# Exercise Sheet: Bayesian Statistics

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

## Part I

# Matrix & vector calculus

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

---

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

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

**Solution.** It is

$$\begin{aligned}(A+B)^{-1} &= A^{-1}(I+A^{-1}B)^{-1} \\ &= A^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}\end{aligned}$$

---

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

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

if $A$ and $C$ are non-singular.

**Solution.**

By checking that $(A+UCV)(A+UCV)^{-1} = I$

$$\begin{aligned}(A+UCV)\times &\left[A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right] \\ &= I + UCVA^{-1} - (U+UCVA^{-1}U)(C^{-1}+VA^{-1}U)^{-1}VA^{-1} \\ &= I + UCVA^{-1} - UC(C^{-1}+VA^{-1}U)(C^{-1}+VA^{-1}U)^{-1}VA^{-1} \\ &= I + UCVA^{-1} - UCVA^{-1} = I.\end{aligned}$$

So

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

---

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

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

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

**Solution.**

$$
\begin{aligned}
(A+uv^T)(A+uv^T)^{-1} &= (A+uv^T)\left(A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}\right) \\
&= AA^{-1} + uv^TA^{-1} - \frac{AA^{-1}uv^TA^{-1} + uv^TA^{-1}uv^TA^{-1}}{1+v^TA^{-1}u} \\
&= I + uv^TA^{-1} - \frac{uv^TA^{-1} + uv^TA^{-1}uv^TA^{-1}}{1+v^TA^{-1}u} \\
&= I + uv^TA^{-1} - \frac{u\cancel{(1+v^TA^{-1}u)}v^TA^{-1}}{\cancel{1+v^TA^{-1}u}} \\
&= I + uv^TA^{-1} - uv^TA^{-1} \\
&= I
\end{aligned}
$$

---

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

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

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

$$
\begin{aligned}
A_{11}^{-1} &= B_{11} = B_{12}B_{22}^{-1}B_{21} \\
A_{11}^{-1}A_{12} &= -B_{12}B_{22}^{-1}
\end{aligned}
$$

**Hint:** Start by noticing that

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

**Solution.** It is

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

So

$$
\begin{aligned}
A_{11}B_{12} + A_{12}B_{22} = 0 &\iff \\
A_{11}^{-1}\left(A_{11}B_{12} + A_{12}B_{22}\right)B_{22}^{-1} = 0 &\iff \\
B_{12}B_{22}^{-1} + A_{11}^{-1}A_{12} = 0 &
\end{aligned}
$$

So

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

Also

$$\begin{aligned} A_{11}B_{12} + A_{12}B_{22} = 0 \iff \\ (A_{11}B_{12} + A_{12}B_{22})B_{22}^{-1}B_{21} = 0 \iff \\ A_{11}B_{12}B_{22}^{-1}B_{21} + A_{12}B_{21} = 0 \\ A_{12}B_{21} = -A_{11}B_{12}B_{22}^{-1}B_{21} \end{aligned}$$

Then, we plug in the above in $A_{11}B_{11} + A_{12}B_{21} = I$ we get

$$\begin{aligned} A_{11}B_{11} + A_{12}B_{21} = I \iff \\ A_{11}B_{11} - A_{11}B_{12}B_{22}^{-1}B_{21} = I \iff \\ B_{11} - B_{12}B_{22}^{-1}B_{21} = A_{11}^{-1} \end{aligned}$$

So

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

# Part II

# Random variables

---

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

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

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

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

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

**Solution.** It is $z = 1 - \exp(-\lambda y) \iff y = -\frac{1}{\lambda}\log(1-z)$, and $z \in [0,1]$. So $h^{-1}(z) = -\frac{1}{\lambda}\log(1-z)$. Then

$$\begin{aligned} f_z(z) =& f_{\text{Ex}(\lambda)}(h^{-1}(z)) \times \left|\det\left(\frac{\mathrm{d}}{\mathrm{d}z}h^{-1}(z)\right)\right| = f_{\text{Ex}(\lambda)}\left(-\frac{1}{\lambda}\log(1-z)\right) \times \left|\det\left(\frac{\mathrm{d}}{\mathrm{d}z}\frac{-1}{\lambda}\log(1-z)\right)\right| \\ =& \exp\left(-\lambda\frac{-1}{\lambda}\log(1-z)\right)\ 1(-\frac{1}{\lambda}\log(1-z) \geq 0)\ \times \left|-\frac{1}{\lambda}\frac{1}{1-z}\right| = 1(z \in [0,1]) \end{aligned}$$

From the density, we recognize that $z \sim \mathrm{U}(0,1)$ follows a uniform distribution.

