Bayesian Statistics III/IV

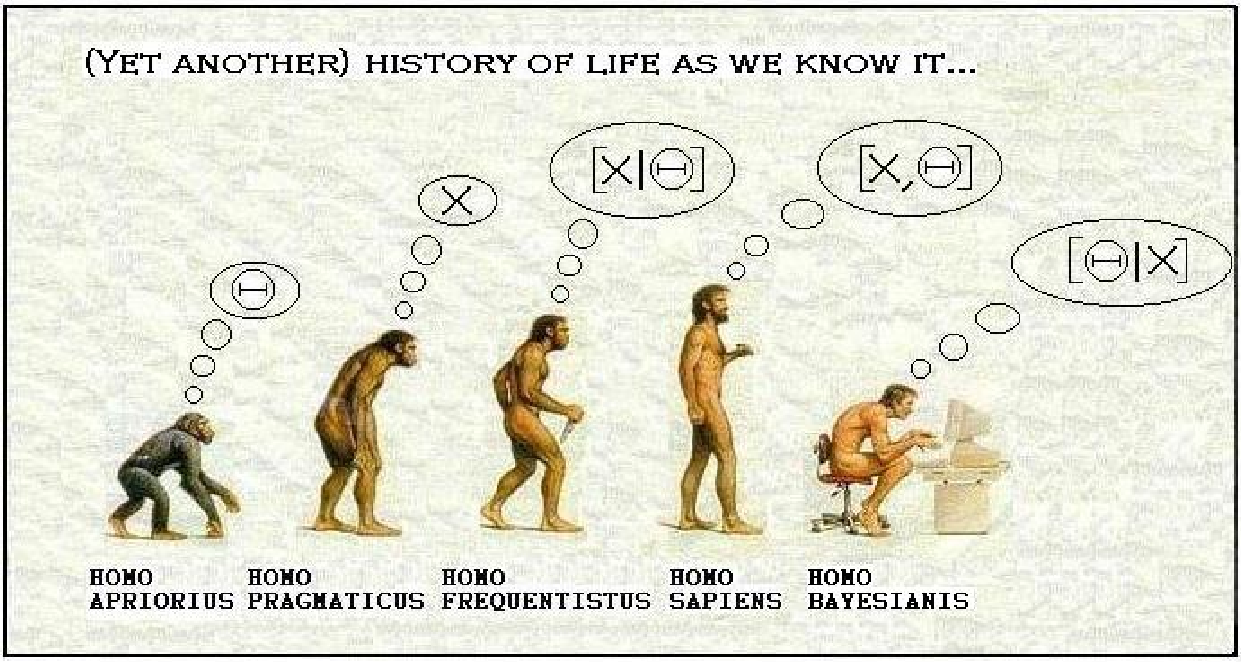
Term 1

Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Department of Mathematical Sciences (Office CM126b)
Durham University
Stockton Road Durham DH1 3LE UK

2019/12/15 at 16:11:28



Reading list

Primary:

- Robert, C. (2007). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
  - Mainly for Bayesian methods

Secondary:

- Berger, J. O. (2013). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
  - Mainly for Bayesian methods
- Raiffa, H., & Schlaifer, R. (1961). Applied statistical decision theory.
  - Mainly for foundations
- DeGroot, M. H. (2005). Optimal statistical decisions (Vol. 82). John Wiley & Sons.
  - Mainly for foundations
- O'Hagan, A., & Forster, J. J. (2004). Kendall's advanced theory of statistics, volume 2B: Bayesian inference (Vol. 2). Arnold.
  - Mainly for Bayesian methods

These lecture Handouts have been derived based on the above reading list.

Shiny applets

Here is a list of interactive shiny applications related that can be used to understand as a suplamentary material.

How to run the Web Applets from the server

You can click on the following links:

- For the demo presenting standard distributions
  - https://georgios-stats.shinyapps.io/demo_distributions/
- For the demo presenting Multivariate distributions
  - https://georgios-stats-3.shinyapps.io/demo_MultivariateNormalDistribution/
- For the demo presenting Central Limit Theorem
  - https://georgios-stats.shinyapps.io/demo_clt/
- For the demo presenting the Weak Law of Large Numbers
  - https://georgios-stats.shinyapps.io/demo_wlln/
- For the demo presenting the conjugate priors
  - https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/
- For the demo comparing Conjugate Jeffreys and Laplace priors
  - https://georgios-stats-1.shinyapps.io/demo_conjugatejeffreyslaplacepriors/
- For the demo presenting the Mixture priors
  - https://georgios-stats-1.shinyapps.io/demo_mixturepriors/
- For the demo presenting standard parametric/predictive Bayes point estimators
  - https://georgios-stats-1.shinyapps.io/demo_PointEstimation/
- For the demo presenting Credible intervals
  - https://georgios-stats-1.shinyapps.io/demo_CredibleSets/

These applications are currently uploaded on non-Durham Univertity server, which means that we have only 25 active hours per mounth. If we exceed this limit, you will be able to run these applications localy on your computer by dowlnoaded them. (see below.)

How to download the Web Applets and run them localy

In order to download, edit, run the Web Applets to your computer, do the following:

1. Run rstudio
2. In the console run
   - install.packages("rmarkdown")
3. Go to File>New Project>Version Control>Git

4. In the section "Repository URL" type:
   - https://github.com/georgios-stats/Shiny_applets.git

5. Then you can run the applications either by clicking and running each ‘name’.Rmd script in the demo_‘name’, or by running the commands:
    - For the demo presenting standard univariate distributions
        - rmarkdown::run("./demo_distributions/demo_distributions.Rmd")
    - For the demo presenting standard multivariate distributions
        - rmarkdown::run("./demo_MultivariateNormalDistribution/demo_MultivariateNormalDistribution.Rmd")
    - For the demo presenting Central Limit Theorem
        - rmarkdown::run("./demo_CLT/demo_CLT.Rmd")
    - For the demo presenting the Weak Law of Large Numbers
        - rmarkdown::run("./demo_WLLN/demo_WLLN.Rmd")
    - For the demo presenting the conjugate priors
        - rmarkdown::run("./demo_ConjugatePriors/demo_ConjugatePriors.Rmd")
    - For the demo comparing Conjugate Jeffreys and Laplace priors
        - rmarkdown::run("./demo_ConjugateJeffreysLaplacePriors/demo_ConjugateJeffreysLaplacePriors.Rmd")
    - For the demo presenting the Mixture priors
        - rmarkdown::run("./demo_MixturePriors/demo_MixturePriors.Rmd")
    - For the demo presenting standard parametric/predictive Bayes point estimators
        - rmarkdown::run("./demo_PointEstimation/demo_PointEstimation.Rmd")
    - For the demo presenting Credible intervals
        - rmarkdown::run("./demo_CredibleSets/demo_CredibleSets.Rmd")

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Lecture handout 1: Random variables [a]

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

Aim

To revise a bit, linear algebra, random variables and probabilities

Linear algebra Cholesky decomposition

Probability theory Random variables, probabilities, expected values, covariance/variance matrices, characteristic function, compound distribution function

Reading list:

- DeGroot, M. H. (1970, or 2005). Optimal statistical decisions (Vol. 82). John Wiley & Sons.
  - Part one: Survey of probability theory. Chapters 1-5

[a] Author: Georgios P. Karagiannis.

1 Linear Algebra

Proposition 1. *[Cholesky decomposition] Every symmetric positive definite matrix $A \in \mathbb{R}^d \times \mathbb{R}^d$ can be decomposed into a product of a unique lower triangular matrix $L \in \mathbb{R}^d \times \mathbb{R}^d$ and its transpose $L^\top$, i.e.*

$$A = LL^T$$

Matrix L is called lower triangular factor of the Cholesky decomposition, and it is often denoted as $A^{1/2} = L$.

2 Probability distributions

Definition 2. A collection $\mathscr{F} = \{A, A_1, A_2, ...\}$ of sets $A, A_1, A_2, ...$, each of which are subsets of set Ω, is called a σ-algebra if and only if

1. $\Omega \in \mathscr{F}$
2. If $A \in \mathscr{F}$, then $A^{\complement} \in \mathscr{F}$
3. If $A_1, A_2, ... \in \mathscr{F}$ is an infinite sequence of sets in $\mathscr{F}$ then $\cup_{i=1}^{\infty} A_i \in \mathscr{F}$

Definition 3. Probability distribution P on $(\Omega, \mathscr{F})$ is called a non-negative function if and only if

1. $P(\Omega) = 1$
2. If $A, B \in \mathscr{F}$ and $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$
3. If $A_1, A_2, ... \in \mathscr{F}$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$ then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Notation 4. We call the triple $(\Omega, \mathscr{F}, P)$ as probability space.

3 Random variables

Definition 5. A d-dimensional random variable y on a probability space $(\Omega, \mathscr{F}, P)$ is a function $y : \Omega \rightarrow \mathbb{R}^d$ such as, for any subset $A \subseteq \mathbb{R}^d$, it is $\{\omega \in \Omega\, :\, y(\omega) \in A\} \in \mathscr{F}$.

Definition 6. A d-dimensional random variable y on a probability space $(\Omega, \mathscr{F}, P)$, induces a probability $P_y(\cdot)$ such that, for all subsets $A \subseteq \mathbb{R}^d$,

$$P_y(y \in A) = P(\{\omega \in \Omega\, :\, y(\omega) \in A\})$$

Essentially, it induces a probability space $(\mathbb{R}^d, \mathfrak{B}, P_y)$, with $\mathfrak{B}$ a σ-algebra[1] containing sub-sets of $\mathbb{R}^d$.

Definition 7. The (cumulative) distribution function (CDF) of a d-dimensional random variable $y \in \mathcal{Y}$ is the function $F_y : \mathbb{R}^d \rightarrow [0, 1]$ such that

$$F_y(y) := F_y(y'_1, ..., y'_d) = P_y(y \in (-\infty, y'_1] \times ... \times (-\infty, y'_d])$$

Notation 8. The distribution function defines the distribution of the random variable. As $y \sim F_y$, we will denote that the random variable y follows a distribution with distribution function F_y.

Definition 9. The d-dimensional random variable $y : \Omega \rightarrow \mathcal{Y}$ with distribution F_y is discrete, if $\mathcal{Y}$ is a countable set and the distribution can be described by its Probability Mass Function (PMF)

$$f_y(y') := f_y(y'_1, ..., y'_d) = P(\{\omega \in \Omega\, :\, y(\omega) = y'\})$$

Definition 10. The d-dimensional random variable $y : \Omega \rightarrow \mathcal{Y}$ with distribution F_y is absolutely continuous, if $\mathcal{Y}$ is a uncountable set and the distribution can be described by its Probability Density Function (PDF) $f_y(y)$ such that

$$P_y(y \in A) = \underbrace{\int \cdots \int}_{A} f_y(y'_1, ..., y'_d)\mathrm{d}y'_1 \cdots \mathrm{d}y'_d, \text{ for any } A \subseteq \mathbb{R}^d.$$

or briefly $P_y(A) = \int_A f_y(y')\mathrm{d}y'$, where $\mathrm{d}y' = \prod_{j=1}^{d} \mathrm{d}y'_j$.

Fact 11. *The PDF of d-dimensional random variable $y : \Omega \rightarrow \mathcal{Y}$ with CDF F_y can be computed by the partial derivative as*

$$f_y(y) = \frac{d}{dt_1 \cdots dt_d} F_y(t_1, ..., t_d)\Big|_{t_1=y_1, \cdots t_d=y_d} \qquad \textit{if } F_y \textit{ is differential.}$$

4 Transforming

Fact 12. *Let $y \in \mathcal{Y}$ be a d-dimensional random variable with PDF $f_y(\cdot)$. Consider a bijective function $h : \mathcal{Y} \rightarrow \mathcal{Z}$ with $z = h(y)$, and h^{-1} its inverse. The PDF of z is*

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

Example 13. Let $y \sim \text{Ex}(\lambda)$ r.v. with Exponential distribution with rate parameter $\lambda > 0$, and $f_{\text{Ex}(\lambda)}(y) = \lambda \exp(-\lambda y)1(y \geq 0)$. Let $z = 1 - \exp(-\lambda y)$. Calculate the PDF of z, and recognize its distribution.

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

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

[1] ...this is not a rigorous definition; a more rigorous definition is out of the scope of the course.

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

5 Marginalizing & Integrating out

Fact 14. *Let $(n+d)$-dimensional random variable $y \in \mathcal{Y}$ with distribution $F_y(\cdot)$. Consider a partition $y=(x,\theta)$ where $x \in \mathcal{X}$ is n-dimensional and $\theta \in \Theta$ is d-dimensional . Then*

1. *the marginal CDF of x results by setting θ as ∞*

$$F_x(x) = \lim_{\theta \to \infty} F_y(x,\theta) \qquad = \lim_{\theta_1 \to \infty, \ldots, \theta_d \to \infty} F_y(x_1, \ldots x_n, \theta_1, \ldots, \theta_d)$$

2. *the marginal PDF/PMF of x results by integrating out θ (the dimensions we marginalize)*

$$f_x(x) = \begin{cases} \int_{\mathbb{R}^d} f_y(x,\theta) d\theta & \textit{if } \theta \textit{ is cont.} \\ \\ \sum_{\forall \theta \in \mathbb{R}^d} f_y(x,\theta) & \textit{if } \theta \textit{ is discr.} \end{cases} = \begin{cases} \int_{\mathbb{R}^d} f_y(x_1, \ldots x_n, \theta_1, \ldots, \theta_d) d\theta_1 \ldots d\theta_d & \textit{if } \theta \textit{ is cont.} \\ \\ \sum_{\forall \theta \in \mathbb{R}^d} f_y(x_1, \ldots x_n, \theta_1, \ldots, \theta_d) & \textit{if } \theta \textit{ is discr.} \end{cases}$$

6 Independence

Definition 15. Given a probability space $(\Omega, \mathfrak{B}, P)$, events $A, B \in \mathfrak{B}$ are independent if and only if

$$P(A \cap B) = P(A)P(B)$$

Fact 16. *Let $(n+m)$-dimensional random variable $y \in \mathcal{Y}$. Consider a partition $y=(x,z)$ where $x \in \mathcal{X}$ is n-dimensional and $z \in \mathcal{Z}$ is m- dimensional.*

- *The r.v. x and y are independent if and only if for any $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Z}$*

$$P(\{x \in A\} \cap \{z \in B\}) = P(x \in A)P(z \in B)$$

- *The r.v. x and y are independent if and only if*

$$F(x,z) = F(x)F(z), \ \forall x \in \mathcal{X}, \forall z \in \mathcal{Z}$$

where $F(\cdot)$ denotes the CDF.

- *This implies that r.v. x and z are independent if and only if*

$$f(x,z) = f(x)f(z)$$

where $f(\cdot)$ denotes the PDF/PMF.

7 Expected value

Definition 17. Expected value of the d-dimensional random variable $y \in \mathcal{Y}$ with distribution F is the d-dimensional quantity

$$\mathrm{E}(y) = \int y \mathrm{d}F(y) = \begin{cases} \int_{y \in \mathcal{Y}} y f(y) \mathrm{d}y, & \text{if } y \text{ is cont.} \\ \\ \sum_{y \in \mathcal{Y}} y f(y), & \text{if } y \text{ is discr.} \end{cases}$$

whose ith element is

$$[\mathrm{E}(y)]_i = \begin{cases} \int y_i f_{y_i}(y_i)\mathrm{d}y_i, & \text{if } y_i \text{ is cont.} \\ \\ \sum_{\forall y_i} y_i f_{y_i}(y_i), & \text{if } y_i \text{ is discr.} \end{cases}$$

for $i = 1, ..., d$. Here $f_{y_i}(y_i) = \int_{y \in \mathcal{Y}} f(y)\mathrm{d}y_1 \cdots \mathrm{d}y_{i-1}\mathrm{d}y_{i+1} \cdots \mathrm{d}y_d$ is the marginal PMF/PDF of y_i.

Fact 18. *If $y \in \mathcal{Y}$ is a d-dimensional random variable with PDF/PMF $f_y(\cdot)$, and $\psi : \mathcal{Y} \to \mathbb{R}^d$ is an integrate function with $\psi(\cdot) := (\psi_1(\cdot), ..., \psi_d(\cdot))$, then*

$$E(\psi(y)) = \begin{cases} \int \psi(y) f_y(y) dy, & \textit{if } y \textit{ is cont.} \\ \\ \sum_{\forall y} \psi(y) f_y(y), & \textit{if } y \textit{ is discr.} \end{cases}$$

with elements

$$[E(\psi(y))]_i = \begin{cases} \int \psi_i(y) f_y(y) dy, & \textit{if } y \textit{ is cont.} \\ \\ \sum_{\forall y} \psi_i(y) f_y(y), & \textit{if } y \textit{ is discr.} \end{cases}$$

Example 19. If $(q \times k)$-dimensional random variable $y \in \mathcal{Y}$ is a matrix, then its expectation $\mathrm{E}(y) = \int y \mathrm{d}F(y)$ is a matrix too

$$\mathrm{E}(y) = \begin{bmatrix} \mathrm{E}(y_{1,1}) & \cdots & \mathrm{E}(y_{1,j}) & \cdots & \mathrm{E}(y_{1,m}) \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \mathrm{E}(y_{i,1}) & \cdots & \mathrm{E}(y_{i,j}) & \cdots & \mathrm{E}(y_{i,m}) \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ \mathrm{E}(y_{n,1}) & \cdots & \mathrm{E}(y_{n,j}) & \cdots & \mathrm{E}(y_{n,m}) \end{bmatrix}$$

whose (i, j)-th element is

$$[\mathrm{E}(y)]_{i,j} = \mathrm{E}(y_{i,j}) = \begin{cases} \int y_{i,j} f_{y_{i,j}}(y_{i,j})\mathrm{d}y_{i,j}, & \text{if } y_{i,j} \text{ is cont.} \\ \\ \sum_{\forall y_{i,j}} y_{i,j} f_{y_{i,j}}(y_{i,j}), & \text{if } y_{i,j} \text{ is discr.} \end{cases}$$

for $i = 1, ..., n$, and $j = 1, ..., m$. Here $f_{y_{i,j}}(\cdot)$ is the marginal PMF/PDF of $y_{i,j}$.

Proposition 20. *The following properties are valid*

1. *Let fixed matrix/vectors A, c, and $z = c + Ay$ with suitable dimensions then*

$$E(z) = E(c + Ay) = c + AE(y)$$

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

$$E(\psi_1(z) + \psi_2(y)) = E(\psi_1(z)) + E(\psi_2(y))$$

3. *Random variables $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$ are independent if and only if*

$$E(\psi_1(z)\psi_2(y)) = E(\psi_1(z))E(\psi_2(y))$$

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

Proof. Given as Exercise 3 in the Exercise Sheet. □

8 Covariance matrix

Definition 21. The covariance matrix between random variable $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ and random variable $y \in \mathcal{Y} \subseteq \mathbb{R}^q$ is defined as the $d \times q$ matrix

$$\mathrm{Cov}(z, y) = \mathrm{E}\left((z - \mathrm{E}(z))(y - \mathrm{E}(y))^\top\right)$$

Proposition 22. *The following properties are the direct analogues of the $1D$ cases*

1. *$Cov(z, y) = E(zy^\top) - E(z)\,(E(y))^\top$*
2. *$Cov(z, y) = (Cov(y, z))^\top$*
3. *$Cov(c_1 + A_1 z, c_2 + A_2 y) = A_1 Cov(z, y) A_2^\top$, 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 $Cov(z, y) = 0$*

Proof. (1)-(3) result from the definition. (4) results from Prop 20, as $\mathrm{E}(zy^\top) = \mathrm{E}(z)\,(\mathrm{E}(y))^\top$. □

Proposition 23. *It can be seen that*

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

for all $i = 1, ..., d$, and $j = 1, ..., q$. Namely, 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.

Definition 24. The covariance matrix of random vector $y \in \mathcal{Y} \subseteq \mathbb{R}^d$ is defined as the $d \times d$ matrix $\mathrm{Var}(y)$

$$\mathrm{Var}(y) = \mathrm{Cov}(y, y) = \mathrm{E}\left((y - \mathrm{E}(y))(y - \mathrm{E}(y))^\top\right)$$

Proposition 25. *It can be seen that*

$$[Var(y)]_{i,j} = Cov(y_i, y_j) \text{ for all } i, j = 1, ..., d$$

and

$$[Var(y)]_{i,i} = Var(y_i)$$

for all $i = 1, ..., d$

Proposition 26. *The following properties are the direct analogues of the $1D$ cases*

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

Proof. Given as Exercise 6 in the Exercise Sheet. □

9 Characteristic function

Characteristic functions (CF) provide an alternative way to the probability function for describing a random variable.

Definition 27. The characteristic function of a d dimensional random variable X is

$$\varphi_x(t) = \mathrm{E}(e^{it^T x}) = \int e^{it^T x} \mathrm{d}F(x)$$

for $t \in \mathbb{R}^d$, where $e^{it^T x} = \cos(t^T x) + i \sin(t^T x)$.

Proposition 28. *Some properties of characteristic functions*

1. *$\varphi_x(t)$ exists for all $t \in \mathbb{R}^d$ and is absolutely continuous*
2. *$\varphi_x(0) = 1$ and $|\varphi_x(t)| \le 1$ for all $t \in \mathbb{R}^d$*
3. *$\varphi_{A+Bx}(t) = e^{it^TA}\varphi_x(B^Tt)$ if $A \in \mathbb{R}^d$ and $B \in \mathbb{R}^{k\times d}$ are constants*
4. *$\varphi_{x+y}(t) = \varphi_x(t)\varphi_y(t)$ if and only if x and y are independent*
5. *if $M_x(t) = E(e^{t^Tx})$ is the moment generating function, then $M_x(t) = \varphi_x(-it)$*

Proof. Given as Exercise 7 in the exercise sheet. □

Fact 29. *Two random variables correspond have equal characteristic functions if and only if they follow the same distribution. AKA: CF completely determines the probability distribution of the random variable*

Fact 30. *If $\varphi_x(t)$ is absolutely integrable, then x has PDF*

$$f(x) = \frac{1}{(2\pi)^d}\int_{-\infty}^{+\infty} e^{-it^Tx}\varphi_x(t)dt$$

Example 31. Address the following:

1. If $z \sim \mathrm{N}(0,1)$ then $\varphi_z(t) = \exp(-\frac{1}{2}t^2)$
2. If $x \sim \mathrm{N}(\mu,\sigma^2)$ then $\varphi_x(t) = \exp(i\mu - \frac{1}{2}t^2\sigma^2)$
3. If $\varphi_z(t) = \exp(-\frac{1}{2}t^2)$ then $f_{\mathrm{N}(0,1)}(z) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2)$

Solution. It is

1. It is

$$\begin{aligned}\varphi_z(t) &= \mathrm{E}(e^{itz}) = \int e^{itz}\mathrm{d}F_{\mathrm{N}(0,1)}(z) = \int e^{itz}f_{\mathrm{N}(0,1)}(z)\mathrm{d}z = \int \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2 + itz)\mathrm{d}z \\ &= \int \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}z^2 + \frac{2}{2}itz \pm \frac{1}{2}(it)^2z)\mathrm{d}z = \overbrace{\int \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(z-it)^2)\mathrm{d}z}^{=1} \times \exp(\frac{1}{2}(it)^2) \\ &= \exp(-\frac{1}{2}t^2)\end{aligned}$$

2. It is $\varphi_x(t) = \varphi_{\mu+\sigma z}(t) = \exp(i\mu)\varphi_z(\sigma t) = \exp(i\mu - \frac{1}{2}t^2\sigma^2)$.

3. It is

$$\begin{aligned}f(z) &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-itz}\varphi_z(t)\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\exp(-itz)\exp(-\frac{1}{2}t^2)\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty}\exp(-\frac{1}{2}t^2 + i^2tz)\mathrm{d}t = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}\left(t^2 - i^2z\right)^2 - z^2\right)\mathrm{d}t \\ &= \sqrt{\frac{1}{2\pi}}\overbrace{\int_{-\infty}^{+\infty}\sqrt{\frac{1}{2\pi}}\exp\left(-\frac{1}{2}\left(t - i^2z\right)^2\right)\mathrm{d}t}^{=1}\exp\left(-\frac{1}{2}z^2\right) = \sqrt{\frac{1}{2\pi}}\exp\left(-\frac{1}{2}z^2\right) = f_{\mathrm{N}(0,1)}(z)\end{aligned}$$

Theorem 32. *The distribution of a d-dimensional random variable $x \in \mathbb{R}^d$ is completely determined be the set of all 1--dimensional distributions of of linear combinations $a^\top x$, for any $a \in \mathbb{R}^d$.*

Proof. Let $y = a^\top x$, for any $a \in \mathbb{R}^d$. Then for any $s \in \mathbb{R}$

$$\varphi_y(s) = \mathrm{E}(e^{is^T y}) = \mathrm{E}(e^{is^\top a^\top x}) = \mathrm{E}(e^{i(as)^\top x}) = \mathrm{E}(e^{i\tilde{t}^\top x}) = \varphi_x(\tilde{t})$$

where $\tilde{t} = as$ is any d-dimensional vector. □

10 Conditioning

Definition 33. Assume a probability space $(\Omega, \mathscr{F}, P)$. For any sets $A, B \in \mathscr{F}$, the conditional probability of A given B is defined as

$$P(A|B) = \frac{P(B \cap A)}{P(B)} \qquad \text{if } P(B) \neq 0.$$

Definition 34. Let $y \in \mathcal{Y}$ be a random variable. Consider a partition $y = (x, \theta)$ with $x \in \mathcal{X}$ and $\theta \in \Theta$. The expected value of θ conditional that random variable $x \in B \subseteq \mathcal{X}$ is

$$\mathrm{E}(\theta|x \in B) = \frac{\mathrm{E}(\theta 1(x \in B))}{P(x \in B)}, \qquad \text{if } P(x \in B) > 0.$$

Fact 35. *Let a random variable $y \in \mathcal{Y}$ with PDF/PMF $f(\cdot)$. Consider a partition $y = (x, \theta)$ with $x \in \mathcal{X}$and $\theta \in \Theta$.*

1. *The conditional MPF/PDF and CDF of random variable θ given the random variable x*

$$f_{\theta|x}(\theta|x) = \frac{f(\theta, x)}{f(x)}, ; \quad F_{\theta|x}(\theta|x) = \begin{cases} \int_{-\infty}^{\theta_1} \cdots \int_{-\infty}^{\theta_d} f_{\theta|x}(\vartheta|x) d\vartheta, & \theta, \textit{ cont.} \\ \\ \sum_{\vartheta_1=-\infty}^{\theta_1} \cdots \sum_{\vartheta_d=-\infty}^{\theta_d} f_{\theta|x}(\vartheta|x), & \theta, \textit{ discr.} \end{cases}$$

   provided that $f(x) > 0$.

2. *The expected value of θ given the random variable x*

$$E(\theta|x) = \int \theta dF_{\theta|x}(\theta|x) = \begin{cases} \int \theta f_{\theta|x}(\theta|x) d\theta & \textit{, if } \theta \textit{ is cont.} \\ \\ \sum_{\forall \theta} \theta f_{\theta|x}(\theta|x) & \textit{, if } \theta \textit{ is discr.} \end{cases} \qquad \textit{provided that } f(x) > 0$$

Example 36. Let a random variable $y \in \mathcal{Y}$ with distribution $F(\cdot)$. Consider a partition $y = (x, \theta)^\top$ with $x \in \mathcal{X}$ and $\theta \in \Theta$. Then

1. $\mathrm{E}(\theta) = \mathrm{E}\left(\mathrm{E}(\theta|x)\right)$
2. $\mathrm{Var}(\theta) = \mathrm{E}\left(\mathrm{Var}(\theta|x)\right) + \mathrm{Var}\left(\mathrm{E}(\theta|x)\right)$

Solution.

1. It is

$$\begin{aligned} \mathrm{E}\left(\mathrm{E}(\theta|x)\right) &= \int \left(\int \theta \mathrm{d}F(\theta|x) \right) \mathrm{d}F(x) = \int\int \theta \mathrm{d}F(x, \theta) = \int\int \theta \mathrm{d}F(\theta|x) \mathrm{d}F(x) \\ &= \int \theta \left(\int \mathrm{d}F(x|\theta) \right) F(\theta) = \int \theta F(\theta) = \mathrm{E}(\theta) \end{aligned}$$

2. It is

$$\begin{aligned}
\text{Var}(\theta) &= \text{E}\left(\text{E}(\theta\theta^{\top})\right) - \text{E}\left(\theta\right)\text{E}\left(\theta\right)^{\top} \;= \text{E}\left(\text{E}(\theta\theta^{\top}|x)\right) - \text{E}\left(\text{E}(\theta|x)\right)\text{E}\left(\text{E}(\theta|x)\right)^{\top}\\
&= \text{E}\left(\text{E}(\theta\theta^{\top}|x)\right) - \text{E}\left(\text{E}(\theta|x)\text{E}(\theta|x)^{\top}\right) + \text{E}\left(\text{E}(\theta|x)\text{E}(\theta|x)^{\top}\right) - \text{E}\left(\text{E}(\theta|x)\right)\text{E}\left(\text{E}(\theta|x)\right)^{\top}\\
&= \text{E}\left(\text{E}(\theta\theta^{\top}|x) - \text{E}(\theta|x)\text{E}(\theta|x)^{\top}\right) + \text{E}\left(\text{E}(\theta|x)\text{E}(\theta|x) - \text{E}\left(\text{E}(\theta|x)\right)\text{E}\left(\text{E}(\theta|x)\right)^{\top}\right)\\
&= \text{E}\left(\text{Var}(\theta|x)\right) + \text{Var}\left(\text{E}(\theta|x)\right)
\end{aligned}$$

Conditional independence

Definition 37. Given a probability space $(\Omega, \mathfrak{B}, P)$, and events $A, B, C \in \mathfrak{B}$, A and B are conditionally independent given C if and only if

$$P(A \cap B|C) = P(A|C)P(B|C), \text{ for } P(C) > 0.$$

Fact 38. *Let $(n+m+d)$-dimensional random variable $y \in \mathcal{Y} \subseteq \mathbb{R}^{n+m+d}$. Consider a partition $y = (x, z, \theta)$ where $x \in \mathcal{X} \subseteq \mathbb{R}^n$, $z \in \mathcal{Z} \subseteq \mathbb{R}^m$, and $\theta \in \Theta \subseteq \mathbb{R}^d$.*

- *The r.v. x and y are independent given θ if and only if*

  $$F(y, z|\theta) = F(y|\theta)F(z|\theta), \; \forall x \in \mathbb{R}^n, \, \forall z \in \mathbb{R}^m$$

  where $F(\cdot|\theta)$ denotes the conditional CDF.

