as Homework 2.

Exercise 44. ($\star\star$)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.

Solution.

1. The likelihood is

$$f(x|\lambda) = \prod_{i=1}^{n} \text{Pn}(x_i|\lambda) \ = \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \lambda^{\sum_{i=1}^n x_i} \exp(-n\lambda) \tag{37}$$

2. The k parameter exponential family of distributions has the form

$$\text{Ef}_k(x|u, g, h, \phi, \theta, c) = u(x)g(\theta)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(x)); \qquad x \in \mathcal{X}$$

and if sampling space $\mathcal{X}$ does not depend on θ it is also called regular. So I just need to bring the sampling density distribution in this form. It is

$$\text{Pn}(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)1_{\mathbb{N}}(x) \ = \frac{1}{x!}\exp(-\lambda)\exp(x\log(\lambda))1_{\mathbb{N}}(x)$$

So $\text{Pn}(\lambda)$ is member of the regular 1-parameter exponential family with

$$u(x) = \frac{1}{x!}1_{\mathbb{N}-\{0\}}(x), \qquad g(\lambda) = \exp(-\lambda), \qquad h_1(x) = x, \qquad \phi_1(\lambda) = \log(\lambda), \qquad c_1 = 1.$$

The sampling space $\mathcal{X}$ does not depend on the uncertain parameter λ and hence it is a regular exponential family of distributions.

3. There are two ways to derive the conjugate prior. I will present both.

Way-1 (Theorem 20 from the Handout)

The sampling distribution is member of the 1- regular exponential distribution family, as the density of the sampling density distribution $\text{Pn}(x|\lambda)$ can be written in the form

$$\text{Pn}(x|\lambda) = u(x)g(\lambda)\exp(\sum_{j=1}^{k} c_j\phi_j(\lambda)h_j(x)); \qquad x \in \mathcal{X}$$

with

$$u(x) = \frac{1}{x!}1_{\mathbb{N}-\{0\}}(x), \qquad g(\lambda) = \exp(-\lambda), \qquad h_1(x) = x, \qquad \phi_1(\lambda) = \log(\lambda), \qquad c_1 = 1.$$

Since the sampling space $\mathcal{X}$ of the sampling distribution does not depend on the unknown parameter λ, (Theorem 20 from the Handout) the conjugate prior is

$$\begin{aligned}\pi(\lambda) &\propto g(\lambda)^{\tau_0}\exp(c_1\tau_1\phi_1(\lambda)) \\ &= \exp(-\lambda\tau_0)\exp(\tau_1\log(\lambda)) \\ &= \lambda^{\tau_1}\exp(-\lambda\tau_0) \\ &\propto \text{Ga}(\lambda|a,b),\ \text{for } a=\tau_1+1,\ b=\tau_0\end{aligned} \tag{38}$$

So the conjugate prior is $\lambda \sim \text{Ga}(\lambda|a,b)$ with $a>0$ and $b>0$.

Way-2 (Theorem 12 in the Handout)

The likelihood can be written as

$$f(x|\lambda)=\prod_{i=1}^{n}\text{Pn}(x_i|\lambda) \ = \underbrace{\lambda^{\sum_{i=1}^n x_i}\exp(-n\lambda)}_{=k(t(x)|\lambda)}\underbrace{\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)}_{=\rho(x)} \tag{39}$$

where a kernel of the likelihood is $k(t(x)|\lambda)=\lambda^{\sum_{i=1}^n x_i}\exp(-n\lambda)$, with sufficient statistics $t(x)=(n,\sum_{i=1}^n x_i)$, and $\rho(x)=\left(\prod_{i=1}^n \frac{1}{x_i!}\right)$ is the residual term of it. The dimensionality of the sufficient statistic $t(x)$ does not depend on the sample size n, and the observables are iid. Hence, (Theorem 12 in the Handout) the conjugate prior results as the aforesaid likelihood kernel from (39) where the sufficient statistics are replaced by a priori hyper-parameters $\tau=(\tau_0,\tau_1)$, such as

$$\pi(\lambda)\propto k(\tau|\lambda)=\lambda^{\tau_1}\exp(-\tau_0\lambda)\propto \text{Ga}(\lambda|a,b),\ \text{for } a=\tau_1+1,\ b=\tau_0 \tag{40}$$

where I recognize the kernel of the Gamma distribution. So the conjugate prior is $\lambda\sim\text{Ga}(a,b)$ with $a>0$ and $b>0$.

In (38) and (40) , as strength of the prior information can be considered the parameter τ_0 (and hence b) because it substitutes the sample size n in the likelihood (37). In (38) and (40) , as prior information summary can be considered the parameter τ_1 (and hence a) because it substitutes the summary $\sum_{i=1}^n x_i$ in the likelihood (37).

4. According to the definition, the posterior PDF can be computed via the Bayes theorem

$$\begin{aligned}\pi(\lambda|x)\propto f(x|\lambda)\pi(\lambda) \quad &\propto \prod_{i=1}^{n}\text{Pn}(x_i|\lambda)\text{Ga}(\lambda|a,b) \\ &\propto \left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)\lambda^{\sum_{i=1}^n x_i}\exp(-n\lambda)\times\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-\lambda b) \\ &\propto \lambda^{\sum_{i=1}^n x_i+a-1}\exp(-\lambda(n+b)) \\ &\propto \text{Ga}(\lambda|\sum_{i=1}^{n}x_i+a, n+b)\end{aligned}$$

So the posterior distribution is $\lambda|x\sim\text{Ga}(\tilde{a},\tilde{b})$, $\tilde{a}=\sum_{i=1}^n x_i+a$, $\tilde{b}=n+b$.

- Alternatively, we could use the Theorem in the Lecture notes stating the properties of the Conjugate priors... I.e. $\lambda|x\sim\text{Ga}(\sum_{i=1}^n x_i+(\tau_1+1), n+(\tau_0))$ –It is up to you...

5. According to the definition, the predictive PMF is

$$
\begin{aligned}
g(y|x) &= \int_{(0,\infty)} f(y|\lambda)\pi(\lambda|x)\mathrm{d}\lambda \quad = \int_{(0,\infty)} \mathrm{Pn}(y|\lambda)\mathrm{Ga}(\lambda|\tilde{a},\tilde{b})\mathrm{d}\lambda \\
&= \int_{(0,\infty)} \frac{1}{y!}\lambda^y \exp(-\lambda)1_{\mathbb{N}-\{0\}}(y)\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\lambda^{\tilde{a}-1}\exp(-\lambda\tilde{b})\mathrm{d}\lambda \\
&= \frac{1}{y!}\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}1_{\mathbb{N}-\{0\}}(y)\int_{(0,\infty)}\lambda^{y+\tilde{a}-1}\exp(-\lambda(\tilde{b}+1))\mathrm{d}\lambda \\
&= \frac{1}{y!}\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\frac{\Gamma(y+\tilde{a})}{(\tilde{b}+1)^{y+\tilde{a}}}1_{\mathbb{N}-\{0\}}(y) \quad = \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{\Gamma(y+\tilde{a})}{\Gamma(\tilde{a})}1_{\mathbb{N}-\{0\}}(y) \\
&= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{(y+\tilde{a}-1)(y+\tilde{a}-2)\cdots(\tilde{a})\cancel{\Gamma(\tilde{a})}}{\cancel{\Gamma(\tilde{a})}}1_{\mathbb{N}-\{0\}}(y) \\
&= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y(y+\tilde{a}-1)(y+\tilde{a}-2)\cdots(\tilde{a})1_{\mathbb{N}-\{0\}}(y) \\
&= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{(y+\tilde{a}-1)!}{(\tilde{a}-1)!}1_{\mathbb{N}-\{0\}}(y) \quad = \frac{(y+\tilde{a}-1)!}{(\tilde{a}-1)!y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y1_{\mathbb{N}-\{0\}}(y) \\
&= \binom{y+\tilde{a}-1}{\tilde{a}-1}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(1-\frac{\tilde{b}}{\tilde{b}+1})^y1_{\mathbb{N}-\{0\}}(y) \quad = \mathrm{Nb}(y|\tilde{a},\frac{\tilde{b}}{\tilde{b}+1})
\end{aligned}
$$

where $\tilde{a}=\sum_{i=1}^n x_i + a$, $\tilde{b} = n + b$.

Exercise 45. ($\star\star$)Consider the Bayesian model

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

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

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

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

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

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

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.

Solution.

1. There are two alternative ways to derive the conjugate prior here.

(a) [Way (a)] I can factorize the likelihood in a form that the likelihood kernel is a function of a sufficient statistic whose dimension is independent on the sample size n, and then derive the conjugate by substituting the sufficient statistic elements by prior hyper-parameters.
There are $k-1$ independent parameters in $\text{Mu}_k(\theta)$ because $\sum_{j=1}^k \theta_j = 1$. I consider as parameters $(\theta_1, ..., \theta_{k-1})$ and the last one is a function of them as $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$.
The likelihood is

$$f(x_{1:n}|\theta) = \prod_{i=1}^n \text{Mu}_k(x_i|\theta) = \prod_{i=1}^n \left[\prod_{j=1}^k \theta_j^{x_{i,j}}\right] = \prod_{j=1}^k \theta_j^{\sum_{i=1}^n x_{i,j}} = \prod_{j=1}^k \theta_j^{x_{*,j}} = \prod_{j=1}^{k-1} \theta_j^{x_{*,j}} \theta_k^{n-x_{*,k}}$$

where $x_{*,j} = \sum_{i=1}^n x_{i,j}$. So

$$f(x_{1:n}|\theta) = \prod_{j=1}^{k-1} \theta_j^{x_{*,j}} \left(1 - \sum_{j=1}^{k-1} \theta_j\right)^{n-x_{*,k}} = (1 - \sum_{j=1}^{k-1} \theta_j)^n \exp\left(\sum_{j=1}^{k-1} x_{*,j} \log(\frac{\theta_j}{1 - \sum_{j=1}^{k-1} \theta_j})\right)$$

and the sufficient statistic is

$$t_n = (n, x_{*,1}, ..., x_{*,k-1})$$

(b) [Way (b)] Alternatively, we can observe that the sampling space $\mathcal{X}_k$ does not depend on the parameters. So we can show that the sampling distribution is an exponential family of distributions, identify its components, and then derive the conjugate prior.
There are $k-1$ independent parameters in $\text{Mu}_k(\theta)$ because $\sum_{j=1}^k \theta_j = 1$. I consider as parameters $(\theta_1, ..., \theta_{k-1})$ and the last one is a function of them as $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$.
It is

$$\text{Mu}_k(x|\theta) = \prod_{j=1}^k \theta_j^{x_j} = \prod_{j=1}^{k-1} \theta_j^{x_j} (1 - \sum_{j=1}^{k-1} \theta_j)^{1-\sum_{j=1}^{k-1} x_j} = (1 - \sum_{j=1}^{k-1} \theta_j) \exp(\sum_{j=1}^{k-1} x_j \log(\frac{\theta_j}{1 - \sum_{j=1}^{k-1} \theta_j}))$$

This is the $k-1$ exponential family PDF with

$$u(x) = 1; \qquad g(\theta) = (1 - \sum_{j=1}^{k-1} \theta_j); \qquad c = (1, ..., 1)$$

$$h(x) = (x_1, ...x_{k-1}); \qquad \phi(\theta) = (\log(\frac{\theta_1}{1 - \sum_{j=1}^{k-1} \theta_j}), ..., \log(\frac{\theta_{k-1}}{1 - \sum_{j=1}^{k-1} \theta_j})),$$

Then either by substituting the sufficient statistics in way (a), or by using the components of the exponential family of distributions in way (b) Let $\tau = (\tau_0, ..., \tau_{k-1})$. It is

$$\pi(\theta|\tau) \propto (1 - \sum_{j=1}^{k-1} \theta_j)^{\tau_0} \exp(\sum_{j=1}^{k-1} \tau_j \log(\frac{\theta_j}{1 - \sum_{j=1}^{k-1} \theta_j}))$$
$$\propto \prod_{j=1}^{k-1} \theta_j^{\tau_j} (1 - \sum_{j=1}^{k-1} \theta_j)^{\tau_0 - \sum_{j=1}^{k-1} \tau_j} \propto \prod_{j=1}^{k-1} \theta_j^{\tau_j} \theta_k^{\tau_0 - \sum_{j=1}^{k-1} \tau_j}$$

Here, I recognize the Dirichlet distribution with $a_j = \tau_j$ for $j = 1, ..., k-1$ and $a_k = \tau_0 - \sum_{j=1}^{k-1} \tau_j$.