---

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

1. Let matrix $A \in \mathbb{R}^{q\times d}$, $c \in \mathbb{R}^q$, and $z = c + Ay$ then
$$\mathrm{E}(z) = \mathrm{E}(c + Ay) = c + A\mathrm{E}(y)$$
2. Let random variables $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$, and let functions $\psi_1$ and $\psi_2$ 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 $\psi_1$ and $\psi_2$ 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 9.** ($\star$)Prove the following properties of the covariance matrix

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

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

**Solution.**

1. It is

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

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

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

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

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

**Solution.**

It is

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

---

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

1. $\text{Var}(y) = \text{E}(yy^\top) - \text{E}(y)\left(\text{E}(y)\right)^\top$
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 12.** ($\star$)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 13.** ($\star$)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 14.** ($\star$)

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

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

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

**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 $\Theta$; 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 $\Theta$; 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 16.** ($\star$)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 17.** ($\star$)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 III

# Probability calculus

---

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

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

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

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

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

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

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

2. Show that

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

3. Show that

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

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

**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:

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

2. It is

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

3. It is

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

---

**Exercise 20.** ($\star\star$)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 21.** ($\star\star$)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

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

---

The Following exercise is part of Homework 1

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

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

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

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.**

---

**Exercise 23.** (⋆⋆⋆)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 IV

# Bayesian paradigm and calculations

---

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

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

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

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

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

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

**Solution.** It is

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

1. By using the Bayes theorem

$$\begin{aligned}
\pi(\theta|y) \propto & f(y|\theta)\pi(\theta) \qquad \propto \prod_{i=1}^n f(y_i|\theta)\pi(\theta) \;= \prod_{i=1}^n \frac{\theta e^{-y_i}}{(1+e^{-y_i})^{\theta+1}}\,\frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}1(\theta\in(0,\infty)) \\
\propto & \cancel{\prod_{i=1}^n \frac{e^{-y_i}}{(1+e^{-y_i})}}\theta^n \prod_{i=1}^n \exp(-\theta\log(1+e^{-y_i}))\,\cancel{\frac{b^a}{\Gamma(a)}}\theta^{a-1}e^{-b\theta}1(\theta\in(0,\infty)) \\
\propto & \theta^{n+a-1}\exp\left(-\theta\left[\sum_{i=1}^n \log(1+e^{-y_i}) + b\right]\right)1(\theta\in(0,\infty)) \propto \text{Ga}(\theta|a+n, b+\sum_{i=1}^n\log(1+e^{-y_i}))
\end{aligned}$$

So

$$\theta|y \sim \text{Ga}\left(\underbrace{a+n}_{=a^*}, \underbrace{b+\sum_{i=1}^n \log(1+e^{-y_i})}_{=b^*}\right)$$

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

$$\begin{aligned}
f(z|y) &= \int_{\mathbb{R}} f(z|\theta)\pi(\theta|y)\mathrm{d}\theta \\
&= \int_{\mathbb{R}_+} \frac{e^{-z}}{(1+e^{-z})}\theta\exp(-\theta\log(1+e^{-z}))\frac{(b^*)^{a^*}}{\Gamma(a^*)}\theta^{a^*-1}\exp(-\theta b^*)\mathrm{d}\theta \\
&= \frac{(b^*)^{a^*}}{\Gamma(a^*)}\frac{e^{-z}}{(1+e^{-z})}\int_{\mathbb{R}_+}\theta^{a^*+1-1}\exp(-\theta(b^*+\log(1+e^{-y})))\mathrm{d}\theta \\
&= \frac{(b^*)^{a^*}}{\Gamma(a^*)}\frac{e^{-z}}{(1+e^{-z})}\frac{\Gamma(a^*+1)}{(b^*+\log(1+e^{-z}))^{a^*+1}} = \frac{e^{-z}}{(1+e^{-z})}\frac{(b^*)^{a^*}}{(b^*+\log(1+e^{-z}))^{a^*+1}}a^*
\end{aligned}$$

