

## Exercise Sheet Handout: Bayesian Statistics

Lecturer: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Author: Georgios P. Karagiannis

### Part I

## Random variables

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

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

**Exercise 2.** (\*) Let  $y \in \mathcal{Y} \subseteq \mathbb{R}$  be a univariate random variable with PDF  $f_Y(\cdot)$ . Consider a bijective function  $h : \mathcal{Y} \rightarrow \mathcal{Z} \subseteq \mathbb{R}$  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

$$f_Z(z) = f_Y(y) \left| \det\left(\frac{dy}{dz}\right) \right| = f_Y(h^{-1}(z)) \left| \det\left(\frac{d}{dz} h^{-1}(z)\right) \right|$$

**Exercise 3.** (\*) Prove the following properties

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

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

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

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

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

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

for any functions  $\psi_1$  and  $\psi_2$  defined on  $\mathcal{Z}$  and  $\mathcal{Y}$ .

**Exercise 4.** (\*) Prove the following properties of the covariance matrix

1.  $\text{Cov}(z, y) = \mathbb{E}(zy^\top) - \mathbb{E}(z)(\mathbb{E}(y))^\top$

$$2. \text{Cov}(z, y) = (\text{Cov}(y, z))^{\top}$$

$$3. \text{Cov}_{\pi}(c_1 + A_1 z, c_2 + A_2 y) = A_1 \text{Cov}_{\pi}(x, y) A_2^{\top}, \text{ for fixed matrices } A_1, A_2, \text{ and vectors } c_1, c_2 \text{ with suitable dimensions.}$$

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

---

**Exercise 5. (★)** Prove that the  $(i, j)$ -th element of the covariance matrix between vector  $z$  and  $y$  is the covariance between their elements  $z_i$  and  $y_j$ :

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


---

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

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

$$2. \text{Var}(c + Ay) = A \text{Var}(y) A^{\top}, \text{ for fixed matrix } A, \text{ and vectors } c \text{ with suitable dimensions.}$$

$$3. \text{Var}(y) \geq 0; \text{ (semi-positive definite)}$$


---

**Exercise 7. (★)** Prove the following properties of characteristic functions

$$1. \varphi_{A+Bx}(t) = e^{it^T A} \varphi_x(B^T t) \text{ if } A \in \mathbb{R}^d \text{ and } B \in \mathbb{R}^{k \times d} \text{ are constants}$$

$$2. \varphi_{x+y}(t) = \varphi_x(t) \varphi_y(t) \text{ if and only if } x \text{ and } y \text{ are independent}$$

$$3. \text{ if } M_x(t) = \mathbb{E}(e^{t^T x}) \text{ is the moment generating function, then } M_x(t) = \varphi_x(-it)$$


---

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

---

**Exercise 9. (★)**

$$1. \text{ Find } \varphi_X(t) \text{ if } X \sim \text{Br}(p).$$

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


---

**Exercise 10. (★★)** Prove the following statement related to the Bayesian theorem:

Assume a probability space  $(\Omega, \mathcal{F}, P)$ . Let a random variable  $y : \Omega \rightarrow \mathcal{Y}$  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

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)dF(\theta)} \quad (1)$$

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

$$\int_A f(x)g(x)dx = f(\xi) \int_A g(x)dx$$

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

59

---

60 **Exercise 11.** (★) Prove that:61 1. if  $Z \sim N(0, I)$  then  $\varphi_Z(t) = \exp(-\frac{1}{2}t^T t)$ , where  $Z \in \mathbb{R}^d$ 62 2. if  $X \sim N(\mu, \Sigma)$  then  $\varphi_X(t) = \exp(it^T \mu - \frac{1}{2}t^T \Sigma t)$ , where  $X \in \mathbb{R}^d$ 63 **Hint:** Assume as known that if  $Z \sim N(0, 1)$  then  $\varphi_Z(t) = \exp(-\frac{1}{2}t^2)$ , where  $Z \in \mathbb{R}$ 64

---

65 **Exercise 12.** (★) Show the following properties of the Characteristic Function66 1.  $\varphi_x(0) = 1$  and  $|\varphi_x(t)| \leq 1$  for all  $t \in \mathbb{R}^d$ 67 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 constants68 3.  $x$  and  $y$  are independent then  $\varphi_{x+y}(t) = \varphi_x(t) \varphi_y(t)$  (we do not prove the other way around)69 4. if  $M_x(t) = E(e^{t^T x})$  is the moment generating function, then  $M_x(t) = \varphi_x(-it)$ 70