2. Well, the posterior is Dirichlet too. It is

$$\pi(\theta|x_{1:n}) = \prod_{i=1}^n \text{Mu}_k(x_i|\theta) \text{Di}_k(\theta|a) \propto \prod_{j=1}^k \theta_j^{x_{*,j}} \prod_{j=1}^k \theta_j^{a_j - 1} = \prod_{j=1}^k \theta_j^{x_j + a_{*,j} - 1} \propto \text{Di}_k(\theta|\tilde{a})$$

where $\tilde{a} = (\tilde{a}_1, ..., \tilde{a}_k)$, with $\tilde{a}_j = a_j + x_{*,j}$ for $j = 1, ..., k$. So the posterior is $\theta|x_{1:n} \sim \text{Di}_k(\tilde{a})$.

Exercise 46. ($\star\star$)Consider the Bayesian model

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

where $\text{Ga}(a, \theta)$ is the Gamma distribution with expected value a/θ. Specify a Jeffrey's prior density for θ.

Solution.

It is

$$\begin{aligned} \text{Ga}(x|a, \theta) = \frac{\theta^a}{\Gamma(a)} x^{a-1} \exp(-\theta x) \implies \\ \log(\text{Ga}(x|a, \theta)) = a \log(\theta) - \log(\Gamma(a)) + (a-1)\log(x) - \theta x \implies \\ \frac{\mathrm{d}}{\mathrm{d}\theta} \log(\text{Ga}(x|a, \theta)) = \frac{a}{\theta} - x \implies \\ \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log(\text{Ga}(x|a, \theta)) = -\frac{a}{\theta^2} \implies \\ \underbrace{-\text{E}(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log(\text{Ga}(x|a, \theta)))}_{=I(\theta)} = \frac{a}{\theta^2} \implies \end{aligned}$$

so $\pi(\theta) \propto \sqrt{I(\theta)} \propto \frac{1}{\theta}$.

Exercise 47. ($\star\star$)Find the Jeffreys prior density for the parameter θ of the Maxwell distribution

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

Solution.

It is

$$\begin{aligned} f(x|\theta) = \sqrt{\frac{2}{\pi}} \theta^{3/2} x^2 \exp(-\frac{1}{2}\theta x^2) \implies \\ \log(f(x|\theta)) = \log(\sqrt{\frac{2}{\pi}} x^2) + \frac{3}{2}\log(\theta) - \frac{1}{2}\theta x^2 \implies \\ \frac{\mathrm{d}}{\mathrm{d}\theta} \log(f(x|\theta)) = \frac{3}{2\theta} - \frac{1}{2}x^2 \implies \\ \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log(f(x|\theta)) = -\frac{3}{2}\theta^2 \implies \\ \underbrace{-\text{E}(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log(f(x|\theta)))}_{=I(\theta)} = \frac{3}{2}\theta^2 \implies \end{aligned}$$

Hence, we take $\pi(\theta) \propto 1/\theta$.

Exercise 48. ($\star\star$)Consider the trinomial distribution

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

Specify a Jeffreys' prior for (π,ρ).

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

Solution.

It is

$$\partial^2 L/\partial\pi^2 = -x/\pi^2 - z/(1-\pi-\rho)^2$$
$$\partial^2 L/\partial p^2 = -y/p^2 - z/(1-\pi-\rho)^2$$
$$\partial^2 L/\partial\pi\partial p = -z/(1-\pi-\rho)^2$$

and

$$I(\pi,\rho\,|\,x,y,z) = -\mathrm{E}\begin{pmatrix} -x/\pi^2 - z/(1-\pi-\rho)^2 & -z/(1-\pi-\rho)^2 \\ -z/(1-\pi-\rho)^2 & -y/p^2 - z/(1-\pi-\rho)^2 \end{pmatrix}$$
$$= \begin{pmatrix} n/\pi + n/(1-\pi-\rho) & n/(1-\pi-\rho) \\ n/(1-\pi-\rho) & n/\rho + n/(1-\pi-\rho) \end{pmatrix}$$

Because $\mathrm{E}(x)=n\pi$, $\mathrm{E}(y)=np$, $\mathrm{E}(z)=n(1-\pi-\rho)$, and

$$\det I(\pi,\rho\,|\,x,y,z) = (n/\pi + n/(1-\pi-\rho))(n/\rho + n/(1-\pi-\rho)) - (n/(1-\pi-\rho))^2$$
$$= \ldots$$
$$= n\{\pi\rho(1-\pi-\rho)\}^{-1}$$

So the Jeffrey's prior is

$$p(\pi,\rho) \propto \pi^{-\frac{1}{2}}\rho^{-\frac{1}{2}}(1-\pi-\rho)^{-\frac{1}{2}}$$

Exercise 49. ($\star\star$)Suppose that x has a Pareto distribution $\mathrm{Pa}(\xi,\gamma)$ 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.

Solution.

It is $\partial p(x|\gamma)/\partial\gamma = 1/\gamma + \log\xi - \log x$ and $\partial^2 p(x|\gamma)/\partial\gamma^2 = -1/\gamma^2$. So $I(\gamma|x) = 1/\gamma^2$ Hence the Jeffrey's prior is $p(\gamma)\propto 1/\gamma$.

Exercise 50. ($\star\star$)Consider the Bayesian model

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

where $\mathrm{Ga}(a,\beta)$ 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,

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

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

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,
  - $F^{(0)}(\alpha) = \frac{\text{d}}{\text{d}\alpha}\log(\Gamma(a))$
  - $F^{(1)}(\alpha) = \frac{\text{d}^2}{\text{d}\alpha^2}\log(\Gamma(a))$
- You may leave your answer in terms of function $F^{(1)}(\alpha)$.

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

$$\frac{\text{d}}{\text{d}\alpha}\log(f(x|\alpha,\beta)) = \log(\beta) - \frac{\text{d}}{\text{d}\alpha}\log(\Gamma(\alpha)) + \log(x)$$
$$\frac{\text{d}}{\text{d}\alpha^2}\log(f(x|\alpha,\beta)) = -\frac{\text{d}^2}{\text{d}\alpha^2}\log(\Gamma(\alpha)) = -F^{(1)}(\alpha)$$
$$\frac{\text{d}}{\text{d}\beta}\log(f(x|\alpha,\beta)) = \frac{\alpha}{\beta} - x$$
$$\frac{\text{d}^2}{\text{d}\beta^2}\log(f(x|\alpha,\beta)) = -\frac{\alpha}{\beta^2}$$
$$\frac{\text{d}^2}{\text{d}\alpha\text{d}\beta}\log(f(x|\alpha,\beta)) = \frac{1}{\beta}$$

and

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

Hence

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

Therefore

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

Exercise 51. (⋆⋆)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.

Solution.

We will check the properness condition. It is

$$f(x) = \int_{\mathbb{R}_+} \mathrm{N}(x|0,\sigma^2)\mathrm{Ex}(\sigma|\lambda)\mathrm{d}\sigma \;\propto\; \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x-0)^2)\sigma^{-1}\exp(-a\sigma^{-2})\mathrm{d}\sigma$$
$$= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(x^2+2a))\mathrm{d}\sigma$$
$$= \int_0^\infty \frac{1}{\sqrt{\xi}}\exp(-\frac{\xi}{2}(x^2+2a))\mathrm{d}\xi$$

for $\xi = 1/\sigma^2$. It is

$$f(x) \propto \int_0^\infty \underbrace{\frac{1}{\sqrt{\xi}}\exp(\underbrace{-\frac{\xi}{2}(x^2+2a)}_{<0})}_{\in(0,1)}\mathrm{d}\xi \leq \int_0^\infty \frac{1}{\sqrt{\xi}}\mathrm{d}\xi < \infty$$

So the posterior is defined.

Exercise 52. (⋆⋆⋆)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: $\frac{\mathrm{d}}{\mathrm{d}x} x^\top A x = 2Ax$, $\frac{\mathrm{d}}{\mathrm{d}x}(c + Ax) = A$, 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 $\mathrm{E}_{y|\beta,\sigma^2 \sim \mathrm{N}_n(\Phi\beta, I\sigma^2)}\left((y - \Phi\beta)^\top (y - \Phi\beta)\right) = n\sigma^2$.