---

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

Assume You specify a Bayesian model <-set-up

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

1. Show that
$$\sum_{i=1}^{n}(y_i - \theta)^2 = n(\bar{y} - \theta)^2 + ns^2,$$
where $s^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

2. Show that the joint posterior distribution $\Pi(\mu, \sigma^2|y)$ is such as
$$\mu|y, \sigma^2 \sim \mathrm{N}(\mu_n, \sigma^2 \frac{1}{\tau_n})$$
$$\sigma^2|y \sim \mathrm{IG}(a_n, k_n)$$
with
$$\mu_n = \frac{n\bar{y} + \tau_0\mu_0}{n + \tau_0}; \qquad \tau_n = n + \tau_0; \qquad a_n = a_0 + n$$
$$k_n = k_0 + \frac{1}{2}ns_n^2 + \frac{1}{2}\frac{\tau_0 n(\mu_0 - \bar{y})^2}{n + \tau_0}$$
**Hint:** It is
$$-\frac{1}{2}\frac{(\mu - \mu_1)^2}{v_1} - \frac{1}{2}\frac{(\mu - \mu_2)^2}{v_2} ... - \frac{1}{2}\frac{(\mu - \mu_n)^2}{v_n} = -\frac{1}{2}\frac{(\mu - \hat{\mu})^2}{\hat{v}} + C$$
where
$$\hat{v} = \left(\sum_{i=1}^{n}\frac{1}{v_i}\right)^{-1}; \quad \hat{\mu} = \hat{v}\left(\sum_{i=1}^{n}\frac{\mu_i}{v_i}\right); \quad C = \frac{1}{2}\frac{\hat{\mu}^2}{\hat{v}} - \frac{1}{2}\sum_{i=1}^{n}\frac{\mu_i^2}{v_i}$$

3. Show that the marginal posterior distribution $\Pi(\mu|y)$ is such as
$$\mu|y \sim \mathrm{T}_1\left(\mu_n, \frac{k_n}{a_n}\frac{1}{\tau_n}, 2a_n\right)$$
**Hint-1:** If $x \sim \mathrm{IG}(a, b)$, $y = cx$, then $y \sim \mathrm{IG}(a, cb)$.
**Hint-2:** The definition of Student T is considered as known

4. Show that the predictive distribution $\Pi(z|y)$ is Student T such as
$$z|y \sim \mathrm{T}_1\left(\mu_n, \frac{k_n}{a_n}(\frac{1}{\tau_n} + 1), 2a_n\right)$$
**Hint-1:** Consider that
$$\mathrm{N}(x|\mu_1, \sigma_1^2)\,\mathrm{N}(x|\mu_2, \sigma_2^2) \;=\; \mathrm{N}(x|m, v^2)\,\mathrm{N}(\mu_1|\mu_2, \sigma_1^2 + \sigma_2^2)$$

where

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

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

**Solution.**

1. It is

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\theta)^2 &= \sum_{i=1}^{n}\left[(y_i-\bar{y})-(\theta-\bar{y})\right]^2\\ &= \sum_{i=1}^{n}\left[(y_i-\bar{y})^2+(\theta-\bar{y})^2-2(y_i-\bar{y})(\theta-\bar{y})\right]\\ &= ns^2+n(\bar{y}-\theta)^2, \text{ where } s^2=\frac{1}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2\end{aligned}$$

2. I use the Bayes theorem

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

It is

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

where

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

So

$$\pi(\mu,\sigma^2|y) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{1}{2}+\frac{n}{2}+a_0+1} \exp\left(\frac{1}{\sigma^2}\left[-\frac{1}{2}\frac{(\mu-\mu_n)^2}{1/\tau_n}+C_n\right]-\frac{1}{\sigma^2}k_0\right)$$

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

$$\propto \mathrm{N}(\mu|\mu_n,\sigma^2/\tau_n)\mathrm{IG}(\sigma^2|a_n,k_n)$$

where

$$\mu_n = \frac{n\bar{y}+\tau_0\mu_0}{n+\tau_0}\,; \qquad a_n = \frac{n}{2}+a_0\,;$$

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

3. 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

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

4. It is

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

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.