- *This implies that r.v. x and z are independent given θ if and only if*

  $$f(y, z|\theta) = f(y|\theta)f(z|\theta)$$

  where $f(\cdot|\theta)$ denotes the conditional PDF/PMF.

11 Inverting / updating

This offers a probabilistic mechanism for (1.) inversion $(B|A) \longmapsto (A|B)$, or (2.) updating $(A) \longmapsto (A|B)$.

Fact 39. *[Bayesian theorem with sets] Assume a probability space $(\Omega, \mathscr{F}, P)$. For any sets $A, B \in \mathscr{F}$, it is*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \qquad \text{provided that } P(B) \neq 0.$$

The extension of the Bayesian theorem to the random variables is not straightforward.

Proposition 40. *[Bayesian theorem with random variables] Let a random variable $y \in \mathcal{Y}$. Consider a partition $y = (x, \theta)$ with $x \in \mathcal{X}$ and $\theta \in \Theta$. Then the PDF/PMF of $\theta|x$ is*

$$f(\theta|x) \quad = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)dF(\theta)} \qquad = \begin{cases} \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} & \text{, if } \theta \text{ is cont.}\\[2ex] \frac{f(x|\theta)f(\theta)}{\sum_{\forall\theta} f(x|\theta)f(\theta)} & \text{, if } \theta \text{ is discr.}\end{cases}$$

Proof. Given as Exercise 10, in the Exercise Sheet. □

12 Practice

Question 41. *Try the Exercises 8, 9, 10, from the Exercise sheet.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 2: Probability calculations & Known distributions[a]

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

Aim

To practice on probability calculations. To become familiar with distributions, Inverted Gamma, multivariate Normal, and multivariate Student T distributions.
It is not required to memorize the formulas in Equations: 2, 3, 4, 6 7, 8, 9, and 10.

References:

- DeGroot, M. H. (1970, or 2005). Optimal statistical decisions (Vol. 82). John Wiley & Sons.
  - Part one: Survey of probability theory. Chapters 1-5 ; However the treatment of the Normal and Student T distributions is different than ours.
- Raiffa, H., & Schlaifer, R. (1961). Applied statistical decision theory.
  - Chapters 8.2, 8.3 ; However the treatment of the Normal and Student T distributions is different than ours.

Web-applets

- Multivariate Normal and Student T distributions:
  `https://georgios-stats-3.shinyapps.io/demo_multivariatenormaldistribution/`
  `https://github.com/georgios-stats/Shiny_applets/tree/master/demo_MultivariateNormalDistribution`

[a] Author: Georgios P. Karagiannis.

1 Inverted Gamma distribution $x|a, b \sim \text{IG}(a, b)$

Definition 1. The random variable $x \in (0, +\infty)$ follows an Inverted Gamma distribution $x \sim \text{IG}(a, b)$, if and only if $x = \frac{1}{y}$ follows a Gamma distribution, $y \sim \text{Ga}(a, b)$, with $a > 0$ and $b > 0$.

Example 2. Let $x \sim \text{IG}(a, b)$, then the PDF of x is

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

Solution. It is

$$f_{\text{IG}(a,b)}(x) = f_{\text{G}(a,b)}(\frac{1}{x}) \left| \frac{\text{d}}{\text{d}x}(\frac{1}{x}) \right| = \frac{b^a}{\Gamma(a)} \left(\frac{1}{x}\right)^{a-1} \exp(-\frac{b}{x}) 1_{(0,+\infty)} \left(\frac{1}{x}\right) \left| -\frac{1}{x^2} \right|$$

Example 3. Let a random variable $x \sim \text{IG}(a, b)$, then

$$\text{E}_{\text{IG}(a,b)}(x) = \frac{b}{a-1};\ a > 1 \qquad \text{and} \qquad \text{Var}_{\text{IG}(a,b)}(x) = \frac{b^2}{(a-1)^2(a-2)};\ a > 2$$

Solution. It is

$$\mathrm{E}_{\mathrm{IG}(a,b)}(x) = \int x f_{\mathrm{IG}(a,b)}(x)\mathrm{d}x \quad = \int_{(0,+\infty)} x\frac{b^{a}}{\Gamma(a)}x^{-a-1}\exp(-\frac{b}{x})\mathrm{d}x$$

Assume that $a > 1$. Then

$$\begin{aligned}\mathrm{E}_{\mathrm{IG}(a,b)}(x) = \int_{(0,+\infty)} \frac{b^{a-1}}{\Gamma(a)}b\frac{\Gamma(a-1)}{\Gamma(a-1)}x^{-a+1-1}\exp(-\frac{b}{x})\mathrm{d}x \quad &= b\frac{\Gamma(a-1)}{\Gamma(a)}\int_{(0,+\infty)}\frac{b^{a-1}}{\Gamma(a-1)}x^{-a+1-1}\exp(-\frac{b}{x})\mathrm{d}x \\ = b\frac{\Gamma(a-1)}{\Gamma(a)}\int_{(0,+\infty)}\frac{b^{a-1}}{\Gamma(a-1)}x^{-a+1-1}\exp(-\frac{b}{x})\mathrm{d}x = b\frac{\Gamma(a-1)}{(a-1)\Gamma(a-1)}\int f_{\mathrm{IG}(a-1,b)}(x)\mathrm{d}x \quad &= \frac{b}{a-1}\end{aligned}$$

Similarly

$$\begin{aligned}\mathrm{E}_{\mathrm{IG}(a,b)}(x^{2}) &= \int_{(0,+\infty)} x^{2}\frac{b^{a}}{\Gamma(a)}x^{-a-1}\exp(-\frac{b}{x})\mathrm{d}x = ... = b\frac{\Gamma(a-1)}{(a-1)\Gamma(a-1)}\int x f_{\mathrm{IG}(a-1,b)}(x)\mathrm{d}x \\ &= \frac{b}{a-1}\frac{b}{a-2};\ a > 2\end{aligned}$$

So

$$\mathrm{Var}_{\mathrm{IG}(a,b)}(x) = \mathrm{E}_{\mathrm{IG}(a,b)}(x^{2}) - \left(\mathrm{E}_{\mathrm{IG}(a,b)}(x)\right)^{2} = \frac{b^{2}}{(a-1)^{2}(a-2)}$$

2 Multivariate Normal distribution[1] $x|\mu,\Sigma \sim \mathbf{N}_{d}(\mu,\Sigma)$

Definition 4. A d-dimensional random variable $x \in \mathbb{R}^{d}$ is said to have a multivariate Normal (Gaussian) distribution, if for every d-dimensional fixed vector $\alpha \in \mathbb{R}^{d}$, the random variable $\alpha^{\top}x$ has a univariate Normal (Gaussian) distribution.

Proposition 5. *A random vector $x \in \mathbb{R}^{d}$ has a d-dimensional Normal distribution with mean $\mu = E(x)$ and covariance matrix $\Sigma = Var(x)$ if and only if random vector $x \in \mathbb{R}^{d}$ has a characteristic function*

$$\varphi_{x}(t) = \exp(it^{\top}\mu - \frac{1}{2}t^{\top}\Sigma t) \tag{2}$$

Hence: the d-dimensional Normal distribution is uniquely defined by the mean and the covariance matrix.

Proof. ($\Longrightarrow$) If x has a d-dimensional distribution then the characteristic function is $\varphi_{x}(t) = \varphi_{t^{\top}x}(1)$. Since x has a d-dimensional Normal distribution with mean $\mu = \mathrm{E}(x)$ and covariance matrix $\Sigma = \mathrm{Var}(x)$, $t^{\top}x$ has a Normal distribution with mean $\mathrm{E}\left(t^{\top}x\right) = t^{\top}\mu$ and variance $\mathrm{Var}\left(t^{\top}x\right) = t^{\top}\Sigma t$. Then

$$\varphi_{x}(t) = \varphi_{t^{\top}x}(1) = \exp\left(i\mathrm{E}\left(t^{\top}x\right) - \frac{1}{2}\mathrm{Var}\left(t^{\top}xt\right)\right) = \exp\left(it^{\top}\mathrm{E}\left(x\right) - \frac{1}{2}t^{\top}\mathrm{Var}\left(x\right)t\right) = \exp\left(it^{\top}\mu - \frac{1}{2}t^{\top}\Sigma t\right)$$

($\Longleftarrow$) If random vector $x \in \mathbb{R}^{d}$ has a characteristic function $\varphi_{x}(t) = \exp(it^{\top}\mu - \frac{1}{2}t^{\top}\Sigma t)$, then for every d-dimensional fixed vector $\alpha \in \mathbb{R}^{d}$ the characteristic function of $\alpha^{\top}x$ is

$$\varphi_{\alpha^{\top}x}(t) = \varphi_{x}(t\alpha) = \exp\left(it\alpha^{\top}\mu - \frac{1}{2}t\alpha^{\top}\Sigma\alpha t\right) = \exp\left(it\left(\alpha^{\top}\mu\right) - \frac{1}{2}\left(\alpha^{\top}\Sigma\alpha\right)t^{2}\right)$$

which defines that $\alpha^{\top}x$ has a univariate Normal distribution with mean $\alpha^{\top}\mu$ and variance $\alpha^{\top}\Sigma\alpha$. □

Notation 6. We denote the d-dimensional Normal distribution with mean μ and covariance matrix $\Sigma \geq 0$ as $\mathrm{N}_{d}(\mu,\Sigma)$.

Notation 7. The d-dimensional standardized Normal distribution is $\mathrm{N}_{d}(0,I)$.

[1] Try the applet: `https://georgios-stats-3.shinyapps.io/demo_multivariatenormaldistribution/`

Proposition 8. *Let random variable* $x \sim N_d(\mu, \Sigma)$, fixed vector $c \in \mathbb{R}^q$ and fixed matrix $A \in \mathbb{R}^q \times \mathbb{R}^d$. The random vector $y = c + Ax$ *has distribution* $y \sim N_q(c + A\mu, A\Sigma A^\top)$.

Proof. First I show that y is Normally distributed. Let $\alpha \in \mathbb{R}^q$ any fixed vector. Then $\alpha^\top y = \tilde{\alpha}^\top x + \alpha^\top c$ where $\tilde{\alpha} = A^\top b$. Because x is multivariate Normal, then $\tilde{\alpha}^\top x$ is univariate Normal (by Definition 4), then $\alpha^\top y$ is univariate Normal. So y is q-variate Normal. Also, $\text{E}(y) = \text{E}(c + Ax) = c + A\text{E}(x)$, and $\text{Var}(y) = \text{Var}(c + Ax) = A\text{Var}(x)A^\top$. □

Proposition 9. *Let a d-dimensional random vector* $x \sim N_{\text{(any)}}(\mu, \Sigma)$.

1. *Let* $x = (x_1, ..., x_d)^\top$*: The* $x_1, ..., x_d$ *are mutually independent if and only if the corresponding off diagonal parts of the* Σ *are zero.*
2. *Let* $y = Ax$ *and* $z = Bx$*, where* $A \in \mathbb{R}^{q\times d}$ *and* $B \in \mathbb{R}^{k\times d}$*: The vectors* $y = Ax$ *and* $z = Bx$ *are independent if and only if* $A\Sigma B^\top = 0$.

Proof. In both cases, the CF (2) factorizes as $\varphi_x(t) = \prod_j \varphi_{x_j}(t_j)$ only when the corresponding of diagonal parts of Σ are zero. □

Proposition 10. *Any sub-vector of a vector with multivariate Normal distribution has a multivariate Normal distribution.*

Proof. Let $x \sim \text{N}_d(\mu, \Sigma)$. Any sub-vector y of x can be expressed as $y = 0 + Px$, where $P \in \mathbb{R}^{q\times d}$ is a suitable projection matrix. Then $y \sim \text{N}_d(P\mu, P\Sigma P^\top)$. □

Proposition 11. *[Marginalization & conditioning]* [2] *Let* $x \sim \text{N}_d(\mu, \Sigma)$. 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 $x_2 \in \mathbb{R}^{d_2}$ Then:

1. For the marginal, it is $x_1 \sim \text{N}_{d_1}(\mu_1, \Sigma_1)$.
2. For $x_{2.1} = x_2 - \Sigma_{21}\Sigma_1^{-1}x_1$, with $\Sigma_1 > 0$, it is $x_{2.1} \sim \text{N}_{d_2}(\mu_{2.1}, \Sigma_{2.1})$ where
$$\mu_{2.1} = \mu_2 - \Sigma_{21}\Sigma_1^{-1}\mu_1 \text{ and } \Sigma_{2.1} = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top \tag{3}$$
3. Random variables x_1 and $x_{2.1}$ are independent.
4. For the conditional, if $\Sigma_1 > 0$, it is
$$x_2|x_1 \sim \text{N}_{d_2}(\mu_{2|1}, \Sigma_{2|1})$$
where
$$\mu_{2|1} = \mu_2 - \Sigma_{21}\Sigma_1^{-1}(x_1 - \mu_1) \text{ and } \Sigma_{2|1} = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top \tag{4}$$

Hint: *If that was a Homework it will be given as a hint to use , in (1.):* $x_1 = Ax$ *with* $A = [I, 0]$*, and in (2.):* $x_{2.1} = Bx$ *with* $[-\Sigma_{21}\Sigma_1^{-1}, I]$.

Solution.

[2] It is good (although not required) to memorize the formulas in (1) and (4) as they are important in Statistics.

1. It is $x_1 = Ax$ with $A = [I, 0]$. Then $x_1 \sim \text{N}(A\mu, A\Sigma A^\top)$ where

$$A\mu = [I, 0] \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \mu_1 \; ; \qquad\qquad A\Sigma A^\top = [I, 0] \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} = \Sigma_1$$

2. It is $x_{2.1} = Bx$ with $[-\Sigma_{21}\Sigma_1^{-1}, I]$. Then $x_{2.1} \sim \text{N}(B\mu, B\Sigma B^\top)$ where

$$\begin{aligned} B\mu &= \left[-\Sigma_{21}\Sigma_1^{-1}, I\right] \left[\mu_1, \mu_2\right]^\top = -\Sigma_{21}\Sigma_1^{-1}\mu_1 + \mu_2; \\ B\Sigma B^\top &= \left[-\Sigma_{21}\Sigma_1^{-1}, I\right] \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix} \begin{bmatrix} -\Sigma_1^{-1}\Sigma_{21}^\top \\ I \end{bmatrix} = \left[0, -\Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top + \Sigma_2\right] \begin{bmatrix} -\Sigma_{21}\Sigma_1^{-1} \\ I \end{bmatrix} \\ &= -\Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top + \Sigma_2 \end{aligned}$$

3. x_1 and $x_{2.1}$ are independent, because (i.) x_1 and x_2 are Normally distributed and (ii.) for $x_1 = Ax$ with $A = [I, 0]$ and $x_{2.1} = Bx$ with $[\Sigma_{21}\Sigma_1^{-1}, 0]$ are

$$\text{Cov}(x_1, x_{2.1}) = \text{Cov}(Ax, Bx) = A\Sigma B^\top = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \Sigma_{21}^\top \\ \Sigma_{21} & \Sigma_2 \end{bmatrix} \begin{bmatrix} -\Sigma_1^{-1}\Sigma_{21}^\top \\ I \end{bmatrix} =$$

$$= \begin{bmatrix} \Sigma_1, & \Sigma_{21}^\top \end{bmatrix} \begin{bmatrix} -\Sigma_1^{-1}\Sigma_{21}^\top \\ I \end{bmatrix} = -\Sigma_{21}^\top + \Sigma_{21}^\top = 0$$

4. From the above, $x_{2.1}$ is independent on x_1; hence the conditional distribution of $x_2|x_1$ (x_2 given x_1 is known) is the same as the marginal distribution of $x_{2.1}$ aka Normal. Namely $\text{d}F(x_{2.1}|x_1) = \text{d}F(x_{2.1}) \in$ Normal. From the above I observe that it is

$$x_{2.1} = x_2 - \Sigma_{21}\Sigma_1^{-1}x_1 \Longleftrightarrow x_2 = x_{2.1} + \Sigma_{21}\Sigma_1^{-1}x_1;$$

hence if I condition x_2 on a given value for x_1, the term $\Sigma_{21}\Sigma_1^{-1}x_1$ is a constant, namely I have $x_2|x_1 = x_{2.1} + \text{const.}$, which implies that the conditional distribution of $x_2|x_1$ is Normal. Now, about the moments

$$\begin{aligned} \text{E}(x_2|x_1) &= \text{E}(x_{2.1} + \Sigma_{21}\Sigma_1^{-1}x_1|x_1) = \text{E}(x_{2.1}|x_1) + \text{E}(\Sigma_{21}\Sigma_1^{-1}x_1|x_1) = \left[\mu_2 - \Sigma_{21}\Sigma_1^{-1}\mu_1\right] + \left[\Sigma_{21}\Sigma_1^{-1}x_1\right] \\ \text{Var}(x_2|x_1) &= \text{Var}(x_{2.1} + \Sigma_{21}\Sigma_1^{-1}x_1|x_1) = \text{Var}(x_{2.1}|x_1) = \Sigma_2 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{21}^\top \end{aligned}$$

Proposition 12. *The density function of the d-dimensional Normal distribution with mean μ and covariance matrix Σ, when Σ is symmetric positive definite matrix ($\Sigma > 0$), exists and it is equal to*

$$f(x) = (2\pi)^{-\frac{d}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right) \tag{5}$$

Proof. Let $x \sim \text{N}(\mu, \Sigma)$. Because $\Sigma > 0$, we use Cholesky decomposition to define L such that $\Sigma = LL^\top$. Let $z = L^{-1}(x - \mu)$. It is $\text{E}(z) = 0$, $\text{Var}(z) = I$, $z \sim \text{N}_d(0, I)$, and hence $z_1, ..., z_d$ are mutually independent So

$$f_z(z) = \prod_{i=1}^{d} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}z_i^2\right) = (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2}z^\top z\right)$$

Then

$$
\begin{aligned}
f_x(x) &= f_z(z)\left|\frac{\mathrm{d}z}{\mathrm{d}x}\right| = f_z(L^{^{-1}}(x-\mu))\left|\det\left(\frac{\mathrm{d}}{\mathrm{d}x}L^{^{-1}}(x-\mu)\right)\right| \\
&= (2\pi)^{-\frac{d}{2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\left(L^{-1}\right)^\top L^{-1}(x-\mu)\right)\det(L^{-1}) \\
&= (2\pi)^{-\frac{d}{2}}\det(\Sigma)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\right)\det(\Sigma)^{-\frac{1}{2}}
\end{aligned}
$$

□

Fact 13. *[In exercises they will be given as a Hint.] Useful formulas about the PDF of the multivariate Normal distribution that we may use.*

1. *If $\Sigma_1 > 0$ and $\Sigma_2 > 0$ symmetric*

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

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. *If $f_{N_d(\mu,\Sigma)}(x)$ denotes the PDF of $N_d(\mu,\Sigma)$, then*

$$
f_{N_d(\mu_1,\Sigma_1)}(x)\,f_{N_d(\mu_2,\Sigma_2)}(x) \;=\; f_{N_d(m,V)}(x)\,f_{N_d(\mu_2,\Sigma_1+\Sigma_2)}(\mu_1)
$$

where

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

3. *If $\Sigma_i > 0$ symmetric for $i = 1, ..., n$*

$$
-\frac{1}{2}\sum_{i=1}^{n}\left(x-\mu_i\right)\Sigma_i^{-1}\left(x-\mu_i\right)^\top = -\frac{1}{2}\left(x-m\right)V^{-1}\left(x-m\right)^\top + C \tag{6}
$$

where

$$
V^{-1} = \sum_{i=1}^{n}\Sigma_i^{-1};\quad m = V\left(\sum_{i=1}^{n}\Sigma_i^{-1}\mu_i\right);\quad C = \frac{1}{2}mV^{-1}m^\top - \frac{1}{2}\left(\sum_{i=1}^{n}\mu_i\Sigma_i^{-1}\mu_i^\top\right) \tag{7}
$$

Proof. (1.) is derived by ±ing terms and doing matrix calculations. (2.) is derived by exponentiation and completing the associated constants. (3.) is shown by induction from the (1.). □

3 Multivariate Student's T distribution[3] $x \sim \mathbf{T}_d(\mu,\Sigma,v)$

Definition 14. A d-dimensional random variable $x \in \mathbb{R}^d$ is said to have a multivariate Student's T distribution with location parameter μ, scale matrix Σ, and degrees of freedom v, and it is denoted as $x \sim \mathrm{T}_d(\mu,\Sigma,v)$, if and only if

$$
x = \mu + y\sqrt{v\xi}
$$

where $y \sim \mathrm{N}_d(0,\Sigma)$ and $\xi \sim \mathrm{IG}(\frac{v}{2},\frac{1}{2})$ are independent random variables.

[3] Try the applet: `https://georgios-stats-3.shinyapps.io/demo_multivariatenormaldistribution/`

Example 15. If $x \sim \mathrm{T}_d(\mu, \Sigma, v)$ and $\Sigma > 0$ then

1. The PDF of x is

$$f_X(x|\mu,\Sigma,v) = \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})\nu^{\frac{d}{2}}\pi^{\frac{d}{2}}\det(\Sigma)^{\frac{1}{2}}}\left(1+\frac{1}{v}(t-\mu)^{\mathrm{T}}\Sigma^{-1}(t-\mu)\right)^{-\frac{\nu+d}{2}} \tag{8}$$

2. The expected value is

$$\mathrm{E}_{\mathrm{T}_d(\mu,\Sigma,\nu)}(X) = \mu \tag{9}$$

3. The covariance matrix is

$$\mathrm{Var}_{\mathrm{T}_d(\mu,\Sigma,\nu)}(X) = \begin{cases} \frac{\nu}{\nu-2}\Sigma & \text{, if } \nu > 2 \\ \\ \text{undefined} & \text{, else} \end{cases} \tag{10}$$

Hint: Use that: $x = \mu + y\sqrt{v\xi}$ where $y \sim \mathrm{N}_d(0,\Sigma)$ and $\xi \sim \mathrm{IG}(\frac{v}{2},\frac{1}{2})$ independent.

Solution. Given Definition 14, $x \sim \mathrm{T}_d(\mu,\Sigma,v)$ results as the marginal distribution of (x,ξ) where $x|\xi \sim \mathrm{N}_d(\mu,\Sigma\xi v)$ and $\xi \sim \mathrm{IG}(\frac{v}{2},\frac{1}{2})$.

1. So it is

$$\begin{aligned}
f_x(x) &= \int f_{x|\xi}(x|\xi)f_\xi(\xi)\mathrm{d}\xi = \int f_{\mathrm{N}_d(\mu,\Sigma v\xi)}(x|\xi)f_{\mathrm{IG}(\frac{v}{2},\frac{1}{2})}(\xi)\mathrm{d}\xi \\
&= \int \underbrace{\left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\frac{1}{\sqrt{\det(\Sigma v\xi)}}\exp\left(-\frac{1}{2}(x-\mu)^\top\frac{\Sigma^{-1}}{v\xi}(x-\mu)\right)}_{=\mathrm{N}_d(x|\mu,\Sigma v\xi)}\underbrace{\frac{\frac{1}{2}^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\xi^{-\frac{v}{2}-1}\exp\left(-\frac{1}{\xi}\frac{1}{2}\right)1_{(0,\infty)}(\xi)\mathrm{d}\xi}_{=\mathrm{IG}(\xi|\frac{v}{2},\frac{1}{2})} \\
&= \left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\frac{1}{\sqrt{\det(\Sigma v)}}\frac{\frac{1}{2}^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\underbrace{\int \xi^{-\frac{v}{2}-\frac{d}{2}-1}\exp\left(-\frac{1}{\xi}\left[\frac{1}{2v}(x-\mu)^\top\Sigma^{-1}(x-\mu)+\frac{1}{2}\right]\right)\mathrm{d}\xi}_{=\Gamma\left(\frac{v}{2}+\frac{d}{2}\right)\left[\frac{1}{2v}(x-\mu)^\top\Sigma^{-1}(x-\mu)+\frac{1}{2}\right]^{-\left(\frac{v}{2}+\frac{d}{2}\right)}} \qquad (11) \\
&= \left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\frac{1}{\sqrt{\det(\Sigma v)}}\frac{\frac{1}{2}^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\Gamma\left(\frac{v}{2}+\frac{d}{2}\right)\left[\frac{1}{2v}(x-\mu)^\top\Sigma^{-1}(x-\mu)+\frac{1}{2}\right]^{-\left(\frac{v}{2}+\frac{d}{2}\right)} \\
&= \left(\frac{1}{2\pi}\right)^{\frac{d}{2}}\frac{1}{\sqrt{\det(\Sigma v)}}\frac{\frac{1}{2}^{\frac{v}{2}}}{\Gamma(\frac{v}{2})}\Gamma\left(\frac{v}{2}+\frac{d}{2}\right)\left(\frac{1}{2}\right)^{-\frac{(v+d)}{2}}\left[\frac{1}{v}(x-\mu)^\top\Sigma^{-1}(x-\mu)+1\right]^{-\frac{v+d}{2}} \\
&= \left(\frac{1}{\pi}\right)^{\frac{d}{2}}\frac{1}{\sqrt{\det(\Sigma)}}\left(\frac{1}{v}\right)^{\frac{d}{2}}\frac{1}{\Gamma(\frac{v}{2})}\Gamma\left(\frac{v+d}{2}\right)\left[\frac{1}{v}(x-\mu)^\top\Sigma^{-1}(x-\mu)+1\right]^{-\frac{v+d}{2}}
\end{aligned}$$

where the integral in (11) was calculated by recognizing the IG density from (1).

2. It is

$$\mathrm{E}_{\mathrm{t}_d(\mu,\Sigma,\nu)}(x) = \mathrm{E}_{\mathrm{IG}(\frac{v}{2},\frac{1}{2})}\left(\mathrm{E}_{\mathrm{N}_d(\mu,\Sigma\xi v)}(x|\xi)\right) = \mathrm{E}_{\mathrm{IG}(\frac{v}{2},\frac{1}{2})}\left(\mu\right) = \mu$$

3. It is

$$
\begin{aligned}
\mathrm{Var}_{\mathrm{t}_d(\mu,\Sigma,\nu)}(x) =& \mathrm{E}_{\mathrm{IG}(\frac{v}{2},\frac{1}{2})}\left(\mathrm{Var}_{\mathrm{N}_d(\mu,\Sigma\xi v)}(x|\xi)\right) + \mathrm{Var}_{\mathrm{IG}(\frac{v}{2},\frac{1}{2})}\left(\mathrm{E}_{\mathrm{N}_d(\mu,\Sigma\xi v)}(x|\xi)\right) \\
\\
=& \mathrm{E}_{\mathrm{IG}(\frac{v}{2},\frac{1}{2})}\left(\Sigma\xi v\right) + \cancelto{=0}{\mathrm{Var}_{\mathrm{IG}(\frac{v}{2},\frac{1}{2})}\left(\mu\right)} = \Sigma v \mathrm{E}_{\mathrm{IG}(\frac{v}{2},\frac{1}{2})}\left(\xi\right) + 0 \\
\\
=& \begin{cases} \Sigma v \frac{\frac{1}{2}}{\frac{v}{2}-1} & \text{, if } \frac{v}{2} > 1 \\ \\ \text{undefined} & \text{, else} \end{cases}
\end{aligned}
$$

4 Practice

Question 16. *For practice try the Exercises 13, 14, and, 16, from the Exercise Sheet.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 3: The Bayesian paradigm & Subjective probability[a]

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

Aim

To understand some foundation concepts of the Bayesian statistics: Subjective probability, an Axiomatic system, Bayesian paradigm. From Section 2, you are can read what is the subjective probability interpretation and the general philosophy of its construction-you are not required to memorize these Theorems & axioms.

Reading list:

- DeGroot, M. H. (1970, or 2005; Chapter 6). Optimal statistical decisions (Vol. 82). John Wiley & Sons.
- O'Hagan, A., & Forster, J. J. (2004; Paragraphs 4.1-1.16). Kendall's advanced theory of statistics, volume 2B: Bayesian inference (Vol. 2). Arnold.
- Robert, C. (2007; Sections 1.1, 1.2, 1.4). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

[a] Author: Georgios P. Karagiannis.

1 Schools of Statistics

Probability is a well defined mathematical quantity, however it has several interpretations; e.g. Frequetist, Subjective, etc. Probability is a fundamental concept in Statistics; different interpretations of Probability lead to different schools of statistics. Below we give more details about the main two (Frequentist and Bayesian) schools of statistics.