Solution. Let's set $\xi = \sigma^2$ to simplify notation... The log likelihood is

$$\log\left(f(y|\beta, \xi)\right) = -\frac{n}{2}\log(\xi) - \frac{1}{2}\frac{1}{\xi}(y - \Phi\beta)^\top (y - \Phi\beta)$$

Let's compute the derivatives

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\xi}\log\left(f(y|\beta, \xi)\right) =& -\frac{n}{2}\frac{1}{\xi} + \frac{1}{2}\frac{1}{\xi^2}(y - \Phi\beta)^\top (y - \Phi\beta) \\
\frac{\mathrm{d}^2}{\mathrm{d}\xi^2}\log\left(f(y|\beta, \xi)\right) =& \frac{\mathrm{d}}{\mathrm{d}\xi}\left(-\frac{n}{2}\frac{1}{\xi} + \frac{1}{2}\frac{1}{\xi^2}(y - \Phi\beta)^\top (y - \Phi\beta)\right) = \frac{n}{2}\frac{1}{\xi^2} - \frac{1}{\xi^3}(y - \Phi\beta)^\top (y - \Phi\beta) \\
\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta, \xi)\right) =& -\frac{1}{\xi}\Phi^\top (y - \Phi\beta) \\
\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log\left(f(y|\beta, \xi)\right) =& \frac{\mathrm{d}}{\mathrm{d}\beta}\left(-\frac{1}{\xi}\Phi^\top (y - \Phi\beta)\right) = -\frac{1}{\xi}\Phi^\top\Phi \\
\frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta, \xi)\right) =& \frac{\mathrm{d}}{\mathrm{d}\xi}\left(-\frac{1}{\xi}\Phi^\top (y - \Phi\beta)\right) = \frac{1}{\xi^2}\Phi^\top (y - \Phi\beta)
\end{aligned}$$

Lets compute the components of the Fisher information. The sampling distribution of y is $y|\beta, \xi \sim \mathrm{N}(\Phi\beta, I\sigma^2)$ with $\mathrm{E}(y|\beta, \xi) = \Phi\beta$ and $\mathrm{Var}(y|\beta, \xi) = I\sigma^2$. So

$$\begin{aligned}
\mathrm{E}_{\mathrm{N}(\Phi\beta, I\sigma^2)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\xi^2}\log\left(f(y|\beta, \xi)\right)\right) =& \mathrm{E}_{\mathrm{N}(\Phi\beta, I\sigma^2)}\left(\frac{n}{2}\frac{1}{\xi^2} - \frac{1}{\xi^3}(y - \Phi\beta)^\top (y - \Phi\beta)\right) = +\frac{n}{2}\frac{1}{\xi^2} - \frac{1}{\xi^3}n\xi = -\frac{n}{2}\frac{1}{\xi^2} \\
\mathrm{E}_{\mathrm{N}(\Phi\beta, I\sigma^2)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log\left(f(y|\beta, \xi)\right)\right) =& -\frac{1}{\xi}\Phi^\top\Phi \\
\mathrm{E}_{\mathrm{N}(\Phi\beta, I\sigma^2)}\left(\frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta, \xi)\right)\right) =& \mathrm{E}_{\mathrm{N}(\Phi\beta, I\sigma^2)}\left(\frac{1}{\xi^2}\Phi^\top (y - \Phi\beta)\right) = \frac{1}{\xi^2}\Phi^\top\left(\mathrm{E}_{\mathrm{N}(\Phi\beta, I\sigma^2)}(y) - \Phi\beta\right) = 0
\end{aligned}$$

Then

$$\begin{aligned}
\mathscr{I} =& -\mathrm{E}_{\mathrm{N}(\Phi\beta, I\sigma^2)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(f(y|\theta))\right) = -\mathrm{E}_{\mathrm{N}(\Phi\beta, I\sigma^2)}\left(\begin{bmatrix} \frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\log\left(f(y|\beta, \xi)\right) & \frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta, \xi)\right) \\ \frac{\mathrm{d}}{\mathrm{d}\xi}\frac{\mathrm{d}}{\mathrm{d}\beta}\log\left(f(y|\beta, \xi)\right) & \frac{\mathrm{d}^2}{\mathrm{d}\xi^2}\log\left(f(y|\beta, \xi)\right) \end{bmatrix}\right) \\
=& \begin{bmatrix} \frac{1}{\xi}\Phi^\top\Phi & 0 \\ 0 & \frac{n}{2}\frac{1}{\xi^2} \end{bmatrix}
\end{aligned}$$

So the Jeffreys prior for (β, ξ) has density such as

$$\pi^{(\mathrm{JP})}(\beta, \xi) \propto \sqrt{\det\left(\mathscr{I}\right)} = \sqrt{\det\left(\frac{1}{\xi}\Phi^\top\Phi\right)\det\left(\frac{n}{2}\frac{1}{\xi^2}\right)} = \left(\frac{1}{\xi}\right)^{\frac{d}{2}+1}\sqrt{\det\left(\Phi^\top\Phi\right)\det\left(\frac{n}{2}\right)} \propto \left(\frac{1}{\xi}\right)^{\frac{d}{2}+1}$$

This is given as an Homework 3

Exercise 53. ($\star\star$)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 $\text{Pn}(\theta)$ is the Poisson distribution with expected value θ. Specify a Jeffreys' prior for θ.

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

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

Solution.

It is

$$\begin{aligned}
\text{Pn}(x|\theta) &= \exp(-\theta)\frac{\theta^x}{x!} \implies \\
\log(\text{Pn}(x|\theta)) &= -\theta + x\log(\theta) - \log(x!) \implies \\
\frac{\text{d}}{\text{d}\theta}\log(\text{Pn}(x|\theta)) &= -1 + x\frac{1}{\theta} \implies \\
\frac{\text{d}^2}{\text{d}\theta^2}\log(\text{Pn}(x|\theta)) &= -x\frac{1}{\theta^2} \implies \\
\underbrace{-\text{E}(\frac{\text{d}^2}{\text{d}\theta^2}\log(\text{Pn}(x|\theta)))}_{=I(\theta)} &= \frac{1}{\theta} \implies
\end{aligned}$$

so $\pi(\theta) \propto \sqrt{I(\theta)} \propto \frac{1}{\theta^{1/2}}$.

This is given as an Homework 3

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

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

where $\text{Pn}(\theta)$ is the Poisson distribution with expected value θ. Specify a Maximum entropy prior under the constrain $\text{E}(\theta) = 2$ 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: $x \sim \text{Pn}(\theta)$ has PMF

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

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

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

Solution.

It is $\int_{\mathbb{R}_+} \pi_0(\theta)\text{d}\theta = +\infty$. The maximum entropy prior is such that

$$\pi(\theta|\lambda) \propto \frac{1}{\sqrt{\theta}} \exp(\lambda\theta) = \theta^{-\frac{1}{2}} \exp(\lambda\theta)$$

where I can recognize $\theta|\lambda \sim \text{Ga}(\frac{1}{2}, -\lambda)$.

The expected value of the Gamma distribution $\text{Ga}(a, b)$ is

$$E_{\text{Ga}(a,b)}(x) = \int x \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-\beta x) 1(x > 0) = \frac{a}{b} \int \frac{b^{a+1}}{\Gamma(a+1)} x^{a+1-1} \exp(-\beta x) 1(x > 0) = \frac{a}{b}$$

Because $2 = \text{E}(\theta|\lambda) = -\frac{1}{2\lambda}$, it is $\lambda = -\frac{1}{4}$. Hence,

$$\theta|\lambda \sim \text{Ga}(\frac{1}{2}, \frac{1}{4})$$

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

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

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

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

- 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 55. ($\star\star$)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 $x = 0$.

Solution.

It is

$$f(x) \propto \int_{\mathbb{R}_+} \mathrm{N}(x|0, \sigma^2)\mathrm{Ex}(\sigma|\lambda)\mathrm{d}\sigma = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-0)^2)\lambda \exp(-\sigma\lambda)\mathrm{d}\sigma$$

$$f(x=0) \propto \int_0^\infty \frac{1}{\sigma} \exp(-\sigma\lambda)\mathrm{d}\sigma$$

We will use a convergence criteria in order to check if $\int_0^\infty \frac{1}{\sigma} \exp(-\sigma\lambda)\mathrm{d}\sigma = \infty$.

I will use the Limit Comparison Test to check if $\int_0^\infty \frac{1}{\sigma} \exp(-\sigma\lambda)\mathrm{d}\sigma = \infty$. Consider $h(\sigma) = \frac{1}{\sigma} \exp(-\sigma\lambda)$. The function $h(\sigma)$ has an improper behavior at 0, as it is not bounded there. Let $g(\sigma) = \frac{1}{\sigma}$. According to the Limit Comparison Test, it is

$$\lim_{\sigma \to 0^+} \frac{h(\sigma)}{g(\sigma)} = \lim_{\sigma \to 0^+} \frac{\frac{1}{\sigma} \exp(-\sigma\lambda)}{\frac{1}{\sigma}} = 1 \neq 0$$

and

$$\int_0^\infty g(\sigma)\mathrm{d}\sigma = \int_0^\infty \frac{1}{\sigma}\mathrm{d}\sigma = \infty.$$

Therefore, it will be

$$\underbrace{\int_0^\infty h(\sigma)\mathrm{d}\sigma}_{=f(x=0)} = \infty$$

as well.

This is given as an Homework 3

Exercise 56. ($\star\star$)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: $x \sim \text{Pn}(\theta)$ has PMF

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

Solution. The prior with $\pi(\theta) \propto \frac{1}{\theta}$ is improper because

$$\int \pi(\theta) \mathrm{d}\theta \propto \int \frac{1}{\theta} \mathrm{d}\theta = \infty$$

So I need to check the proneness condition,

$$\underbrace{\int_{\mathbb{R}_+} \text{Pn}(x|\theta)\pi(\theta)\mathrm{d}\theta}_{\propto f(x)} \quad \begin{cases} < \infty & \text{posterior distribution is defined} \\ = \infty & \text{posterior distribution is not defined} \end{cases}$$

I will show that the posterior distribution is not defined given that I have collected a single observation $x = 0$. So I need to show that

$$\underbrace{\int_{\mathbb{R}_+} \text{Pn}(x = 0|\theta)\pi(\theta)\mathrm{d}\theta}_{\propto f(x=0)} = \infty$$

It is

$$f(x) \propto \int_{\mathbb{R}_+} \text{Pn}(x|\theta) \frac{1}{\theta} \mathrm{d}\theta = \int_0^\infty \exp(-\theta) \frac{\theta^x}{x!} \frac{1}{\theta} \mathrm{d}\theta$$

$$f(x = 0) \propto \int_{\mathbb{R}_+} \exp(-\theta) \frac{\theta^0}{0!} \frac{1}{\theta} \mathrm{d}\theta = \int_0^\infty \exp(-\theta) \frac{1}{\theta} \mathrm{d}\theta$$

We will use a convergence criteria in order to check if $\int_0^\infty \exp(-\theta) \frac{1}{\theta} \mathrm{d}\theta = \infty$.

Consider $h(\theta) = \exp(-\theta) \frac{1}{\theta}$. The function $h(\theta)$ has an improper behavior at 0, as it is not bounded there. Let $g(\theta) = \frac{1}{\theta}$. According to the Limit Comparison Test, it is

$$\lim_{\theta \to 0^+} \frac{h(\theta)}{g(\theta)} = \lim_{\theta \to 0^+} \frac{\frac{1}{\theta} \exp(-\theta)}{\frac{1}{\theta}} = 1 \neq 0$$

and

$$\int_0^\infty g(\theta) \mathrm{d}\theta = \int_0^\infty \frac{1}{\theta} \mathrm{d}\theta = \infty.$$

Therefore, it will be

$$\underbrace{\int_0^\infty h(\theta) \mathrm{d}\theta}_{= f(x=0)} = \infty$$

as well.

Part VII

Decision theory

Exercise 57. ($\star\star\star$) 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.

Solution.

- δ_1 is not admissible if there exists an estimator δ_0 such that $R(\theta, \delta_1) \geq R(\theta, \delta_0)$. Let $\delta_0 = \frac{\delta_1 + \delta_2}{2}$. It is

$$\begin{aligned} R(\theta, \delta_0) &\leq \mathrm{E}_F(\ell(\theta, \delta_0)) = \mathrm{E}_F(\theta - \frac{\delta_1(x) + \delta_2(x)}{2})^2 \\ &= \frac{1}{4}\mathrm{E}_F((\theta - \delta_1(x)) + (\theta - \delta_2(x)))^2 \\ &= \frac{1}{4}(\mathrm{E}_F(\theta - \delta_1(x))^2 + \mathrm{E}_F(\theta - \delta_2(x))^2 + 2\underbrace{\mathrm{E}_F(\theta - \delta_1(x))(\theta - \delta_2(x)))}_{=0 \text{ (assuming independence \& unbias)}} \\ &= \frac{1}{4}(\mathrm{E}_F(\theta - \delta_1(x))^2 + \mathrm{E}_F(\theta - \delta_2(x))^2 \\ &= \frac{1}{4}(R(\theta, \delta_1) + R(\theta, \delta_2)) = \frac{1}{2}R(\theta, \delta_1) \end{aligned}$$

  thus $R(\theta, \delta_1) \geq R(\theta, \delta_0)$.