---

## Part II

# Probability calculus

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

**Exercise 14.** (★★★)

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

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

1. This dependency is denoted as

$$\begin{aligned}x|z &\sim F(x|z) \\ z &\sim \Pi(z)\end{aligned}$$

2. Distribution  $F(x|z)$  is called conditional or parametrized distribution
3. Distribution  $\Pi(z)$  is called mixing or latent distribution;  $z$  is called mixing or latent variable.
4. Distribution  $G(x)$  that results by marginalizing (integrating) the conditional distribution  $F(x|z)$  with respect to  $\Pi(z)$  as

$$G(x) = \int F(x|z) d\Pi(z)$$

is called the compound (mixture) distribution of  $x$ . The compound distribution  $G(x)$  has PDF/PMF

$$g(x) = \int f(x|z) d\Pi(z) = \begin{cases} \int g(x|z) \pi(z) dz & , z \text{ cont.} \\ \sum_{\forall z} g(x|z) \pi(z) & , z \text{ discr.} \end{cases}$$

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

**Proposition.** From the Lecture notes, we get

$$\begin{aligned}E_G(x) &= E_{\Pi}(E_F(x|z)) \\ \text{Var}_G(x) &= E_{\Pi}(\text{Var}(x|z)) + \text{Var}_{\Pi}(E_F(x|z))\end{aligned}$$

---

<sup>a</sup>From WIKIPEDIA

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

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

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

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

2. Show that

$$E_{F(n, m)}(x) = \frac{m}{m-2}$$

3. Show that

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

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

---

**Exercise 15. (★★)** Prove the following statement:

Let  $x \sim N_d(\mu, \Sigma)$ ,  $x \in \mathbb{R}^d$ , and  $y = (x - \mu)^\top \Sigma^{-1} (x - \mu)$ . Then

$$y \sim \chi_d^2$$

---

**Exercise 16. (★★)** Let

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

with PDF

$$\begin{aligned} f_{N_d(\mu, \Sigma\xi)}(x|\xi) &= (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu)\right) \\ f_{\text{IG}(a, b)}(\xi) &= \frac{b^a}{\Gamma(a)} \xi^{-a-1} \exp\left(-\frac{b}{\xi}\right) \mathbf{1}_{(0, \infty)}(\xi) \end{aligned}$$

Show that the marginal PDF of  $x$  is

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

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

---

The Following one will be given as Homework

**Exercise 17. (★★★)**

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

$$x|\xi \sim \mathbf{N}_d(\mu, \Sigma\xi v)$$

$$\xi \sim \mathbf{IG}(\frac{v}{2}, \frac{1}{2})$$

Consider partition such that

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

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

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

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

2. Show that

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

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

**Hint:** The PDF of  $y \sim \mathbf{N}_d(\mu, \Sigma)$  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  $y \sim \mathbf{IG}(a, b)$  is

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

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

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

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

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

where

$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$$

$$\dot{\Sigma}_{2|1} = \frac{\nu + (\mu_1 - x_1)^\top \Sigma_1^{-1}(\mu_1 - x_1)}{\nu + d_1} \Sigma_{2|1}$$

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top$$

$$\nu_{2|1} = \nu + d_1$$

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

**Exercise 18.** (★★) Show that

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

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

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

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

## Part III

# Exchangeability

We work on the proofs of the following theorems:

- Marginal distributions of finite exchangeable sequences  $y_1, y_2, \dots, y_k$  are invariant under permutations; i.e.:

$$dF(y_{p(1)}, y_{p(2)}, \dots, y_{p(k)}) = dF(y_1, y_2, \dots, y_k) \text{ for all } p \in \mathfrak{P}_n. \quad (3)$$

In particular, for  $k = 1$ , it follows that all  $y_i$  are identically distributed (but not necessarily independently, as stated in the Lecture notes)

- (Marginal) Expectations of finite exchangeable sequences  $y_1, y_2, \dots, y_k$  are all identical:

$$E(g(y_i)) = E(g(y_1)) \text{ for all } i = 1, \dots, k \text{ and all functions } g: \mathcal{Y} \rightarrow \mathbb{R} \quad (4)$$

- (Marginal) Variances of finite exchangeable sequences  $y_1, y_2, \dots, y_k$  are all identical:

$$\text{Var}(y_i) = \text{Var}(y_1). \quad (5)$$

- Covariances between elements of finite exchangeable sequences  $y_1, y_2, \dots, y_k$  are all identical:

$$\text{Cov}(y_i, y_j) = \text{Cov}(y_1, y_2) \text{ whenever } i \neq j. \quad (6)$$