The Frequentist school of statistics

The Frequentist school of statistics uses the 'Frequency interpretation of probability' which asserts that the probability $P(A)$ of an event A is the limiting relative frequency of occurrence of the event in an infinite sequence of random trials. Recall that classical rules of inference are judged on their long-run behavior in repeated sampling.

Frequentist statisticians presume that the observations have been generated from a (idealized) model which would presumably run along the lines "out of infinitely many worlds one is selected at random...". Often that model has quantities (parameters) whose values are unknown to You, but You are interested in learning. These parameters are presumed to be constants quantities in the sense that they are equal to an ideal/real value which You want to discover.

The Subjective Bayesian school of statistics

The Subjective Bayesian school of statistics uses the 'Subjective interpretation of probability' which asserts that the probability $P(A)$ represents a degree of belief in a proposition A, based on all the available information Ω. In Subjective Bayesian statistics all probabilities and distributions are subjective, or personalistic; they represent Your (investigator's) degrees of belief.

Subjective probability concerns Yours judgments about uncertain events or propositions. Eg., $P(A)$ measures the strength of Your degree of belief that A will occur. $P(A) = 1$ describes that You are certain that A will occur, and $P(A) = 0$ describes that You are certain that A will not occur. As $P(A)$ increases from 0 to 1, it describes an increasing degree of belief in the occurrence of A. Different researchers may have a different degree of belief in the same proposition, and these different researchers can assign a different probabilities to that proposition based on their own judgments. The only constraint is that a Your probabilities should not be inconsistent, and therefore they should obey the Kolmogorov axioms of probability.

2 Subjective probability and its construction

Acceptance of the Bayesian method as the natural and proper approach to statistical inference has become almost synonymous with the adoption of a subjective interpretation of probability. For instance, a fully subjective interpretation of probability, allows the (Subjective) Bayesian analysis to avoid to produce controversial results, such as violation of the Likelihood Principle.

2.1 Axiomatic formulation of relative likelihood

Consider $(\Omega, \mathscr{F})$ where Ω is a sample space and $\mathscr{F}$ is a σ-algebra of events. Consider events $A, B \in \mathscr{F}$ as sets containing one or more elements from the set Ω. We think of A, B, Ω as a more general propositions. The intersection $A \cap B$ corresponds to the logical conjunction 'A and B'; the union $A \cup B$ corresponds to the logical dis-junction 'A or B'; the complement $A^{\complement}$ corresponds to the logical negation 'not A'; and $A \supset B$ corresponds to the logical expression 'B implies A'. The empty set $\emptyset$ corresponds to a proposition that is certainly false, and the universal set Ω (aka the sampling space) to a proposition that is certainly true.

Definition 1. Let $A \precsim B$ denote the judgment that A is not more likely to occur than B; $A \sim B$ denote the judgment that A and B are equally likely to occur; $A \prec B$ denote the judgment that A is less likely to occur than B given a common underlying (initial) information base.

Consider the following set of (reasonable) axioms:

Axiom-LA1 For any A, B, only one of $A \prec B$, $B \prec A$, $A \sim B$ can occur.

Axiom-LA2 If A_1, A_2, B_1, B_2 are events such that $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$, $A_1 \precsim B_1$, $A_2 \precsim B_2$ then $A_1 \cup A_2 \precsim B_1 \cup B_2$. Additionally, if either $A_1 \prec B_1$, $A_2 \prec B_2$ then $A_1 \cup A_2 \prec B_1 \cup B_2$.

Axiom-LA3 For any A, $\emptyset \precsim A$. Furthermore, $\emptyset \prec \Omega$.

Axiom-LA4 If $A_1 \supset A_2 \supset$ ... is a decreasing sequence of events with limit $\cap_{i=1}^{\infty} A_i$, and B is some fixed event such that $A_i \succsim B$ for all $i = 1, 2...$ then $\cap_{i=1}^{\infty} A_i \succsim B$.

Axiom-LA5 There exists a random variable $u \in [0, 1]$ such that if A_1 and A_2 are the events that u falls in given sub-intervals of $[0, 1]$ with lengths ℓ_1 and ℓ_2 respectively, then $A_1 \precsim A_2$ if and only if $\ell_1 \leq \ell_2$.

(LA1) ensures that all events may be compared. (LA2) and (LA3) ensure that the comparisons are made in a logically consistent way, and simply reflect some obvious properties of any notion of 'not more probable than'. (LA4) is an assumption stronger than (LA2) and (LA3). (LA5), essentially defines the Uniform distribution, and allows the construction of the subjective probability suggesting how much more likely an event is.

2.2 Construction of probability

Theorem 2. *Let $\mu[a, b]$ denote the event that a random variable, from Axiom-LA5, lies in $[a, b]$. For any event $A \in \mathscr{F}$ satisfying the axioms LA1-LA5, there exists a unique number $a^{\star} \in [0, 1]$ such that $A \sim \mu[0, a^{\star}]$*

Definition 3. Subjective probability (Your degree of believe) of an event $A \in \mathscr{F}$ satisfying the axioms LA1-LA5, we define the unique number $a^{\star}$ from Theorem 2. It is symbolized as $P(A)$, and hence satisfies

$$A \sim \mu[0, P(A)] \quad \text{, for all A}\in\mathscr{F}. \tag{1}$$

Theorem 4. *Let two events $A, B \in \mathscr{F}$. Then $A \precsim B$* if and only if $P(A) \leq P(B)$

Theorem 5. *Given Axioms LA1-LA5, the quantity $P(\cdot)$ (Definition 3) is the probability as it satisfies the usual probability axioms:*

P1 $P(A) \geq 0$ *and* $P(\Omega) = 1$

P2 *If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$*

P3 *If $\{A_1, A_2, ...\}$ an infinite sequence of events such that $A_i \cap A_j = \emptyset$ for all i, j then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.*

2.3 Extension to the conditional likelihoods & probability

Consider events $A, D \in \mathscr{F}$; $(A|D)$ corresponds to the conjunction 'A when D is known'.

Definition 6. Given a common underlying (initial) information base, $(A|D) \precsim (B|D)$ denotes the judgment that A is not more likely to occur than B, when D is known.

Consider the following additional (reasonable) axiom needed to :

Axiom-LA6 For any $A, B, D \in \mathscr{F}$, $(A|D) \precsim (B|D)$, if and only if $A \cap D \precsim B \cap D$.

Theorem 7. *Given [LA1] and [LA6], for any $A, B, D \in \mathscr{F}$, exactly one of the following three relations can occur: $(A|D) \prec (B|D)$, $(A|D) \succ (B|D)$, or $(A|D) \sim (B|D)$.*

The following theorem relates the standard probability to the conditional likelihoods.

Theorem 8. *If relations $(\prec, \succ, \sim)$ satisfy assumptions [LA1]-[LA5], and [LA6] then quantity P in Definition 3 is the unique probability distribution which has the property: For any $A, B, D \in \mathscr{F}$ such that $P(D) > 0$,*

$$(A|D) \lesssim (B|D) \qquad \textit{if and only if} \qquad P(A|D) \leq P(B|D)$$

Remark 9. All subjective probabilities are conditional. We recognize some initial information Ω. $P(A|\Omega)$ expresses Your degree of belief about A based on the totality of his current information Ω. If You observe the occurrence of another event B, Your probability for A becomes $P(A|B, \Omega)$ because B has been added to Your information. As Ω is common, it is convenient to simplify the notation by suppressing Ω, and writing $P(A) := P(A|\Omega)$ and $P(A|B) := P(A|B, \Omega)$. In Bayesian paradigm language: the background information Ω may reflect all one knows before the collection of the data; $P(A)$ is your prior information; B is the new information from an experiment; $P(A|B)$ is Your degree of believe about A after observing B; and the Bayesian theorem is the mechanism performing this update.

3 The Bayesian paradigm

Assume that a specific experiment $e \in \mathcal{E}$ has been performed; $\mathcal{E}$ denotes the family of potential experiments. Assume there is available a sequence of observations (or data) $y_{1:n} = (y_1, ..., y_n)$, where $y_i \in \mathcal{Y}$ for $i = 1, ..., n$, generated as outcomes of the performed experiment $e \in \mathcal{E}$.

Let $\mathrm{d}F(\cdot|\theta)$ denote the sampling distribution with PDF/PMF $f(y|\theta)$, which models the data generating process of the performed experiment e. For simplicity, we suppress conditioning on e from $\mathrm{d}F(\cdot|\theta, e)$ although we shouldn't. $\mathrm{d}F(\cdot|\theta)$ is a statement of the Your subjective views about the data generating process.

Definition 10. A **parametric statistical model** consists of a sequence of observations $y_{1:n} = (y_1, ..., y_n)$ of a random variable y, and the sampling distribution $\mathrm{d}F(\cdot|\theta)$ where only the parameter $\theta \in \Theta$ is unknown.

- Symbol. $y_{1:n} \sim \mathrm{d}F(\cdot|\theta)$ or $(y_{1:n}, \mathrm{d}F(\cdot|\theta))$.

Definition 11. The **likelihood function** $L(y_{1:n}|\theta)$ of $y_{1:n}$ and e given θ is defined as $L(y_{1:n}|\theta) = f(y_{1:n}|\theta)$, and contains all the information available from the observed data $y_{1:n}$ and the experiment performed.

Definition 12. Priori distribution $\mathbf{d\Pi(\theta)}$ of the unknown parameter $\theta \in \Theta$, with PDF/PMF $\pi(\theta)$, quantifies Your believes or judgments about parameter θ before You perform the experiment and get the observations.

- symbol. $\theta \sim \mathrm{d}\Pi(\theta)$,

Remark 13. We say 'To account for the uncertainty of the unknown parameter $\theta \in \Theta$, we assign an a priori distribution $\mathrm{d}\Pi(\theta)$ with PDF/PMF $\pi(\theta)$'. It is specified in an entirely subjective manner based on Your judgments.

Definition 14. (Bayesian model) A Bayesian statistical model is made of a parametric statistical model, $y_{1:n} \sim \mathrm{d}F(\cdot|\theta)$, and a prior distribution (or prior model) $\theta \sim \mathrm{d}\Pi(\theta)$, which admits PDF/PMF $\pi(\theta)$, on the unknown parameters θ. It is denoted as a hierarchical model:

$$\begin{cases} y_{1:n}|\theta & \sim \mathrm{d}F(y_{1:n}|\theta) \\ \theta & \sim \mathrm{d}\Pi(\theta) \end{cases} \tag{2}$$

Remark 15. (2) implies that observations $y_{1:n}$ have been generated from a distribution $\mathrm{d}F(y_{1:n}|\theta)$ parameterized by θ, where θ is a latent variable that follows a distribution $\mathrm{d}\Pi(\theta)$ prior to the generation of $y_{1:n}$.

Definition 16. The joint distribution $\mathrm{d}P(\theta, y_{1:n})$ of $(\theta, y_{1:n})$ can be defined from (2)

$$\mathrm{d}P(\theta, y_{1:n}) = \mathrm{d}F(y_{1:n}|\theta)\mathrm{d}\Pi(\theta) \quad \text{with PDF/PMF} \quad p(\theta, y_{1:n}) = f(y_{1:n}|\theta)\pi(\theta)$$

Definition 17. The prior predictive distribution $\mathrm{d}F(y_{1:n})$ of $y_{1:n}$ results by integrating out $\mathrm{d}F(y_{1:n}|\theta)$ with respect to $\mathrm{d}\Pi(\theta)$. We are interested in its PDF/PMF called marginal likelihood

$$f(y_{1:n}) = \int_\Theta f(y_{1:n}|\theta)\mathrm{d}\Pi(\theta) = \begin{cases} \int_\Theta f(y_{1:n}|\theta)\pi(\theta)\mathrm{d}\theta & \text{, if } \theta \text{ cont} \\ \\ \sum_{\forall\theta\in\Theta} f(y_{1:n}|\theta)\pi(\theta) & \text{, if } \theta \text{ disc} \end{cases} \tag{3}$$

Definition 18. The posterior distribution $\mathrm{d}\Pi(\theta|y_{1:n})$ of $\theta \in \Theta$ given $y_{1:n}$ and e is defined by the Bayes Theorem

$$\mathrm{d}\Pi(\theta|y_{1:n}) = \frac{L(y_{1:n}|\theta)\mathrm{d}\Pi(\theta)}{\int_\Theta L(y_{1:n}|\theta)\mathrm{d}\Pi(\theta)} \quad \text{with PDF/PMF} \quad \pi(\theta|y_{1:n}) = \frac{L(y_{1:n}|\theta)\pi(\theta)}{\int_\Theta L(y_{1:n}|\theta)\mathrm{d}\Pi(\theta)} \tag{4}$$

Remark 19. Posterior expectation of any integrate function $h(\cdot)$ defined on Θ is

$$\mathrm{E}_\Pi(h(\theta)|y_{1:n}) = \frac{\int_\Theta h(\theta)L(y_{1:n}|\theta)\mathrm{d}\Pi(\theta)}{\int_\Theta L(y_{1:n}|\theta)\mathrm{d}\Pi(\theta)}$$

The posterior probability that $\theta \in A$ where $A \subseteq \Theta$ is

$$\Pi(\theta \in A|y_{1:n}) = \mathrm{E}_\Pi(1(\theta \in A)|y_{1:n}) = \frac{\int_A L(y_{1:n}|\theta)\mathrm{d}\Pi(\theta)}{\int_\Theta L(y_{1:n}|\theta)\mathrm{d}\Pi(\theta)}$$

Remark 20. The posterior distribution $\mathrm{d}\Pi(\theta|y_{1:n})$ represents Your degree of believe about θ after You performed the experiment e and saw the data $y_{1:n}$ generated by e. Elaborating on (4), the Bayesian Theorem can be seen as a reasonable probabilistic mechanism to

- combine Your a priori information about θ (exclusively incorporated in $\mathrm{d}\Pi(\theta)$) and the experimental information about θ (exclusively incorporated in likelihood $L(y_{1:n}|\theta)$

- update Your degree of believe about θ from the a priori distribution $\pi(\theta)$ to the a posteriori distribution $\pi(\theta|y_{1:n})$ in the light of new information from experiment e.

- perform the inversion $(y_{1:n}|\theta) \mapsto (\theta|y_{1:n})$ in a subjective probabilistic manner. Usually, the experiment, as described by the parametric model $\mathrm{d}F(y_{1:n}|\theta)$, is a process there we observe the effect $y_{1:n}$ but we are interested in learning the unknown cause θ .

Definition 21. The predictive distribution of a sequence $z_{1:m} = (y_{n+1}, ..., y_{n+m})$ of m future outcomes given a sequence of observations $y_{1:n}$ has PDF/PMF

$$f(z_{1:m}|y_{1:n}) = \mathrm{E}_{\Pi}(f(z_{1:m}|y_{1:n}, \theta)|y_{1:n}) = \int_{\Theta} f(z_{1:m}|y_{1:n}, \theta)\mathrm{d}\Pi(\theta|y_{1:n})$$

It the case that $z_{1:m}$ and $y_{1:n}$ are conditionally independent given θ; i.e. $f(z_{1:m}|y_{1:n}, \theta) = f(z_{1:m}|\theta)$ it is

$$g(z_{1:m}|y_{1:n}) = \mathrm{E}_{\Pi}(f(z_{1:m}|\theta)|y_{1:n}) = \int_{\Theta} f(z_{1:m}|\theta)\mathrm{d}\Pi(\theta|y_{1:n})$$

Remark 22. Specification of the subjective probability should be the result of very careful weighing of all the available information, in order to avoid the case that the probability reflects a person's prejudices and without any scientific basis at all. For a given problem, usually the researcher specifies (i.) $\mathrm{d}F(y_{1:n}|\theta)$ and $\mathrm{d}\Pi(\theta)$, or (ii.) $\mathrm{d}P(y_{1:n}, \theta)$ and derives the rest distributions, by probability calculations, since

$$\mathrm{d}P(y_{1:n}, \theta) = \mathrm{d}F(y_{1:n}|\theta)\mathrm{d}\Pi(\theta) = \mathrm{d}\Pi(\theta|y_{1:n})\mathrm{d}F(y_{1:n})$$

Notation 23. If the likelihood of $y_{1:n}$ given θ is factorized as

$$L(y_{1:n}|\theta) = k(y_{1:n}|\theta)\rho(y_{1:n}),$$

then $k(y_{1:n}|\theta)$ is called kernel of the likelihood of $y_{1:n}$ given θ, and $\rho(y_{1:n})$ is called residue of this likelihood.

Notation 24. If a PDF/PMF $\pi(\theta)$ can be written as

$$\pi(\theta) = \frac{K(\theta)}{\int K(\theta)\mathrm{d}\theta},$$

then then $K(\theta)$ is called kernel of the density $\pi(\theta)$. We can write $\pi(\theta) \propto K(\theta)$ as the normalizing constant $\int K(\theta)\mathrm{d}\theta$ is implied by the specification of $K(\theta)$.

Example 25. [Bernoulli-Beta model] Let e be a Bernoulli experiment with unknown probability of success $\theta \in [0, 1]$. Specify a Bayesian model

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

where $\theta \in [0, 1]$, and $a > 0$, $b > 0$ are known (fixed) hyper-parameters.

1. State the sampling distribution of $y_{1:n}$, the likelihood function of θ, the prior distribution of θ, and the joint PMF/PDF of $(y_{1:n}, \theta)$.

2. Calculate the marginal PMF of $y_{1:n}$

3. Calculate the posterior PDF of θ given $y_{1:n}$ and recognize the distribution family.

4. Calculate the predictive PMF of a sequence of future outcomes $z_{1:m} = (z_1, .., z_m)$ given $y_{1:n}$.

Hint: Consider PDF/PMF:

$$f_{\text{Br}(\theta)}(y) = \theta^y(1-\theta)^{1-y}1(y \in \{0, 1\}) ; \qquad \pi_{\text{Be}(a,b)}(\theta) = \frac{1}{B(a, b)}\theta^{a-1}(1-\theta)^{b-1}1(\theta \in [0, 1])$$

Solution.

1. The sampling distribution is a Bernoulli distribution with parameter θ; i.e. $\text{Br}(\theta)$. The Likelihood function is

$$L(y_{1:n}|\theta) = f(y_{1:n}|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}.$$

The prior is the Beta distribution with parameters $a > 0$ and $b > 0$; i.e. $\theta \sim \text{Be}(a, b)$:

$$\pi(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}1(\theta \in [0,1])$$

The joint PMF/PDF of $(y_{1:n}, \theta)$ is

$$\begin{aligned} p(y_{1:n},\theta) &= f(y_{1:n}|\theta)\pi(\theta) = \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta) \\ &= \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i}1(y_{1:n} \in \{0,1\}^n)\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}1(\theta \in [0,1]) \\ &= \theta^{\sum_{i=1}^{n} y_i+a-1}(1-\theta)^{n-\sum_{i=1}^{n} y_i+b-1}1(y_{1:n} \in \{0,1\}^n)1(\theta \in [0,1]) \end{aligned}$$

2. The marginal likelihood of the observations $y_{1:n}$ is

$$\begin{aligned} f(y_{1:n}) &= \int p(y_{1:n},\theta)\mathrm{d}\theta = \int f(y_{1:n}|\theta)\pi(\theta)\mathrm{d}\theta = \int \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta)\mathrm{d}\theta \\ &= \int_0^1 \theta^{\sum_{i=1}^{n} y_i+a-1}(1-\theta)^{n-\sum_{i=1}^{n} y_i+b-1}\mathrm{d}\theta \; 1(y_{1:n} \in \{0,1\}^n) \\ &= B(\sum_{i=1}^{n} y_i + a, n - \sum_{i=1}^{n} y_i + b)1(y_{1:n} \in \{0,1\}^n) \end{aligned}$$

3. The posterior of θ given the observations $y_{1:n}$ has PDF

$$\begin{aligned} \pi(\theta|y_{1:n}) &= \frac{f(y_{1:n}|\theta)\pi(\theta)}{\int_\Theta f(y_{1:n}|\theta)\mathrm{d}\Pi(\theta)} \propto f(y_{1:n}|\theta)\pi(\theta) \\ &= \prod_{i=1}^{n} f(y_i|\theta)\pi(\theta|a,b) \;\; = \theta^{\sum_{i=1}^{n} y_i+a-1}(1-\theta)^{n-\sum_{i=1}^{n} y_i+b-1} \\ &\propto \text{Be}(\theta|a^*,b^*) \end{aligned}$$

where $a^* = \sum_{i=1}^{n} y_i + a$, $b^* = n - \sum_{i=1}^{n} y_i + b$. Hence the posterior distribution of θ is a Beta distribution with parameters $a^* = \sum_{i=1}^{n} y_i + a$, and $b^* = n - \sum_{i=1}^{n} y_i + b$.

4. The predictive distribution of $z_{1:m}$ from parametric model $\text{Br}(\theta)$ given $y_{1:n}$ has PMF

$$\begin{aligned} g(z_{1:m}|y_{1:n}) &= \int_\Theta f(z_{1:m}|\theta)\pi(\theta|y_{1:n})\mathrm{d}\theta \;\; = \int_\Theta \prod_{i=1}^{m} f(z_i|\theta)\pi(\theta|y_{1:n})\mathrm{d}\theta \\ &= \int_0^1 \left[\theta^{\sum_{i=1}^{m} z_i}(1-\theta)^{m-\sum_{i=1}^{m} z_i}1(z_{1:m} \in \{0,1\}^m)\right]\left[\frac{1}{B(a^*,b^*)}\theta^{a^*-1}(1-\theta)^{b^*-1}\right]\mathrm{d}\theta \\ &= \frac{1}{B(a^*,b^*)}\int_0^1 \theta^{\sum_{i=1}^{m} z_i+a^*-1}(1-\theta)^{m-\sum_{i=1}^{m} z_i+b^*-1}\mathrm{d}\theta \; 1(z_{1:m} \in \{0,1\}^m) \\ &= \frac{B(\sum_{i=1}^{m} z_i + a^*, m - \sum_{i=1}^{m} z_i + b^*)}{B(a^*,b^*)}1(z_{1:m} \in \{0,1\}^m) \end{aligned}$$

Example 26. Consider an i.i.d. sample $y_1, \ldots, y_n$ from the skew-logistic distribution with PDF

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

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

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

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

1. Derive the posterior distribution of θ.

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

Solution. It is

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

1. By using the Bayes theorem

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

So

$$\theta | y_{1:n} \sim \text{Ga}(\underbrace{a+n}_{=a^*}, \underbrace{b + \sum_{i=1}^{n} \log(1+e^{-x_i}))}_{=b^*}$$

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

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

4 Sequential processing of data via Bayes theorem

Note 27. Bayesian paradigm enjoys a coherence property according to which updating the prior one observation at a time, or all observations together does not matter, it leads to the same posterior inference.

Note 28. Let $y = (y_1, y_2)$ be a partition of the observables.

- Consider the Learning Procedure 1 for θ where the Prior $\Pi(\theta)$ is updated to a posterior $\Pi(\theta|y_1, y_2)$ in the light of the full data (y_1, y_2) observed at once; namely

  $$\text{Learning Procedure 1: } \Pi(\theta) \underset{f(y_1,y_2|\theta)}{\overset{(y_1,y_2)}{\longmapsto}} \Pi(\theta|y_1, y_2)$$

  where

  $$\Pi(\theta|y_1, y_2) \quad \text{with pdf } \pi(\theta|y_1, y_2) = \frac{f(y_1, y_2|\theta)\pi(\theta)}{f(y_1, y_2)}; \text{ where } f(y) = \int_\Theta f(y|\theta)\mathrm{d}\Pi(\theta)$$

- Consider Learning Procedure 2 for θ where at first stage the prior $\Pi(\theta)$ is updated to a posterior $\Pi(\theta|y_1)$ in the light of data y_1 and at second stage $\Pi(\theta|y_1)$ is updated to $\Pi(\theta|y_1, y_2)$ in the light of new data y_2. Here, $\Pi(\theta|y_1)$ is the distribution of θ posterior to observing $y_1$and prior to observing y_2! Similarly the likelihood should be conditional on all the observables incorporated so far as $f(y_2|y_1, \theta)$. Learning Procedure 2 for θ is

  $$\text{Learning Procedure 2: } \Pi(\theta) \underset{f(y_1|\theta)}{\overset{y_1}{\longmapsto}} \Pi'(\theta|y_1) \underset{f(y_2|y_2,\theta)}{\overset{y_2}{\longmapsto}} \Pi'(\theta|y_1, y_2)$$

  with pdf/pmfs

  $$\pi(\theta|y_1) = \frac{f(y_1|\theta)\pi(\theta)}{f(y_1)}; \text{ where } f(y_1) = \int_\Theta f(y_1|\theta)\mathrm{d}\Pi(\theta)$$

  $$\begin{aligned}\pi'(\theta|y_1, y_2) &= \frac{f(y_2|y_1, \theta)\pi(\theta|y_1)}{f(y_2|y_1)} = \frac{f(y_2|y_1, \theta)\frac{f(y_1|\theta)\pi(\theta)}{f(y_1)}}{f(y_2|y_1)} = \frac{f(y_2|y_1, \theta)f(y_1|\theta)\pi(\theta)}{f(y_2|y_1)f(y_1)} \\ &= \frac{f(y_1, y_2|, \theta)\pi(\theta)}{f(y_1, y_2)} = \pi(\theta|y_1, y_2)\end{aligned}$$

- We observe the two Learning Scenarios are equivalent in the sense that they lead to the same posterior $\Pi(\theta|y_1, y_2)$ at the end posterior $\Pi(\theta|y_1, y_2)$ in a single application of Bayes theorem with the full data $y = (y_1, y_2)$.

Note 29. By induction, this result is extended to the case where several data are collected sequentially as $y_1, y_2, y_3, y_4, ...$

5 Practice

Question 30. *Feel free to work on the Exercise 33, in the Exercise sheet.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 4: The exchangeable model [a]

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

Aim

Get familiar with the concept of exchangeability, and its relation to Subjective probability and Bayesian paradigm.

Reading list:

- Bernardo, J. M., & Smith, A. F. (2009, Section 4.3). Bayesian theory (Vol. 405). John Wiley & Sons.
- Berger, J. O. (2013, Section 3.5.7). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

[a] Author: Georgios P. Karagiannis.

1 The exchangeable model

As mentioned in Handout 3, one way to specify the Bayesian model is by subjectively specifying the probability distributions $\mathrm{d}F(y|\theta)$ and $\mathrm{d}\Pi(\theta)$, or the joint distribution $\mathrm{d}P(y,\theta)$, that enables You to derive the rest distributions.

Alternatively, You can specify a probability distribution on the data generating process $\mathrm{d}G(y_{1:n})$ describing the actual sequence of the data $y_{1:n}$. On approach is to subjectively set certain invariance assumptions on the observables $y_{1:n}$ involving probabilistic believes of invariant with respect to some aspect of the observable quantities.

A reasonable invariance assumption about $y_{1:n} = (y_1, ..., y_n)$ is the Exchengeability: The 'labels' identifying the individual observable quantities are 'uninformative', in the sense the information that the y_i's provide is independent of the order in which they are collected. Exchangeability, although a simple asssumption, it accurately describes a large class of experimental setups.

Definition 1. A sequence of random quantities $y_{1:n} = (y_1, ..., y_n)$ is finitely exchangeable under a probability distribution G if all permutations of $\{y_1, ..., y_n\}$ have the same joint distribution G. Namely; if

$$\mathrm{d}G(y_1, ..., y_n) = \mathrm{d}G(y_{\mathfrak{p}(1)}, ..., y_{\mathfrak{p}(n)})$$

for all permutations $\mathfrak{p}$ defined on the set $\{1, ..., n\}$.

Definition 2. An infinite sequence of random quantities $y_1, y_2...$ is infinitely exchangeable under a probability distribution G if every finite sub-sequence is finitely exchangeable under G.

Example 3. If a sequence of random quantities $y_{1:n} = (y_1, ..., y_n)$ is mutually independent, then it is exchangeable.

Solution. Obviously, as $\mathrm{d}G(y_1, ..., y_n) = \prod_{i=1}^{d} \mathrm{d}G(y_i)$ which is invariant to permutations of the indexes.

The assumption of exhcengeability leads to the following development, which theoretically justifies (to some extend) the existence of the Prior distribution, and the Bayesian paradigm.

Theorem 4. *(General representation theorem) If $y_1, y_2, ...$ is an infinitely exchangeable sequence of random quantities with probability distribution F, there exists a probability measure Π over $\mathcal{F}$, the space of all distribution functions on*

$\mathcal{Y}^n \subseteq \mathbb{R}^n$ for $n \geq 1$, such that the joint distribution function of $y_{1:n} = (y_1, y_2, ...)$ has CDF

$$G(y_1, ..., y_n) = \int_{\mathcal{F}} \prod_{i=1}^{n} F(y_i) d\Pi(F) \tag{1}$$

where F is an unknown/unobservable distribution function, $\Pi(F) = \lim_{n\rightarrow\infty} \mathsf{P}(\hat{F}_n)$ is a probability distribution on the space of functions $\mathcal{F}$, which is defined as a limit distribution on the empirical distribution function $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \leq x)$ defined by $y_1, ..., y_n$ (as $n \rightarrow \infty$), and $F(x) = \lim_{n\rightarrow\infty} \hat{F}_n(x)$.

Remark 5. The Representation Theorem shows that if $y_1, y_2, ...$ is infinitely exchangeable, then the elements of $y_{1:n}$ are i.i.d. conditional on the empirical distribution of $y_{1:n}$.

Interpretation The general representation Theorem 4 says

- $y_{1:n}$ are considered to be an i.i.d. sample generated from an unknown (i.e., random) distribution function F, (conditional on F); i.e. $y_i|F \sim F(\cdot)$.
- F is an the unknown CDF, which follows itself a probability distribution Π representing (prior) believes about F
- and F has the operational role of what You believe the empirical distribution function would look like for a large sample.