- Let us consider a strictly convex loss function $\ell(\theta, \delta)$ then

$$R(\theta, \delta_1) = \mathrm{E}_F(\ell(\theta, \delta_1)) \geq \ell(\theta, \mathrm{E}_F(\delta_1))$$

  because of Jensen's inequality. If $\delta' = \mathrm{E}_F(\delta_1)$ then $R(\theta, \delta_1) \geq R(\theta, \delta')$ and hence δ_1 is not admissible.

Part VIII

Point estimation

Exercise 58. ($\star\star$)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.

Solution.

- For the mean loss functions, the estimator is the posterior mean

$$\begin{aligned} \hat{\delta} &= \int \sigma^2 \text{IG}(\sigma^2|a + \frac{1}{2}n, b + \frac{n}{2}s^2)\text{d}\sigma^2 \\ &= \int \sigma^2 \frac{(b + \frac{n}{2}s^2)^{a+\frac{1}{2}n}}{\Gamma(a + \frac{1}{2}n)} (\sigma^2)^{-(a+\frac{1}{2}n)-1} \exp(-\frac{1}{\sigma^2}(b + \frac{n}{2}s^2))\text{d}\sigma^2 \\ &= \frac{(b + \frac{n}{2}s^2)^{a+\frac{1}{2}n}}{\Gamma(a + \frac{1}{2}n)} \int (\sigma^2)^{-(a+\frac{1}{2}n-1)-1} \exp(-\frac{1}{\sigma^2}(b + \frac{n}{2}s^2))\text{d}\sigma^2 \\ &= \frac{(b + \frac{n}{2}s^2)^{a+\frac{1}{2}n}}{\Gamma(a + \frac{1}{2}n)} \frac{\Gamma(a + \frac{1}{2}n - 1)}{(b + \frac{n}{2}s^2)^{a+\frac{1}{2}n-1}} \\ &= \frac{b + \frac{n}{2}s^2}{a + \frac{1}{2}n - 1} \end{aligned}$$

- For the zero-one loss function, the estimator is the posterior mode

$$\hat{\delta} = \arg\max_{\forall \sigma^2} (\log(\text{IG}(\sigma^2|a + \frac{1}{2}n, b + \frac{n}{2}s^2))).$$

So $\log(\pi(\sigma^2|x)) \propto -(a + \frac{1}{2}n - 1)\log(\sigma^2) - \frac{1}{\sigma^2}(b + \frac{n}{2}s^2)$.

$$\begin{aligned} 0 =& \frac{\text{d}}{\text{d}\sigma^2} \log(\pi(\sigma^2|x))|_{\sigma^2=\hat{\delta}} \\ \hat{\delta} =& \frac{b + \frac{n}{2}s^2}{a + \frac{1}{2}n + 1}. \end{aligned}$$

Exercise 59. ($\star\star$)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 $f(x) = b \exp(-bx) 1_{(0,+\infty)}(x)$,

Solution.

1. The posterior distribution of θ is $\text{Ga}(a+n, b+n\bar{x})$, where $\bar{x}$ is the sample mean, because

$$\begin{aligned}\pi(\theta|x) &\propto \prod_{i=1}^{n} \text{Ex}(x_i|\theta) \text{Ga}(\theta|a,b) \\ &\propto \theta^{(a+n)-1} \exp(-\theta(b+n\bar{x}))\end{aligned}$$

   For the squared loss functions, the estimator is

$$\begin{aligned}\delta(x) &= \int \theta \text{Ga}(\theta|a+n, b+n\bar{x}) \text{d}\theta \\ &= \int \theta \frac{(b+n\bar{x})^{a+n}}{\Gamma(a+n)} \theta^{(a+n)-1} \exp(-\theta(b+n\bar{x})) \text{d}\theta \\ &= \frac{(b+n\bar{x})^{a+n}}{\Gamma(a+n)} \int \theta^{(a+n+1)-1} \exp(-\theta(b+n\bar{x})) \text{d}\theta \\ &= \frac{(b+n\bar{x})^{a+n}}{\Gamma(a+n)} \frac{\Gamma(a+n+1)}{(b+n\bar{x})^{a+n+1}} \\ &= \frac{a+n}{b+n\bar{x}},\end{aligned}$$

   since $\Gamma(x+1) = x\Gamma(x)$.

   For the zero-one loss function, the estimator is $\delta(x) = \arg\max_{\forall \sigma^2} (\log(\text{Ga}(\theta|a+n, b+n\bar{x})))$. So $\log(\pi(\theta|x_{1:n})) \propto (a+n-1)\log(\theta) - \theta(b+n\bar{x})$.

$$\begin{aligned}0 &= \frac{\text{d}}{\text{d}\theta} \log(\pi(\theta|x))|_{\theta=\delta(x)} \\ \delta(x) &= \frac{a+n-1}{b+n\bar{x}}.\end{aligned}$$

2. For the squared loss functions, the estimator is the predictive expected value

$$\begin{aligned}\delta(x) &= \int y g(y|x) \text{d}y \\ &= \int y \frac{(b+n\bar{x})^{a+n}}{\Gamma(a+n)} \frac{\Gamma(a+n+1)}{\Gamma(1)} (b+n\bar{x}+y)^{-(a+n)-1} \text{d}y \\ &= \frac{(b+n\bar{x})^{a+n}}{\Gamma(a+n)} \frac{\Gamma(a+n+1)}{\Gamma(1)} \int y^{2-1} (b+n\bar{x}+y)^{-(a+n-1)-2} \text{d}y \\ &= \frac{(b+n\bar{x})^{a+n}}{\Gamma(a+n)} \frac{\Gamma(a+n+1)}{\Gamma(1)} \frac{\Gamma(a+n-1)}{(b+n\bar{x})^{a+n-1}} \frac{\Gamma(2)}{\Gamma(a+n-1+2)} \\ &= \frac{b+n\bar{x}}{a+n-1}.\end{aligned}$$

For the zero-one loss function, the estimator is the predictive mode. But, the predictive distribution does not have a mode because

$$\frac{\mathrm{d}}{\mathrm{d}y}\log(g(y|x)) < 0$$

for all $y \geq 0$.

Exercise 60. ($\star\star$)Consider observables $x = (x_1, ..., x_n)$. Consider the Bayesian model

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

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 61. (⋆⋆)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:

$$\ell(\hat{\vec{x}}, \vec{x}) = \sum_{i=1}^n (1 - \delta(\hat x_i, x_i)) \tag{41}$$

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

$$E\left(\ell(\hat{\vec{x}}, \vec{X})|y\right) = E\left(1 - \sum_{i=1}^{n} \delta(\hat{x}_i, X_i)|y\right) \tag{42}$$

$$= n - \sum_{i=1}^{n} E\left(\delta(\hat{x}_i, X_i)|y\right) = n - \sum_{i=1}^{n} \sum_{x_i \in \mathcal{X}_i} \delta(\hat{x}_i, x_i) p\left(x_i|y\right) = n - \sum_{i=1}^{n} p\left(\hat{x}_i|y\right) \tag{43}$$

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:

$$\hat{x}_i^* = \arg\max_{x_i \in \mathcal{X}_i} p\left(x_i|y\right) \tag{44}$$

[The x_i in (44) 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:

$$\hat{\vec{x}}^* = \arg\max_{\vec{x} \in \vec{\mathcal{X}}} p\left(\hat{x}_i|y\right) \tag{45}$$

2. The individual terms of the sum in (43) 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

$$\ell(\hat{\vec{x}}, \vec{x}) = 1 - \delta(\hat{\vec{x}}, \vec{x}) = 1 - \prod_{i=1}^{n} \delta(\hat{x}_i, x_i) \tag{46}$$

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

The following exercise is given as a Homework 4

Exercise 62. ($\star\star$) [3]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.

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

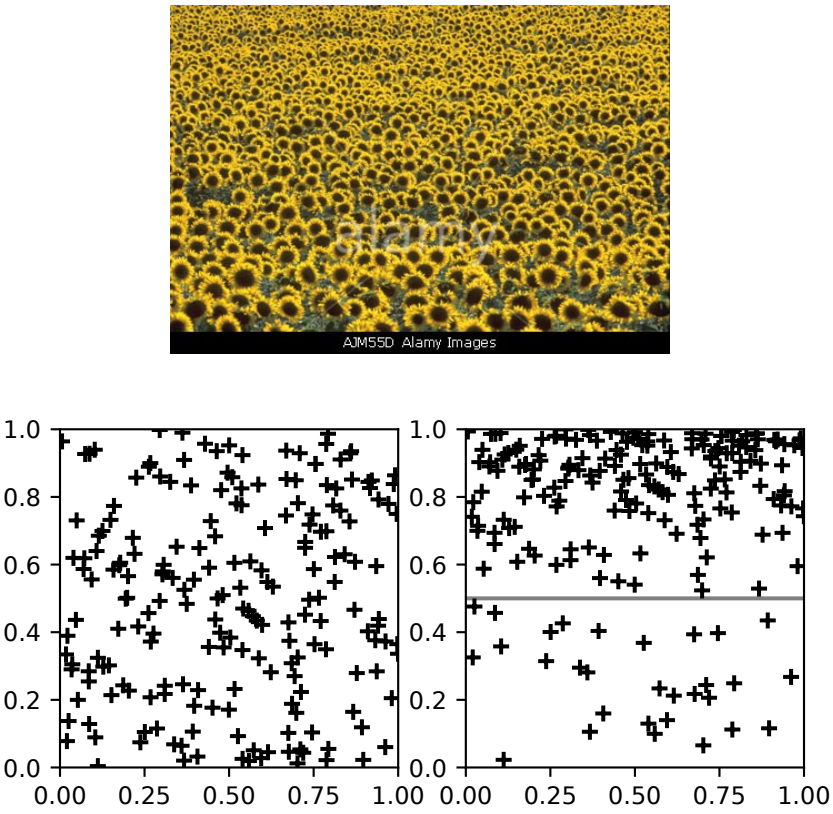

Figure 1: Two sampled point patterns. (Sunflower image © Soren Breiting / Alamy.)

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:

$$x = u \tag{47}$$

$$y = \frac{v(1+t)}{1+tv} \tag{48}$$

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.

$$f(u, v|t) = 1 \tag{49}$$

(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:

$$\mathrm{d}F(u, v) = \mathsf{P}_F(u \in \mathrm{d}u, v \in \mathrm{d}v) = \mathrm{d}u\mathrm{d}v \tag{50}$$

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

$$\mathrm{d}F(u_1, v_1, \ldots, u_n, v_n) = \prod_{i=1}^{n} \mathrm{d}F(u_i, v_i) = \prod_{i=1}^{n} \mathrm{d}u_i\mathrm{d}v_i \tag{51}$$

This does not depend on the data points (u_i, v_i) and so is the same for both patterns.[4]

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:

$$\mathsf{P}(r|n, \text{uniform}) = \binom{n}{r} \frac{1}{2^n} \tag{52}$$

(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.