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

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

$$E(g(Y_{p(1)}, Y_{p(2)}, \dots, Y_{p(k)})) = E(g(Y_1, Y_2, \dots, Y_k)) \quad (7)$$

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

**Exercise 19.** (\*\*\*\*)(Representation Theorem with 0 – 1 quantities). If  $y_1, y_2, \dots$  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  $p(y_1, \dots, y_n)$  for  $y_1, \dots, y_n$  has the form

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

where

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

aka  $\theta$  is the limiting relative frequency of 1s, by SLLN

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

**Exercise 20.** (\*\*) Clearly a set of independent and identically distributed random variables form an exchangeable sequence. Thus sampling with replacement generates an exchangeable sequence. What about sampling without replacement? Prove that sampling  $n$  items from  $N$  distinct objects without replacement (where  $n \leq N$ ) is exchangeable.

**Exercise 21.** (\*\*) Let  $Y_1, \dots, Y_n$  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  $E(g(Y_i))$  does not depend on  $i$ :

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

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

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

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

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

$$\text{Var}(Y_i) = \text{Var}(Y_1) \text{ for all } i \in \{2, \dots, n\} \quad (10)$$

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

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

**Exercise 24.** (\*) What does

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



imply about the correlation of infinite exchangeable sequences?

## Part IV

# Bayesian Calculations

**Exercise 25.** (★) Consider an i.i.d. sample  $y_1, \dots, y_n$  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

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

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  $z = y_{n+1}$ .

**Exercise 26.** (★★)(Nuisance parameters are involved)

<-story

Assume observable quantities  $y = (y_1, \dots, y_n)$  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  $N(\mu, \sigma^2)$  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

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

1. Show that

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

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

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

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

with

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

$$k_n = k_0 + \frac{1}{2}ns_n^2 + \frac{1}{2}\frac{\tau_0n(\mu_0 - \bar{y})^2}{n + \tau_0}$$

**Hint:** It is

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

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

$$\mu|y \sim T_1 \left( \mu_n, \frac{k_n}{a_n} \frac{1}{\tau_n}, 2a_n \right)$$

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

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

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

$$z|y \sim T_1 \left( \mu_n, \frac{k_n}{a_n} \left( \frac{1}{\tau_n} + 1 \right), 2a_n \right)$$

**Hint-1:** Consider that

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

where

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

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

---

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

**Exercise 27.** (★★)(Normal linear regression model with unknown error variance)

<-story

Consider we are interested in recovering the mapping

$$x \xrightarrow{\eta(x)} y$$

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

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

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

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

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

where  $\epsilon_i \sim N(0, \sigma^2)$  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

$$y|\beta, \sigma^2 \sim N(\Phi\beta, I\sigma^2); \text{ the sampling distr}$$

$$\beta|\sigma^2 \sim N(\mu_0, V_0\sigma^2); \text{ prior distr}$$

$$\sigma^2 \sim \text{IG}(a_0, k_0) \text{ prior distr}$$

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

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

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

with

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

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

**Hint-1:**

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

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

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

where

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

2. Show that the marginal posterior of  $\beta$  given  $y$  is

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

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

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

Consider that

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

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

## Part V

# Sufficiency

**Exercise 28.** (\*\*) Consider the Bayesian model

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

**Hint-1:** The PDF of  $x \sim G(a, b)$  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  $x = (x_1, \dots, x_n)$ .
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  $x = (x_1, \dots, x_n)$ , has PDF

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

**Exercise 29.** (\*\*\*) Consider the Bayesian model

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

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

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

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

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

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

1. Show that the parametric model (12) is a member of the  $k - 1$  exponential family.
2. Compute the likelihood  $f(x_{1:n}|\theta)$ , and find the sufficient statistic  $t_n := t_n(x_{1:n})$ .
3. Compute the posterior distribution. State the name of the distribution, and expresses its parameters with respect to the observations and the hyper-parameters of the prior. Justify your answer.
4. Compute the probability mass function of the predictive distribution for a future observation  $y = x_{n+1}$  in closed form.