**Exercise 26.** (⋆⋆⋆)(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 $\sigma^2$.

You are interested in learning $\beta$, but you do not care about $\sigma^2$. 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 $\Phi$ 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 $\beta$ given $y$ is

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

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

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

Consider that

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

where

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

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

**Solution.**

1. I use the Bayes theorem

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

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

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

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

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

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

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

3. It is

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

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.

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

---

**Exercise 27.** ($\star\star$)Let $y = (y_1, ..., y_n)$ be observables drawn iid from sampling distribution $y_i|\theta \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \theta^2)$ for all $i = 1, ..., n$, where $\theta \in \mathbb{R}$ is unknown. Specify a conjugate prior density for $\theta$ 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 28.** ($\star\star$)If the sampling distribution $F(\cdot|\theta)$ is discrete and the prior $\Pi(\theta)$ is proper, then the posterior $\Pi(\theta|y)$ is always proper.

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

**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.

## The Limit Comparison Theorem for Improper Integrals

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

Let

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

Then

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

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

Let

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

Then

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

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

Let

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

Then

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

**Note:** A useful test function is

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

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

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

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

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

**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.

---

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

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

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

**Solution.**

We will check the properness condition. It is

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

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

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

So the posterior is defined.

---

The Following exercise is part of Homework 1

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

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

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

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

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

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

**Solution.**

---

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 $\sigma^2$ and unknown $\mu$.

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

   - 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 $\mu$ as $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$ where $\mu_0, \sigma_0^2$ 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

$$\begin{aligned}\pi(\mu|x_1) &\propto \overbrace{\mathrm{N}(x_1|\mu,\sigma^2)}^{\text{likelihood}}\overbrace{\mathrm{N}(\mu|\mu_0,\sigma_0^2)}^{\text{prior}}\\ &\propto \exp(-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2})\exp(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2})\\ &\propto \exp(-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2} - \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2})\\ &\propto \exp(-\frac{1}{2}\frac{(\mu-\hat{\mu}_1)^2}{\hat{\sigma}_1^2}) \propto \mathrm{N}(\mu|\hat{\mu}_1,\hat{\sigma}_1^2) \end{aligned} \tag{3}$$

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 (3), 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

$$\begin{aligned}\pi(\mu|x_1,x_2) &\propto \overbrace{(x_2|\mu,\sigma^2)}^{\text{likelihood}}\overbrace{\mathrm{N}(\mu|\hat{\mu}_1,\hat{\sigma}_1^2)}^{\text{prior}}\\ &\propto \exp(-\frac{1}{2}\frac{(x_2-\mu)^2}{\sigma^2})\exp(-\frac{1}{2}\frac{(\mu-\hat{\mu}_1)^2}{\hat{\sigma}_1^2})\\ &\propto \exp(-\frac{1}{2}\frac{(x_2-\mu)^2}{\sigma^2} - \frac{1}{2}\frac{(\mu-\hat{\mu}_1)^2}{\hat{\sigma}_1^2})\\ &\propto \exp(-\frac{1}{2}\frac{(\mu-\hat{\mu}_2)^2}{\hat{\sigma}_2^2}) \propto \mathrm{N}(\mu|\hat{\mu}_2,\hat{\sigma}_2^2) \end{aligned} \tag{4}$$