[4] 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.**

(a) Note that

$$u(x,y) = x \tag{53}$$

$$v(x,y) = \frac{y}{1+t-ty} \tag{54}$$

which is an 1-1 transformation. Thus the Jacobian is [5]

$$J = \frac{\mathrm{d}(u,v)}{\mathrm{d}(x,y)} = \det \begin{bmatrix} 1 & 0 \\ 0 & \frac{(1+t-ty)1-y(-t)}{(1+t-ty)^2} \end{bmatrix} = \frac{1+t}{(1+t-ty)^2} \tag{55}$$

Since the pdf on (U,V) is uniform, we find that the pdf is

$$f(x,y|t) = f(u(x,y),v(x,y)|t) \left| \frac{\mathrm{d}(u,v)}{\mathrm{d}(x,y)} \right| = \frac{(1+t)}{(1+t-ty)^2} \tag{56}$$

and hence the distribution is such that

$$\mathrm{d}F(x,y|t) = \underbrace{\frac{1+t}{(1+t-ty)^2}}_{=f(x,y|t)} \mathrm{d}(x,y) \tag{57}$$

[5] Remember how to calculate the PDF or PMF when we perform transformation of random variables (in multivariate case). $\mathrm{d}(u(x,y),v(x,y))$ is an informal way of writing a transformed infinitesimal area, i.e. it denotes the area of the parallelogram covered by $(u(x,y),v(x,y))$ when both x and y are changed infinitesimally over a rectangle at (x,y) with dimensions $(\delta x, \delta y)$:

$$|J| = |\frac{\mathrm{d}(u(x,y),v(x,y))}{\mathrm{d}(x,y)}| = \left| \det \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix} \right|$$

(b) Since the points are sampled independently, the resulting pdf is simply the product of the individual pdfs:

$$f(x_1, y_1, \ldots, x_n, y_n|t) = \prod_{i=1}^{n} \frac{(1+t)}{(1+t-ty_i)^2} = \frac{(1+t)^n}{\prod_{i=1}^{n}(1+t-ty_i)^2} \tag{58}$$

(c) If we have no reason to favour any particular value of α, it would make sense to use a uniform distribution over α, i.e. the probability density with respect to α will be constant. Since $\alpha \in [0, \pi/2]$, we have probability distribution density

$$\pi(\alpha) = \frac{2}{\pi} \tag{59}$$

which is normalized; and hence probability such that

$$\mathrm{d}\Pi(\alpha) = \underbrace{\frac{2}{\pi}}_{=\pi(\alpha)} \mathrm{d}\alpha \tag{60}$$

(d) Since $\alpha(t) = \tan^{-1}(t)$, we have[6]

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t}(t) = \frac{1}{1+t^2} \tag{61}$$

so the pdf is

$$\pi(t) = \pi(\alpha(t)) \left| \frac{\mathrm{d}\alpha}{\mathrm{d}t}(t) \right| = \underbrace{\frac{2}{\pi}\frac{1}{1+t^2}}_{=\pi(t)} \tag{62}$$

and the distribution such that

$$\mathrm{d}\Pi(t) = \underbrace{\frac{2}{\pi}\frac{1}{1+t^2}}_{=\pi(t)} \mathrm{d}t \tag{63}$$

(e) The posterior density of t given the data points is

$$\pi(t|x_1, y_1, \ldots, x_n, y_n) \propto f(x_1, y_1, \ldots, x_n, y_n|t)\pi(t) \tag{64}$$

$$\propto \frac{(1+t)^n}{1+t^2} \prod_{i=1}^{n} \frac{1}{(1+t-ty_i)^2} \tag{65}$$

(f) The log posterior probability density is, up to an additive constant,

$$\log\left(\pi(t|x_1, y_1, \ldots, x_n, y_n)\right) = C + n\ln(1+t) - \ln(1+t^2) - 2\sum_{i=1}^{n} \ln(1+tz_i) \tag{66}$$

with $z_i = 1 - y_i$.

Differentiating with respect to t, and setting the result equal to zero gives the equation satisfied by the MAP estimate:

$$\frac{n}{1+t} - 2\sum_i \frac{z_i}{1+z_i t} = \frac{2t}{1+t^2} \tag{67}$$

(g) Consider that

$$g(t) = C + n\ln(1+t) - \ln(1+t^2) - 2\sum_{i=1}^{n} \ln(1+tz_i)$$

[6] Remember how we can calculate the PDF or the PMF when we perform a transformation of random variables

then

$$\frac{\mathrm{d}}{\mathrm{d}t}g(t) = \frac{n}{1+t} - \frac{2t}{1+t^2} - 2\sum_{i=1}^{n}\frac{z_i}{1+tz_i}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}g(t) = \frac{-n}{(1+t)^2} + \frac{2t^2-2}{(1+t^2)^2} + 2\sum_{i=1}^{n}\frac{z_i^2}{(1+tz_i)^2}$$

Then by using Taylor expansion of 2nd order around 0, it is

$$\tilde{g}(t) + g(0) + \frac{1}{1!}\frac{\mathrm{d}}{\mathrm{d}t}g(t)|_{t=0}(t-0) + \frac{1}{2!}\frac{\mathrm{d}^2}{\mathrm{d}t^2}g(t)|_{t=0}(t-0)^2 + \text{tiny stuff...}$$
$$\approx C + (n-2n\bar{z})t + (-n-2+2n\overline{z^2})t^2 + \text{tiny stuff...}$$

So, I will try to find the MAP estimate $\hat{t}^*$ by using this crude Taylor expansion approximation valid for values of t around zero 0, right... it is

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{g}(t)|_{t=\hat{t}^*} = 0$$
$$(n-2n\bar{z}) - (n+2-2n\overline{z^2})2t = 0$$
$$\hat{t}^* = \frac{\bar{z}-\frac{1}{2}}{\overline{z^2}-\frac{1}{2}-\frac{1}{n}}$$

Hence the MAP estimate is for small t approximately

$$\hat{t}^* = \frac{\bar{z}-\frac{1}{2}}{\overline{z^2}-\frac{1}{2}-\frac{1}{n}} \tag{68}$$

Some interpretation story ... Notice that the $\frac{1}{n}$ term comes from the prior pdf for t with respect to dt. The more data we have, the less significant it is. The rest of the formula measures the deviation of the mean point (note that $\bar{z} = 1-\bar{y}$) from the central $z = \frac{1}{2}$ point, which makes sense. The value of $\hat{t}^*$ will only be small, and hence the approximation will only be consistent, if $\bar{z}$ is closer to $\frac{1}{2}$ than $\overline{z^2}$.
In addition, for those who checked, this is only a maximum (as opposed to a minimum), within the scope of this approximation, when $\overline{z^2} < \frac{1}{2} + \frac{1}{n}$.

(h)

$$\pi(\alpha|x_1,y_1,\ldots,x_n,y_n) \propto f(x_1,y_1,\ldots,x_n,y_n|\alpha)\pi(\alpha) \tag{69}$$
$$\propto (1+t(\alpha))^n \prod_{i=1}^{n}\frac{1}{(1+t(\alpha)-t(\alpha)y_i)^2} \tag{70}$$

where $t(\alpha) = \tan(\alpha)$—this is essentially the same as before up to the Jacobian factor $1/(1+t^2)$.
To find the equation satisfied by the MAP estimate we can now take logarithms:

$$\log\left(\pi(\alpha|x_1,y_1,\ldots,x_n,y_n)\right) = C + n\ln(1+t(\alpha)) - 2\sum_{i=1}^{n}\ln(1+t(\alpha)z_i) \tag{71}$$

and differentiate with respect to α. By the chain rule, the derivative is zero when

$$\left(\frac{n}{1+t(\alpha)} - 2\sum_i \frac{z_i}{1+z_i t(\alpha)}\right)\frac{\mathrm{d}t}{d\alpha}(\alpha) = 0 \tag{72}$$

On first order expansion, now also using first order approximation $t(\alpha) = \alpha$,

$$n(1-\alpha) - 2\sum_i z_i(1 - z_i\alpha) = 0 \tag{73}$$

and thus the MAP estimate is

$$\hat{\alpha}^* = \frac{\bar{z} - \frac{1}{2}}{\overline{z^2} - \frac{1}{2}} \tag{74}$$

Note that $\hat{t}^* \neq \alpha^*$ even though $t = \alpha$ for small t (or equivalently, for small α). In general, the MAP estimate does not respect transformations.

Part IX

Credible regions

Exercise 63. ($\star\star\star\star$)A random sample, $\boldsymbol{x} = (x_1, x_2, ..., x_n)$, of size n is taken from $\mathrm{N}(\theta, \sigma^2)$, 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.

Solution. I more detailed proof can be found in [Lindley, D. (1965; Section 5.2). Introduction to Probability and Statistics from a Bayesian Viewpoint. Cambridge: Cambridge University Press.]

The likelihood (after inserting a convenient constant) is

$$f(\boldsymbol{x}|\theta) = (2\pi\phi/n)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\overline{x}-\theta)^2/(\sigma^2/n)\}.$$

Hence by Bayes' Theorem, within I_α

$$\begin{aligned} Ac(1-\varepsilon)(2\pi\sigma^2/n)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\overline{x}-\theta)^2/(\sigma^2/n)\} \leqslant \pi(\theta|\boldsymbol{x}) \\ \leqslant Ac(1+\varepsilon)(2\pi\sigma^2/n)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\overline{x}-\theta)^2/(\sigma^2/n)\} \end{aligned}$$

and outside I_α

$$0 \leqslant \pi(\theta|\boldsymbol{x}) \leqslant AMc(2\pi\sigma^2/n)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\overline{x}-\theta)^2/(\sigma^2/n)\},$$

where A is a constant equal to $p(\boldsymbol{x})^{-1}$. Using the right hand inequality for the region inside I_α we get

$$\begin{aligned}\int_{I_\alpha} \pi(\theta|\boldsymbol{x})\,\mathrm{d}\theta &\leqslant Ac(1+\varepsilon)\int_{I_\alpha}(2\pi\sigma^2/n)^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\overline{x}-\theta)^2/(\sigma^2/n)\}\,\mathrm{d}\theta \\ &= Ac(1+\varepsilon)\int_{-\lambda_\alpha}^{\lambda_\alpha}(2\pi)^{-\frac{1}{2}}\exp(-\tfrac{1}{2}t^2)\,\mathrm{d}t, \text{ where } t=(\overline{x}-\theta)/\sqrt{\sigma^2/n} \\ &= Ac(1+\varepsilon)[\Phi(\lambda_\alpha)-\Phi(-\lambda_\alpha)] = Ac(1+\varepsilon)(1-\alpha)\end{aligned}$$

since $\Phi(-\lambda_\alpha) = 1-\Phi(\lambda_\alpha)$. Similarly, the same integral exceeds $AQc(1-\varepsilon)(1-\alpha)$, and, if J_α is the outside of I_α,

$$0 \leqslant \int_{J_\alpha} \pi(\theta|\boldsymbol{x})\,\mathrm{d}\theta \leqslant AMc\alpha.$$

Combining these results we have, since $\int_{I_\alpha \cup J_\alpha} \pi(\theta|\boldsymbol{x})\,\mathrm{d}\theta = 1$,

$$Ac(1-\varepsilon)(1-\alpha) \leqslant 1 \leqslant Ac[(1+\varepsilon)(1-\alpha)+M\alpha],$$

and hence

$$\frac{1}{(1+\varepsilon)(1-\alpha)+M\alpha} \leqslant Ac \leqslant \frac{1}{(1-\varepsilon)(1-\alpha)}.$$

The result now follows on remarking that the maximum value of the exponential in J_α occurs at the end-points $\theta \pm \lambda_\alpha\sqrt{\sigma^2/n}$, where it has the value $\exp(-\frac{1}{2}\lambda_\alpha^2)$.