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

---

**Exercise 30.** (\*\*) 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

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

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

---

**Exercise 31.** (\*) Establish the formula

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

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

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

---

The following is a proof of a theorem

**Exercise 32.** (\*\*\*) (This is a theorem in Handout 5) Prove the following statement.

Let  $t : \mathcal{Y} \rightarrow \mathcal{T}$  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

$$L(y; \theta) = k(t(y)|\theta)\rho(y). \quad (13)$$

---

The following is a proof of a theorem

**Exercise 33.** (\*\*\*) Let  $y_1, y_2, \dots$  be an infinitely exchangeable sequence of random quantities. Let  $t = t(y_1, \dots, y_n)$  be a statistic for a finite  $n \geq 1$ . Then  $t$  is predictive sufficient if, and only if, it is parametric sufficient.

## Part VI

# Priors

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

**Exercise 34.** (★★) Assume that observable quantities  $x_1, x_2, \dots$  are generated i.i.d by a process that can be modeled as a sampling distribution  $N(\mu, \sigma^2)$  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  $\mu \sim N(\mu_0, \sigma_0^2)$  where  $\mu_0, \sigma_0^2$  are known.

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

Next assume that you additionally observe 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  $\mu \sim N(\mu_0, \sigma_0^2)$  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

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

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

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

**Exercise 35.** (★★) If the sampling distribution  $F(\cdot|\theta)$  is discrete and the prior  $\Pi(\theta)$  is proper, then the posterior  $\Pi(\theta|y)$  is always proper.

**Exercise 36.** (★★) 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.

**Exercise 37.** (★★) Let  $y = (y_1, \dots, y_n)$  be observable quantities, generated from an exponential family of distributions as

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

with density

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

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

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

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

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

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

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

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

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

---

### Exercise 38. (★★)

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

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

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

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

with hyperparameters  $w_0 := (\alpha_0, \beta_0)$  (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  $x_1, \dots, x_n$  from the skew-logistic distribution, and assuming that the prior is  $\text{Gamma}(\alpha_0, \beta_0)$ , derive the posterior distribution of  $\theta$ .
4. Given an i.i.d. sample  $x_1, \dots, x_n$  from the skew-logistic distribution, and assuming that the prior is  $\text{Gamma}(\alpha_0, \beta_0)$ , 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?

---

The exercise below is theoretical, and very challenging.

**Exercise 39.** (★★) 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 member 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 (!!!)

**Exercise 40.** (★★) 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  $\theta$ ?

**Exercise 41.** (★★) Find a (two-dimensional) sufficient statistic for  $(\alpha, \beta)$  given an  $n$ -sample  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  from the two-parameter gamma distribution

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

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

**Exercise 42.** (★★) Let  $y = (y_1, \dots, y_n)$  be observables drawn iid from sampling distribution  $y_i|\theta \stackrel{\text{iid}}{\sim} N(\theta, \theta^2)$  for all  $i = 1, \dots, n$ , where  $\theta \in \mathbb{R}$  is unknown. Specify a conjugate prior density for  $\theta$  up to an unknown normalizing constant.

**Exercise 43.** (★★) Suppose that your prior for  $\theta$  with PDF

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

that a single observation  $x \sim N(\theta, 1)$  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, :

$$\begin{aligned} -\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu_i)^2}{\sigma_i^2} &= -\frac{1}{2} \frac{(y - \hat{\mu})^2}{\hat{\sigma}^2} + c(\hat{\mu}, \hat{\sigma}^2), \\ c(\hat{\mu}, \hat{\sigma}^2) &= -\frac{1}{2} \sum_{i=1}^n \frac{\mu_i^2}{\sigma_i^2} + \frac{1}{2} \left( \sum_{i=1}^n \frac{\mu_i}{\sigma_i^2} \right)^2 \left( \sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1}; \\ \hat{\sigma}^2 &= \left( \sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1} \quad ; \quad \hat{\mu} = \hat{\sigma}^2 \left( \sum_{i=1}^n \frac{\mu_i}{\sigma_i^2} \right) \end{aligned}$$

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

The Exercise below has been set as Homework 2.



**Exercise 44.** (★★) Let  $x = (x_1, \dots, x_n)$  be observables. Consider a Bayesian model such as

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

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

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

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

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  $\lambda \sim \text{Ga}(a, b)$ , 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.

---

The Exercise below has been set as Homework 2.

**Exercise 45.** (★★) Let  $x = (x_1, \dots, x_n)$  be observables. Consider a Bayesian model such as

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

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

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

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

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  $\lambda \sim \text{Ga}(a, b)$ , 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.