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 (3), 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

$$\begin{aligned}\pi(\mu|x_1,x_2) &\propto \overbrace{\mathrm{N}(x_1|\mu,\sigma^2)\mathrm{N}(x_2|\mu,\sigma^2)}^{\text{likelihood}}\overbrace{\mathrm{N}(\mu|\mu_0,\sigma_0^2)}^{\text{prior}}\\ &\propto \exp(-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2})\exp(-\frac{1}{2}\frac{(x_2-\mu)^2}{\sigma^2})\exp(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2})\\ &\propto \exp(-\frac{1}{2}\frac{(x_1-\mu)^2}{\sigma^2} - \frac{1}{2}\frac{(x_2-\mu)^2}{\sigma^2} - \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2})\\ &\propto \exp(-\frac{1}{2}\frac{(\mu-\hat{\mu})^2}{\hat{\sigma}^2}) \propto \mathrm{N}(\mu|\hat{\mu},\hat{\sigma}^2) \end{aligned} \tag{5}$$

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 (5), we recognized the kernel of the Normal PDF. Hence, $\mu|x_1, x_2 \sim \text{N}(\hat{\mu}, \hat{\sigma}^2)$

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 $\text{d}\Pi(\mu|x_1, x_2)$ at the end posterior $\text{d}\Pi(\mu|x_1, x_2)$ in a single application of Bayes theorem with the full data $x = (x_1, x_2)$.

---

# Part V

# Exchangeability

We work on the proofs of the following theorems:

- Marginal distributions of finite exchangeable sequences $y_1, y_2, \ldots, y_k$ are invariant under permutations; i.e.:
$$\mathrm{d}F(y_{\mathfrak{p}(1)}, y_{\mathfrak{p}(2)}, \ldots, y_{\mathfrak{p}(k)}) = \mathrm{d}F(y_1, y_2, \ldots, y_k) \text{ for all } \mathfrak{p} \in \mathfrak{P}_n. \tag{6}$$
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:
$$\mathrm{E}(g(y_i)) = \mathrm{E}(g(y_1)) \text{ for all } i = 1, ..., k \text{ and all functions } g\colon \mathcal{Y} \to \mathbb{R} \tag{7}$$
- (Marginal) Variances of finite exchangeable sequences $y_1, y_2, \ldots, y_k$ are all identical:
$$\mathrm{Var}(y_i) = \mathrm{Var}(y_1). \tag{8}$$
- Covariances between elements of finite exchangeable sequences $y_1, y_2, \ldots, y_k$ are all identical:
$$\mathrm{Cov}(y_i, y_j) = \mathrm{Cov}(y_1, y_2) \text{ whenever } i \neq j. \tag{9}$$

**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 $\pi \in \Pi_n$,*

$$E(g(Y_{\mathfrak{p}(1)}, Y_{\mathfrak{p}(2)}, \ldots, Y_{\mathfrak{p}(k)})) = E(g(Y_1, Y_2, \ldots, Y_k)) \tag{10}$$

---

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

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

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

where

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

aka $\theta$ 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 $N \geq n \geq t_n \geq 0$,

$$\begin{aligned} p(S_n = t_n) &= \sum_{t_N=0}^{N} p(S_n = t_n | S_N = t_N) p(S_N = t_N) \\ &= \underbrace{\sum_{t_N=0}^{t_n-1} p(S_n = t_n | S_N = t_N) p(S_N = t_N)}_{=0} \qquad (11) \\ &\quad + \sum_{y_N=y_n}^{N-(n-y_n)} p(S_n = t_n | S_N = t_N) p(S_N = t_N) \\ &\quad + \underbrace{\sum_{t_N=N-(n-t_n)+1}^{N} p(S_n = t_n | S_N = t_N) p(S_N = t_N)}_{=0} \qquad (12) \\ &= \sum_{y_N=y_n}^{N-(n-y_n)} p(S_n = t_n | S_N = t_N) p(S_N = t_N) \end{aligned}$$

The terms in (11, 12) 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

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

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

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

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,

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

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

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

$$\lim_{N\to\infty} \frac{(\theta N)((1-\theta)N)_{n-t_n}}{(N)_n} = \theta^{t_n}(1-\theta)^{n-t_n} = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \tag{13}$$

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 $\Pi$ is a distribution function.