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

$$\begin{cases} x_i|\sigma^2 & \overset{\text{IID}}{\sim} \mathrm{N}(\mu,\sigma^2), \qquad i=1,...,n \\ \sigma^2 & \sim \mathrm{IG}(a,b) \end{cases} \tag{75}$$

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

1. Calculate the posterior distribution of σ^2 of Bayesian model (75). Justify your calculations with brief comments.

   Hint: It is

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

   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 (75), is Student T such as

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

   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 $Var(x) = \frac{a}{a-2} v$.

Solution. Denote the sequence of observables as $x_{1:n} = (x_1, ..., x_n)$.

1. By using Bayes theorem, it is

$$\begin{aligned}
\pi(\sigma^2|x_{1:n}) &\propto \prod_{i=1}^{n} \text{N}(x_i|\mu, \sigma^2) \text{IG}(\sigma^2|a, b) \;\propto\; (\frac{1}{\sigma^2})^{\frac{n}{2}} \exp(-\frac{1}{2} \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^2}) (\frac{1}{\sigma^2})^{a+1} \exp(-\frac{1}{\sigma^2} b) \\
&\propto\; (\frac{1}{\sigma^2})^{\frac{n}{2}+a+1} \exp(-\frac{1}{\sigma^2}(b + \frac{1}{2} n s^2)) \\
&\propto \text{IG}(\sigma^2|a + \frac{n}{2}, b + \frac{1}{2} n s^2)
\end{aligned}$$

   So it is

$$\sigma^2|x_{1:n} \sim \text{IG}(a^*, b^*)$$

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

2. By using the definition of the predictive distribution, it is

$$\begin{aligned}
p(y|x_{1:n}) &= \int_{\mathbb{R}_+} \text{N}(y|\mu, \sigma^2) \text{IG}(\sigma^2|a^*, b^*) \text{d}\sigma^2 \\
&\propto \int_{\mathbb{R}_+} (\frac{1}{\sigma^2})^{\frac{1}{2}} \exp(-\frac{1}{2} \frac{(y - \mu)^2}{\sigma^2}) (\frac{1}{\sigma^2})^{a^*+1} \exp(-\frac{1}{\sigma^2} b^*) \\
&\propto \int_{\mathbb{R}_+} (\frac{1}{\sigma^2})^{a^*+\frac{1}{2}+1} \exp(-\frac{1}{\sigma^2}(b^* + \frac{1}{2}(y - \mu)^2)) \\
&= \Gamma(a^* + \frac{1}{2}) \, (b^* + \frac{1}{2}(y - \mu)^2)^{-(a^*+\frac{1}{2})} \;\propto (b^* + \frac{1}{2}(y - \mu)^2)^{-(a^*+\frac{1}{2})} \\
&\propto (1 + \frac{1}{2a^*} \frac{a^*}{b^*}(y - \mu)^2)^{-\frac{2a^*+1}{2}} \;\propto \text{T}(y|\mu, \frac{b + \frac{1}{2} n s^2}{a + \frac{1}{2} n}, 2a + n)
\end{aligned}$$

   Here, once again, to compute the integral, we used the fact that the Gamma PDF integrates to 1.

3. We use the theorem from the notes to compute the credible interval C_γ . However, IG is not symmetric, and hence the theorem leads to the following system which we can solve with a computer:

$$\begin{aligned}
0 &= F_{\text{IG}(a^*,b^*)}(U) - F_{\text{IG}(a^*,b^*)}(L) - (1 - \gamma) \\
0 &= f_{\text{IG}(a^*,b^*)}(\sigma^2 = U) - f_{\text{IG}(a^*,b^*)}(\sigma^2 = L)
\end{aligned}$$

   with $a^* = \frac{n}{2} + a$ and $b^* = b + \frac{1}{2} n s^2$, and f^{IG} and F^{IG} denote the PDF and CDF.

- Please notice that this way to compute HPD interval makes sense only in the case that the posterior $\pi(\sigma^2|x_{1:n})$ has a mode. For example, in the case of the exponential distribution (Ex$(\xi|\lambda)$) whose PDF monotonically decreases, the produced Credible Interval would be of the form $(0, U]$ where $U = q^*_{,\lambda,1-\gamma}$ where $q^*_{,\lambda,1-\gamma}$ is the $1-\gamma$ quantile of the Ex$(\xi|\lambda)$ distribution. To better understand this, please draw the plot of IG PDF when the left tail goes to infinity, and apply the mnemonic rule with the horizontal line that we discussed int he classroom.

4. We use the theorem from the notes to compute the HPD credible interval. Moreover the predictive distribution is symmetric. Hence,

$$\int_L^U p(y|x_{1:n})\mathrm{d} = 1-\gamma$$
$$\mathsf{P}_{T\left(\mu,\frac{b+\frac{1}{2}ns^2}{a+\frac{1}{2}n},2a+n\right)}(y<U) = 1-\frac{\gamma}{2},$$
$$\mathsf{P}_{T(0,1,2a+n)}(\tilde{y} < \frac{U-\mu}{\sqrt{\frac{b+\frac{1}{2}ns^2}{a+\frac{1}{2}n}}}) = 1-\frac{\gamma}{2}.$$

Then $U = \mu + q^*\sqrt{\frac{b+\frac{1}{2}ns^2}{a+\frac{1}{2}n}}$ where $q^* = t_{2a+n,1-\frac{\gamma}{2}}$ is the quantile of the standard Student t distribution with $2a+n$ degrees of freedom such that $F_{\text{St}}(q^*|0,1,2a+n) = 1-\frac{\gamma}{2}$. By symmetry $L = \mu - q^*\sqrt{\frac{b+\frac{1}{2}ns^2}{a+\frac{1}{2}n}}$. So

$$C_\gamma = (\mu - t_{2a+n,1-\frac{\gamma}{2}}\sqrt{\frac{b+\frac{1}{2}ns^2}{a+\frac{1}{2}n}}, \mu + t_{2a+n,1-\frac{\gamma}{2}}\sqrt{\frac{b+\frac{1}{2}ns^2}{a+\frac{1}{2}n}})$$

5. It is $a^* = \frac{n}{2} + a = 30$, $b^* = b + \frac{1}{2}ns^2 = 45$, $\mu = 1$, and $q^* = t_{\text{df}=2a+n=60\,,\,1-\frac{\gamma}{2}=0.975} \approx 2$. By substituting the values, we get, $C_{\gamma=0.05} = [1-2\sqrt{3/2}\,,\,1+2\sqrt{3/2}] = [-1.45\,,\,3.45]$

Part X

Hypothesis tests

Exercise 65. ($\star\star$)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-2 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 $Var(\lambda) = 1$.

Solution.

1. It is

$$\begin{aligned}\text{Pn}(x|\lambda) &= \frac{1}{x!}\lambda^x \exp(-\lambda)1_{\mathbb{N}}(x)\\ &= \frac{1}{x!}\exp(-\lambda)\exp(x\log(\lambda))1_{\mathbb{N}}(x)\end{aligned}$$

   So $\text{Pn}(\lambda)$ is member of the regular 1-parameter exponential family with

$$u(x) = \frac{1}{x!}, \qquad g(\lambda) = \exp(-\lambda), \qquad h_1(x) = x, \qquad \phi_1(\lambda) = \log(\lambda), \qquad c_1 = 1.$$

2. The likelihood is

$$f(x_{1:n}|\lambda) = \prod_{i=1}^{n}\text{Pn}(x_i|\lambda) \;\; = \prod_{i=1}^{n}\frac{1}{x_i!}\lambda^{\sum_{i=1}^n x_i}\exp(-n\lambda)$$

3. The parametric model is member of the exponential distribution family. The likelihood can be written in form (via the exponential distribution family):

$$f(x_{1:n}|\lambda) = \prod_{i=1}^{n}\text{Pn}(x_i|\lambda) \;\; = \prod_{i=1}^{n}\frac{1}{x_i!}(\exp(-\lambda))^n\exp((\sum_{i=1}^{n}x_i)\log(\lambda))$$

   Based on the Theorem in the notes (as well as the informal definition), the PDF of the conjugate prior has the form

$$\begin{aligned}\pi(\lambda) &\propto g(\lambda)^{\tau_0}\exp(c_1\tau_1\phi_1(\lambda))\\ &= \exp(-\lambda\tau_0)\exp(\tau_1\log(\lambda))\\ &= \lambda^{\tau_1}\exp(-\lambda\tau_0)\\ &\propto \text{Ga}(\lambda|a,b), \text{ for } a = \tau_1 + 1,\ b = \tau_0\end{aligned}$$

   So the e conjugate prior is $\lambda \sim \text{Ga}(\lambda|a,b)$, where $a > 0$, and $b > 0$.

4. According to the definition, the posterior PDF can be computed via the Bayes theorem

$$\begin{aligned}\pi(\lambda|x_{1:n}) \propto f(x_{1:n}|\lambda)\pi(\lambda) \quad &\propto \prod_{i=1}^{n} \text{Pn}(x_i|\lambda)\text{Ga}(\lambda|a,b)\\ &\propto \prod_{i=1}^{n}\frac{1}{x_i!}\lambda^{\sum_{i=1}^n x_i}\exp(-n\lambda)\times\frac{b^a}{\Gamma(a)}\lambda^{a-1}\exp(-\lambda b)\\ &\propto \lambda^{\sum_{i=1}^n x_i+a-1}\exp(-\lambda(n+b))\\ &\propto \text{Ga}(\lambda|\sum_{i=1}^{n}x_i+a, n+b)\end{aligned}$$

So the posterior distribution is $\lambda|x_{1:n}\sim\text{Ga}(\tilde{a},\tilde{b})$, $\tilde{a}=\sum_{i=1}^n x_i+a$, $\tilde{b}=n+b$.

- Alternatively, we could use the Theorem in the Lecture notes stating the properties of the Conjugate priors... I.e. $\lambda|x_{1:n}\sim\text{Ga}(\sum_{i=1}^n x_i+(\tau_1+1), n+(\tau_0))$ –It is up to you...