The parametric model form

Fact 6. *Given additional (problem specific) invariance assumptions (subjunctive judgments) regarding the generating process of $y_1, y_2, ...$, the unknown sampling distribution $F(y_i)$ in (1) can be written as a parametric model $F(y_i|\theta)$ labeled by an unknown parameter $\theta \in \Theta$, which is the limit of some function of $y_{1:n}$ (as $n \rightarrow \infty$), and there exists a probability distribution $d\Pi$ for θ such that*

$$G(y_1, ..., y_n) = \int_{\Theta} \prod_{i=1}^{n} F(y_i|\theta) d\Pi(\theta) \tag{2}$$

In PDF/PMF, (2) is written as

$$g(y_1, ..., y_n) = \int_{\Theta} \prod_{i=1}^{n} f(y_i|\theta) d\Pi(\theta) \tag{3}$$

Hereafter, we consider the concept of exchangeability by using the parametric form 2 and 3 for convenience.

Remark 7. Regarding the Bayesian paradigm, (2) provides a rational for the consideration of the uncertain parameter θ as a random variable and the subjective prior $\mathrm{d}\Pi(\theta)$. If $y_1, y_2, ...$ is an exchangeable sequence of real-valued random quantities, then any finite subset of them is an i.i.d. random sample from parametric model $\mathrm{d}F(\cdot|\theta)$ labeled by some uncertain parameter $\theta \in \Theta$, and there exists a (prior) probability distribution $\Pi(\theta)$ for θ which has to describe the initially available information about the parameter which labels the model.

The following 'representation Theorem with $0 - 1$ quantities' is a special case.

Theorem 8. *(Representation Theorem with $0 - 1$ quantities). If $y_1, y_2, ...$ is an infinitely exchangeable sequence of $0 - 1$ random quantities with probability measure G, there exists a distribution function Π such that the joint mass function $g(y_1, ..., y_n)$ for $y_1, ..., y_n$ has the form*

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

where

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

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

Proof. It is given in the exercise sheet to read. □

Remark 9. The representation of exchangeable sequence of $0-1$ random quantities $y_1, ..., y_n$ can be interpreted as follows: (1.): the x_i are considered to be conditional independent Bernoulli random quantities given the random quantity θ; i.e. $x_i|\theta \overset{iid}{\sim} \text{Br}(\theta)$. (2.): θ is itself assigned a probability distribution Π which can be interpreted as its prior distribution, (3.): by the SLLN, θ is defined as $\theta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} x_i$, and hence Π can be interpreted as beliefs about the limiting relative frequency of 1's.

The following Example 10 can justify the the notion of prior, posterior and predictive distribution in the context of the exchangeability.

Example 10. Let $y_1, y_{2,}...$ be an infinitely exchangeable sequence of random quantities under distribution G admitting a PDF/PMF g. Then from the representation theorem in (2), the conditional distribution $G(y_{n+1:n+m}|y_{1:n})$ has PDF/PMF

$$g(y_{n+1:n+m}|y_{1:n}) = \int_{\Theta} \prod_{i=m+1}^{n} f(y_i|\theta)\mathrm{d}\Pi(\theta|y_{1:n}) \quad \text{where} \quad \mathrm{d}\Pi(\theta|y_{1:n}) = \frac{\prod_{i=1}^{n} f(y_i|\theta)\mathrm{d}\Pi(\theta)}{\int_{\Theta}\prod_{i=1}^{n} f(y_i|\theta)\mathrm{d}\Pi(\theta)}$$

Solution. It can be shown that

$$\begin{aligned} g(y_{n+1:n+m}|y_{1:n}) &= \frac{g(y_{1:n}, y_{n+1:n+m})}{g(y_{1:n})} = \frac{\int_{\Theta}\prod_{i=1}^{n} f(y_i|\theta)\prod_{i=n+1}^{n+m} f(y_i|\theta)\mathrm{d}\Pi(\theta)}{\int_{\Theta}\prod_{i=1}^{n} f(y_i|\theta)\mathrm{d}\Pi(\theta)} \\ &= \int_{\Theta}\prod_{i=n+1}^{m} f(y_i|\theta)\underbrace{\frac{\prod_{i=1}^{n} f(y_i|\theta)\mathrm{d}\Pi(\theta)}{\int_{\Theta}\prod_{i=1}^{n} f(y_i|\theta)\mathrm{d}\Pi(\theta)}}_{=\mathrm{d}\Pi(\theta|y_{1n})} \end{aligned}$$

Remark 11. Subjectively specifying $\mathrm{d}G(y_{1:n})$ and then deriving $\mathrm{d}F(y_{1:n}|\theta)$ and $\mathrm{d}\Pi(\theta)$ is philosophically interesting. It can suggest useful sampling-model & prior ($\mathrm{d}F(y_{1:n}|\theta)$, $\mathrm{d}\Pi(\theta)$) decompositions, that allow the design of new meaningful models. On the other hand, it is often easier to subjectively specify the Bayesian model by $\mathrm{d}F(y_{1:n}|\theta)$ and $\mathrm{d}\Pi(\theta)$.

Example 12. Let $y_1, y_{2,}...$ be an infinitely exchangeable sequence of real valued random quantities with $y_i \in \mathbb{R}$ for any i.

1. Show that $\text{Corr}(y_i, y_j) \ge 0$, for $i \ne j$.

2. Find the condition under which (i.) I have $\text{Corr}(y_i, y_j) > 0$, for $i \ne j$ (ii.) I have $\text{Corr}(y_i, y_j) = 0$, for $i \ne j$

Hint: Consider the parametric form (2) of the general representation theorem.

Solution. Since sequence $y_1, y_{2,}...$ is infinitely exchangeable, I use the general representation theorem (the parametric form for simplicity). Hence, for a given θ, it is $x_i|\theta \overset{\text{iid}}{\sim} \mathrm{d}F(\cdot|\theta)$ for all i.

1. So

$$\begin{aligned} \text{Cov}_G(y_i, y_j) &= \text{E}_G(x_i^\top x_j) - \text{E}_G(x_i)^\top \text{E}_G(x_j) = \text{E}_G\left(\text{E}_F(x_i^\top x_j|\theta)\right) - \text{E}_G\left(\text{E}_F(x_i|\theta)\right)^\top \text{E}_G\left(\text{E}_F(x_j|\theta)\right) \\ &= \text{E}_\Pi\left(\text{E}_F(x_i^\top|\theta)\text{E}_F(x_j|\theta)\right) - \text{E}_\Pi\left(\text{E}_F(x_i|\theta)\right)^\top \text{E}_\Pi\left(\text{E}_F(x_j|\theta)\right) \\ &= \text{Var}_\Pi(\mu(\theta)) \ge 0 \end{aligned}$$

where $\mu(\theta) = \mathrm{E}_F(x_i|\theta)$ for all i. Also

$$\mathrm{Var}_G(y_i) = \mathrm{Var}_\Pi(\mathrm{E}_F(x_j|\theta)) + \mathrm{E}_\Pi(\mathrm{Var}_F(x_j|\theta)) = \mathrm{Var}_\Pi(\mu(\theta)) + \mathrm{E}_\Pi(\sigma^2(\theta))$$

where $\sigma^2(\theta) = \mathrm{Var}_F(x_i|\theta)$ for all i. Lets consider the 1-D case, that is requested; $y_i \in \mathbb{R}$ for any i. It is

$$\mathrm{Corr}(y_i, y_j) = \frac{\mathrm{Var}_\Pi(\mu(\theta))}{\mathrm{Var}_\Pi(\mu(\theta)) + \mathrm{E}_\Pi(\sigma^2(\theta))} \geq 0$$

2. For $\mathrm{Var}_\Pi(\mathrm{E}_F(x_i|\theta)) > 0$, I have $\mathrm{Corr}(y_i, y_j) > 0$. For $\mathrm{Var}_\Pi(\mathrm{E}_F(x_i|\theta)) = 0$, I have $\mathrm{Corr}(y_i, y_j) = 0$; NB correlation does not necessarily imply independence.

Remark 13. Example 12 shows that elements of infinite exchangeable sequence cannot be negatively correlated.

2 Practice

Question 14. *For practice try the Exercises 21, 22, 23, and 24 from the Exercise Sheet.*

Handout 5: Sufficiency and Exponential family of distributions[a] [b]

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

Aim

To explain, extend, and apply sufficiency concepts in the Bayesian framework, as well as the exponential family of distributions.

References:

- Raiffa, H., & Schlaifer, R. (1961; Chapter 2). Applied statistical decision theory.
- Casella, G., & Berger, R. L. (2002; Section 3.4, Chapter 6). Statistical inference (Vol. 2). Pacific Grove, CA: Duxbury.

[a] Author: Georgios P. Karagiannis.

[b] Acknowledgments to students in 2017 for spotting typos in the handouts.

1 Sufficiency

It is often of interest to simplify the Bayesian model by reducing the complexity of the observed quantity $y = (y_1, ..., y_n)$, where $y \in \mathcal{Y}$. For computational purposes: the dataset may involve large sample sizes (large n) or high-dimensional observable (high d); or for theoretical purposes: study how the information for the experiment affects the Bayesian model.

Sufficient statistic aims at summarizing the whole of relevant information supplied by the sample.

1.1 Summary statistics

Summary statistics (or statistics) are functions of observations summarizing the main features of a sequence of observable quantities, $y = (y_1, ..., y_n)$.

Definition 1. Let $y = (y_1, ..., y_n)$ be observable quantities such that $y \in \mathcal{Y}$, and let function $t : \mathcal{Y} \to \mathbb{R}^k$ with $k \leq n$. The quantity $t = t(y)$ is called a Statistic. The function $t(\cdot)$ (or mapping $t_n : \mathcal{Y} \to \mathbb{R}^k$) will be called statistic too.

Notation 2. Let $\mathcal{T} = \{t : t = t(y), \text{ for some } y \in \mathcal{Y}\}$ be the image of $\mathcal{Y}$ under $t(\cdot)$.

Notation 3. Let $\mathcal{Y}(t) = \{y : t(y) = t\}$ is the set comprising all y's such that the (the uncertain) $t(\cdot)$ assumes value t. Essentially, $t(\cdot)$ partitions the sample space $\mathcal{Y}$ into $\mathcal{X}_t$ for all $t \in \mathcal{T}$.

Note 4. The definition of statistic $t(\cdot)$ as a function of the observable quantity y induces a probability distribution $\mathrm{d}F(t|\theta)$ (of course it depends on the experiment $e \in \mathcal{E}$ but conditioning is omitted here) labeled by unknown parameter $\theta \in \Theta$, which is determined by sampling distribution $\mathrm{d}F(y|\theta)$. Given a prior distribution $\mathrm{d}\Pi(\theta)$ on θ, the posterior distribution of θ given $t = t(y)$ can be calculated as $\mathrm{d}\Pi(\theta|t)$ from the Bayesian theorem. Hence:

$$t|\theta \sim \mathrm{d}F(t|\theta); \qquad \theta \sim \mathrm{d}\Pi(\theta)$$

1.2 Bayesian sufficiency

Note 5. Sufficient statistic is a statistic that aims at summarizing the whole of relevant information supplied by the sample. We extend the concept of sufficiency and sufficient statistic (learned in SC2) to the Bayesian statistics.

Recall from SC2 that:

Definition 6. In the Frequentist statistics: A statistic $t : \mathcal{Y} \to \mathcal{T}$ is efficient statistic for θ (in the Frequentist sense) if the conditional distribution $\mathrm{d}F(y|t, \theta)$ does not depend on θ. I.e., ...iff the PDF/PMF is $f(y|t, \theta)$ does not depend on θ.

Note 7. In the Bayesian framework, it is reasonable to assume that, given the same prior info in θ, the coarser information from t and the richer information in y (regarding the outcome of an experiment) will lead to the some believes about θ if they lead to identical posterior probabilities.

Example 8. (Bernoulli model) For instance, in the Example with the Bernoulli-Beta model [HLN-3], the posterior of θ given the data $y = (y_1, ..., y_n)$ was

$$\theta|y \sim \text{Be}(\sum_{i=1}^{n} y_i + a, n - \sum_{i=1}^{n} y_i + b).$$

This is equivalent to

$$\theta|(n, \bar{y}) \sim \text{Be}(n\bar{y} + a, n - n\bar{y} + b)$$

Hence suffices to know $t = (n, \bar{y})$. A benefit is that $t = (n, \bar{y})$ has lower dimensionality (just 2 numbers) compared to $(n, y = (y_1, ..., y_n))$ $(1 + n$ numbers). So $t = (n, \bar{y})$ is easier/cheaper to store in the computer.

Definition 9. The statistic $t : \mathcal{Y} \to \mathcal{T}$ is (parametric) sufficient for θ if and only if for any prior distribution $\mathrm{d}\Pi(\theta)$ with pdf/pmf $\pi(\theta)$ we get

$$\mathrm{d}\Pi(\theta|t = t') = \mathrm{d}\Pi(\theta|y = y'), \quad \text{where} \quad t' = t(y')$$

for some observed data y' and $t' = t(y')$. Both the quantity t and the function/mapping $t(\cdot)$, as well as their realizations/values, will be called sufficient statistics.

Example 10. (Bernoulli model, cont. Example 8) The statistic $t = (n, \bar{y})$ is a parametric sufficient statistic.

Note 11. The following Theorem 12 provides a manner to identify a sufficient statistic in the sense of Definition 9. It examines the kernel-residue factorization of the likelihood function, and understands that the derived posterior can be determined by the kernel ok the likelihood while it is invariant to the residue.

Theorem 12. *Let $t : \mathcal{Y} \to \mathcal{T}$ be a statistic. Then t is a parametric sufficient statistic for theta in the sense of Definition 9 if and only if the likelihood function $L(\cdot|\cdot)$ on $\mathcal{Y} \times \Theta$ can be factorized as the product of a kernel function k on $\mathcal{Y} \times \Theta$ and a residue function ρ on Θ as*

$$L(\theta|y) = k(t(y)|\theta)\rho(y). \tag{1}$$

Proof. Exercise 31 in Exercise sheet □

Example 13. Let $y_i|\theta \sim \text{U}(\theta_1, \theta_2)$, iid for $i = 1, ..., n$. Then by factorization criterion

$$\begin{aligned} f(y|\theta) &= \prod_{i=1}^{n} \text{U}(y_i|\theta_1, \theta_2) = \prod_{i=1}^{n} \left[\frac{1}{\theta_2 - \theta_1} 1\left(y_i \in [\theta_1, \theta_2]\right) \right] = (\frac{1}{\theta_2 - \theta_1})^n \prod_{i=1}^{n} 1\left(y_i \in [\theta_1, \theta_2]\right) \\ &= (\frac{1}{\theta_2 - \theta_1})^n 1\left(\min_{\forall i=1:n} (y_i) \in [\theta_1, \theta_2]\right) 1\left(\max_{\forall i=1:n} (y_i) \in [\theta_1, \theta_2]\right) \end{aligned}$$

Then the sufficient statistic is $t := (n, \min_{\forall i=1:n}(y_i), \max_{\forall i=1:n}(y_i))$.

Note 14. The following Proposition suggests that; the concepts of Bayesian parametric sufficiency and Frequentist sufficiency are equivalent.

Proposition 15. *Let $t : \mathcal{Y} \to \mathcal{T}$ be a statistic. Then t is a parametric sufficient statistic in the sense of Definition 9 (in the Bayesian sense) if and only if t is sufficient statistic in the sense of Definition 6 (Frequentist sense).*

Proof. This is straightforward. According to the Neyman factorization theorem: t is sufficient statistic of θ in the Frequentist if and only if the likelihood can be decomposed as in (1) of Theorem 12. □

Definition 16. The statistic $t : \mathcal{Y} \to \mathcal{T}$ is predictive sufficient for the next outcome y_f given the Bayesian model ($y = (y_1, ..., y_n)$, $y \sim \mathrm{d}F(y|\theta)$, $\theta \sim \mathrm{d}\Pi(\theta)$) if and only if the predictive distribution $\mathrm{d}G(y_f|t = t')$ of a future outcome y_f given t and the predictive distribution $\mathrm{d}G(y_f|t = t')$ of a future outcome y_f given the whole data $y = (y_1, ..., y_n)$ are equal; i.e.

$$\mathrm{d}G(y_f|t = t') = \mathrm{d}G(y_f|y = y'), \quad \text{where} \quad t' = t(y').$$

Definition 17. A sufficient statistic $t = t(y)$ is called minimal sufficient statistic if for any other statistic $\tilde{t} = \tilde{t}(y)$, $t = t(y)$ is a function of $\tilde{t} = \tilde{t}(y)$.

Proposition 18. *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.*

Proof. Exercise 32 in the Exercise sheet. □

Example 19. (Bernoulli model, cont. Examples 8&10) Statistic $t = (n, \bar{y})$ is both predictive and parametric statistic in the Bernoulli model in Example 8&10), as the Bayesian model considered in exchangeable. It is also minimal sufficient statistic.

2 Exponential family of distributions

An important family of distributions which admits a reduction by means of sufficient statistics is the exponential family

Definition 20. A probability distribution $\mathrm{d}F(y|\theta)$ with pmf/pdf, $f(y|\theta)$ labeled by $\theta \in \Theta$, is said to belong to the k-parameter exponential family of distributions if it is of the form

$$f(y|\theta) = \mathrm{Ef}_k(y|u, g, h, \phi, \theta, c) = u(y)g(\theta)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(y)) \tag{2}$$

for $y \in \mathcal{Y}$ where $h := (h_1, ..., h_k)$, $\phi(\theta) = (\phi_1, ..., \phi_k)$ and given the functions u, h, ϕ, and constants $\{c_j\}$,

$$g(\theta)^{-1} = \begin{cases} \int_{\mathcal{X}} u(y)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(y))\mathrm{d}x < \infty & \text{, if } y \text{ is cont} \\ \sum_{x \in \mathcal{X}} u(y)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)h_j(y)) < \infty & \text{, if } y \text{ is disc} \end{cases}$$

Definition 21. The Exponential family of distributions is called regular if $\mathcal{Y}$ does not depend on θ; otherwise it is called non-regular.

Definition 22. If $\eta_j = c_j\phi_j(\theta)$ is taken to be the parameter in (2), so that

$$f(y|\eta) = \mathrm{Ef}_k(y|u, g, h, \eta) = u(y)\tilde{g}(\eta)\exp(\sum_{j=1}^{k} \eta_j h_j(y))$$

with normalizing constant $\tilde{g}(\eta)^{-1} < \infty$, we say that the exponential family has been given the natural particularization.

Theorem 23. *If $y = (y_1, y_2, ..., y_n)$ are generated from a regular k-parameter exponential family of distributions*

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

then $t := t(y) = (n, \sum_{i=1}^{n} h_1(y_i), ..., \sum_{i=1}^{n} h_k(y_i))$ is a sufficient statistic.

Proof. The likelihood is

$$\begin{aligned} f(y|\theta) = \prod_{i=1}^{n} \text{Ef}(y_i|u, g, h, \phi, \theta, c) &= \prod_{i=1}^{n} u(y_i)g(\theta)\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\sum_{i=1}^{n} h_j(y_i)) \\ &= (\prod_{i=1}^{n} u(y_i))(g(\theta))^n \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\sum_{i=1}^{n} h_j(y_i)) \end{aligned}$$

and from Neyman factorization criterion it is implied that $t := (n, \sum_{i=1}^{n} h_1(y_i), ..., \sum_{i=1}^{n} h_k(y_i))$ □

Example 24. (Exponential distribution) Let $y_i|\theta \sim \text{Ex}(\theta)$. Then,

$$f(y|\theta) = \text{Ex}(y|\theta) = \theta\exp(-\theta y),\ y \in \mathcal{Y} \equiv \mathbb{R}_+,\ \ \theta \in \mathbb{R}_+$$

with

$$u(y) = 1, \qquad g(\theta) = \theta, \qquad h(y) = y, \qquad \phi(\theta) = \theta, \qquad c = -1.$$

It is in the regular exponential family because $\mathcal{Y}$ does not depend on θ. From Theorem 23 the sufficient statistic is $t_n := (n, \sum_{i=1}^{n} y_i)$ or equiv. $t = (n, n\bar{y})$.

Example 25. (Bernoulli distribution) Let $y_i|\theta \sim \text{Br}(\theta)$. Then,

$$f(y|\theta) = \text{Br}(y|\theta) = \theta^y(1-\theta)^{1-y} \ = \exp(y\log(\theta) + (1-y)\log(1-\theta)) \ = (1-\theta)\exp(y\log(\frac{\theta}{1-\theta}))$$

with $y \in \mathcal{Y} \equiv \{0, 1\}$, $\theta \in [0, 1]$, and

$$u(y) = 1, \qquad g(\theta) = 1-\theta, \qquad h(y) = y, \qquad \phi(\theta) = \log(\frac{\theta}{1-\theta}), \qquad c = 1$$

It is in the regular exponential family because $\mathcal{Y}$ does not depend on θ. From Theorem 23 the sufficient statistic is $t_n := (n, \sum_{i=1}^{n} y_i)$ or equiv. $t = (n, n\bar{y})$.

Example 26. (Uniform distribution) Let $y_i|\theta \sim \text{U}(0, \theta)$. Then

$$f(y_i|\theta) = \text{U}(y_i|0, \theta) = \frac{1}{\theta}1\left(y_i \in [0, \theta]\right),\ y_i \in \mathcal{Y} \equiv [0, \theta],\ \ \theta \in \mathbb{R}_+$$

with

$$u(y_i) = 1, \qquad g(\theta) = 1/\theta, \qquad h(y) = 0, \qquad \phi(\theta) = \theta, \qquad c = 1.$$

It is in the non-regular exponential family because $\mathcal{Y}$ depends on θ. I cannot use Theorem 23 to find the sufficient statistic because Uniform distribution is a non-regular exponential distribution family. Instead, I can use Neyman factorization criterion

$$f(y|\theta) = \prod_{i=1}^{n} \text{U}(y_i|0, \theta) = (\frac{1}{\theta})^n \prod_{i=1}^{n} 1\left(y_i \in [0, \theta]\right) \ = (\frac{1}{\theta})^n 1\left(\max_{\forall i=1:n}(y_i) \in [0, \theta]\right)$$

Then the sufficient statistic is $t := (n, \max_{\forall i=1:n}(y_i))$.

Example 27. (Normal distribution) Let $y_i|\theta = \mathrm{N}(\mu, \sigma^2)$. Then

$$\begin{aligned} f(y_i|\theta) &= \mathrm{N}(y_i|\mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - \mu)^2) \\ &= (\frac{1}{2\pi})^{\frac{1}{2}} (\frac{1}{\sigma^2})^{\frac{1}{2}} \exp(-\frac{1}{2}\frac{1}{\sigma^2}y_i^2 + \frac{\mu}{\sigma^2}y_i - \frac{1}{2}\frac{\mu^2}{\sigma^2}) = (\frac{1}{2\pi})^{\frac{1}{2}} (\frac{1}{\sigma^2})^{\frac{1}{2}} \exp(-\frac{1}{2}\frac{\mu^2}{\sigma^2}) \exp(-\frac{1}{2}\frac{1}{\sigma^2}y_i^2 + \frac{\mu}{\sigma^2}y_i) \end{aligned}$$

$$u(y_i) = (\frac{1}{2\pi})^{\frac{1}{2}}, \quad g(\theta) = (\frac{1}{\sigma^2})^{\frac{1}{2}} \exp(-\frac{1}{2}\frac{\mu^2}{\sigma^2}), \quad h(y_i) = (y_i, y_i^2), \quad \phi(\theta) = (\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}), \quad c = (1, -\frac{1}{2}); \quad k = 2,$$

It is in the 2 parameter regular exponential family because $\mathcal{Y}$ does not depend on θ. From theorem 23, the sufficient statistic is $t := (n, \sum_{i=1}^n y_i, \sum_{i=1}^n y_i^2)$ or equiv $t = (n, \bar{y}, s_y^2)$.

Example 28. (Uniform distribution)Let $y_i|\theta \sim \mathrm{U}(\theta_1, \theta_2)$. Then

$$f(y|\theta) = \mathrm{U}(y|\theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1} 1\left(y \in [\theta_1, \theta_2]\right), \ y \in \mathcal{Y} \equiv [\theta_1, \theta_2], \ \theta_1, \theta_2 \in \mathbb{R}_+$$

$$u(y) = 1, \qquad g(\theta) = \frac{1}{\theta_2 - \theta_1}, \qquad h(x) = 0, \qquad \phi(\theta) = (\theta_1, \theta_2), \qquad c = (0, 0), \qquad k = 2$$

It is a 2-parameter non-regular exponential family because $\mathcal{Y}$ depends on θ. I cannot use Theorem 23 to find the sufficient statistic because it is a non-regular exponential distribution family. I can use Neyman factorization criterion.

Theorem 29. *Let $y_i \sim Ef_k(u, g, h, \phi, \theta, c)$ for $i = 1, ..., n$ an i.i.d. sample. The distribution of $s = (s_1(y), ..., s_k(y))$ with $s_j(y) = \sum_{i=1}^n h_j(y_i)$ has pdf/pmf of the form*

$$f(s|\theta) = \tilde{u}(s) g(\theta) \exp(\sum_{j=1}^k c_j \phi_j(\theta) s_j)$$

3 Likelihood principle

Consider two experiments e_1 and e_2, one yielding data y_1 and the other yielding data y_2. If the two likelihoods $L_1(y_1|\theta)$ and $L_2(y_2|\theta)$ are identical up to multiplication by arbitrary functions of y_1 or y_2, then they contain identical information about θ and lead to identical posterior distributions. The experiments might be very different in other respects, but those differences are irrelevant for inference about θ.

Likelihood Principle In making inferences or decisions about θ after y is observed, all relevant experimental information is contained in the likelihood function $L(y_{1:n}|\theta)$ for the observed $y_{1:n}$. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ), e.g. $L_1(y_1|\theta) = cL_2(y_2|\theta)$ for every $\theta \in \Theta$; hence they must lead to identical inferences for θ.

Remark 30. Likelihood Principle:

- ... implies that in order to draw any conclusion from an experiment only the actual observation matters and not the other possible outcomes that might have occurred but were not.
- ... does not say that all information about θ is contained in likelihood function $L(y_{1:n}|\theta)$; but just the experimental information. Other information relevant to the statistical analysis, such as prior information may exist.

Remark 31. Bayesian methods always satisfy the Likelihood principle. This is because, posterior knowledge about θ is expressed from the posterior distribution $\Pi(\theta|y)$ derived by the Bayesian theorem where all the knowledge regarding the experiment is expressed in the likelihood $L(y|\theta)$, and all the prior knowledge is expressed in the prior distribution $\Pi(\theta)$ exclusively.

Theorem 32. *The likelihood principle is satisfied in the Bayesian framework.*

Proof. Your belief about the uncertain parameter θ is represented by the posterior distribution. Consider two experiments e_1 and e_2, one yielding data y_1 and the other yielding data y_2, and assume that $L_1(y_1|\theta) = cL_2(y_2|\theta)$. Then if $\Pi_1(\theta|y_1)$ and $\Pi_2(\theta|y_2)$ have PDF/PMFs $\pi_1(\theta|y_1)$ and $\pi_2(\theta|y_2)$:

$$\pi(\theta|y_1) = \frac{L_1(y_1|\theta)\pi(\theta)}{\int_\Theta L_1(y_1|\theta)\mathrm{d}\Pi(\theta)} = \frac{\cancel{c}L_2(y_2|\theta)\pi(\theta)}{\int_\Theta \cancel{c}L_2(y_2|\theta)\mathrm{d}\Pi(\theta)} = \pi(\theta|y_2)$$

□

Example 33. (Binomial vs Negative Binomial experiment) We are given a coin and are interested in the success frequency θ of having it come up heads when flipped. An experiment is conducted by flipping the coin (independently) in a series of trials, the result of which is the observation of 3 heads and 9 tails (hence 12 flips in total). This is not yet enough information to specify $f(x|\theta)$, since the 'series of trials' was not explained. Two possibilities are:

M_1**:** the experiment consisted of a predetermined $n = 12$ flips, so that the number of heads $r \sim \text{Bn}(n = 12, \theta)$ with observed $r = 3$. (Binomial experiment)

$$\text{Bn}(r|n,\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}1(r \in \{0, ..., n\})$$

M_2**:** the experiment consisted of flipping the coin until $r = 3$ heads were observed, so that the number of trials $n \sim \text{Nb}(r = 3, \theta)$ with observed $n = 12$. (Negative binomial experiment)

$$\text{Nb}(n|r,\theta) = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}1(n \in \{r, r+1, ...\})$$

The two models/experiments, the likelihoods are such as

$$\text{Bn}(r|n,\theta) \propto \text{Nb}(n|r,\theta) \propto \theta^r(1-\theta)^{n-r}$$

In the Bayesian framework, one could assign a prior $\theta \sim \text{Be}(a = 1, b = 1) \equiv \text{U}(0,1)$ and get the following Bayesian models

$$M_1 : \begin{cases} x & \sim \text{Bn}(n = 12, \theta) \\ \theta & \sim \text{Be}(a = 1, b = 1) \end{cases} \qquad M_2 : \begin{cases} n & \sim \text{Nb}(r = 3, \theta) \\ \theta & \sim \text{Be}(a = 1, b = 1) \end{cases}$$

The two Bayesian models lead to the same posterior inference, since the posterior PDFs of θ are

$$\begin{aligned}
\pi(\theta|n, r, M_1) &\propto \text{Bn}(r|n,\theta)\text{Be}(\theta|a,b) \\
&\propto \theta^r(1-\theta)^{n-r}\theta^{a-1}(1-\theta)^{b-1} \propto \theta^{r+a-1}(1-\theta)^{n-r+b-1} \\
&\propto \text{Be}(\theta|r+a, n-r+b) = \text{Be}(\theta|4, 10) \\
\pi(\theta|n, r, M_2) &\propto \text{Nb}(n|r,\theta)\text{Be}(\theta|a,b) \propto ... \\
&\propto \text{Be}(\theta|r+a, n-r+b) = \text{Be}(\theta|4, 10)
\end{aligned}$$

meaning that we learn the same thing from both experiments.