---

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

**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:

$$f(y_1, ..., y_n) = \frac{1}{N^{\underline{n}}} = \frac{(N-n)!}{N!} \tag{14}$$

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

---

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

$$\mathrm{E}(g(Y_i)) = \mathrm{E}(g(Y_1)) \text{ for all } i \in \{2, \ldots, n\} \tag{15}$$

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.

$$\mathrm{E}(g(Y_1)) = \sum_{(y_1,y_2,y_3,\ldots,y_n)\in\mathcal{Y}^n} g(y_1) f(y_1, y_2, y_3, \ldots, y_n) \tag{16}$$

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

$$= \sum_{(y_1,y_2,y_3,\ldots,y_n)\in\mathcal{Y}^n} g(y_1) f(y_2, y_1, y_3, \ldots, y_n) \tag{17}$$

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

$$= \sum_{(y_2,y_1,y_3,\ldots,y_n)\in\mathcal{Y}^n} g(y_2) f(y_1, y_2, y_3, \ldots, y_n) = \mathrm{E}(g(Y_2)) \tag{18}$$

---

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

$$\mathrm{E}(g(Y_i)) = \mathrm{E}(g(Y_1)) \text{ for all } i \in \{2, \ldots, n\} \tag{19}$$

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

$$\mathrm{Var}(Y_i) = \mathrm{Var}(Y_1) \text{ for all } i \in \{2, \ldots, n\} \tag{20}$$

**Solution.** By the usual properties of variance,

$$\text{Var}(Y_i) = \text{E}(Y_i^2) - \text{E}(Y_i)^2 \tag{21}$$

and now applying19 twice

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

---

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

$$cov(Y_i, Y_j) \geq -\frac{var(Y_1)}{n-1} \tag{22}$$

**Solution.** It is

$$0 \leq var\left(\sum_{k=1}^{n} Y_k\right) = \sum_{k=1}^{n} var(Y_k) + 2\sum_{k=1}^{n-1}\sum_{\ell=k+1}^{n} cov\left(Y_k, Y_\ell\right) \tag{23}$$

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

$$= n\,\text{var}\left(Y_1\right) + (n^2 - n)\text{cov}\left(Y_i, Y_j\right) \tag{24}$$

where the $n^2 - n$ factor can be derived as follows: note that the pairs of indices $(k, \ell)$ 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,

$$\text{Cov}\left(Y_i, Y_j\right) \geq -\frac{n\text{var}\left(Y_1\right)}{n^2 - n} = -\frac{\text{var}\left(Y_1\right)}{n-1} \tag{25}$$

---

**Exercise 39.** ($\star$)What does

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

imply about the correlation of infinite exchangeable sequences?

**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 VI

# Sufficiency

---

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

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

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

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

1. Show that the parametric model is member of the Exponential family, and the sufficient statistic for a sample of observables $x = (x_1, ..., x_n)$.
2. Show that the posterior distribution $\theta$ 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 41.** ($\star\star\star$)Consider the Bayesian model

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

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

**Hint-1:** $\text{Mu}_k$ denotes the Multinomial probability distribution with PMF

$$\text{Mu}_k(x|\theta) = \begin{cases} \prod_{j=1}^k \theta_j^{x_j} & \text{, if } x \in \mathcal{X}_k \\ 0 & \text{, otherwise} \end{cases} \tag{26}$$

**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 (26) 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

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

so $p(y|x_{1:n})=\frac{1}{n+a_*}(a_{j'}+x_{*,j'})$, where $j'$ such that $y_{j'}=1$.

---

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

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

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

**Solution.** It is

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

---

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

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

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

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

**Solution.** Elementary manipulation gives

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

---

The following is a proof of a theorem

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

**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

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

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 $\mathrm{d}\Pi(\theta)$.

---

The following is a proof of a theorem

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

Let $t : \mathcal{Y} \to \mathcal{T}$ be a statistic. Then $t$ is a parametric sufficient statistic for $\theta$ 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 $\rho$ on $\Theta$ as

$$L(y;\theta) = k(t(y)|\theta)\rho(y). \tag{27}$$

**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 $\Theta'$ 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

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

($\Longrightarrow$) 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

$$\kappa(y|\theta) = \frac{f(y|\theta)}{f(y)}\,; \qquad\qquad \rho(y) = f(y)\,; \tag{28}$$

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 (28), so I get

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

# Part VII

# Priors

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

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

with density

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

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

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

1. Show that the posterior $\Pi(\theta|y)$ of $\theta$ 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 $\theta$ is a continuous random quantity, it is

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

For the case where $\theta$ is a discrete random quantity, the proof is similar.