5. According to the definition, the predictive PMF is

$$\begin{aligned}f(y|x_{1:n}) &= \int_{(0,\infty)} f(y|\lambda)\pi(\lambda|x_{1:n})\text{d}\lambda \quad = \int_{(0,\infty)}\text{Pn}(y|\lambda)\text{Ga}(\lambda|\tilde{a},\tilde{b})\text{d}\lambda\\ &= \int_{(0,\infty)}\frac{1}{y!}\lambda^y\exp(-\lambda)\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\lambda^{\tilde{a}-1}\exp(-\lambda\tilde{b})\text{d}\lambda\\ &= \frac{1}{y!}\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\int_{(0,\infty)}\lambda^{y+\tilde{a}-1}\exp(-\lambda(\tilde{b}+1))\text{d}\lambda\\ &= \frac{1}{y!}\frac{\tilde{b}^{\tilde{a}}}{\Gamma(\tilde{a})}\frac{\Gamma(y+\tilde{a})}{(\tilde{b}+1)^{y+\tilde{a}}} \quad = \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{\Gamma(y+\tilde{a})}{\Gamma(\tilde{a})}\\ &= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{(y+\tilde{a}-1)(y+\tilde{a}-2)\cdots(\tilde{a})\cancel{\Gamma(\tilde{a})}}{\cancel{\Gamma(\tilde{a})}}\\ &= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y(y+\tilde{a}-1)(y+\tilde{a}-2)\cdots(\tilde{a})\\ &= \frac{1}{y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\frac{(y+\tilde{a}-1)!}{(\tilde{a}-1)!} \quad = \frac{(y+\tilde{a}-1)!}{(\tilde{a}-1)!y!}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(\frac{1}{\tilde{b}+1})^y\\ &= \binom{y+\tilde{a}-1}{\tilde{a}-1}(\frac{\tilde{b}}{\tilde{b}+1})^{\tilde{a}}(1-\frac{\tilde{b}}{\tilde{b}+1})^y \quad = \text{Nb}(y|\tilde{a},\frac{\tilde{b}}{\tilde{b}+1})\end{aligned}$$

6.

(a)

- The pair of hypotheses for this test is

$$\begin{cases}\text{H}_0: \quad \cancel{\lambda=2}\\ \cancel{\text{H}_1: \quad \lambda>0}\end{cases}$$

...or more seriously...

$$\begin{cases}\text{H}_0: \quad x_i \overset{\text{IID}}{\sim} \text{Pn}(\lambda_0=2), \text{ for all } i=1,...,n\\ \text{H}_1: \quad x_i \overset{\text{IID}}{\sim} \text{Pn}(\lambda),\ \lambda>0 \text{ for all } i=1,...,n\end{cases}$$

where H_0 is a single hypothesis, and H_1 is the general alternative.

- The overall prior can be specified as

$$\pi(\lambda)=\pi_0 1_{\{\lambda_0\}}(\lambda)+(1-\pi_0)\text{Ga}(\lambda|a,b)$$

for $\pi_0 > 0$, which in this case is $\pi_0 = 1/2$, and $\lambda_0 = 2$.

- The Bayes factor is

$$B_{01}(x_{1:n}) = \frac{p_0(x_{1:n})}{p_1(x_{1:n})} = \frac{\prod_{i=1}^n \text{Pn}(x_i|\lambda_0)}{\int \prod_{i=1}^n \text{Pn}(x_i|\lambda)\text{Ga}(\lambda|a,b)\text{d}\lambda}$$

where

$$p_0(x_{1:n}) = \prod_{i=1}^n \text{Pn}(x_i|\lambda_0) = \frac{1}{\prod_{i=1}^n x_i!}\lambda_0^{n\bar{x}}\exp(-n\lambda_0)$$

and

$$\begin{aligned} p_1(x_{1:n}) &= \int \prod_{i=1}^n \text{Pn}(x_i|\lambda)\text{Ga}(\lambda|a,b)\text{d}\lambda = \frac{1}{\prod_{i=1}^n x_i!}\frac{b^a}{\Gamma(a)}\int_0^\infty \lambda^{n\bar{x}+a-1}\exp(-(n+b)\lambda)\text{d}\lambda \\ &= \frac{1}{\prod_{i=1}^n x_i!}\frac{\Gamma(n\bar{x}+a)}{\Gamma(a)}\frac{b^a}{(n+b)^{n\bar{x}+a}} \end{aligned}$$

So

$$\begin{aligned} B_{01}(x_{1:n}) &= \frac{\lambda_0^{n\bar{x}}\exp(-n\lambda_0)}{\frac{b^a\Gamma(n\bar{x}+a)}{\Gamma(a)(n+b)^{n\bar{x}+a}}} = \lambda_0^{n\bar{x}}(n+b)^{n\bar{x}+a}\exp(-n\lambda_0)\frac{1}{b^a}\frac{\Gamma(a)}{\Gamma(n\bar{x}+a)} \\ &= \lambda_0^{n\bar{x}}\exp(-n\lambda_0)\frac{(n+b)^{n\bar{x}+a}}{b^a}\frac{\Gamma(a)}{(n\bar{x}+a-1)\cdots a\Gamma(a)} \\ &= \frac{\lambda_0^{n\bar{x}}\exp(-n\lambda_0)}{(n\bar{x}+a-1)\cdots a}\frac{(n+b)^{n\bar{x}+a}}{b^a} \end{aligned}$$

(b) Obviously, for the posterior probability that $\pi(\lambda = \lambda_0|x_{1:n})$, it is

$$\begin{aligned} \pi(\lambda = \lambda_0|x_{1:n}) = \pi(\text{H}_0|x_{1:n}) &= (1 + \frac{1-\pi_0}{\pi_0}\frac{p_1(x_{1:n})}{p_0(x_{1:n})})^{-1} \\ &= (1 + \frac{1-\pi_0}{\pi_0}\frac{b^a(n\bar{x}+a-1)\cdots a}{\lambda_0^{n\bar{x}}(n+b)^{n\bar{x}+a}\exp(-n\lambda_0)})^{-1} \\ &= \frac{\pi_0\lambda_0^{n\bar{x}}(n+b)^{n\bar{x}+a}\exp(-n\lambda_0)}{\pi_0\lambda_0^{n\bar{x}}(n+b)^{n\bar{x}+a}\exp(-n\lambda_0) + (1-\pi_0)b^a(n\bar{x}+a-1)\cdots a} \end{aligned}$$

(c) This is actually the aforesaid hypothesis test with $\lambda_0 = 2$.

- Based on the prior information, it is $\pi_0 = 0.5$, and $a = 4$, and $b = 2$ because

$$\left\{\begin{array}{c} E^{\text{Ga}(a,b)}(\lambda) = 2 \\ Var^{\text{Ga}(a,b)}(\lambda) = 1 \end{array}\right. \Leftrightarrow \left\{\begin{array}{c} a/b = 2 \\ a/b^2 = 1 \end{array}\right. \Leftrightarrow \left\{\begin{array}{c} a/b = 2 \\ 2/b = 1 \end{array}\right. \Leftrightarrow \left\{\begin{array}{c} a = 4 \\ b = 2 \end{array}\right.$$

- Based on the sample I have $n\bar{x} = 2 + 3 = 5$, $n = 2$
- Hence,

$$\begin{aligned} B_{01}(x_{1:n}) &= \frac{\lambda_0^{n\bar{x}}(n+b)^{n\bar{x}+a}\exp(-n\lambda_0)}{b^a(n\bar{x}+a-1)\cdots a} \\ &= \frac{2^5(2+2)^{5+4}\exp(-2\times 2)}{2^4(5+4-1)\cdots 4} = \frac{2^5\times 4^9\times\exp(-4)}{16\times 8\times 7\times 6\times 5\times 4} \\ &\approx 1.42 \end{aligned}$$

- Then $B_{01}(x_{1:n}) \approx 1.42$, and $\log_{10}(\text{B}_{01}(x_{1:n})) \approx 0.15$. According to Jeffrey's scaling rule, H_0 is supported

Exercise 66. ($\star\star\star$) Let $y = (y_1, ..., y_n)$ observables and consider the Bayesian hypothesis test

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

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

1. Let $\mathrm{B}_{01}(y)$ denote the Bayes factor $\mathrm{B}_{01}(y)$ 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

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

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

Solution. The overall prior is

$$\pi(\mu) = \pi_0 1_{\{\theta_0\}}(\mu) + (1-\pi_0)\mathrm{N}(\mu|\mu_0, \sigma_0^2)$$

with $\pi_0 = 1/2$ (although the value does not play any role here), and known μ_0 and σ_0^2.

1. The Bayes factor is

$$\mathrm{B}_{01}(y) = \frac{f_0(y)}{f_1(y)}$$

So

$$f_0(y) = f(y|\theta_0) = \prod_{i=1}^n \mathrm{N}(y_i|\theta_0, \sigma^2) = (\frac{1}{2\pi\sigma^2})^{\frac{n}{2}} \exp(-\frac{1}{2}\sum_{i=1}^n \frac{(y_i-\theta_0)^2}{\sigma^2})$$

$$= (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp(-\frac{1}{2}\frac{1}{\sigma^2}(\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{=ns^2} + n(\bar{y}-\theta_0)^2))$$

$$= (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp(-\frac{1}{2}\frac{1}{\sigma^2/n}s^2) \exp(-\frac{1}{2}\frac{1}{\sigma^2/n}(\bar{y}-\theta_0)^2)$$

$$f_1(y) = \int_{\mathbb{R}} f(y|\theta)\mathrm{d}\Pi_1(\theta) = \int_{\mathbb{R}} \prod_{i=1}^{n} \mathrm{N}(y_i|\mu,\sigma^2)\mathrm{N}(\mu|\mu_0,\sigma_0^2)\mathrm{d}\mu$$

$$= \int_{\mathbb{R}} (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp(-\frac{1}{2}\frac{1}{\sigma^2}(\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{=ns^2} + n(\bar{y}-\mu)^2))\times$$

$$\times (2\pi)^{-\frac{1}{2}}(\sigma_0^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}\frac{1}{\sigma_0^2}(\mu-\mu_0)^2)\mathrm{d}\mu$$

$$= (2\pi)^{-\frac{n}{2}-\frac{1}{2}}(\sigma^2)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}\frac{1}{\sigma^2/n}s^2)$$

$$\times \int_{\mathbb{R}} \exp(\underbrace{-\frac{1}{2}\frac{1}{\sigma^2/n}(\bar{x}-\mu)^2 - \frac{1}{2}\frac{1}{\sigma_0^2}(\mu-\mu_0)^2}_{=A(\mu)})\mathrm{d}\mu$$

$$A(\mu) = -\frac{1}{2}\frac{1}{\sigma^2/n}(\bar{y}-\mu)^2 - \frac{1}{2}\frac{1}{\sigma_0^2}(\mu-\mu_0)^2 = -\frac{1}{2}\frac{(\mu-\hat{\mu})^2}{\hat{\sigma}^2} - \frac{1}{2}(\frac{\bar{y}^2}{\sigma^2/n} + \frac{\mu_0^2}{\sigma_0^2}) + \frac{1}{2}\frac{(\frac{\bar{y}}{\sigma^2/n}+\frac{\mu_0}{\sigma_0^2})^2}{\frac{1}{\sigma^2/n}+\frac{1}{\sigma_0^2}}$$

$$= -\frac{1}{2}\frac{(\mu-\hat{\mu})^2}{\hat{\sigma}^2} - \frac{1}{2}\frac{(\bar{y}-\mu_0)^2}{\frac{\sigma^2}{n}+\sigma_0^2}$$

where

$$\hat{\sigma}^2 = (\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2})^{-1}$$

and

$$\int_{\mathbb{R}} \exp\left(A(\mu)\right)\mathrm{d}\mu = \exp\left(-\frac{1}{2}\frac{(\bar{y}-\mu_0)^2}{\frac{\sigma^2}{n}+\sigma_0^2}\right)\int_{\mathbb{R}} \exp\left(-\frac{1}{2}\frac{(\mu-\hat{\mu})^2}{\hat{\sigma}^2}\right)\mathrm{d}\mu = \exp\left(-\frac{1}{2}\frac{(\bar{y}-\mu_0)^2}{\frac{\sigma^2}{n}+\sigma_0^2}\right)(2\pi)^{\frac{1}{2}}(\hat{\sigma}^2)^{\frac{1}{2}}$$