**Exercise 46.** (★★) Consider the Bayesian model

$$\begin{cases} x_i | \theta & \stackrel{\text{iid}}{\sim} \text{Mu}_k(\theta) \\ \theta & \sim \Pi(\theta) \end{cases}$$

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

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

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

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

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

1. Derive the conjugate prior distribution for  $\theta$ , 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.

**Exercise 47.** (★★) Consider the Bayesian model

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

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

**Exercise 48.** (★★) Find the Jeffreys prior density for the parameter  $\theta$  of the Maxwell distribution

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

**Exercise 49.** (★★) Consider the trinomial distribution

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

Specify a Jeffreys' prior for  $(\pi, \rho)$ .

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

**Exercise 50.** (★★) Suppose that  $x$  has a Pareto distribution  $\text{Pa}(\xi, \gamma)$  where  $\xi$  is known but  $\gamma$  is unknown, that is,

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

Use Jeffreys' rule to find a suitable reference prior for  $\gamma$ .

**Exercise 51.** (★★) Consider the Bayesian model

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

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

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

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

$$\begin{aligned} \bullet F^{(0)}(\alpha) &= \frac{d}{d\alpha} \log(\Gamma(a)) \\ \bullet F^{(1)}(\alpha) &= \frac{d^2}{d\alpha^2} \log(\Gamma(a)) \end{aligned}$$

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

**Exercise 52.** (★★) Consider the Bayesian model

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

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.

**Exercise 53.** (★★★) Consider the the model of Normal linear regression where the observables are pairs  $(\phi_i, y_i)$  for  $i = 1, \dots, n$ , assumed to be modeled according to the sampling distribution

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

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

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

where  $y = (y_1, \dots, y_n)$ , and  $\Phi$  is the design matrix. Here  $\beta$  is  $d$ -dimensional. Find the Jeffreys' priors for  $(\beta, \sigma^2)$ .

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

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

This is given as an Homework 3

**Exercise 54.** (\*\*) Consider the Bayesian model

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

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

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

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

This is given as an Homework 3

**Exercise 55.** (\*\*) Consider the Bayesian model

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

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

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

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

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

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

### 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

$$\begin{aligned} \int_a^\infty g(x) dx < \infty &\implies \int_a^\infty f(x) dx < \infty \\ \int_a^\infty f(x) dx = \infty &\implies \int_a^\infty g(x) dx = \infty \end{aligned}$$

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

Let

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

Then

- If  $c \in (0, \infty)$  :

$$\int_a^\infty g(x)dx < \infty \iff \int_a^\infty f(x)dx < \infty$$

- If  $c = 0$  :

$$\int_a^\infty g(x)dx < \infty \implies \int_a^\infty f(x)dx < \infty$$

- If  $c = \infty$  :

$$\int_a^\infty f(x)dx = \infty \implies \int_a^\infty g(x)dx = \infty$$

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

Let

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

Then

- If  $c \in (0, \infty)$  :

$$\int_a^\infty g(x)dx < \infty \iff \int_a^\infty f(x)dx < \infty$$

- If  $c = 0$  :

$$\int_a^\infty g(x)dx < \infty \implies \int_a^\infty f(x)dx < \infty$$

- If  $c = \infty$  :

$$\int_a^\infty f(x)dx = \infty \implies \int_a^\infty g(x)dx = \infty$$

**Note:** A useful test function is

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

**Exercise 56.** (★★) Consider the Bayesian model

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

where  $\text{Ex}(\lambda)$  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  $x = 0$ .

This is given as an Homework 3

**Exercise 57.** (★★) Let  $x$  be an observation. Consider the Bayesian model

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

where  $\text{Pn}(\theta)$  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:  $x \sim \text{Pn}(\theta)$  has PMF

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

---

## Part VII

# Decision theory

**Exercise 58.** (★★) Show that, under the squared error loss, if two unbiased and independent real estimators  $\delta_1$  and  $\delta_2$  are distinct and satisfy

$$R(\theta, \delta_1) = E_F(\theta - \delta_1(x))^2 = R(\theta, \delta_2) = E_F(\theta - \delta_2(x))^2,$$

the estimator  $\delta_1$  is not admissible.