---

**Exercise 47.** ($\star\star$)

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

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

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

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 $(\tau_0, \tau_1)$ to the standard parameters $(\alpha_0, \beta_0)$ 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$(\alpha_0, \beta_0)$, derive the posterior distribution of $\theta$.

4. Given an i.i.d. sample $x_1, \ldots, x_n$ from the skew-logistic distribution, and assuming that the prior is Gamma$(\alpha_0, \beta_0)$, derive the predictive PDF for a future $y = x_{n+1}$ up to a normalising constant.

5. Give a minimal sufficient statistic for $\theta$ under i.i.d. sampling from the skew-logistic distribution. Would this statistic still be sufficient if we had chosen a prior for $\theta$ 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

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

since $\int_{\mathbb{R}_+}\text{Ga}(\theta|\tau_0+1,\tau_1)\text{d}\theta = 1$. 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

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

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

Consider a PDF/PMF $f(x|\theta)$ with $x \in \mathcal{X}$ and $\theta \in \Theta$ where $\mathcal{X}$ does not depend on $\theta$. 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 $\theta$.*

**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 $\theta$. 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 49.** ($\star\star$)Suppose that you have a prior distribution for the probability $\theta$ 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 $\theta$?

**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 $\theta$, 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 $\theta$. 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 $\theta$ is 0.4, so $\text{E}(\theta) = \frac{a}{a+b} = 0.4$. 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

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

---

**Exercise 50.** ($\star\star$)Find a (two-dimensional) sufficient statistic for $(\alpha, \beta)$ 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 $\alpha$ and $\beta$ 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 51.** ($\star\star$)Suppose that your prior for $\theta$ 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

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

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

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

- 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

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

---

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

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

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

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

**Hint-2** Negative Binomial distribution $x \sim \mathrm{Nb}(r, \theta)$ has PMF: $\mathrm{Nb}(x|r, \theta) = \binom{r+x-1}{r-1}\theta^r(1-\theta)^x 1_{\mathbb{N}}(x)$ 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 $\lambda$, and identify the parametric family of distributions as $\lambda \sim \mathrm{Ga}(a, b)$, with $a > 0$, and $b > 0$. While you are deriving the conjugate prior distribution of $\lambda$, 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 $\lambda$, 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 $\theta$ 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

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

The sampling space $\mathcal{X}$ does not depend on the uncertain parameter $\lambda$ 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 $\lambda$, (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

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

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 $\tau_0$ (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 $\tau_1$ (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 53.** ($\star\star$)Let $x = (x_1, ..., x_n)$ be observables. Consider a Bayesian model such as

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

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

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

**Hint-2** Negative Binomial distribution $x \sim \mathrm{Nb}(r, \theta)$ has PMF: $\mathrm{Nb}(x|r, \theta) = \binom{r+x-1}{r-1}\theta^r(1-\theta)^x1_{\mathbb{N}}(x)$ 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 $\lambda$, and identify the parametric family of distributions as $\lambda \sim \mathrm{Ga}(a, b)$, with $a > 0$, and $b > 0$. While you are deriving the conjugate prior distribution of $\lambda$, 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 $\lambda$, identify the posterior distribution as a Gamma distribution $\mathrm{Ga}(\tilde{a}, \tilde{b})$, 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 $\theta$ 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

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

The sampling space $\mathcal{X}$ does not depend on the uncertain parameter $\lambda$ 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 $\lambda$, (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

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

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 $\tau_0$ (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 $\tau_1$ (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 54.** ($\star\star$)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:** $\mathrm{Mu}_k$ 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 $\theta$, 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 $\theta | x_{1:n} \sim \text{Di}_k(\tilde{a})$.

---