- [The derivation of hypothesis test in this bullet is out of the scope; You are not required to know it]. In the Frequentist framework, if we wish to do the hypothesis test for $\text{H}_0 : \theta = 0.5$ vs. $\text{H}_1 : \theta < 0.5$, (i.) in case M_1: we get p-value $= \text{Bn}(r \leq 3|n = 12, \theta = 1/2) = \sum_{i=0}^{3}\text{Bn}(r|n = 12, \theta = 1/2) = 0.073 > 5\%$ (I do not reject H_0) (ii.) while in case M_2: we get p-value $= \text{Nb}(n \geq 12|\theta = 1/2) = 1 - \sum_{n=0}^{11}\text{Nb}(n|r = 9, \theta =$

$1/2) = 0.032 < 5\%$ (I reject H_0) !!! The two models lead to quite different conclusions (in Frequentist stats), and are in contradiction to the Likelihood Principle.

4 Practice

Question 34. *For practice try the Exercises 28, 29 from the Exercise Sheet.*

Handout 6: Conjugate priors[a]

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

Aim: Explain the prior distribution. Explain, theorize, and construct conjugate and conditional conjugate prior distribution.

References:

- Raiffa, H., & Schlaifer, R. (1961; Sections 3.1-3.3). Applied statistical decision theory.
- Berger, J. O. (2013; Sections 4.2.2). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- Robert, C. (2007; Sections 3.1 & 3.3). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

Web applets: `https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/`

[a] Author: Georgios P. Karagiannis.

1 Proper & improper priors

Note 1. Priors do not necessarily need to be probability distributions but they need to lead to posterior probability distributions.

Definition 2. The prior $\Pi(\theta)$ with pdf/pmf $\pi(\theta) > 0$ for $\theta \in \Theta$, is called proper prior if

$$\int_{\Theta} \pi(\theta)\mathrm{d}\theta < \infty \text{ when } \theta \text{ is continuous;} \qquad \text{and} \qquad \sum_{\forall \theta \in \Theta} \pi(\theta) < \infty, \text{ when } \theta \text{ is discrete}$$

and hence it is a probability distribution ; and improper prior if

$$\int_{\Theta} \pi(\theta)\mathrm{d}\theta = \infty \text{ when } \theta \text{ is continuous;} \qquad \text{and} \qquad \sum_{\forall \theta \in \Theta} \pi(\theta) = \infty, \text{ when } \theta \text{ is discrete}$$

and hence it is a not a probability distribution.

Note 3. An improper prior $\Pi(\theta)$ can only be used for inference if it leads to a well defined posterior probability distribution (aka proper posterior); namely if the 'Properness condition'

$$\int_{\Theta} f(y|\theta)\pi(\theta)\mathrm{d}\theta < \infty \tag{1}$$

is satisfied for the observable sequence y at hand. If (1) is not satisfied, posterior quantities like mean, median, variance have no meaning.

Proposition 4. *If the sampling distribution $F(\cdot|\theta)$ is discrete and the prior $d\Pi(\theta)$ is proper, then the posterior $\Pi(\theta|y)$ is always proper.*

Proof. Provided as an Exercise 34 in the Exercise sheet. □

Proposition 5. *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.*

Proof. Provided as an Exercise 35 in the Exercise sheet. □

2 Conjugate priors

Note 6. We aim at specifying a prior distribution family, which (i.) leads to a tractable (to some extend) posterior distribution, (ii.) is rich enough to allow us to quantify prior believe, and (iii.) has a reasonable interpretation.

Definition 7. Let $\mathcal{F} = \{F(y|\theta);\, \forall\theta \in \Theta\}$ be a family of sampling distributions. A family of prior distributions $\mathcal{P}$ on Θ is said to be (natural) conjugate for $\mathcal{F}$ if the posterior $\Pi(\theta|y)$ belongs to $\mathcal{P}$ for all prior $\Pi(\theta) \in \mathcal{P}$ and all $F(y|\theta) \in \mathcal{F}$; i.e.

$$\Pi(\theta|y) \in \mathcal{P},\ \ \forall F(y|\theta) \in \mathcal{F}\,\text{and}\,\Pi(\theta){\in}\mathcal{P}.$$

Note 8. By specifying a tractable conjugate prior distribution $\mathrm{d}\Pi(\theta)$, we can achieve tractability for the posterior $\mathrm{d}\Pi(\theta|y)$ since it belongs to the same distribution family as the prior.

2.1 General derivation

Note 9. Let $y = (y_1, ..., y_n)$ be observables. We restrict the derivation of conjugate priors in cases where y_i are drawn from $F(\cdot|\theta)$ conditionally independent of θ, and there exists a parameteric sufficient statistic $t : \mathcal{Y} \to \mathbb{R}^k$ with $t(y_1, ..., y_n) = t \in \mathbb{R}^k$ where its dimension k is independent on the number of observables n.

What is in the box describes the rational of the approach and can be skipped.

Fact 10. [a]*Let $t^{(1)} = t(y_1, ..., y_q)$ and $t^{(2)} = t(y_{q+1}, ..., y_n)$ be sufficient statistics of two data-sets. Then, under the conditions of Note 9, there exists a binary operator $*$ such that*

$$y^{(1)} * y^{(2)} = y^* := (y_1^*, ..., y_k^*) \tag{2}$$

such that

$$f(y_1, ..., y_n|\theta) \propto k(y^*|\theta) \text{ and } \quad k(y^*|\theta) \propto k(y^{(1)}|\theta)k(y^{(2)}|\theta)$$

Note 11. Let statistic $t : \mathcal{Y} \to \mathbb{R}^k$ with $t(y_1, ..., y_n) = t \in \mathbb{R}^k$ be parametric sufficient, and let its dimension k be independent from data size n. Then from Neyman factorization theorem, the likelihood can be factorized as

$$f(y|\theta) = k(t(y)|\theta)\rho(y)\ \propto k(t(y)|\theta), \tag{3}$$

where $\rho(y)$ is the residual term of a likelihood kernel $k(t(y)|\theta)$ of θ. By Bayes theorem, the posterior distribution is

$$\mathrm{d}\Pi(\theta|y) = \frac{f(\theta|y)\mathrm{d}\Pi(\theta)}{\int f(\theta|y)\mathrm{d}\Pi(\theta)} = \frac{k(t(y)|\theta)\mathrm{d}\Pi(\theta)}{\int k(t(y)|\theta)\mathrm{d}\Pi(\theta)} \tag{4}$$

Assuming fictitious observables y' quantifying Your prior believe about θ (prior to getting the experimental info from observable data y), such that the sufficient statistic is $t' = t(y') \in \mathbb{R}^k$, I could possibly choose prior $\mathrm{d}\Pi(\theta)$ such as

$$\mathrm{d}\Pi(\theta) = \underbrace{\frac{1}{N(\tau)}k(t' = \tau|\theta)}_{=\pi(\theta)}\mathrm{d}\theta \text{ with } \pi(\theta) \propto k(\tau|\theta) \tag{5}$$

by assigning some researcher specified fixed hyper-parameters $\tau = (\tau_0, ..., \tau_{k-1})$ on $t' = t(y')$ such that the normalising constant is finite

$$N(\tau) = \begin{cases} \int k(t(y) = \tau|\theta)\mathrm{d}\theta < \infty & \text{cont.} \\ \sum_{\forall \theta \in \Theta} k(t(y) = \tau|\theta) < \infty & \text{discr.} \end{cases}$$

Then from Fact 10, the posterior (4) could get a form

$$\mathrm{d}\Pi(\theta|y) = \frac{k(t(y)|\theta)\mathrm{d}\Pi(\theta)}{\int k(t(y)|\theta)\mathrm{d}\Pi(\theta)} = \frac{k(t(y)|\theta)k(\tau|\theta)\mathrm{d}\theta}{\int k(t(y)|\theta)k(\tau|\theta)\mathrm{d}\theta} = \frac{k(t(y) * \tau|\theta)\mathrm{d}\theta}{\int k(t(y) * \tau|\theta)\mathrm{d}\theta} = \frac{1}{N(t(y) * \tau)}k(t(y) * \tau|\theta)\mathrm{d}\theta \tag{6}$$

where $*$ is the binary operator (2) that combines the two kernels $k(t(y)|\theta)k(\tau|\theta) = k(t(y) * \tau|\theta)$. Essentially the prior $\mathrm{d}\Pi(\theta)$ (5) and the posterior $\mathrm{d}\Pi(\theta|y)$ (6) belong to the same distribution family $\mathcal{P}$. The only difference is in the hyper-parameter values. In the posterior distribution the hyper-parameters combine both the prior info quantified in τ and the experimental info quantified in $t(y)$ according to the binary operator $*$.

[a] Raiffa, H., & Schlaifer, R. (1961; Sections 3.1-3.3). Applied statistical decision theory.

Theorem 12. *Let $y = (y_1, ..., y_n)$ be observable quantities drawn from $F(y|\theta)$ independently conditional on θ, and let $f(y|\theta)$ be the likelihood with sufficient statistic $t := t(y)$ of a fixed dimension k independent from n. The conjugate prior $\Pi(\theta)$ with hyper-parameter τ of the likelihood $f(y|\theta)$ can be specified by setting its pdf/pmf as*

$$\pi(\theta) := \tilde{\pi}(\theta|\tau) = \frac{1}{N(\tau)}k(\tau|\theta) \propto k(\tau|\theta) \tag{7}$$

where $k(\cdot|\theta)$ is a kernel of the likelihood from the Neyman factorization

$$f(y|\theta) = k(t(y)|\theta)\rho(y) \propto k(t(y)|\theta),$$

and τ are hyper-parameters such that $N(\tau) = \int k(\tau|\theta)d\theta < \infty$.

Note 13. Essentially, in Theorem 12, and Note 11, we expect that since the likelihood kernel $k(\tau|\theta)$ is tractable, it may lead to a tractable conjugate prior, and hence to a tractable posterior distribution.

Note 14. Once the conjugate family of prior distributions $\Pi(\theta)$ has been specified, You can assign values on the hyper-parameters τ based on your a priori information. In fact, the values assigned on τ do not necessarily need to lie in the support of the sufficient statistics $\mathcal{T}$; the only restriction is that τ has to lead to a proper posterior $N(t(y) * \tau) < \infty$.

Example 15. Let $y = (y_1, ..., y_n)$ be observables drawn iid from sampling distribution $y_i \overset{\text{iid}}{\sim} \mathrm{U}(0, \theta)$ for all $i = 1, ..., n$. Specify the conjugate prior for θ.

Pareto distribution: If $x \sim \mathrm{Pa}(a, b)$, then it has a pdf $f(x) = ab^a(\frac{1}{\theta})^{a+1}1(b < \theta)$

Solution. The likelihood $f(y|\theta)$ can be factorized as

$$f(y|\theta) = \prod_{i=1}^{n} \mathrm{U}(y_i|0, \theta) = (\frac{1}{\theta})^n \prod_{i=1}^{n} 1\left(y_i \in [0, \theta]\right) = \underbrace{(\frac{1}{\theta})^n 1\left(\max_{\forall i=1:n}(y_i) \in [0, \theta]\right)}_{=k(t(y)|\theta)}$$

with sufficient statistic $t = (n, \max_{\forall i=1:n}(y_i))$. Hence, I set

$$\pi(\theta) := \pi(\theta|\tau) \propto (\frac{1}{\theta})^{\tau_0} 1\left(\tau_1 \in [0, \theta]\right) \propto \mathrm{Pa}(\theta|a = \tau_0 - 1, b = \tau_1)$$

By Bayes theorem the posterior is

$$\pi(\theta|y) \propto \prod_{i=1}^{n} \text{Un}(y_i|0,\theta)\text{Pa}(\theta|\tau_0-1,\tau_1) \propto \overbrace{(\frac{1}{\theta})^n \prod_{i=1}^{n} 1(y_i < \theta)}^{=\prod_{i=1}^{n}\text{Un}(y_i|0,\theta)} \times \overbrace{(\frac{1}{\theta})^{\tau_0} 1(\theta > \tau_1)}^{\propto \text{Pa}(\theta|\tau_0-1,\tau_1)}$$

$$\propto (\frac{1}{\theta})^{n+\tau_0} \underbrace{\prod_{i=1}^{n} 1(\theta > x_i) 1(\theta > \tau_1)}_{=1(\theta > \max(\tau_1, x_{(n)}))} \propto \text{Pa}(\theta|a^* = n+\tau_0-1, b^* = \max(\tau_1, x_{(n)})).$$

where $\theta > \max(\tau_1, x_{(n)})$.

Example 16. Consider the model of Normal linear regression where the observables are pairs (ϕ_i, y_i) for $i = 1, ..., n$, assumed to be modeled according to the sampling distribution $y_i|\beta, \sigma^2 \sim \text{N}(\phi_i^\top \beta, \sigma^2)$ for $i = 1, ..., n$ with unknown $(\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$. Find the conjugate prior for (β, σ^2).

Hint: $(y - \Phi\beta)^\top (y - \Phi\beta) = (\beta - \hat{\beta}_n)^\top \left[\Phi^\top \Phi\right] (\beta - \hat{\beta}_n) + (n + d - 2)\hat{\sigma}_n^2$;

$$\hat{\beta}_n = (\Phi^\top \Phi)^{-1} \Phi y \, ; \qquad \hat{\sigma}_n^2 = \frac{(y - \Phi\hat{\beta}_n)^\top (y - \Phi\hat{\beta}_n)}{n + d - 2}$$

Solution. The likelihood is

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

$$\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta}_n)^\top \left[\Phi^\top \Phi\right] (\beta - \hat{\beta}_n) - \frac{1}{2\sigma^2}(n + d - 2)\hat{\sigma}_n^2\right)}_{=k(t(y)|\beta, \sigma^2)}$$

where Φ is the design matrix and the sufficient statistic is $t = (n, \Phi y, \Phi^\top \Phi)$. Then, given prior hyper-parameters $\tau = (\tau_0, \tau_1, \tau_2, \tau_3)$, I set as a conjugate prior

$$\pi(\beta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\tau_0}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \tau_1)^\top \tau_2 (\beta - \tau_1) - \frac{1}{\sigma^2}\tau_3\right)$$

$$\propto \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \tau_1)^\top \tau_2 (\beta - \tau_1)\right)}_{\propto \text{N}(\beta|\tau_1, \tau_2^{-1}\sigma^2)} \times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{\tau_0 - d}{2} - 1 + 1} \left(-\frac{1}{\sigma^2}\tau_3\right)}_{=\text{IG}(\sigma^2|\frac{\tau_0 - d}{2} - 1, \tau_3)}$$

but as $\tau = (\tau_0, \tau_1, \tau_2, \tau_3)$ are just arbitrary parameters set by the researcher, I can use a friendlier parametrization[1]

$$\beta|\sigma^2 \sim \text{N}(\mu_0, V_0\sigma^2);\text{prior distr}$$
$$\sigma^2 \sim \text{IG}(a_0, \kappa_0)\text{prior distr}$$

[1] In Exercise 26 of theExercise sheet, the posterior $\pi(\beta, \sigma^2|y)$ is derived as

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

with some hyper-parameters $\mu_n, V_n, a_n, \kappa_n$ computed there.

2.2 Conjugate priors for Exponential families [2]

Note 17. Exponential family of distributions cover a large range of distributions satisfying the conditions in Note 9.

Fact 18. *(Pitman-Koopman-Lemma) If a distribution family $\{F(y|\theta),\ \forall\theta \in \Theta\}$ is such that there exists a sufficient statistic whose dimension is independent on the number of observations and the support $y \in \mathcal{Y}$ of $F(y|\theta)$ does not depend on θ, then it is an exponential family.*

Note 19. When the parametric model is member of the Exponential family, a conjugate prior distribution on its uncertain parameters can be specified.

Theorem 20. *Let $y = (y_1, ..., y_n)$ be observable quantities generated from an exponential family distribution as*

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

with pdf/pmf

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

Then the conjugate prior distribution $\mathrm{d}\Pi(\theta)$ *for the likelihood has pdf/pmf of the form*

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

for $\theta \in \Theta$, where $\tau = (\tau_0, \tau_1, ..., \tau_k)$ are hyper-parameters is such that

$$K(\tau) = \begin{cases} \int_\Theta g(\theta)^{\tau_0}\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j) < \infty & \textit{cont.} \\ \sum_{\forall\theta\in\Theta} g(\theta)^{\tau_0}\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j) < \infty & \textit{discr.} \end{cases}$$

Proof. The likelihood is

$$f(y|\theta) = \prod_{i=1}^{n}\mathrm{Ef}(y_i|u, g, h, c, \phi, \theta, c) = \prod_{i=1}^{n} u(y_i)\underbrace{g(\theta)^n\exp(\sum_{j=1}^{k} c_j\phi_j(\theta)(\sum_{i=1}^{n} h_j(y_i)))}_{=k(t(y)|\theta)}.$$

with sufficient statistic for θ

$$t(y) = (n, \sum_{i=1}^{n} h_1(y_i), ..., \sum_{i=1}^{n} h_k(y_i)) = (t_0, ..., t_k)$$

Let $\tau = (\tau_0, \tau_1, ..., \tau_n)$. The conjugate prior form has the form

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

□

Note 21. Intuition about the prior in Theorem 20: τ_0 replaces the sample size n, and hence τ can be thought of as being the weight of prior info or 'quality of prior info'; i.e. the larger the value the stronger the prior info. The rest $\tau_1, ...\tau_k$ can be thought of as summarizing the prior info.

Example 22. Let $y = (y_1, ..., y_n)$ be observable quantities, generated from an exponential family of distributions as

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

[2] https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/

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

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

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

$$\pi(\theta|y) = \pi(\theta|\tau^*) \ \propto g(\theta)^{\tau^*} \exp(\sum_{j=1}^{k} c_j\phi_j(\theta)\tau_j^*) \qquad (8)$$

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

Solution. According to the Bayes theorem, where

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

Example 23. Let $y = (y_1, ..., y_n)$ be observables drawn iid from a Bernoulli sampling distribution $y_i \overset{\text{iid}}{\sim} \text{Br}(\theta)$ for all $i = 1, ..., n$ where $\theta \in [0, 1]$. Specify a conjugate prior distribution for θ.

Hint: Beta distribution: if $x \sim \text{Be}(a, b)$, then $f(x) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}1(x \in [0, 1])$

Solution. The sampling distribution $f(x|\theta)$ is the Bernoulli distribution which belongs to the exponential family as

$$f(y_i|\theta) = \text{Br}(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i} = (1-\theta)\exp(\log(\frac{\theta}{1-\theta})y_i)$$

with $u(y_i) = 1$, $g(\theta) = (1-\theta)$, $c_1 = 1$, $\phi_1(\theta) = \log(\frac{\theta}{1-\theta})$, $h_1(y_i) = y_i$. The corresponding conjugate prior has pdf such as

$$\pi(\theta) \propto g(\theta)^{\tau_0} \exp(c_1\phi_1(\theta)\tau_j) = (1-\theta)^{\tau_0} \exp(\log(\frac{\theta}{1-\theta})\tau_1) \ = \theta^{(\tau_1+1)-1}(1-\theta)^{(\tau_0-\tau_1+1)-1}$$

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

$$\theta \sim \text{Be}(a, b)$$

where $a = \tau_1 + 1 > 0$, $b = \tau_0 - \tau_1 + 1 > 0$.

Example 24. Let $y = (y_1, ..., y_n)$ be observables drawn iid from sampling distribution $y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$ for all $i = 1, ..., n$, where $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, +\infty)$ is unknown. Specify a conjugate prior distribution for $\theta = (\mu, \sigma^2)$.

Solution. The sampling distribution $f(x|\mu, \sigma^2)$ is Normal distribution which is member of the regular 2-parameter exponential family, since

$$f(y_i|\mu, \sigma^2) = \text{N}(y_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(y_i - \mu)^2) = (\frac{1}{2\pi})^{\frac{1}{2}}(\frac{1}{\sigma^2})^{\frac{1}{2}}\exp(-\frac{1}{2}\frac{\mu^2}{\sigma^2})\exp(-\frac{1}{2}\frac{1}{\sigma^2}y_i^2 + \frac{\mu}{\sigma^2}y_i)$$

with $\quad u(y_i) = (\frac{1}{2\pi})^{\frac{1}{2}}, \quad g(\theta) = (\frac{1}{\sigma^2})^{\frac{1}{2}} \exp(-\frac{1}{2}\frac{\mu^2}{\sigma^2}), \quad h(y_i) = (y_i, y_i^2), \quad \phi(\theta) = (\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}), \quad c = (1, -\frac{1}{2})$

The corresponding conjugate prior has pdf such as

$$\begin{aligned}
\pi(\mu,\sigma^2) &\propto \left(\sqrt{\frac{1}{\sigma^2}} \exp(-\frac{1}{2}\frac{1}{\sigma^2}\mu^2)\right)^{\tau_0} \exp(\mu\frac{1}{\sigma^2}\tau_1 - \frac{1}{2}\frac{1}{\sigma^2}\tau_2) \\
&\propto \underbrace{(\frac{1}{\sigma^2/\tau_0})^{\frac{1}{2}} \exp(-\frac{1}{2}\frac{1}{\sigma^2/\tau_0}(\mu - \frac{\tau_1}{\tau_0})^2)}_{\propto \text{N}(\mu|\frac{\tau_1}{\tau_0}, \frac{\sigma^2}{\tau_0})} \underbrace{(\frac{1}{\sigma^2})^{\frac{(\tau_0-3)}{2}+1} \exp(-\frac{1}{\sigma^2}\frac{1}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0}))}_{\propto \text{IG}(\sigma^2|\frac{\tau_0-3}{2}, \frac{1}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0}))}
\end{aligned}$$

where $\tau = (\tau_0, \tau_1, \tau_2)$. I recognize that the prior distribution is of standard form $\pi(\theta|\mu_0, n_0, a_0, \kappa_0) = \text{N}(\mu|\mu_0, \frac{\sigma^2}{\lambda_0})\text{IG}(\sigma^2|a_0, \kappa_0)$, with $\mu_0 = \frac{\tau_1}{\tau_0}$, $\lambda_0 = \tau_0$, $a_0 = \frac{\tau_0 - 3}{2}$, and $b_0 = \frac{1}{2}(\tau_2 - \frac{\tau_1^2}{\tau_0})$.

3 Conditional conjugate

Note 25. In some problems involving more realistic/complicated statistical models, certain computational tools, (e.g., the Gibbs sampler in Term 2), require the availability of tractable posterior conditionals instead of that of the full joint posterior. Specifying, conditional conjugate priors is a way to achieve this.

Definition 26. Let $\mathcal{F} = \{F(y|\theta_1, \theta_2);\, \forall\theta_1 \in \Theta_1, \forall\theta_2 \in \Theta_2\}$ be a family of sampling distributions. A family of prior distributions $\mathcal{P}_{\theta_1}$ for θ_2 conditional on θ_1 is said to be conditional conjugate for $\mathcal{F}$ if the posterior $\Pi(\theta_2|y, \theta_1)$ belongs to $\mathcal{P}_{\theta_1}$ for all prior $\Pi(\theta_2|\theta_1) \in \mathcal{P}_{\theta_1}$ and all $F(y|\theta_1, \theta_2) \in \mathcal{F}$; i.e.

$$\Pi(\theta_1, \theta_2|y) \in \mathcal{P}_{\theta_1}, \;\; \forall F(y|\theta_1, \theta_2) \in \mathcal{F} \,\text{and}\, \Pi(\theta_2|\theta_1) {\in}\, \mathcal{P}_{\theta_1}.$$

Note 27. The conditional conjugate prior for θ_2 conditional θ_1 is specified from Theorem 12 as the conjugate prior of θ_2 on $F(y|\theta_1, \theta_2)$ given that parameter θ_1 is known/fixed. Based on this, Neyman factorization is applied as

$$f(y|\theta_1, \theta_2) = k_1(t(y)|\theta_1)\rho(y|\theta_1) \;\propto k(t(y)|\theta_1),$$

and the prior is specified according to

$$\pi(\theta_1) \propto k_1(\tau_1|\theta_1) \tag{9}$$

for some researchers specified prior hyper-parameter vector τ_1. Likewise, I get the conditional conjugate prior for θ_1 conditional θ_2 as $\pi(\theta_2) \propto k_2(\tau_2|\theta_2)$. The join prior $\Pi(\theta)$ satisfying conditional conjugation for θ_1 and θ_2 is

$$\pi(\theta) = \pi(\theta_1)\pi(\theta_2) \propto k_1(\tau_1|\theta_1)k_2(\tau_2|\theta_2).$$

This derivation is extendable to any number of blocks $\theta = (\theta_1, ..., \theta_B)$.

Example 28. Consider the the model of Normal linear regression where the observables are pairs (ϕ_i, y_i) for $i = 1, ..., n$, assumed to be modeled according to the sampling distribution $y_i|\beta, \sigma^2 \sim \text{N}(\phi_i^\top\beta, \sigma^2)$ for $i = 1, ..., n$ with unknown $(\beta, \sigma^2) \in \mathbb{R}^d \times \mathbb{R}_+$. Find the conditional conjugate priors for (β, σ^2).

Solution. The likelihood kernel is

$$f(y|\beta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta}_n)^\top \left[\Phi^\top\Phi\right](\beta - \hat{\beta}_n) - \frac{1}{2\sigma^2}(n + d - 2)\hat{\sigma}_n^2\right) \tag{10}$$

To find the conditional conjugate $\pi(\beta)$: I consider σ^2 as fixed/known/nuisance, and hence the kernel in 9 is

$$f(y|\beta, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta}_n)^\top \left[\Phi^\top\Phi\right](\beta - \hat{\beta}_n)\right)$$

leading to a conjugate prior

$$\pi(\beta) \propto \exp\left(-\frac{1}{2}(\beta-\mu_0)^\top \overbrace{V_0^{-1}}^{\text{absorbs constant } \sigma^2} (\beta-\mu_0)\right) \propto \mathrm{N}(\beta|\mu_0, V_0)$$

To find the conditional conjugate $\pi(\sigma^2)$: I consider β as fixed/known/nuisance, and hence the likelihood kernel in 9 is

$$f(y|\beta,\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\overbrace{n}^{\text{data}}}{2}} \exp\left(-\frac{1}{\sigma^2}\frac{\overbrace{(\beta-\hat{\beta}_n)^\top \left[\Phi^\top\Phi\right](\beta-\hat{\beta}_n)+(n+d-2)\hat{\sigma}_n^2}^{\text{data/constants}}}{2}\right)$$

leading to a conjugate conditional prior

$$\pi(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{a_0+1} \exp\left(-\frac{1}{\sigma^2}\kappa_0\right) \propto \mathrm{IG}(\sigma^2|a_0,\kappa_0)$$

Then the conditional conjugate $\pi(\beta,\sigma^2)=\pi(\beta)\pi(\sigma^2)$ is

$$\begin{cases}\beta \sim \mathrm{N}(\mu_0, V_0); \\ \sigma^2 \sim \mathrm{IG}(a_0,\kappa_0)\end{cases} \tag{11}$$

The full posterior distribution $\Pi(\beta|y,\sigma^2)$ is $\beta|y,\sigma^2 \sim \mathrm{N}(\mu_n, V_n)$, computed by Bayesian theorem as

$$\pi(\beta|y,\sigma^2) \propto f(y|\beta,\sigma^2)\pi(\beta) \propto \exp\left(-\frac{1}{2}(\beta-\hat{\beta}_n)^\top\left[\frac{\Phi^\top\Phi}{\sigma^2}\right](\beta-\hat{\beta}_n)-\frac{1}{2}(\beta-\mu_0)^\top V_0^{-1}(\beta-\mu_0)\right) \propto \mathrm{N}(\mu_n, V_n')$$

$$\text{with} \quad V_n' = \left[\frac{\Phi^\top\Phi}{\sigma^2}+V_0^{-1}\right]^{-1} \quad \text{and} \quad \mu_n = V_n'\left[\frac{\Phi^\top\Phi}{\sigma^2}\hat{\beta}_n + V_0^{-1}\mu_0\right]$$

and $\Pi(\sigma^2|y,\beta)$ is $\sigma^2|y,\beta \sim \mathrm{IG}(a_n,\kappa_n)$, computed by Bayesian theorem as

$$\pi(\sigma^2|y,\beta) \propto f(y|\beta,\sigma^2)\pi(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a_0+1} \exp\left(-\frac{1}{\sigma^2}\left[\frac{1}{2}(\beta-\hat{\beta}_n)^\top\left[\Phi^\top\Phi\right](\beta-\hat{\beta}_n)+\kappa_0\right]\right) \propto \mathrm{IG}(\sigma^2|a_n,\kappa_n)$$

with $\kappa_n = \frac{1}{2}(\beta-\hat{\beta}_n)^\top\left[\Phi^\top\Phi\right](\beta-\hat{\beta}_n)+\kappa_0$ and $a_n = \frac{n}{2}+a_0$. Hence, according to Definition 26, we verified conditional conjugation of (11) with the associated full conditional posteriors

$$\begin{cases}\beta|y,\sigma^2 \sim & \mathrm{N}(\mu_n, V_n) \\ \sigma^2|y,\beta \sim & \mathrm{IG}(a_n,\kappa_n)\end{cases}$$

4 Practice

Question 29. *For practice try Exercises 36, 33, and 37 from the Exercise sheet.*

Question 30. *Consider a Normal regression problem,*

$$y_i = \phi_i^\top\beta+\epsilon, \;\; \epsilon \sim N(0,\sigma^2)$$

where let's say y_i denotes the length (in cm), and $\phi_i = (1, x_i)$ with x_i denoting the temperature (in Celsius degrees) of water the i-th fish swims. Between the priors specified in Example 16 and Example 28, which one (and why) is more reasonable from the modeling point of view?