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

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

$$g(E_F(X)) \leq E_F(g(X))$$

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

---

## Part VIII

# Point estimation

**Exercise 59.** (★★) Consider a Bayesian model

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

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

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

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

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

---

**Exercise 60.** (★★) Consider a Bayesian model

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

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

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

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

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

**Hint** The predictive pmf of  $y$  is

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

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

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

---

**Exercise 61.** (\*\*) Consider observables  $x = (x_1, \dots, x_n)$ . Consider the Bayesian model

$$\begin{cases} x_i | \theta & \stackrel{\text{iid}}{\sim} \text{N}(\theta, 1), & i = 1, \dots, n \\ \theta & \sim \Pi(\theta) \end{cases}$$

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

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

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

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

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

**Hint-3:** It is

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

where

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

---

**Exercise 62.** (\*\*) Suppose we wish to estimate the values of a collection of discrete random variables  $\vec{X} = X_1, \dots, X_n$ . We have a posterior joint probability mass function for these variables,  $p(\vec{x}|y) = p(x_1, \dots, x_n|y)$  based on some data  $y$ . We decide to use the following loss function:

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

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

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

The following exercise is given as a Homework 4

**Exercise 63.** (★★) <sup>1</sup>This exercise is based on a problem that arises in image processing. Look at the first row of Fig 1. If we were to observe the sunflower field from above, the sunflowers would be spread uniformly over it. Viewed from an angle, the sunflowers cluster at the top of the picture due to the effect of perspective. We would like to be able to tell from this clustering at what angle the camera was pointing and its height above the ground. We will not solve this problem here (it is rather difficult in general), but instead look at an idealized and simplified version of it.

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

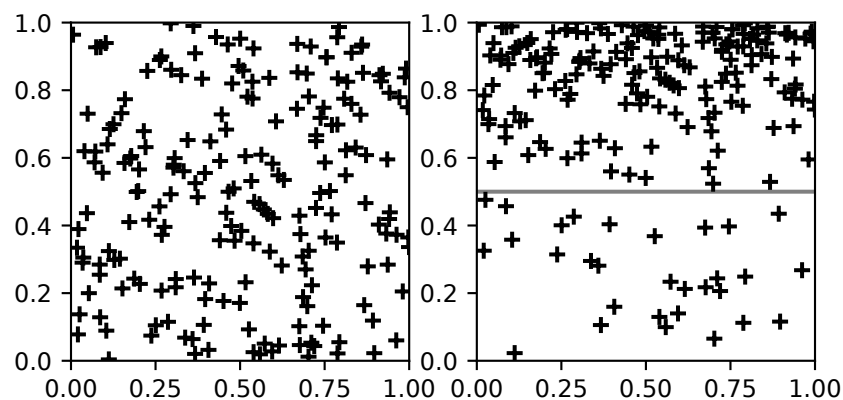


Figure 1: Two sampled point patterns. (Sunflower image © Soren Breiting / Alamy.)

If we were asked whether the right hand image had been sampled from a uniform distribution on the unit square, I am sure we would all say ‘no’. The question is how we can justify this response. The first part of the exercise is

<sup>1</sup>This exercise is a modified version of the one from Dr Jermyn’s lecture notes in Bayesian statistics 2015-2016



an alternative to the examples given in lectures, showing the need to use Bayesian and not ad hoc methods to obtain sensible answers in many inference problems.

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

### 1. Classical treatment.

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

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

- (a) If we assume that both point patterns were sampled from a uniform distribution on the unit square, which of the two sets of points is more probable? Does this help to justify the inference that the right hand image was not sampled from a uniform distribution?
- (b) If the total number of points is  $n$ , what is the probability distribution for  $r$  (the 'sampling distribution') under the assumption of sampling from the uniform distribution on the unit square?
- (c) What are the mean and variance of this distribution?
- (d) For the right hand image in the bottom row of Fig 1, by visual inspection, extract (approximately) the value of the statistic  $r$ .
- (e) Using a normal approximation to the sampling distribution, perform a significance test under the null hypothesis  $H_0$  that the sampling was uniform. What is your conclusion?
- (f) How do we reconcile the answer to Sub-question 1a with the result of the hypothesis test? When we calculate the probability of observing  $r$  points in the bottom half of the unit square, what are we actually calculating?
- (g) What would be the result of the hypothesis test if we defined  $r$  as the number of points in the left half of the unit square?
- (h) Is this a reasonable thing to do? Why?

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

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

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

### 2. Bayesian treatment.

To disallow these extreme possibilities (if indeed they are unreasonable), we have to assign probabilities to the possible non-uniformities. One way to do this is via a choice of a parameterized family of non-

uniformities. Any parameterized model implicitly assigns probability zero to any non-family member, and hence is Bayesian by default.