So

$$f_1(y) = (2\pi)^{-\frac{n}{2}-\frac{1}{2}}(\sigma^2)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}\frac{1}{\sigma^2/n}s^2)\int_{\mathbb{R}} \exp\left(A(\mu)\right)\mathrm{d}\mu$$

$$= (2\pi)^{-\frac{n}{2}-\frac{1}{2}}(\sigma^2)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{1}{\sigma^2/n}s^2\right)(2\pi\hat{\sigma}^2)^{1/2}\exp\left(-\frac{1}{2}\frac{(\bar{y}-\mu_0)^2}{\frac{\sigma^2}{n}+\sigma_0^2}\right)$$

It is

$$\mathrm{B}_{01}(y) = \frac{f_0(y)}{f_1(y)} = \frac{\left(\frac{\sigma^2}{n}\right)^{-\frac{1}{2}}}{\left(\frac{\sigma^2}{n}+\sigma_0^2\right)^{-\frac{1}{2}}}\frac{\exp\left(-\frac{1}{2}\frac{(\bar{y}-\theta_0)^2}{\frac{\sigma^2}{n}}\right)}{\exp\left(-\frac{1}{2}\frac{(\bar{y}-\mu_0)^2}{\frac{\sigma^2}{n}+\sigma_0^2}\right)}$$

2. It is

$$\mathsf{P}_\Pi(\mathrm{H}_0|y) = \left(1+\frac{1-\pi_0}{\pi_0}\mathrm{B}_{01}(y)^{-1}\right)^{-1} = \left(1+\frac{1-\pi_0}{\pi_0}\frac{\left(\frac{\sigma^2}{n}+\sigma_0^2\right)^{-\frac{1}{2}}}{\left(\frac{\sigma^2}{n}\right)^{-\frac{1}{2}}}\frac{\exp\left(-\frac{1}{2}\frac{(yx-\mu_0)^2}{\frac{\sigma^2}{n}+\sigma_0^2}\right)}{\exp\left(-\frac{1}{2}\frac{(\bar{y}-\theta_0)^2}{\frac{\sigma^2}{n}}\right)}\right)^{-1}$$

Exercise 67. ($\star\star\star$)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:

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

Solution. For simplicity, we suppress the indexing denoting the sub-set of the regressors.

1. For simplicity, we suppress the indexing denoting the sub-set of the regressors. It is

$$f(y) = \int \mathrm{N}\left(y|\Phi\beta, I\sigma^2\right)\mathrm{N}\left(\beta|\mu, V\sigma^2\right)\mathrm{IG}\left(\sigma^2|a,k\right)\mathrm{d}\beta\mathrm{d}\sigma^2$$

$$= \int \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(y-\Phi\beta)^\top(y-\Phi\beta)\right)\times\left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\left(\frac{1}{|\sigma^2 V|}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(\beta-\mu)^\top V^{-1}(\beta-\mu)\right)$$
$$\times\frac{k^a}{\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\left(-\frac{k}{\sigma^2}\right)\mathrm{d}\beta\mathrm{d}\sigma^2 \;=\; ...$$

$$
\begin{aligned}
f(y) &= \left(\frac{1}{2\pi}\right)^{\frac{n+d}{2}} \left(\frac{1}{|V|}\right)^{\frac{1}{2}} \frac{k^a}{\Gamma(a)} \\
&\quad \times \int \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1} \exp\left(-\frac{1}{2\sigma^2}(y-\Phi\beta)^\top(y-\Phi\beta) - \frac{1}{2\sigma^2}(\beta-\mu)^\top V^{-1}(\beta-\mu) - \frac{k}{\sigma^2}\right) \mathrm{d}\beta\mathrm{d}\sigma^2 \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n+d}{2}} \left(\frac{1}{|V|}\right)^{\frac{1}{2}} \frac{k^a}{\Gamma(a)} \\
&\quad \times \int \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1} \exp\left(-\frac{1}{\sigma^2}\left[\frac{(y-\Phi\beta)^\top(y-\Phi\beta)+(\beta-\mu)^\top V^{-1}(\beta-\mu)}{2}+k\right]\right) \mathrm{d}\beta\mathrm{d}\sigma^2 \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n+d}{2}} \left(\frac{1}{|V|}\right)^{\frac{1}{2}} \frac{k^a}{\Gamma(a)} \\
&\quad \times \int \left[\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1} \exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}S+k\right)\right)\right]\left[\int \exp\left(-\frac{1}{2}\frac{1}{\sigma^2}(\beta-v)^\top\left(V^*\right)^{-1}(\beta-\mu^*)\right)\mathrm{d}\beta\right] \mathrm{d}\sigma^2 \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}+\frac{d}{2}} \left(\frac{1}{|V|}\right)^{\frac{1}{2}} \frac{k^a}{\Gamma(a)} \int \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1} \exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}S+k\right)\right) \times \left[(2\pi)^{\frac{d}{2}}\left(\sigma^2\right)^{\frac{d}{2}}|V^*|^{\frac{1}{2}}\right] \mathrm{d}\sigma^2 \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \left(\frac{|V^*|}{|V|}\right)^{\frac{1}{2}} \frac{k^a}{\Gamma(a)} \frac{\Gamma\left(\frac{n}{2}+a\right)}{\left(\frac{1}{2}S+k\right)^{\frac{n}{2}+a}} = \quad \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \left(\frac{|V^*|}{|V|}\right)^{\frac{1}{2}} \frac{k^a}{\Gamma(a)} \frac{\Gamma\left(\frac{n}{2}+a\right)}{\left(k^*\right)^{\frac{n}{2}+a}}
\end{aligned}
$$

Where

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

$$
k^* = k + \frac{1}{2}S; \qquad V^* = \left(V^{-1}+\Phi^\top\Phi\right)^{-1}; \qquad \mu^* = V^*\left(V^{-1}\mu+\Phi^\top y\right)
$$

For simplicity, we use the indexing $\cdot_0$ and $\cdot_1$ instead of $\cdot_{\mathscr{M}_0}$ and $\cdot_{\mathscr{M}_1}$ in what follows. So the Bayes factor is

$$
B_{01}(y) = \frac{f_0(y)}{f_1(y)} = \sqrt{\frac{|V_1|}{|V_0|}}\sqrt{\frac{|V_0^*|}{|V_1^*|}}\left(\frac{k_0^*}{k_1^*}\right)^{-\frac{n}{2}-a}
$$

2. So

$$
\mathsf{P}_\Pi\left(\mathrm{H}_0|y\right) = \left(1+\frac{1-\pi_0}{\pi_0}\mathrm{B}_{01}(y)^{-1}\right)^{-1}
$$

Part XI

Inference under model uncertainty

Exercise 68. ($\star\star$)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

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

Solution. It is straightforward by considering

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

$$\Longleftrightarrow \pi\left(\mathscr{M}_k|y\right) = \frac{\pi(\mathscr{M}_k)}{\pi(\mathscr{M}_0)} \frac{\pi\left(\mathscr{M}_k|y\right)/\pi(\mathscr{M}_k)}{\pi\left(\mathscr{M}_0|y\right)/\pi(\mathscr{M}_0)} \Bigg/ \sum_{k'=0}^{K} \frac{\pi(\mathscr{M}_{k'})}{\pi(\mathscr{M}_0)} \frac{\pi\left(\mathscr{M}_{k'}|y\right)/\pi(\mathscr{M}_{k'})}{\pi\left(\mathscr{M}_0|y\right)/\pi(\mathscr{M}_0)}$$

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

Exercise 69. ($\star\star$)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 $\forall k, i, j \in \mathcal{K}$.

Solution. It is

$$B_{k,j}(y) = \frac{\pi\left(\mathscr{M}_k|y\right)/\pi\left(\mathscr{M}_k\right)}{\pi\left(\mathscr{M}_j|y\right)/\pi\left(\mathscr{M}_j\right)} = \frac{\pi\left(\mathscr{M}_k|y\right)/\pi\left(\mathscr{M}_k\right)}{\pi\left(\mathscr{M}_i|y\right)/\pi\left(\mathscr{M}_i\right)} \frac{\pi\left(\mathscr{M}_i|y\right)/\pi\left(\mathscr{M}_i\right)}{\pi\left(\mathscr{M}_j|y\right)/\pi\left(\mathscr{M}_j\right)} = B_{k,i}(y)B_{i,j}(y)$$

Exercise 70. ($\star\star$)Consider the Bayesian model

$$\begin{cases} y|\theta_k, \mathscr{M}_k \sim F(y|\theta_k, k) \\ \theta_k|\mathscr{M}_k \sim \Pi(\theta_k|k) \\ \mathscr{M}_k \sim \Pi(k) \end{cases} \tag{76}$$

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

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

Solution.

1. It is

$$\text{E}_G\left(z|y\right)=\sum_{k\in\mathcal{K}}\int_{\Theta_k}z\text{d}G\left(z|y,k\right)\pi(k|y)=\sum_{k\in\mathcal{K}}\text{E}_G\left(z|y,k\right)\pi(k|y)=\text{E}_\Pi\left(\text{E}_G\left(z|y,k\right)|y\right)$$

2. It is

$$\begin{aligned}\text{Var}_G\left(z|y\right)&=\text{E}_G\left(z|y\right)-\left(\text{E}_G\left(z|y\right)\right)^2=\text{E}_\Pi\left(\text{E}_G\left(z^2|y,k\right)|y\right)-\left(\text{E}_G\left(z|y\right)\right)^2\\&=\text{E}_\Pi\left(\text{Var}_G\left(z|y,k\right)+\left(\text{E}_G\left(z|y,k\right)\right)^2|y\right)-\left(\text{E}_G\left(z|y\right)\right)^2\\&=\sum_{k\in\mathcal{K}}\left(\text{Var}_G\left(z|y,k\right)+\left(\text{E}_G\left(z|y,k\right)\right)^2\right)\pi(k|y)-\left(\text{E}_G\left(z\,|y\right)\right)^2\end{aligned}$$