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 7: Mixture priors[a]

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

Aim: Explain the mixture distribution. Explain, theorize, and construct conjugate mixture prior distribution.

References:

- Berger, J. O. (2013; Sections 4.2.2). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- Robert, C. (2007; Sections 3, pp. 105-123, & pp. 127-141). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

Web applets: `https://georgios-stats-1.shinyapps.io/demo_mixturepriors/`

[a] Author: Georgios P. Karagiannis.

1 Finite mixture distributions

Definition 1. Let $\mathcal{P}_m = \{\Pi_l(\theta|\chi_l);\ l = 1, ..., m\}$ be a collection of probability distributions where $\{\chi_l\}_{l=1}^m$ are parameters of the l-th component $\Pi_l(\theta|\chi_l)$. Let $\{\varpi_l\}_{l=1}^m$ be a set of weights where $\varpi_l > 0$ and $\sum_{l=1}^m \varpi_l = 1$. The mixture distribution derived from the aforementioned collections is

$$\Pi(\theta|\varpi, \chi) = \sum_{l=1}^m \varpi_l \Pi_l(\theta|\chi_l), \qquad \theta \in \Theta \tag{1}$$

where $\chi := (\chi_l,\ l = 1 : m)$ and $\varpi := (\varpi_l,\ l = 1, ..., m)$. $\mathrm{d}\Pi_l(\theta|\chi_l)$ is called l-th mixture component with mixture weight ϖ_l.

Example 2. Let $h(\cdot)$ be a function defined on Θ. The expectation of $h(\cdot)$ with respect to (1) is

$$\mathrm{E}_\Pi(h(\theta)|\varpi, \chi) = \int \sum_{l=1}^m \varpi_l h(\theta)\mathrm{d}\Pi_l(\theta|\chi_l) = \sum_{l=1}^m \varpi_l \int h(\theta)\mathrm{d}\Pi_l(\theta|\chi_l) = \sum_{l=1}^m \varpi_l \mathrm{E}_{\Pi_l}(h(\theta)|\chi_l)$$

where $\mathrm{E}_{\Pi_l}(h(\theta)|\chi_l) = \int h(\theta)\mathrm{d}\Pi_l(\theta|\chi_l)$.

Example 3. A mixture of probability distributions (1) is a probability distribution; i.e.

$$\int_\Theta \pi(\theta|\varpi, \chi)\mathrm{d}\theta = \sum_{l=1}^m \varpi_l \underbrace{\int_\Theta \mathrm{d}\pi_l(\theta|\chi_l)\mathrm{d}\theta}_{=1} = \sum_{l=1}^m \varpi_l = 1$$

Definition 4. The mixture is called finite mixture if $m < \infty$, and countably infinite mixture if $m \to \infty$. Here, we focus on finite mixtures.

Definition 5. A mixture model is called is called parametric mixture model if its components are members of the same parametric family of distributions eg. $\mathcal{P}_m = \{\Pi(\theta|\chi_l);\ l = 1, ..., m\}$, and hence

$$\Pi(\theta|\varpi, \chi) = \sum_{l=1}^{m} \varpi_l \Pi(\theta|\chi_l)$$

Example 6. An example of a parametric mixture model is the Normal mixture model

$$\pi(\theta|\varpi, \mu, \sigma^2) = \sum_{l=1}^{m} \varpi_l \mathrm{N}(\theta|\mu_l, \sigma_l^2),$$

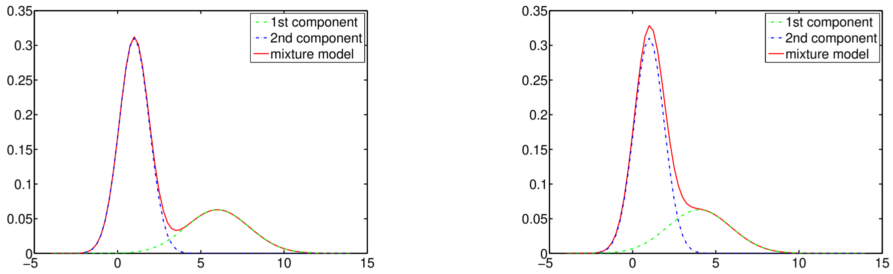
where all components belong to the Normal distribution family.

Note 7. Mixture distributions are useful because (among others):

- they can approximate complicate other distributions by using a combination of simpler distributions $\{\Pi_l(\theta|\chi_l)\}$ which lead to more convenient computations.
- they can naturally model heterogeneity. E.g. consider a population which is heterogeneous in the sense that there are multiple sub-groups labeled by $\ell \in \{1, ..., m\}$, each group represented in the population with proportion ϖ_ℓ, and distributed as $\Pi_l(\theta|\chi_l)$. Then $y \sim \Pi(\theta|\chi)$ can be realized by drawing ℓ with probability $P(\ell = l) = \varpi_l$ and drawing θ from $\Pi_\ell(\theta|\chi_\ell)$ given ℓ, aka

$$\begin{cases} \theta|\ell & \sim \Pi_\ell(\theta|\chi_\ell) \\ \ell & \sim P(\ell) \end{cases} \quad \text{which implies } \theta \sim \Pi(\theta|\chi) \quad \text{since} \quad \Pi(\theta|\chi) = \sum_{\ell=1}^{m} \Pi(\theta|\chi_\ell)P(\ell) = \sum_{l=1}^{m} \varpi_l \Pi(\theta|\chi_l)$$

Example 8. A bimodal, or a right skewed (non-symmetric) distribution can be approximated by a Mixture of (unimodal and symmetric) Normal distributions with different parameter values. Also it describes a population with two groups Normally distributed with different parameters.



(a) Bimodal PDF: $\pi(\theta) = 0.7\mathrm{N}(\theta|1, 0.9^2) + 0.3\mathrm{N}(\theta|6, 1.9^2)$

(b) Right skewed PDF: $\pi(\theta) = 0.7\mathrm{N}(\theta|1, 0.9^2) + 0.3\mathrm{N}(\theta|4, 1.9^2)$

Figure 1: Normal mixture model

2 Mixture prior distributions

Note 9. Mixture models can be used to specify priors distributions, either as a mean to approximate Your actual prior distribution with simpler & tractable distributions, or as a mean to represent heterogeneous prior believes.

Theorem 10. *Let $y := (y_1, ..., y_n)$ be observables generated from the sampling distribution $F(y|\theta)$.* Prior mixture distribution $\Pi(\theta|\varpi)$ is called the prior with pdf/pmf

$$\pi(\theta|\varpi) = \sum_{l=1}^{m} \varpi_l \pi_l(\theta), \tag{2}$$

where, $\mathcal{P}_m = \{\pi_l(\theta),\, \theta \in \Theta\}_{l=1}^{m}$ is a collection of distributions, and $\{\varpi_l\}$ are weights such that $\sum_{l=1}^{m} \varpi_l = 1$ and $\varpi_l > 0$. Then:

1. the posterior distribution $\Pi(\theta|y, \varpi)$ has pdf/pmf

$$\pi(\theta|y, \varpi) = \sum_{l=1}^{m} \varpi_l^{*} \pi_l(\theta|y)\,, \tag{3}$$

2. the predictive distribution of a future outcome z has pdf/pmf

$$g(z|y, \varpi) = \sum_{l=1}^{m} \varpi_l^{*} g_l(z|y)$$

where

$$\begin{aligned}
\varpi_l^{*} &= \frac{\varpi_l f_l(y)}{\sum_{l=1}^{m} \varpi_l f_l(y)} \ \propto \ \varpi_l f_l(y) \\
f_l(y) &= \int_{\Theta} f(y|\theta) \mathrm{d}\Pi_l(\theta) \\
\pi_l(\theta|y) &= \frac{f(y|\theta)\pi_l(\theta)}{\int_{\Theta} f(y|\vartheta)\mathrm{d}\Pi_l(\vartheta)} \ \propto \ f(y|\theta)\pi_l(\theta) \\
g_l(z|y) &= \int_{\Theta} f(z|\theta)\pi_l(\theta|y)\mathrm{d}\theta
\end{aligned}$$

Proof. From Bayes theorem, we have:

$$\begin{aligned}
\mathrm{d}\Pi(\theta|y) &= \frac{f(y|\theta)\mathrm{d}\Pi(\theta)}{\int_{\Theta} f(y|\vartheta)\mathrm{d}\Pi(\theta)} = \frac{f(y|\theta)\sum_{l=1}^{m} \varpi_l \mathrm{d}\Pi_l(\theta)}{\int_{\Theta} f(y|\vartheta)\sum_{l'=1}^{m} \varpi_{l'} \mathrm{d}\Pi_{l'}(\vartheta)} = \frac{\sum_{l=1}^{m} \varpi_l f(y|\theta)\mathrm{d}\Pi_l(\theta)}{\sum_{l'=1}^{m} \int_{\Theta} \varpi_{l'} f(y|\vartheta)\mathrm{d}\Pi_{l'}(\vartheta)} \\
&= \frac{\sum_{l=1}^{m} \varpi_l f_l(y) \overbrace{\frac{f(y|\theta)\mathrm{d}\Pi_l(\theta)}{f_l(y)}}^{=\mathrm{d}\Pi_l(\theta|y)}}{\sum_{l'=1}^{m} \int_{\Theta} \varpi_{l'} f_l'(y) \underbrace{\frac{f(y|\vartheta)\mathrm{d}\Pi_{l'}(\vartheta)}{f_{l'}(y)}}_{=\mathrm{d}\Pi_{l'}(\vartheta|y)}} = \frac{\sum_{l=1}^{m} \varpi_l f_l(y)\mathrm{d}\Pi_l(\theta|y)}{\sum_{l'=1}^{m} \varpi_{l'} f_{l'}(y) \underbrace{\int_{\Theta} \mathrm{d}\Pi_{l'}(\vartheta|y)}_{=1}} \\
&= \sum_{l=1}^{m} \underbrace{\frac{\varpi_l f_l(y)}{\sum_{l'=1}^{m} \varpi_{l'} f_{l'}(y)}}_{=\varpi_l^{*}} \mathrm{d}\Pi_l(\theta|y) = \sum_{l=1}^{m} \varpi_l^{*} \mathrm{d}\Pi_l(\theta|y).
\end{aligned}$$

Also

$$g(z|y, \varpi) = \int_{\Theta} f(z|\theta)\mathrm{d}\Pi_l(\theta|y) = \int_{\Theta} f(y|\theta) \sum_{l=1}^{m} \varpi_l^{*} \mathrm{d}\Pi_l(\theta|y) \quad = \sum_{l=1}^{m} \varpi_l^{*} g_l(z|y)$$

□

Remark 11. Theorem 10 shows that the posterior distribution $\Pi(\theta|y)$ (derived by a mixture prior) is a mixture of 'individual posterior distributions' $\Pi_l(\theta|y)$ weighted by ϖ_l^*. It is determined not only by the observables but also by the weights of the individual distributions. Note that, for $l = 1, ..., m$, the prior weights ϖ_l and posterior weights ϖ_l^* my differ a lot, however they can be close each other.

Remark 12. Mixture priors $\Pi(\theta)$ whose components $\{\Pi_l(\theta)\}$ are conjugate to the likelihood $f(y|\theta)$ can facilitate tractable Bayesian inference. In Theorem 10, if each $\Pi_l(\theta)$ is conjugate to the likelihood $f(y|\theta)$, then obviously each $\Pi_l(\theta|y)$ will belong to the same distribution family as the corresponding $\Pi_l(\theta)$ because $\pi_l(\theta|y) \propto f(y|\theta)\pi_l(\theta)$. Then, provided that components $\pi_l(\theta)$ are tractable, the components $\pi_l(\theta|y)$ will be tractable too (as they belong to the same distr. family). Likewise, the posterior weights can be calculated in closed form i.e., $\varpi_l^* \propto \varpi_l f_l(y)$ since the integral $f_l(y) = \int_\Theta f(y|\theta)\mathrm{d}\Pi_l(\theta)$ will be tractable. Hence, the produced posterior mixture pdf/pmf will be tractable.

Remark 13. Mixtures of conjugate priors are as easy to manipulate as regular conjugate distributions, while leading to a greater freedom in the modeling of the prior information.

Example 14. Let $y = (y_1, ..., y_n)$ observable quantities, generated iid from a Bernoulli sampling distribution with unknown parameter θ; aka $y_i|\theta \overset{\text{iid}}{\sim} \text{Br}(\theta),\ i = 1, ..., n.$

1. Find the likelihood function
2. Find the PDF of the conjugate prior mixture prior $\pi(\theta) = \sum_{l=1}^m \varpi_l \pi_l(\theta)$, with m components $\{\pi_l(\theta)\}_{l=1}^m$.
3. Compute the pdf of the posterior distribution, and recognize it.
4. Compute the predictive pdf for the next outcome $z = (y_{n+1}, ..., y_{n+m})$ given we have observed y? What do you observe?

Hint: Beta distribution: $x \sim \text{Be}(a, b)$ has pdf

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}1(x \in [0,1]); \quad \text{where} \qquad B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, a > 0, b > 0$$

Solution.

1. The likelihood function is

$$f(y|\theta) = \prod_{i=1}^n f(y_i|\theta) = \prod_{i=1}^n \text{Br}(x_i|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n y_i}$$

2. In previous examples, we found that

   - the sampling distribution $\mathrm{d}F(\cdot|\theta)$ is a Bernoulli distribution which is a member of the Exponential family.
   - the conjugate prior is Beta distribution, aka $\theta \sim \text{Be}(a, b)$ $a > 0$ and $b > 0$.

   Therefore, the components $\{\pi_l(\theta)\}_{l=1}^m$ in the prior mixture distribution will be from the family of Beta distributions $\mathcal{P}_m = \{\text{Be}(\theta|a_l, b_l); l = 1, ..., m\}$.

   Therefore, the conjugate mixture prior has pdf

$$\pi(\theta) = \sum_{l=1}^m \varpi_l \text{Be}(\theta|a_l, b_l)$$

   where $\text{Be}(\theta|a_l, b_l)$ is the pdf of $\text{Be}(a_l, b_l)$, and $\{(a_l, b_l)\}$ are fixed prior hyper-parameters.

3. The posterior mixture posterior has PDF

$$\pi(\theta|y) = \sum_{l=1}^{m} \varpi_l^* \pi_l(\theta|y)$$

According to Theorem 10, the l-th components of the mixture posterior is

$$\pi_l(\theta|y) \propto f(y|\theta)\pi_l(\theta) = \prod_{i=1}^{n} \text{Br}(y_i|\theta) \times \text{Be}(\theta|a_l, b_l) \; \propto \prod_{i=1}^{n} [\theta^{y_i}(1-\theta)^{1-y_i}] \times \theta^{a_l-1}(1-\theta)^{b_l-1}$$
$$\overset{r_n=\sum_{i=1}^n y_i}{\propto} \theta^{r_n+a_l-1}(1-\theta)^{n-r_n+b_l-1} \; \propto \; \text{Be}(\theta|a_l^*, b_l^*)$$

with $a_l^* = r_n + a_l$, $b_l^* = n - r_n + b_l$, and $r_n = \sum_{i=1}^n y_i$.

According to Theorem 10, the posterior weights can be calculated as

$$\begin{aligned}\varpi_l^* \; &\propto \varpi_l f_l(y) \; \propto \; \varpi_l \int_{(0,\infty)} \prod_{i=1}^{n} f(x_i|\theta)\pi_l(\theta)\text{d}\theta = \varpi_l \int \prod_{i=1}^{n} \text{Br}(y_i|\theta)\text{Be}(\theta|a_l, b_l)\text{d}\theta \\ &= \varpi_l \int_{(0,\infty)} \theta^{r_n}(1-\theta)^{n-r_n} \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \theta^{a_l-1}(1-\theta)^{b_l-1}\text{d}\theta \\ &= \varpi_l \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \int_{(0,1)} \theta^{r_n+a_l-1}(1-\theta)^{n-r_n+b_l-1}\text{d}\theta \; = \varpi_l \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(a_l^*)\Gamma(b_l^*)}{\Gamma(a_l^*+b_l^*)} \end{aligned} \tag{4}$$

namely,

$$\varpi_l^* = \frac{\varpi_l \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(a_l^*)\Gamma(b_l^*)}{\Gamma(a_l^*+b_l^*)}}{\sum_{l=1}^m \varpi_l \frac{\Gamma(a_l+b_l)}{\Gamma(a_l)\Gamma(b_l)} \frac{\Gamma(a_l^*)\Gamma(b_l^*)}{\Gamma(a_l^*+b_l^*)}}$$

4. It is ...

$$g(z|y) = \int_\Theta f(z|\theta)\text{d}\Pi(\theta|y) = \sum_{i=1}^{m} \varpi_l^*(y) \underbrace{\int_\Theta f(z|\theta)\text{d}\Pi_l(\theta|y)}_{=g_l(z|y)}$$

where the following is just copy-paste from Handout 3...

$$\begin{aligned} g_l(z|y) &= \int_\Theta f(z|\theta)\pi_l(\theta|y)\text{d}\theta \; = \int_\Theta \prod_{i=1}^{m} f(z_i|\theta)\pi_l(\theta|y)\text{d}\theta \; = \int_{(0,\infty)} \prod_{i=1}^{m} \text{Br}(z_i|\theta)\text{Be}(\theta|a_l^*, b_l^*)\text{d}\theta \\ &= \int_0^1 \left[\theta^{\sum_{i=1}^m z_i}(1-\theta)^{m-\sum_{i=1}^m z_i}\right] \left[\frac{\theta^{a_l^*-1}(1-\theta)^{b_l^*-1}}{B(a_l^*, b_l^*)}\right] \text{d}\theta \; 1(z \in \{0,1\}^m) \\ &= \frac{1}{B(a_l^*, b_l^*)} \int_0^1 \theta^{\sum_{i=1}^m z_i + a_l^* - 1}(1-\theta)^{m-\sum_{i=1}^m z_i + b_l^* - 1}\text{d}\theta \; 1(z \in \{0,1\}^m) \\ &= \frac{B(\sum_{i=1}^m z_i + a_l^*, m - \sum_{i=1}^m z_i + b_l^*)}{B(a_l^*, b_l^*)} 1(z \in \{0,1\}^m) \end{aligned}$$

3 Practice

Question 15. *Try the Exercise 43 from the Exercise sheet.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 8: Non-informative priors[a]

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

Aim: Explain the non-informative priors. Explain, theorize, and construct Laplace and Jeffreys' priors.

References:

- Robert, C. (2007; Sections 3.5.1 - 3.5.3.). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- Berger, J. O. (2013; Sections 3.3, & 4.2.3). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

Web applets: `https://georgios-stats-1.shinyapps.io/demo_conjugatejeffreyslaplacepriors/`

[a] Author: Georgios P. Karagiannis.

1 Non-informative priors

Note 1. Non-informative priors (or objective priors) are often specified (in practice) when no prior information is available.

Note 2. Non-informative prior distributions can be hardly justified in the subjective Bayesian stats because probabilities are considered to be subjective, and hence every researcher has some personal believe abut the unknowns. Non-informative priors are used in the Subjective Bayes framework as a last resort when no prior information exist or when the specific application requires them.

Note 3. Objective Bayes and objective probability are variants of the subjective Bayesian stats and probability, where the probability is assumed to represent degree of believe about a proposition (similar to Subjective framework) but this is not a matter of an individual's personal degree of believe (in contrast to Subjective framework). There is no room for personal belief, hence everyone should assign the same prior probabilities, hence the posterior probability should be the same for different researchers. This Bayesian philosophical variation can be hardly justified, no?

Note 4. Objective Bayes dogmatically requires only the specification of non-informative priors as a mean to eliminate subjective/individual believes to the priors, and assign priors generally accepted by everybody. So these non-informative priors are also called objective priors.

Note 5. Objective Bayes is an arguable variation of the Bayesian framework. Different people may have different prior degrees of believe and hence the use of a prior accepted by everybody is arguable. Also, prior degree of believe of a group or several groups of researchers can still be expressed by subjective priors and be justified in the Subjective Bayesian framework.

Note 6. Some tools used to specify non-informative (objective) priors often break the Bayesian paradigm, and may produce unreasonable results, e.g. violation of the likelihood principle.

Note 7. In most applications it is almost impossible to specify non-informative priors representing exactly total ignorance about the problem at hand. They should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing. More realistically, non-informative priors about specific aspects of the problem or features of the statistical model (e.g., transformations, rotations) can be derived, and justified in the Subjective and Objective Bayes framework.

2 Laplace non-informative priors

Definition 8. The principle of insufficient prior states that we do not have any reason to think that one value of the unknown quantity is more likely than any other, so we should use a Uniform prior.

Definition 9. Laplace priors about $\theta \in \Theta$ is specified as

$$\mathrm{d}\Pi(\theta) \propto \underbrace{1}_{\propto \pi(\theta)} \mathrm{d}\theta \quad \text{with pdf/pmf} \quad \pi(\theta) \propto 1 \tag{1}$$

and builds upon the principle of insufficient prior.

Remark 10. Laplace prior (1) places the same degree of believe at each value of θ if θ discrete, and at each interval $\mathrm{d}\theta$ of the same length if θ is continuous.

Remark 11. Laplace priors (1) can be improper. In such cases the properness condition has to be checked. Laplace priors are improper when Θ is unbounded parametric space but are proper when Θ is bounded.

Example 12. Consider the Bayesian model

$$\begin{cases} y_i|\mu & \sim \mathrm{N}(\mu, 1), \ \forall i = 1, ..., n \\ \mu & \sim \mathrm{d}\Pi(\mu) \end{cases}$$

Laplace prior $\mathrm{d}\Pi(\mu) \propto 1\mathrm{d}\mu$ is an improper prior since $\int_\Theta \mathrm{d}\Pi(\mu) = \int_\mathbb{R} \pi(\mu)\mathrm{d}\mu = \int_\mathbb{R} 1\mathrm{d}\mu = +\infty$. However, it can be used as a prior because it satisfies the properness condition

$$\int_\mathbb{R} \prod_{i=1}^{n} \mathrm{N}(y_i|\mu, 1)1\mathrm{d}\mu \propto 2^{-\frac{n}{2}}(\pi)^{-\frac{n-1}{2}} \exp(-\frac{1}{2}(\sum x_i^2) + \frac{1}{2}(\sum x_i)^2) < \infty$$

and hence it leads to a well defined posterior probability distribution.

Remark 13. Laplace prior (1) are not invariant under transformations. Assume Laplace (non-informative) prior for θ, with pdf $\pi(\theta) \propto 1$. Consider a random quantity ψ where $\psi = g(\theta)$, such that $g : \Theta \to \Psi$ is an 1-1 transformation. Then it is

$$\pi_\psi(\psi) \propto \pi_\theta(g^{-1}(\psi))|\frac{\mathrm{d}}{\mathrm{d}\psi} g^{-1}(\psi)|$$

which is not necessarily flat, and hence not necessarily non-informative. This is strange because it means that we a priori know nothing about θ but we a priori know something about $\psi = g(\theta)$...

Example 14. Consider an experiment with sampling distribution

$$y_i|\theta \sim \mathrm{Br}(\theta), \ \forall i = 1, ..., n$$

The Laplace (non-informative) prior for success frequency $\theta \in [0, 1]$ is $\mathrm{d}\Pi(\theta) \propto 1\mathrm{d}\theta$ and hence $\theta \sim \mathrm{Un}(0, 1)$ which is a proper prior. This implies that odds $\psi = \frac{\theta}{1-\theta}$ have prior

$$\pi_\psi(\psi) \propto \pi_\theta(\frac{\psi}{1+\psi}) \left| \frac{\mathrm{d}}{\mathrm{d}\psi} \frac{\psi}{1+\psi} \right| \propto \frac{1}{(1+\psi)^2}$$

which is informative. So I a priori know nothing about the frequency θ but I know something about the odds ψ...

Note 15. Those using Laplace priors argue that You should parametrize the likelihood $f(y|\theta)$ according to a desired parameterization (e.g., success frequency θ), assign a Laplace prior, and stick with it ignoring reparametrizations...

3 Jeffreys' priors

Note 16. Let $y = (y_1, ..., y_n)$ observables drawn from a sampling distribution $F(y|\theta)$ with density $f(y|\theta)$. I use this notation hereafter. We aim to specify a prior $\Pi(\theta)$ with density $\pi(\theta)$ in the Bayesian model

$$\begin{cases} y|\theta & \sim \mathrm{d}F(y|\theta) \\ \theta & \sim \mathrm{d}\Pi(\theta) \end{cases} \tag{2}$$

so that $\Pi(\theta)$ can be invariant to $1-1$ transformations. Precisely, this is an invariance to a 1-1 transformations, and it is reasonable in certain type of applications (mentioned in the classroom).

Definition 17. The Jeffreys' prior distribution of the unknown parameter $\theta \in \Theta = \mathbb{R}^k$ $k \geq 1$ is defined as < SC2

$$\mathrm{d}\Pi(\theta) \propto \underbrace{\sqrt{\det(\mathscr{I}(\theta))}}_{\propto \pi(\theta)} \mathrm{d}\theta \quad \text{with pdf/pmf} \quad \pi(\theta) \propto \sqrt{\mathscr{I}(\theta)},$$

where $\mathscr{I}(\theta)$ is the Fisher Information.

Definition 18. Let $y = (y_1, ..., y_n)$ observables drawn from a sampling distribution $F(y|\theta)$ with density $f(y|\theta)$ labeled by parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. Let $f(y|\theta)$ denote the likelihood.

Fisher Information $\mathscr{I}(\theta)$ is a $k \times k$ matrix defined as

$$\mathscr{I}(\theta) = \mathrm{E}_{F(\cdot|\theta)}\left(\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f(y|\theta))\right]^\top \left[\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f(y|\theta))\right]\right)$$

where

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(y|\theta) = \left[\tfrac{\mathrm{d}}{\mathrm{d}\theta_1}\log(f(y|\theta)), \quad \dots \quad \tfrac{\mathrm{d}}{\mathrm{d}\theta_j}\log(f(y|\theta)) \quad , \dots \quad \tfrac{\mathrm{d}}{\mathrm{d}\theta_d}\log(f(y|\theta))\right] \in \mathbb{R}^k$$

So the (i, j) element of Fisher Information $\mathscr{I}(\theta)$ is

$$[\mathscr{I}(\theta)]_{i,j} = \mathrm{E}_{F(\cdot|\theta)}\left(\left[\frac{\mathrm{d}}{\mathrm{d}\theta_i}\log(f(y|\theta))\right]\left[\frac{\mathrm{d}}{\mathrm{d}\theta_j}\log(f(y|\theta))\right]\right)$$

In the univariate case $\theta \in \Theta \subseteq \mathbb{R}$, it is

$$\mathscr{I}(\theta) = \mathrm{E}_{F(\cdot|\theta)}\left(\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log(f(y|\theta)\right)^2\right)$$

Fact 19. *Some properties of Fisher information $\mathscr{I}(\theta)$:*

1. *Under regularity conditions, when $\theta \in \Theta = \mathbb{R}^k$, $\mathscr{I}(\theta)$ simplifies to*

$$[\mathscr{I}(\theta)]_{i,j} = -E_{F(\cdot|\theta)}\left(\frac{d^2}{d\theta_i d\theta_j}\log(f(y|\theta))\,|\,\theta\right) = -\int_{\mathcal{X}} \frac{d^2}{d\theta_i d\theta_j}\log(f(y|\theta))dF(y|\theta).$$

   and when $\theta \in \Theta = \mathbb{R}$ (univariate), $\mathscr{I}(\theta)$ simplifies to

$$\mathscr{I}(\theta) = -E_{F(y|\theta)}\left(\frac{d^2}{d\theta^2}\log(f(y|\theta))\,|\,\theta\right) = -\int_{\mathcal{X}} \frac{d^2}{d\theta^2}\log(f(y|\theta))dF(y|\theta),$$

2. *Let $y = (y_1, ..., y_n)$ be observables frown iid from distribution $F(y|\theta)$, and let $\mathscr{I}_n(\theta)$ be the associated Fisher information based on n observables. Then $\mathscr{I}_n(\theta) = n\mathscr{I}_1(\theta)$.*

3. *Let $g : \Theta \to \Psi$ with $\psi = g(\theta)$ be a $1-1$ transformation. Then the Fisher information is*

$$\mathscr{I}(\psi) = J^\top \mathscr{I}(\theta) J, \qquad \textit{and} \qquad \det(\mathscr{I}(\psi)) = \det(\mathscr{I}(\theta))\det(J)^2$$

where $J = \frac{d\theta}{d\psi}$ is the Jacobian of the transformation $\psi = g(\theta)$ whose (i, j) element is $[J]_{i,j} = \frac{\partial \theta_i}{\partial \psi_i}$. < AMV2

In the univariate case where $\theta \in \Theta \subseteq \mathbb{R}$ $\psi \in \Psi \subseteq \mathbb{R}$ it is

$$\mathscr{I}(\psi) = \mathscr{I}(\theta) \left(\frac{d\theta}{d\psi} \right)^2$$

Note 20. The rational of Jeffreys prior is that Fisher information is widely accepted as an indicator of the amount of information brought by the statistics model (or the observation) about θ, and hence the values of θ for which $\mathscr{I}(\theta)$ is larger should be more likely for the prior distribution. E.g, in Fact 19(2), $\mathscr{I}_n(\theta)$ is understood as n unites of observation information if $\mathscr{I}_1(\theta)$ is understood as 1 unite of observational information. Fisher information $\mathscr{I}(\theta)$ can evaluate the ability of the model to discriminate between θ and $\theta + \mathrm{d}\theta$ through the expected slope of $\log(f(y|\theta)$.