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

$$x = u \quad (19)$$

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

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

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

$$f(u, v|t) = 1 \quad (21)$$

- (b) Write down the corresponding probability density, given  $t$ , of a set of points  $(x_1, y_1), \dots, (x_n, y_n)$  sampled independently from the non-uniform density.
- (c) Suppose we have no reason to favour any particular value of  $\alpha \in [0, \pi/2]$  before we see the data. Write down the prior probability density for  $\alpha$ .
- (d) From the answer to Sub-question 2c, derive the prior probability density of  $t$ . [Hint:  $\frac{d}{dt}(\tan^{-1} t) = \frac{1}{1+t^2}$ .]
- (e) Hence write down the posterior probability density for  $t$  up to an overall normalization factor, given the data points  $(x_1, y_1), \dots, (x_n, y_n)$ .
- (f) From this result, derive the equation satisfied by the MAP estimate of  $t$ . (Taking logarithms makes things easier.)
- (g) By expanding the log posterior probability density about 0 to second order in  $t$ , find the MAP estimate for  $t$  when  $t$  is small.
- (h) Find the MAP estimate of  $\alpha$  when  $\alpha$  is small.

## Solution.

### 1. Classical treatment.

- (a) The probability that a single point falls in the infinitesimal element  $du dv$  at point  $(u, v)$  is:

$$dF(u, v) = P_F(u \in du, v \in dv) = du dv \quad (22)$$

The probability that  $n$  points sampled independently fall in the infinitesimal elements  $du_1 dv_1, \dots, du_n dv_n$  at points  $(u_1, v_1), \dots, (u_n, v_n)$  is therefore

$$dF(u_1, v_1, \dots, u_n, v_n) = \prod_{i=1}^n dF(u_i, v_i) = \prod_{i=1}^n du_i dv_i \quad (23)$$

This does not depend on the data points  $(u_i, v_i)$  and so is the same for both patterns.<sup>2</sup>

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

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

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

- (c) The mean of a binomial distribution is  $np$  and the variance is  $np(1 - p)$ , meaning, in this case, that the mean is  $\frac{n}{2}$  and the variance is  $\frac{n}{4}$ . For the given example, these are mean 100 and standard deviation  $\sqrt{50} \simeq 7$ .
- (d) There are about  $r = 30$  points in the bottom half of the square. (The exact number is not so relevant here.)
- (e) The standardized value of the statistic  $r$  is  $z = (r - 100)/7$ , so for  $r = 30$  we get  $z = -10$ , i.e. 10 standard deviations below the mean. The probability of finding this or a more extreme value in the bottom half of the square is then  $2(1 - \Phi(10))$ , a number that is vanishingly small. The null hypothesis is thus convincingly rejected.
- (f) Superficially there seems to be a contradiction between the fact that the probabilities of the two sets of data are the same, but the hypothesis test so strongly rejects the null hypothesis. In fact, of course, the probability of the data and the probability computed in the hypothesis test are completely different. The former is the probability of a particular set of positions for points. The latter is a different in two ways. Remember that the probability of an individual configuration is constant under the uniform hypothesis. Then first, the probability of a particular value of  $r$  is given by the integral of this constant over all possible positions of the points that keeps the same number in the bottom half of the square. Second, the probability in the hypothesis test is then the sum of these probabilities for all values of  $r$  ‘more extreme’, i.e. further from the mean, than the value we observe. For the hypothesis test, then, we calculate the probability of a large, indeed in this case infinite, set of conceivable data sets, none of which we have actually observed.
- Naturally, with such a difference in the probabilities, the conclusions to be drawn are different.
- (g) The test does not justify rejecting the null hypothesis.
- (h) It seems unreasonable because we know or suspect that the non-uniformity is in the vertical direction, and the statistic should be some measure of this non-uniformity. However, this alternative hypothesis has no place in the classical statistical testing methodology: all we have is  $H_0$ .

## 2. Bayesian treatment.

---

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

## Part IX

### Credible regions

**Exercise 64.** (\*\*\*A random sample,  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , of size  $n$  is taken from  $N(\theta, \sigma^2)$ , where  $\sigma^2$  is known. Suppose there exist positive constants;  $a, \epsilon, M$  and  $c$  (small values of  $a$  and  $\epsilon$  are of interest), such that in the interval  $I_a$ , defined by

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

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

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

inside  $I_a$ , and

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

outside  $I_a$ .

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

---

**Exercise 65.** (\*\*\*Consider the Bayesian model

$$\begin{cases} x_i | \sigma^2 \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2), & i = 1, \dots, n \\ \sigma^2 \sim \text{IG}(a, b) \end{cases} \quad (25)$$

where  $\mu$  is known, and the prior hyper-parameters  $a, b$  are known.

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

**Hint:** It is

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

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

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

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

Justify your calculations with brief comments.

3. Find the  $C_\gamma$  (posterior) parametric HPD credible interval for  $\sigma^2$ . In particular, find the system of equations whose roots are the boundaries of the HPD credible interval. Justify your calculations with brief comments.
4. Find the  $C_\gamma$  predictive HPD credible interval for a future  $y = x_{n+1} \in \mathbb{R}$ . Justify your calculations with brief comments.

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

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

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

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

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

## Part X

# Hypothesis tests

**Exercise 66.** (★★) Consider a Bayesian model

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

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

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

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

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

- a priori the probability that  $\{\lambda = 2\}$  is 0.5,
- given  $\{\lambda \neq 2\}$ , the prior distr. of  $\lambda$  is a conjugate one with  $E(\lambda) = 2$ , and  $Var(\lambda) = 1$ .

---

**Exercise 67.** (★★) Let  $y = (y_1, \dots, y_n)$  observables and consider the Bayesian hypothesis test

$$H_0 : y_i | \theta_0 \stackrel{\text{iid}}{\sim} N(\theta_0, \sigma^2), i = 1, \dots, n \quad \text{vs} \quad H_1 : \begin{cases} y_i | \theta \stackrel{\text{iid}}{\sim} N(\theta, \sigma^2), i = 1, \dots, n \\ \theta \sim N(\mu_0, \sigma_0^2) \end{cases}$$

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

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

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

2. Let  $P_\Pi(H_0|y)$  denote the posterior probability of the null hypothesis  $H_0$ . Calculate:

$$P_\Pi(H_0|y) = \left(1 + \frac{1 - \pi_0}{\pi_0} \frac{\left(\frac{\sigma^2}{n} + \sigma_0^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \frac{(\bar{y} - \mu_0)^2}{\frac{\sigma^2}{n} + \sigma_0^2}\right)}{\left(\frac{\sigma^2}{n}\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \frac{(\bar{y} - \theta_0)^2}{\frac{\sigma^2}{n}}\right)}\right)^{-1}$$

**Hint-1:** It is

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

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

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

---

**Exercise 68.** (★★) Recall the linear regression model mapping in a linear form the dependent variable  $y$  with a set of regressors  $\{\Phi_j\}_{j \in \mathcal{M}}$  where  $\mathcal{M}$  is the set of size  $d$  that includes the labels of the available regressors; e.g.

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

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

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

processes ‘better’. I.e. we test

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

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

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

2. Calculate the posterior marginal probability  $P_{\Pi}(\mathbf{H}_0|y)$  as

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

where for  $j = 0, 1$

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

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

**Hint:** You may use the following identity:

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

$$S^* = \mu^\top V^{-1} \mu - (\mu^*)^\top (V^*)^{-1} (\mu^*) + y^\top y; \quad V^* = (V^{-1} + \Phi^\top \Phi)^{-1}; \quad \mu^* = V^* (V^{-1} \mu + \Phi^\top y)$$

## Part XI

# Inference under model uncertainty

**Exercise 69.** (\*\*) Given a finite collection of models  $\{\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_K\}$  the marginal model posterior probability of  $\mathcal{M}_k$  can be calculated from Bayes factors as

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

**Exercise 70.** (\*\*) Let  $B_{k,j}(y)$  be the Bayes factor of model  $\mathcal{M}_k$  against model  $\mathcal{M}_j$ , for all  $\forall k, i, j \in \mathcal{K}$ . . Show that  $B_{k,j}(y) = B_{k,i}(y) B_{i,j}(y)$ , for all  $\forall k, i, j \in \mathcal{K}$ .

816 **Exercise 71.** (★★) Consider the Bayesian model

$$817 \quad \begin{cases} y|\theta_k, \mathcal{M}_k \sim F(y|\theta_k, k) \\ \theta_k|\mathcal{M}_k \sim \Pi(\theta_k|k) \\ \mathcal{M}_k \sim \Pi(k) \end{cases} \quad (26)$$

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

821 Show that

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

$$823 \quad E_G(z|y) = E_\Pi(E_G(z|y, k) | y) = \sum_{k \in \mathcal{K}} E_G(z|y, k) \pi(k|y), \forall k \in \mathcal{K}$$

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

$$825 \quad \text{Var}_G(z|y) = \sum_{k \in \mathcal{K}} \left( \text{Var}_G(z|y, k) + (E_G(z|y, k))^2 \right) \pi(k|y) - (E_G(z|y))^2, \forall k \in \mathcal{K}$$

826

827