- Therefore, to favor the values of θ for which $\mathscr{I}(\theta)$ is large is equivalent to minimizing the influence of the prior distribution and is therefore as noninformative as possible.

Theorem 21. *Jeffreys priors are invariant under 1-1 transformations. Suppose that $d\Pi_\theta(\theta) \propto \sqrt{\mathscr{I}(\theta)}d\theta$ with density $\pi_\theta(\theta) \propto \sqrt{\mathscr{I}(\theta)}$ and $\pi_\psi(d\psi) \propto \sqrt{\mathscr{I}(\psi)}d\psi$ with density $\pi_\psi(\psi) \propto \sqrt{\mathscr{I}(\psi)}$ are Jeffreys priors associated to sampling pdf/pmf $f(y|\theta)$ and $f(y|\psi)$ respectively, where $\psi = g(\theta)$ and $g : \Theta \to \Psi$ is a 1-1 transformation. Then $\pi_\psi(\psi) \propto \pi_\theta(\theta)|\frac{\partial\theta}{\partial\psi}|$.*

Proof. We prove it for the case where $\theta \in \mathbb{R}$, as the extension to the multivatiate setting is straightforward. It is $\mathscr{I}(\psi) = \mathscr{I}(\theta)(\frac{\partial\theta}{\partial\psi})^2$, because

$$\begin{aligned}
\mathscr{I}(\psi) &= \mathrm{E}_{y\sim F(\cdot|\psi)} \left(\frac{\partial}{\partial\psi} \log(f(y|\psi)) \right)^2 = \int_{\mathcal{Y}} \left(\frac{\partial}{\partial\psi} \log(f(y|\psi)) \right)^2 \mathrm{d}F(y|\psi) \\
&= \int_{\mathcal{Y}} \left(\frac{\partial}{\partial\psi} \log(f(y|\theta)) \right)^2 \mathrm{d}F(y|\theta) = \int_{\mathcal{Y}} \left(\frac{\partial\theta}{\partial\psi} \frac{\partial}{\partial\theta} \log(f(y|\theta)) \right)^2 \mathrm{d}F(y|\theta) \\
&= \int_{\mathcal{X}} \left(\frac{\partial}{\partial\theta} \log(f(y|\theta)) \right)^2 \mathrm{d}F(y|\theta)(\frac{\partial\theta}{\partial\psi})^2 = \mathrm{E}_{y\sim F(\cdot|\theta)} \left(\frac{\partial}{\partial\theta} \log(f(y|\theta)) \right)^2 (\frac{\partial\psi}{\partial\theta})^2 \\
&= \mathscr{I}(\theta)(\frac{\partial\psi}{\partial\theta})^2
\end{aligned}$$

Then

$$\pi_\psi(\psi) \ \propto \sqrt{\mathscr{I}(\psi)} \ = \sqrt{\mathscr{I}(\theta)(\frac{\partial\theta}{\partial\psi})^2} \ \propto \sqrt{\mathscr{I}(\theta)}|\frac{\partial\theta}{\partial\psi}| \ \propto \pi_\theta(\theta)|\frac{\partial\theta}{\partial\psi}|$$

□

Example 22. Consider observable r drawn from a Binomial sampling distribution $r|\theta \sim \mathrm{Bn}(n, \theta)$ with pdf $\mathrm{Br}(r|\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$ and mean $\mathrm{E}_{\mathrm{Bn}(n,\theta)}(r) = n\theta$. Find the Jeffreys prior for θ. Compute the posterior of θ given r.

Solution. The likelihood is $f(r|\theta) = \mathrm{Br}(r|\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}$ So

$$\begin{aligned}
\log(f(r|\theta)) &= \log \binom{n}{r} + r\log(\theta) + (n-r)\log(1-\theta) \Longrightarrow \\
\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log(f(r|\theta)) &= -\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2} \Longrightarrow \\
\mathscr{I}(\theta) &= -\mathrm{E}_{\mathrm{Bn}(n,\theta)}(-\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2}) = \frac{n}{\theta(1-\theta)} \Longrightarrow
\end{aligned}$$

$$\text{Jeffreys' prior is} \quad \pi^{(\text{Jef;Bn})}(\theta) \propto \frac{1}{\theta^{1/2}(1-\theta)^{1/2}} \propto \mathrm{Be}(\theta|0.5, 0.5) \tag{3}$$

The posterior is

$$\pi^{(\text{Jef;Bn})}(\theta|r) \propto \text{Bn}(r|n,\theta)\pi^{(\text{Jef;Bn})}(\theta) \propto \theta^{r-1/2}(1-\theta)^{n-r-1/2}$$

Example 23. Consider observable n drawn from a Negative binomial $n|\theta \sim \text{Nb}(r,\theta)$ with pdf $\text{Nb}(n|r,\theta) = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}$and mean $\text{E}_{\text{Nb}(r,\theta)}(n) = r/\theta$. Find the Jeffreys prior for θ. Compute posterior of θ given n.

Solution. So

$$\begin{aligned}
\log(f(r|\theta)) &= \log\binom{n-1}{r-1} + r\log(\theta) + (n-r)\log(1-\theta) \Longrightarrow \\
\frac{\text{d}^2}{\text{d}\theta^2}\log(f(r|\theta)) &= -\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2} \Longrightarrow \\
\mathscr{I}(\theta) &= -\text{E}_{\text{Nb}(r,\theta)}(-\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2}) = \frac{r}{\theta^2(1-\theta)} \Longrightarrow \\
\text{Jeffreys' prior is}\quad \pi^{(\text{Jef;Nb})}(\theta) &\propto \frac{1}{\theta^1(1-\theta)^{1/2}}
\end{aligned} \tag{4}$$

The posterior is $\pi^{(\text{Jef;Nb})}(\theta|n) \propto f(r|\theta)\pi^{(\text{Jef;Nb})}(\theta) \propto \theta^{r-1}(1-\theta)^{n-r-1/2}$.

Remark 24. Jeffreys' prior can violate the Likelihood Principle. This is because its derivation depends on the form of the specific experiment via Fisher information which can differ for two different experiments even though they have proportional likelihoods. E.g.; For two experiments with propositional likelihoods $f_1(\theta|y_1) \propto f_2(\theta|y_2) \propto L(\theta)$, Jeffreys' priors are $\pi^{(\text{jp},1)}(\theta) \propto \mathscr{I}_1(\theta)$, and $\pi^{(\text{jp},2)}(\theta) \propto \mathscr{I}_2(\theta)$, so

$$\pi^{(\text{exp 1})}(\theta|y_1) = \frac{L(\theta)\sqrt{\mathscr{I}_1(\theta)}}{\int L(\theta)\sqrt{\mathscr{I}_1(\theta)}\text{d}\theta}; \quad \pi^{(\text{exp 2})}(\theta|y_2) = \frac{L(\theta)\sqrt{\mathscr{I}_2(\theta)}}{\int L(\theta)\sqrt{\mathscr{I}_2(\theta)}\text{d}\theta}; \quad \overset{\mathscr{I}_1(\theta)\neq\mathscr{I}_2(\theta)}{\Longrightarrow} \quad \pi^{(\text{exp 1})}(\theta|y_1) \neq \pi^{(\text{exp 2})}(\theta|y_2)$$

In the Examples 22 and 23, even though the two likelihoods are equal up to a multiplicative constant, i.e.

$$\text{Bn}(r|n,\theta) \propto \text{Nb}(n|r,\theta) \propto \theta^r(1-\theta)^{n-r} \tag{5}$$

Jeffreys' priors led to different posteriors $\pi^{(\text{Jef;Br})}(\theta|r) \neq \pi^{(\text{Jef;Nb})}(\theta|n)$.

Example 25. Let $y \in \mathbb{R}$ be an observable. Consider the statistical model

$$y|\mu,\sigma^2 \overset{\text{iid}}{\sim} \text{N}(\mu,\sigma^2) \quad \text{where} \quad (\mu,\sigma^2) \in \mathbb{R}\times(0,\infty).$$

1. Specify the Jeffreys' prior for $\theta = \mu$, when σ is known.
2. Specify the Jeffreys' prior for $\theta = \sigma$, when μ is known.
3. Specify the Jeffreys' prior for $\theta = (\mu,\sigma)$.

Solution. It is

$$\log f(y|\theta) = \log(\text{N}(y|\mu,\sigma^2)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}$$

1. It is

$$\begin{aligned}
\frac{\text{d}}{\text{d}\mu}\log(\text{N}(y|\mu,\sigma^2)) &= \frac{(y-\mu)}{\sigma^2} \\
\frac{\text{d}^2}{\text{d}\mu^2}\log(\text{N}(y|\mu,\sigma^2)) &= \frac{\partial}{\partial\mu}\frac{(y-\mu)}{\sigma^2} = -\frac{1}{\sigma^2} \\
\mathscr{I}(\mu) &= -\text{E}_{y\sim\text{N}(\mu,\sigma^2)}\left(\frac{\text{d}^2}{\text{d}\mu^2}\log(\text{N}(y|\mu,\sigma^2))\right) = \frac{1}{\sigma^2}
\end{aligned}$$

and hence $\pi^{(\text{JP})}(\mu) \propto \sqrt{\mathscr{I}(\mu)} \propto 1$

2. It is

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\sigma}\log(\mathrm{N}(y|\mu,\sigma^2)) &= -\frac{1}{\sigma}+\frac{(y-\mu)^2}{\sigma^3}\\
\frac{\mathrm{d}^2}{\mathrm{d}\sigma^2}\log(\mathrm{N}(y|\mu,\sigma^2)) &= \frac{1}{\sigma^2}-3\frac{(y-\mu)^2}{\sigma^4}=\frac{1}{\sigma^2}-3\frac{1}{\sigma^2}\left(\frac{y-\mu}{\sigma^2}\right)^2
\end{aligned}
$$

So

$$
\mathscr{I}(\sigma) \quad = -\mathrm{E}_{y\sim\mathrm{N}(\mu,\sigma^2)}\left(\frac{\partial^2}{\partial\theta^2}\log(\mathrm{N}(y|\mu,\sigma^2))\right) = -\mathrm{E}_{y\sim\mathrm{N}(\mu,\sigma^2)}\left(\frac{1}{\sigma^2}-3\frac{1}{\sigma^2}\left(\frac{y-\mu}{\sigma^2}\right)^2\right) \quad \propto \frac{1}{\sigma^2}
$$

and hence $\pi^{(\mathrm{JP})}(\sigma)\propto\sqrt{\mathscr{I}(\sigma)}\propto\frac{1}{\sigma}$

3. it is

$$
\frac{\mathrm{d}^2}{\mathrm{d}\mu\mathrm{d}\sigma}\log f(y|\theta)=\frac{\mathrm{d}^2}{\mathrm{d}\mu\mathrm{d}\sigma}\log(\mathrm{N}(y|\mu,\sigma^2))=-2\frac{(y-\mu)}{\sigma^3}
$$

So

$$
\mathscr{I}(\mu,\sigma)=-\mathrm{E}_{y\sim\mathrm{N}(\mu,\sigma^2)}\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(\mathrm{N}(y|\mu,\sigma^2))\right)=-\mathrm{E}_{y\sim\mathrm{N}(\mu,\sigma^2)}\begin{bmatrix}-\frac{1}{\sigma^2} & -2\frac{(y-\mu)}{\sigma^3}\\ -2\frac{(y-\mu)}{\sigma^3} & \frac{1}{\sigma^2}-3\frac{(y-\mu)^2}{\sigma^4}\end{bmatrix}=\begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & 2\frac{1}{\sigma^2}\end{bmatrix}
$$

and $\pi^{(\mathrm{JP})}(\mu,\sigma)\propto\sqrt{\det(\mathscr{I}(\mu,\sigma))}\propto\frac{1}{\sigma^2}$.

Example 26. (Just have a look here) Consider the model of Normal linear regression where the observables are pairs (ϕ_i,y_i) for $i=1,...,n$, assumed to be modeled according to the sampling distribution $y_i|\beta,\sigma^2\overset{\text{ind}}{\sim}\mathrm{N}(\phi_i^\top\beta,\sigma^2)$ for $i=1,...,n$ with unknown $(\beta,\sigma^2)\in\mathbb{R}^d\times\mathbb{R}_+$. Namely, the sampling distribution in vector form is

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

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

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

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

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

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

Let's compute the derivatives

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

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

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

$$\text{E}_{\text{N}(\Phi\beta,I\sigma^2)}\left(\frac{\text{d}^2}{\text{d}\beta^2}\log\left(f(y|\beta,\xi)\right)\right) = -\frac{1}{\xi}\Phi^\top\Phi$$

$$\text{E}_{\text{N}(\Phi\beta,I\sigma^2)}\left(\frac{\text{d}}{\text{d}\xi}\frac{\text{d}}{\text{d}\beta}\log\left(f(y|\beta,\xi)\right)\right) = \text{E}_{\text{N}(\Phi\beta,I\sigma^2)}\left(\frac{1}{\xi^2}\Phi^\top\left(y-\Phi\beta\right)\right) = \frac{1}{\xi^2}\Phi^\top\left(\text{E}_{\text{N}(\Phi\beta,I\sigma^2)}(y)-\Phi\beta\right) = 0$$

Then

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

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

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

Namely $\pi^{(\text{JP})}(\beta,\sigma^2) = \left(\sigma^2\right)^{-\frac{d}{2}-1}$.

4 Limiting posterior distributions ()

Note 27. One way to derive a posterior in the absence of prior information, is to specify a non-informative (possibly improper) prior, check the properness condition, and compute the posterior by the Bayes theorem; i.e.

$$\pi(\theta|y) = \frac{f(y|\theta)\pi^{(\text{impr.})}(\theta)}{\int f(y|\theta)\pi^{(\text{impr.})}(\theta)\text{d}\theta}$$

Note 28. An alternative way to derive a posterior distribution in the absence of prior information, is to compute it as a limit of a posterior updated from a proper prior. Namely:

1. specify a proper prior with a specific parametric form $\pi(\theta|\tau) \propto \tilde{\pi}(\theta|\tau)$ (e.g., conjugate prior with hypoer-parameter τ), and compute the kernel $\tilde{\pi}(\theta|\tau)$ of its pdf/pmf as $\pi(\theta|\tau) \propto \tilde{\pi}(\theta|\tau)$
2. find values τ' for the prior hyper parameters such that the limit

$$\tilde{\pi}'(\theta) = \lim_{\tau\to\tau'}\tilde{\pi}(\theta|\tau), \quad \text{when} \quad \tau\to\tau'$$

   can be considered as the kernel of a non-informative prior. E.g., $\tilde{\pi}'(\theta)$ can be a Jeffreys prior.
3. compute the posterior distribution $\pi(\theta|y,\tau)$ updated from the proper prior $\pi(\theta|\tau)$, by Bayes theorem

$$\pi(\theta|y,\tau) = \frac{f(y|\theta)\tilde{\pi}(\theta|\tau)}{\int f(y|\theta)\tilde{\pi}(\theta|\tau)\text{d}\theta}$$

4. compute the limiting posterior as

$$\pi'(\theta|y) = \lim_{\tau\to\tau'}\pi(\theta|y,\tau)$$

The limiting posterior $\pi'(\theta|y)$ is derived based on no prior information,

Note 29. A convenient way is to specify a conjugate prior $\pi(\theta|\tau)$ and let the τ approach values that reduce the strength of the prior information.

Example 30. Let $\pi^{\text{(CP)}}(\theta|\tau)$ denote the conjugate prior, $\pi^{\text{(JP)}}(\theta|\tau)$ denote the Jeffreys' prior, and $\pi^{\text{(LP)}}(\theta|\tau)$ denote the Laplace prior. For Bernoulli statistical model $y_i \overset{\text{iid}}{\sim} \text{Br}(\theta),\ \forall i = 1, ..., n$, it is

Conjugate prior: $\theta|\tau = (a, b) \sim \text{Be}(a, b)$ has pdf kernel $\pi^{\text{(CP)}}(\theta|a, b) \propto \tilde{\pi}^{\text{(CP)}}(\theta|a, b) = \theta^{a-1}(1-\theta)^{b-1}$is updated By Bayes theorem to

$$\pi^{\text{(CP)}}(\theta|y, a, b) = \text{Be}(\sum_{i=1}^{n} y_i + a, n - \sum_{i=1}^{n} y_i + b)$$

Laplace prior: $\theta \sim \text{U}(0, 1) \equiv \text{Be}(1, 1)$ has pdf kernel $\pi^{\text{(LP)}}(\theta) \propto \tilde{\pi}^{\text{(LP)}}(\theta) = 1$. We find point $(a', b') = (1, 1)$ such that

$$\lim_{(a,b)\to(1,1)} \tilde{\pi}^{\text{(CP)}}(\theta|a, b) = 1 = \tilde{\pi}^{\text{(LP)}}(\theta)$$

so

$$\pi^{\text{(LP)}}(\theta|y) = \lim_{(a,b)\to(1,1)} \pi^{\text{(CP)}}(\theta|y, a, b) = \text{Be}(\sum_{i=1}^{n} y_i + 1, n - \sum_{i=1}^{n} y_i + 1)$$

Jeffreys prior: $\theta \sim \text{Be}(\theta|0.5, 0.5)$ has pdf kernel $\pi^{\text{(JP)}}(\theta) \propto \tilde{\pi}^{\text{(LP)}}(\theta) = \theta^{-0.5}(1-\theta)^{-0.5}$. We find point $(a', b') = (0.5, 0.5)$ such that

$$\lim_{(a,b)\to(0.5,0.5)} \tilde{\pi}^{\text{(JP)}}(\theta|a, b) = 1 = \tilde{\pi}^{\text{(JP)}}(\theta)$$

so

$$\pi^{\text{(JP)}}(\theta|y) = \lim_{(a,b)\to(0.5,0.5)} \pi^{\text{(CP)}}(\theta|y, a, b) = \text{Be}(\sum_{i=1}^{n} y_i + 0.5, n - \sum_{i=1}^{n} y_i + 0.5)$$

5 Practice

Question 31. *For practice try to address the Exercises 47, 48, and 52, from the Exercise sheet. You can try the Exercise 53 from the Exercise sheet. which is related to Regression.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 9: Prior elicitation and Prior specification under the presence of partial prior info [a]

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

Aim: To explain and apply elicitation, and to explain, and derive maximum entropy priors.

References:

- Robert, C. (2007; Sections 3; pp. 105-123, & pp. 127-141). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- Berger, J. O. (2013; Sections 3.4). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.

[a] Author: Georgios P. Karagiannis.

1 Prior elicitation

Note 1. In many projects, You (the statistician) will be required to elicit the a priori knowledge of an expert or domain scientist (e.g. engineering, biologist, physicist, etc...) in the form of the prior distribution.

Note 2. The following process can facilitate the specification of the prior $\mathrm{d}\Pi(\theta)$ to reflect as accurately as possible expert's a priori knowledge. Essentially You could possibly discuss with the expert with purpose to do the following:

1. Structure: Determine the structure the distribution in terms of independence, conditional independence, exchangeability, transformations.

2. Elicitation of summaries: Elicitate appropriate summaries (mainly moments such as prior mean, variance, quantiles, etc...) expressing the most important aspects of the expert's a priori knowledge.

3. Fitting: Fit a suitable distribution to the expert's elicited summaries according to the structures determined.

4. Application: Recognize that the prior distribution is an approximation of the expert's a priori knowledge. Update the prior distribution to the posterior in the light of experimental data, to perform inference.

... and good luck at figuring out what the expert has in his/her mind...

Example 3. An illustrative example will be given in the classroom.

2 Maximum entropy priors[1]

Note 4. We aim at specifying the priors under the presence of partial prior info. Often prior information about certain characteristics of the application are available in the form of moments or expectations; e.g., the prior mean, variance, quarantines, etc...

[1] I do not actually demonstrate the concept as it is originally developed, but I give a similar explanation.

2.1 A general framework

Note 5. Our desideratum is to recover a mathematical representation for the pdf/pmf $\pi(\theta)$ of the prior distribution $\Pi(\theta)$ so that it can (1.) be 'as closely as possible' to a reference distribution with pdf/pmf $u(\theta)$, (2.) satisfy the k independent constrains

$$m_j = \mathrm{E}_\Pi(h_j(\theta)), \qquad\qquad j = 1, ..., k \tag{1}$$

where $m_j \in \mathbb{R}$ are specified values by the expert, and (3.) possibly satisfy the normalizing constrain $\int_\Theta \mathrm{d}\Pi(\theta) = 1$.

Note 6. A general measure of 'lack of fit' or difference between a distribution pdf/pmf (assumed to be true) $\pi(\theta)$ and its approximation $u(\theta)$ is the Kullback-Leibler divergence. –However, others exist too.

Definition 7. The Kullback-Leibler divergence (or KL divergence, or relative entropy) between distribution density/mass $\pi(\theta)$ and $u(\theta)$, where $\pi \ll u$, $\theta \in \Theta$, it is defined as

$$\mathrm{KL}(\pi\|u) = \mathrm{E}_\Pi\left(\log\left(\frac{\pi(\theta)}{u(\theta)}\right)\right) = \begin{cases} \int_\Theta \log\left(\frac{\pi(\theta)}{u(\theta)}\right)\pi(\theta)\mathrm{d}\theta & \text{, if } \theta \text{ is cont.} \\ \\ \sum_{\theta\in\Theta} \log\left(\frac{\pi(\theta)}{u(\theta)}\right)\pi(\theta) & \text{, if } \theta \text{ is discr.} \end{cases} \tag{2}$$

and measures how far $u(\cdot)$ is from $u(\cdot)$.

Fact 8. *KL divergence:*

- is non negative $\mathrm{KL}(\pi\|u) \geq 0$, and the equality holds if and only if $\pi(\cdot) = u(\cdot)$ a.s.[2]
- *is not symmetric* $\mathrm{KL}(\pi\|u) \neq \mathrm{KL}(u\|\pi)$, so is not a distance
- *is convex*

$$\mathrm{KL}(\xi\pi_1 + (1-\xi)\pi_2\|u\xi u_1 + (1-\xi)u_2) \leq \xi\mathrm{KL}(\pi_1\|u_1) + (1-\xi)\mathrm{KL}(\pi_2\|u_2),\ \forall\xi \in (0,1)$$

Proposition 9. *The Kullback-Leibler divergence KL($\pi\|u$), cross entropy H(π, u), and entropy H(π), between two probability distributions with densities π and u are associated as*

$$KL(\pi\|u) = H(\pi,u) - H(\pi)$$

$$\Longleftrightarrow$$

$$\overbrace{\underbrace{\int_\Theta \log(\frac{\pi(\theta)}{u(\theta)})d\Pi(\theta)}_{=E_\Pi\left(\log(\frac{\pi(\theta)}{u(\theta)})\right)}}^{KL(\pi\|u)=} = \overbrace{\underbrace{-\int_\Theta \log(u(\theta))d\Pi(\theta)}_{=E_\Pi(\log(u(\theta)))}}^{H(\pi,u)=} - \overbrace{\underbrace{\left[-\int_\Theta \log(\pi(\theta))d\Pi(\theta)\right]}_{=-E_\Pi(\log(\pi(\theta)))}}^{H(\pi)=} \tag{3}$$

Definition 10. The cross-entropy between two probability distributions with densities π and u is defined as

$$\mathrm{H}(\pi,u) = \mathrm{E}_\Pi(-\log(u(\theta))) = \begin{cases} -\int_\Theta \log(u(\theta))\pi(\theta)\mathrm{d}\theta & \text{, if } \theta \text{ is cont.} \\ \\ -\sum_{\theta\in\Theta} \log(u(\theta))\pi(\theta) & \text{, if } \theta \text{ is discr.} \end{cases}$$

Note 11. Cross-entropy $\mathrm{H}(\pi,u)$ measures lack of fit between two densities π and u similar to $\mathrm{KL}(\pi\|u)$. In (3) the third term does not depend on u.

[2] can be proved by log-sum and Jensen's inequalities

Definition 12. The entropy of a random variable $\theta \in \Theta$ with distribution Π admitting density $\pi(\theta)$ is defined as

$$\mathrm{H}(\pi) = \mathrm{E}_{\Pi}(-\log(\pi(\theta))) = \begin{cases} -\int_{\Theta} \log(\pi(\theta))\pi(\theta)\mathrm{d}\theta & \text{, if } \theta \text{ is cont.} \\ \\ -\sum_{\theta\in\Theta} \log(\pi(\theta))\pi(\theta) & \text{, if } \theta \text{ is discr.} \end{cases}$$

Note 13. Entropy of a random variable $\theta \in \Theta$ with distribution Π admitting pdf/pmf $\pi(\theta)$ measures how much $\pi(\theta)$ diverges from the uniform density $u(\theta) = 1/|\Theta|$ on the support of θ, when Θ is bounded. The more $\pi(\theta)$ diverges the lesser its entropy and vice versa:

$$\mathrm{H}(\pi) = \log(|\Theta|) - \int_{\Theta} (\log(\pi(\theta)) - \log(1/|\Theta|))\mathrm{d}\Pi(\theta) = \underbrace{\log(|\Theta|)}_{=\text{constant}} - \mathrm{KL}(\pi\|u)$$

Note 14. The specification of a measure for the difference between two functions allows to set-up the Desiderata on Note 5 and compute π such as

$$\begin{aligned} \text{minimise: } & \mathrm{KL}(\pi\|u) && \text{approximate the reference measure } u(\theta) && (4) \\ \text{subject to: } & \mathrm{E}_{\Pi}(h_1(\theta)) = m_1 && \text{satisfy partial prior info} \\ & \vdots && \vdots \\ & \mathrm{E}_{\Pi}(h_k(\theta)) = m_k && \text{satisfy partial prior info} && (5) \\ & \int_{\Theta} \mathrm{d}\Pi(\theta) = 1 && \text{hopefully redive a ptoper prior} && (6) \end{aligned}$$

Based on the method of Lagrange multipliers, solving (4)-(6) is equivalent to minimizing, < AMV2

$$Q(\pi,\lambda) = \int_{\Theta} \log(\frac{\pi(\theta)}{u(\theta)})\mathrm{d}\Pi(\theta) + \sum_{j=1}^{k} \lambda_j [\int_{\Theta} h_j(\theta)\mathrm{d}\Pi(\theta) - m_j] + \lambda_0 [\int_{\Theta} \mathrm{d}\Pi(\theta) - 1] \quad (7)$$

with respect to π and $\lambda = (\lambda_0, ..., \lambda_k)$ where $\{\lambda_j\}$ are arbitrary constants.

- We will call such priors as Maximum entropy priors.

Theorem 15. *The function* $Q(\pi)$, in *(7), is minimized by the pdf/pmf*

$$\pi(\theta) \qquad = g(\lambda)u(\theta)\exp(\sum_{j=1}^{k} \lambda_j h_j(\theta)) \qquad \propto u(\theta)\exp(\sum_{j=1}^{k} \lambda_j h_j(\theta)). \quad (8)$$

where,

$$g(\lambda)^{-1} = \int_{\Theta} u(\theta)\exp(\sum_{j=1}^{k} \lambda_j h(\theta))d\theta < +\infty$$

and $\lambda = (\lambda_1, ..., \lambda_k)$ *such as* $m_j = -\frac{\partial}{\partial\lambda_j}\log(g(\lambda))$ for all, $j = 1, ..., k$.

Proof. A sketch of the proof is given in the Appendix 3, but it is out the scope and not examinable. □

Remark 16. Maximum entropy priors are exponential family distributions (see Theorem 15). we can refer to (8) as 'the exponential family generated by u and h'.

2.2 Non-informative framework

Note 17. Assume interest lies in specifying a prior distribution Π with density $\pi(\theta)$ which satisfies the constrains in (1), but apart from that it is (in some sense) non-informative. In this case, the reference measure u can be defined in many ways; Eg. Laplace prior $u(\theta) \propto \pi^{(\mathrm{L})}(\theta)$, Jeffreys' prior $u(\theta) \propto \pi^{(\mathrm{J})}(\theta)$, etc... Still the maximum entropy priors can be produced by solving the system (4-5), but without the requirement to integrate to 1.

Note 18. If the reference prior measure is Jeffreys' prior $u(\theta) \propto \pi^{(\mathrm{J})}(\theta)$, we get

$$\pi(\theta) \propto \pi^{(\mathrm{J})}(\theta) \exp(\sum_{j=1}^{k} \lambda_j h_j(\theta)) \tag{9}$$

where $\lambda = (\lambda_1, ..., \lambda_k)$ such as $m_j = -\frac{\partial}{\partial \lambda_j} \log(g(\lambda))$ for all, $j = 1, ..., k$.

Note 19. If the reference prior measure u is very 'vague', in the sense that u is extremely diffusely spread over Θ; in other words: $u(\theta) \propto 1$, e.g. the Laplace prior $u(\theta) \propto \pi^{(\mathrm{L})}(\theta) \propto 1$, we get

$$\pi(\theta) \propto \exp(\sum_{j=1}^{k} \lambda_j h_j(\theta))$$

where $\lambda = (\lambda_1, ..., \lambda_k)$ such as $m_j = -\frac{\partial}{\partial \lambda_j} \log(g(\lambda))$ for all, $j = 1, ..., k$.

Note 20. When the reference measure is 'vague' $u(\theta) \propto 1$, maximizing $\mathrm{KL}(\pi \| u)$ subject to a given set of k constraints is equivalent to maximizing[3] the 'entropy' $\mathrm{H}(\pi)$ subject to a given set of k constraints. In this information-theoretic sense, the maximum entropy prior π is such that the prior information brought through π about θ is minimized.

Example 21. Specify the Maximum entropy prior for $\Theta = \mathbb{R}^+$, subject to constraints $\mathrm{E}_\Pi(\theta) = m_1$ with reference measure $u(\theta) \propto 1$.

Hint: Exponential distribution $x \sim \mathrm{Ex}(r)$ has pdf $f(x) = r\exp(-rx)1(x > 0)$, and mean $\mathrm{E}(x) = 1/r$.

Solution. It is $h(\theta) = \theta$, so the prior is such that

$$\pi(\theta) \propto \exp(\lambda_1 \theta) \propto \mathrm{Ex}(\theta | -\lambda_1)$$

and from the constrains $\mathrm{E}_\Pi(\theta) = -1/\lambda_1$, hence $\lambda_1 = -1/m_1$

Example 22. Specify the Maximum entropy prior for $\Theta = \mathbb{R}$, subject to constraints $\mathrm{E}_\Pi(\theta) = m_1$, $\mathrm{E}_\Pi(\theta^2) = m_2$ with reference measure $u(\theta) \propto 1$.

Solution. It is $h(\theta) = (\theta, \theta^2)$, so the prior is such that

$$\pi(\theta) \propto \exp(\lambda_2 \theta^2 + \lambda_1 \theta) \propto \exp(-\frac{1}{2}\frac{(\theta - (-\frac{\lambda_1}{2\lambda_2}))^2}{-\frac{1}{2\lambda_2}}) \ \propto \ \mathrm{N}(\theta | \mu = -\frac{\lambda_1}{2\lambda_2}, \sigma^2 = -\frac{1}{2\lambda_2})$$

From the constrains, $\mathrm{E}_\Pi(\theta) = m_1$ and $\mathrm{E}_\Pi(\theta^2) = m_2$. So $\mu = \mathrm{E}_\Pi(\theta) = m_1$, and $\sigma^2 = \mathrm{E}_\Pi(\theta^2) - (\mathrm{E}_\Pi(\theta))^2 = m_2 - m_1^2$.

Remark 23. The constraints (1) are not always sufficient to derive a proper prior distribution on θ. Maximum entropy priors may be improper; In that case, the proneness condition has to be checked.

Remark 24. Maximum entropy priors are mainly used as an objective Bayesian treatment, by choosing $u(\theta)$ as a non-informative (and so without any subjective information) measure; eg, Laplace prior, Jeffreys' prior, etc... However, it is unclear how the constrains (1) contribute to this objective story...

Remark 25. Maximum entropy priors are not necessarily invariant under re-parametrisations, e.g. in contrast to Jeffrey's priors.

[3] This is the actual 'Maximum entropy prior' , however we will call the same those in (8) and (9).

3 Appendix

Proof. Sketch of the proof of Theorem 15. It can be skipped and it is not examinable.
Consider the case that θ is continuous random quantity. By using variation calculus arguments, a necessary condition for π to give a stationary value of $Q(\pi)$ is

$$\frac{\partial}{\partial a}Q(\pi(\theta)+a\tau(\theta))|_{a=0}=0,$$

for any function $\tau:\Theta\rightarrow\mathbb{R}$ of sufficiently small norm. It is

$$\begin{aligned}Q(\pi(\theta)+a\tau(\theta))&=\int_{\Theta}\log(\frac{\pi(\theta)+a\tau(\theta)}{u(\theta)})(\pi(\theta)+a\tau(\theta))\mathrm{d}\theta+\sum_{j=1}^{k}\lambda_j[\int_{\Theta}h_j(\theta)(\pi(\theta)+a\tau(\theta))\mathrm{d}\theta-m_j]\\&+\lambda_0[\int_{\Theta}(\pi(\theta)+a\tau(\theta))\mathrm{d}\theta-1]\qquad\Longrightarrow\\\frac{\partial}{\partial a}Q(\pi(\theta)+a\tau(\theta))&=\int_{\Theta}\left(\tau(\theta)+\log(\frac{\pi(\theta)+a\tau(\theta)}{u(\theta)})\tau(\theta)\right)\mathrm{d}\theta+\sum_{j=1}^{k}\lambda_j\int_{\Theta}h_j(\theta)\tau(\theta)\mathrm{d}\theta+\lambda_0\int_{\Theta}\tau(\theta)\mathrm{d}\theta\\&=\int_{\Theta}\left(1+\log(\frac{\pi(\theta)+a\tau(\theta)}{u(\theta)})+\sum_{j=1}^{k}\lambda_jh_j(\theta)+\lambda_0\right)\tau(\theta)\,\mathrm{d}\theta\end{aligned}$$

Hence, the condition

$$\frac{\partial}{\partial a}Q(\pi(\theta)+a\tau(\theta))|_{a=0}=0$$

reduces to

$$\int_{\Theta}(\log(\frac{\pi(\theta)}{u(\theta)})+\sum_{j=1}^{k}\lambda_jh_j(x)+(\lambda_0+1))\tau(\theta)\mathrm{d}\theta=0$$

which implies

$$\pi(\theta)=\underbrace{\left(\frac{1}{\exp(\lambda_0+1)}\right)}_{=g(\lambda)}u(\theta)\exp(\sum_{j=1}^{k}\lambda_jh_j(\theta)),\qquad(10)$$

by setting $\lambda_j\leftarrow-\lambda_j$ for $j=1,...,k$. By applying the normalizing constraints, we get,

$$g(\lambda)^{-1}=\int_{\Theta}u(\theta)\exp(\sum_{j=1}^{k}\lambda_jh_j(\theta))\mathrm{d}\theta.$$

We recognize that (10) is a member of the exponential family. Its canonical form can be recovered for $y_j=h_j(\theta)$, $\psi_j=\lambda_j$, and $b(\psi(\lambda))=-\log(g(\lambda))$ (since . ... $\psi_j=\lambda_j$), we get

$$\begin{aligned}\mathrm{E}(y|\psi)&=\frac{\mathrm{d}}{\mathrm{d}\psi}b(\psi)\Longleftrightarrow\\{}[\mathrm{E}(y|\psi)]_j&=\frac{\partial}{\partial\psi_j}b(\psi),\text{ for all, }j=1,...,k\Longleftrightarrow\\\mathrm{E}(h_j(\theta)|\lambda)&=\frac{\partial\lambda_j}{\partial\psi_j}\frac{\partial}{\partial\lambda_j}(-\log(g(\lambda))),\text{ for all, }j=1,...,k\Longleftrightarrow\\m_j&=-\frac{\partial}{\partial\lambda_j}\log(g(\lambda)),\text{ for all, }j=1,...,k\end{aligned}$$

□

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 10: Decision theory set-up [a]

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

Aim: To explain concepts and elements of decision theory

References:

- Berger, J. O. (2013; Sections 1.3, 4.4, and 4.8.3). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- Robert, C. (2007; Chapter 2, pp. 52-102). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- Raiffa, H., & Schlaifer, R. (1961; Chapter 1). Applied statistical decision theory.
- Ferguson, T. S. (1967; Chapter 1 & 2). Mathematical statistics: A decision theoretic approach (Vol. 1). Academic press.

[a] Author: Georgios P. Karagiannis.

1 Why we need decision theory

Once we have a probability distribution describing our knowledge of propositions of interest, what can we do with it? We take decisions all the time, for example, to take a particular decision rather than another, or, more abstractly, to decide on a best estimate of the true value of a quantity. Is probability theory involved here?

A decision is certainly a function of our knowledge quantified by probabilities. For example, a doctor deciding on the best treatment for a patient will certainly use their knowledge of what illness the patient has. Different illnesses may have different probabilities based on the symptoms, test results, and so on, and it would be foolish to concentrate on the most improbable illness and treat for that. Such questions constitute a decision theory problems.

In statistics, the overall purpose of most inferential studies is to provide the statistician (or a client) with a decision (E.g, what estimator to use? Shall I reject a hypothesis?), it seems reasonable to ask for an evaluation criterion of decision procedures that assesses the consequences of each decision and depends on the parameters of the model, i.e., the true state of the world (or of Nature). Such a criterion can be the loss function (or more rigorously the utility function).

2 Set-up of the decision problem (intro)

Definition 1. The decision problem $(\Theta, \mathcal{D}, \ell)$ involves the following basic elements

- Decision space: $\mathcal{D} = \{d\}$

  The decision maker wishes to select a single decision $d \in \mathcal{D}$ from a space of all possible decisions $\mathcal{D}$.

- State space (Space of the word or Parameter space): $\Theta = \{\theta\}$

The decision process is assumed to be affected by the unknown (uncertain) quantity $\theta \in \Theta$ which signifies the state of the world. The set of all possible states of the world is denoted by Θ. The decision maker perceives that a particular decision $d \in \mathcal{D}$ results in a corresponding state $\theta \in \Theta$.

- Loss function $\ell : \Theta \times \mathcal{D} \to \mathbb{R}_+$

  The decision maker assigns a loss function (or error) $\ell(d, \theta)$ that evaluates the penalty (or error, or suffer, or regret, or etc...) associated with decision d when the parameter takes the value θ.

Definition 2. Statistical decision problem is a decision problem $(\Theta, \mathcal{D}, \ell)$ coupled with an experiment $e \in \mathcal{E}$ ($\mathcal{E}$ denotes the family of experiments) that involves an observable $y \in \mathcal{Y}$ ($\mathcal{Y}$ denotes the sample space) whose sampling distribution $\mathrm{d}F(y|\theta)$ depends on the state $\theta \in \Theta$ chosen by nature.

Note 3. The main aim in a statistical decision problem $(\Theta, \mathcal{D}, \ell)$ is for the decision maker to choose the/an optimal decision $d^* := d^*(y) \in \mathcal{D}$.

Example 4. Statistical inference consists of taking a decision $d = d(y) \in \mathcal{D}$ elated to the parameter $\theta \in \Theta$ based on the observation $y \in \mathcal{Y}$, where y and θ are related by the sampling distribution $F(y|\theta)$. In many cases, the decision $d \in \mathcal{D}$ will be a procedure performing: Point estimation (what type of estimator to use?), Credible regions (how to find the bounds?), Hypothesis tests (how to reject the hypothesis?), model selection (which model is the 'best'?).

Definition 5. Decision rule is a function which maps the observables $y \in \mathcal{Y}$ to an appropriate decision $d \in \mathcal{D}$. It is defined as $\delta : \mathcal{Y} \to \mathcal{D}$, where $\delta(y) \in \mathcal{D}$, and it is such that

$$\int_{\mathcal{Y}} \ell(\theta, \delta(y)) \mathrm{d}F(y|\theta) < \infty$$

Note 6. (Out of the scope) Formally speaking about statistical decision problems,

- The decision maker assigns a utility function $u : \mathcal{E} \times \mathcal{Y} \times \mathcal{D} \times \Theta \to \mathbb{R}$ with $u(e, y, d, \theta)$ which describes the value to perform a particular experiment $e \in \mathcal{E}$, observing a particular observation $y \in \mathcal{Y}$, taking a particular decision $d \in \mathcal{D}$, and then finding that a particular $\theta \in \times$ obtains.
- The existence of the utility function u is introduced in Decision Theory 3, and ignored in this handout.
- Once the utility function u is defined, the loss function ℓ can be specified as

  $$\ell(e, y, d, \theta) = u(e^*, y, d^*, \theta) - u(e, y, d, \theta), \quad \text{or} \quad \ell(e, y, d, \theta) = -u(e, y, d, \theta)$$

  where e^* and d^* are optimal choices of the experiment and decision.
- In this handout, we assume the simplified case where the experiment e is fixed and hence we drop e, y from arguments in u and ℓ; hence

  $$\ell(d, \theta) = u(d^*, \theta) - u(d, \theta), \quad \text{or} \quad \ell(d, \theta) = -u(d, \theta).$$

3 Finding Bayesian optimal decision rules (intro)

Question 7. *Given a statistical decision problem $(\Theta, \mathcal{D}, \ell)$, what is the optimal decision rule $\delta(y) \in \mathcal{D}$?*

Note 8. To get an optimal decision there is a need to derive an effective comparison criterion that orders the possible decisions $d \in \mathcal{D}$ based on the loss function $\ell(\cdot, \cdot)$, (equiv. the utility function $u(\cdot, \cdot)$).

Definition 9. (Frequentist) Risk function $R(\theta, \delta) := R(\theta, \delta(y))$ is defined as

$$R(\theta, \delta) = \mathrm{E}_F(\ell(\theta, \delta(y))|\theta) \quad = \int_{\mathcal{Y}} \ell(\theta, \delta(y)) \mathrm{d}F(y|\theta), \tag{1}$$

where $\delta(\cdot)$ is the decision rule, i.e, the allocation of a decision to each outcome $y \sim F(y|\theta)$ from the random experiment e.

~~*Note* 10. For the derivation of inferential tools for θ, such as estimators $\hat{\theta}$, the frequentist approach relies on comparisons based on the risk function (1) in order to choose the 'optimal' decision $\hat{\theta} = \delta^{(\mathrm{FR})}(y)$. The objections are the following:~~

- ~~Risk function $R(\theta, \delta)$ averages loss function $\ell(\theta, \delta(y))$ over the different values of y proportionally to the density $f(y|\theta)$. Therefore, it seems that the observation y is not taken into account any further. It evaluates procedures on their long-run performance and not directly for the given observation y. To average over all possible values of y, when we know the observed value of y, on one hand may lead to interesting mathematical properties, but on the other hand it is rather a waste of information.~~
- ~~The frequentist approach implicitly assumes that this problem will be met again and again for the frequency evaluation to make sense. However, there are objections against the this sense of repeatability of experiments. Eg, if new observations become available and one wants to make use of them, this could possibly modify the way the experiment is conducted.~~
- ~~Risk function $R(\theta, \delta)$ is a function of the parameter θ. There may not exist, in general, an optimal procedure $\delta^{(\mathrm{F})}(\cdot)$ that uniformly minimises $R(\theta, \delta)$ for all $\theta \in \Theta$. The frequentist approach does not induce a total ordering on the set of procedures. For this reason, frequentist approach, requires additional restrictions (in a rather artificial manner), eg, admissibility.~~

Note 11. To address the decision problem $(\Theta, \mathcal{D}, \ell)$ in the Bayesian paradigm the decision maker assigns a probability distribution $\mathrm{d}P(\theta, y)$ on the possibility space $\Theta \times \mathcal{Y}$. Recall (Handout 3; Remark 22) that the joint distribution $\mathrm{d}P(\theta, y)$ determines probability distributions such as

$$\mathrm{d}P(y, \theta) = \mathrm{d}F(y|\theta)\mathrm{d}\Pi(\theta) = \mathrm{d}\Pi(\theta|y)\mathrm{d}F(y)$$

where $\mathrm{d}\Pi(\theta)$ is the prior distri., $\mathrm{d}\Pi(\theta|y)$ is the posterior distr., and $\mathrm{d}F(y)$ is the prior predictive distr.

Note 12. To derive a quantity able to order decisions in the Bayesian paradigm, it is reasonable to integrate out the loss $\ell(\theta, d)$ on the space Θ, (instead of integrating on the space $\mathcal{Y}$) and condition on the observables y. This is because θ is unknown and y is known/observed.

Definition 13. The posterior expected loss of a decision $d \in \mathcal{D}$ when the posterior distribution is $\Pi(\theta|y)$ is defined as

$$\varrho(\pi, d|y) = \mathrm{E}_{\Pi}(\ell(\theta, d)|y) \quad = \int_{\Theta} \ell(\theta, d)\mathrm{d}\Pi(\theta|y) \tag{2}$$

Note 14. Posterior expected loss $\varrho(\pi, d|y)$ integrates the loss $\ell(\theta, d)$ with respect to the posterior distribution $\mathrm{d}\Pi(\theta|y)$ of the parameter θ, conditionally on the observed value y. It is a function of y but this dependence is not troublesome, because y is known (as it is observed).

Note 15. It is reasonable for a Bayesian Statistician to consider as optimal decision the one that minimizes (2) for a given observable y.

Definition 16. The Bayes risk of a decision $d \in \mathcal{D}$, with respect to a prior $\Pi(\theta)$ on Θ is defined as

$$r(\pi, \delta) = \mathrm{E}_{\Pi}(R(\theta, \delta)) \quad = \int_{\Theta} \int_{\mathcal{Y}} \ell(\theta, \delta(y))\mathrm{d}F(y|\theta)\mathrm{d}\Pi(\theta); \tag{3}$$

It is the risk function $R(\theta, \delta)$ integrated over θ with respect to the prior distribution $\mathrm{d}\Pi(\theta)$

Remark 17. Bayes risk $r(\pi, \delta)$ induces a total ordering on the set of potential decisions, and hence it allows for the direct comparison of decisions unlike the risk function $R(\theta, \delta)$. This is because $r(\pi, \delta)$ associates a real number with every decision d, it is not a function of θ.

Note 18. In the Bayesian framework an optimal decision for the decision problem $(\Theta, \mathcal{D}, \ell)$ can be found in two ways:

1. the extensive form of analysis: it minimizes the posterior expected loss $\varrho(\pi, d|y)$ (2) w.r.t. d
$$\min_{d} \int_{\Theta} \ell(\theta, d) \mathrm{d}\Pi(\theta|y)$$

2. the Normal form of analysis: it minimizes the Bayesian risk $r(\pi, \delta)$ (3) w.r.t. δ
$$\min_{\delta} \int_{\Theta} \int_{\mathcal{Y}} \ell(\theta, \delta(y)) \mathrm{d}F(y|\theta) \mathrm{d}\Pi(\theta)$$

Theorem 19 provides a constructive tool and a reasonable justification that the two ways are equivalent.

Theorem 19. *An decision minimizing the integrated risk $r(\pi, \delta)$ can be obtained by selecting, for every $y \in \mathcal{Y}$, the value $\delta(y)$ which minimizes the posterior expected loss $\varrho(\pi, d|y)$, since*
$$\min_{\forall \delta \in \mathcal{D}} r(\pi, \delta) = \int_{\mathcal{Y}} \min_{\forall d \in \mathcal{D}} \varrho(\pi, d|y) dF(y)$$
and
$$r(\pi, \delta) = \int_{\mathcal{Y}} \varrho(\pi, d|y) dF(y) = \begin{cases} \int_{\mathcal{Y}} \varrho(\pi, d|y) f(y) dy & , y \textit{ cont.} \\ \\ \sum_{\mathcal{Y}} \varrho(\pi, d|y) f(y) & , y \textit{ is disc.} \end{cases}$$

Proof. For this proof assume that the loss function is bounded ($\ell(\theta, d) \geq 0$). Assume that x and θ are continuous for simplicity, e.g.
$$p(\theta, y) = f(y|\theta)\pi(\theta) = \pi(\theta|y) f(y)$$

It is
$$\begin{aligned} r(\pi, \delta) &= \int_{\Theta} \int_{\mathcal{Y}} \ell(\theta, \delta(y)) \mathrm{d}F(y|\theta) \mathrm{d}\Pi(\theta) \\ &= \int_{\Theta} \int_{\mathcal{Y}} \ell(\theta, d) f(y|\theta) \mathrm{d}y \pi(\theta) \mathrm{d}\theta && (4) \\ &= \int_{\mathcal{Y}} \int_{\Theta} \ell(\theta, d) f(y|\theta) \pi(\theta) \mathrm{d}\theta \mathrm{d}y && (5) \\ &= \int_{\mathcal{Y}} \int_{\Theta} \ell(\theta, d) \pi(\theta|y) f(y) \mathrm{d}\theta \mathrm{d}y && (6) \\ &= \int_{\mathcal{Y}} \left[\int_{\Theta} \ell(\theta, d) \mathrm{d}\Pi(\theta|y) \right] \mathrm{d}F(y) && (7) \\ &= \int_{\mathcal{Y}} \varrho(\pi, d|y) \mathrm{d}F(y) && (8) \end{aligned}$$

Hence, it is implied that
$$\min_{\forall \delta \in \mathcal{D}} r(\pi, \delta) = \int_{\mathcal{X}} \min_{\forall d \in \mathcal{D}} \varrho(\pi, d|y) f(y) \mathrm{d}y.$$
Here, we used Fubini's theorem to go from (4) to (5), and Bayes theorem to go from (5) to (6). □

A Bayes rule associated with a prior distribution $\Pi(\theta)$ and a loss function $\ell(\cdot, \cdot)$ is any δ^{π} which minimizes $r(\pi, \delta)$. For every $y \in \mathcal{Y}$, it is given by $\delta^{\pi}(y)$ such that
$$\delta^{\pi}(y) = \arg\min_{d} \varrho(\pi, d|y)$$

Definition 20. Minimum Bayes risk is defined as the value $r(\pi) = r(\pi, \delta^{\pi})$, where δ^{π} is a Bayes rule.

Remark 21. Operationally, to find the Bayes rule $\delta^{\pi}(y)$, we will mainly minimize the posterior expected loss $\varrho(\pi, d|y)$.

Remark 22. From a strictly Bayesian point of view, only the posterior expected loss $\varrho(\pi, d|y)$ is important, because the Bayesian paradigm is based on the conditional approach (conditional on the observations).

Remark 23. The 'reasonable' way to view the situation is that minimizing the posterior expected loss $\varrho(\pi, d|y)$. One should condition on what is known (aka the observables y) and integrate/average out on what is unknown (aka θ). Minimizing $r(\pi, \delta^{\pi})$ that integrates out y seems bizarre from this perspective.

Remark 24. Compared to the Frequentist, the Bayesian approach is sufficiently reductive to reach an effective decision δ^{π} because, by minimizing the posterior expected risk $\varrho(\pi, d|y)$ (w.r.t. δ^{π}), it minimizes the integrated risk $r(\pi, \delta^{\pi})$ (w.r.t. δ^{π}) which in fact allows a direct comparison of estimators.

Remark 25. The Bayesian approach works conditional upon the actual observation y, as well as it incorporates the probabilistic properties of the sampling distribution $F(y|\theta)$. This is in contrast to the frequentist approach (Note 10) where the observed information y seems to be wasted because it averages over all possible values of y instead of conditioning on the given observed y.

Definition 26. Generalized Bayes rule is called any $\delta^{\pi}(y)$ that minimizes $r(\pi, \delta)$ or $\varrho(\pi, d|y)$ for each $y \in \mathcal{Y}$ where $f(y) > 0$ but the prior distribution $\Pi(\theta)$ is improper.

Remark 27. Definition 3, and Theorem 19 are valid for proper and improper priors, provided that $r(\pi) < \infty$, otherwise other treatments exist.

4 Admissibility

Contrary to the Bayesian approach, the frequentist approach is not reductive enough to lead to a single optimal rule/estimator, and for this reason frequentist use additional optimality concepts such as admissibility. Here, we study Bayes rules/estimators with respect to the frequentist optimality criterion of admissibility.

Definition 28. A decision rule δ_0 is inadmissible if there exists a decision rule δ_1 which dominates δ_0, namely:

$$
\begin{aligned}
R(\theta, \delta_0) &\geq R(\theta, \delta_1), \quad \forall \theta \in \Theta \\
R(\theta_0, \delta_0) &> R(\theta_0, \delta_1), \quad \exists \theta_0 \in \Theta
\end{aligned}
$$

Otherwise, δ_0 is said to be admissible.

Remark 29. Well, ... I wouldn't use an inadmissible estimator δ as a decision rule with a better risk can be found...

Admissibility in Bayesian rules

Note 30. One may expect Bayes rule to be admissible because if a rule with better risk $R(\theta, \delta)$ existed, that rule would also have better Bayes risk $r(\pi, \delta) = \mathrm{E}_{\Pi}(R(\theta, \delta))$. This that is true in cases that the π is proper, and we study this case in what follows. Surprisingly, it is not always true when π is improper (non-informative), and unfortunately we do not have time to go study this.

Theorem 31. *If the Bayes rule associated with a prior π is unique, it is admissible.*

Proof. Let δ^{π} be a Bayes rule. Assume that δ^{π} is inadmissible. So let δ^{*} be any decision rule with $R(\theta, \delta^{*}) \leq R(\theta, \delta^{\pi})$, for all $\theta \in \Theta$. Then

$$
\begin{aligned}
r(\pi, \delta^{\pi}) - r(\pi, \delta^{*}) &= \int_{\Theta} R(\theta, \delta^{*}) \mathrm{d}\Pi(\theta) - \int_{\Theta} R(\theta, \delta^{\pi}) \mathrm{d}\Pi(\theta) \\
&= \int_{\Theta} (R(\theta, \delta^{*}) - R(\theta, \delta^{\pi})) \mathrm{d}\Pi(\theta) \leq 0
\end{aligned}
$$

and so δ^* is also Bayes. Because δ^π is unique Bayes by assumption, we must have $\delta^*(y) = \delta^\pi(y)$ for all $y \in \mathcal{Y}$. Therefore, δ^π must be admissible. □

Theorem 32. *If a prior distribution π is strictly positive on Θ, with finite Bayes risk and the risk function, $R(\theta, \delta)$, is a continuous function of θ for every δ, the Bayes estimator δ^π is admissible.*

~~*Proof.* Let δ^π be a Bayes rule. Assume that δ^π is inadmissible. So let δ^* be any decision rule with $R(\theta, \delta^*) \le R(\theta, \delta^\pi)$, for all $\theta \in \Theta$, and $R(\theta_0, \delta^*) < R(\theta_0, \delta^\pi)$, for some $\theta_0 \in \Theta$. Let $R(\theta_0, \delta^\pi) - R(\theta_0, \delta^*) = \eta > 0$ for some $\theta_0 \in \Theta$. By continuity of $R(\theta, \delta)$ in θ, $R(\theta, \delta^\pi) - R(\theta, \delta^*)$ is continues as well, and hence~~

$$(\forall \epsilon > 0)(\exists \zeta > 0)(\forall \theta \in \Theta)(|\theta - \theta_0| < \zeta \Longrightarrow |R(\theta, \delta^\pi) - R(\theta, \delta^*)| < \epsilon)$$

~~which implies $R(\theta, \delta^\pi) - R(\theta, \delta^*) > \eta - \epsilon = \tilde{\eta} > 0$ Let $A = \{\theta \in \Theta \text{ st } |\theta - \theta_0| < \zeta\}$. Then,~~

$$\begin{aligned}
r(\pi, \delta^\pi) - r(\pi, \delta^*) &= \int_\Theta (R(\theta, \delta^\pi) - R(\theta, \delta^*))\mathrm{d}\Pi(\theta) \\
&= \int_A (R(\theta, \delta^\pi) - R(\theta, \delta^*))\mathrm{d}\Pi(\theta) + \int_{A^\complement} (R(\theta, \delta^\pi) - R(\theta, \delta^*))\mathrm{d}\Pi(\theta) \\
&\ge \int_A (R(\theta, \delta^\pi) - R(\theta, \delta^*))\mathrm{d}\Pi(\theta) > \int_A \tilde{\eta}\mathrm{d}\Pi(\theta) = \tilde{\eta}\mathsf{P}_\Pi(\theta \in A) > 0
\end{aligned}$$

~~which contradicts that δ^π is Bayes rule. Therefore, δ^π must be admissible.~~ □

Theorem 33. *Assume that Θ is discrete (say $\Theta = \{\theta_1, ...\}$) and that the prior π gives positive probability to each $\theta_i \in \Theta$, then Bayes rule δ^π with respect to π is admissible.*

Proof. Let δ^π be a Bayes rule, and δ^* be any decision rule with $R(\theta, \delta^*) \le R(\theta, \delta^\pi)$, for all $\theta \in \Theta$, and $R(\theta_k, \delta^*) < R(\theta_k, \delta^\pi)$, for some $\theta_k \in \Theta$. Let $R(\theta_k, \delta^\pi) - R(\theta_k, \delta^*) = \eta > 0$ for some $\theta_0 \in \Theta$. It is

$$\begin{aligned}
r(\pi, \delta^\pi) - r(\pi, \delta^*) &= \sum_{i=1}^{\infty} (R(\theta_i, \delta^\pi) - R(\theta_i, \delta^*))\pi(\theta_i) \\
&= \sum_{i \ne k} (R(\theta_i, \delta^\pi) - R(\theta_i, \delta^*))\pi(\theta_i) + (R(\theta_k, \delta^\pi) - R(\theta_k, \delta^*))\pi(\theta_k) \\
&\ge (R(\theta_k, \delta^\pi) - R(\theta_k, \delta^*))\pi(\theta_k) \\
&> \eta\pi(\theta_k) > 0
\end{aligned}$$

which contradicts that δ^π is Bayes estimator. Therefore, δ^π must be admissible. □

Admissibility in Generalised Bayesian rules

Generalized Bayes rules (Definition 26) may be or may not be admissible.

Remark 34. When the loss is positive and

$$r(\pi, \delta^\pi) = \int_\Theta R(\theta, \delta^\pi)\mathrm{d}\Pi(\theta) < \infty,$$

the generalized Bayes rule δ^π (improper prior) can be easily shown to be admissible, similar to the Bayes rule case (proper prior). In fact in this case, as mentioned in Remark 27, δ^π minimizes $r(\pi, \delta)$ or $\varrho(\pi, d|x)$ and it can be shown (similar to the Bayes rule) that generalized Bayes rule must be admissible under suitable conditions.

Remark 35. When

$$r(\pi, \delta^\pi) = \int_\Theta R(\theta, \delta^\pi)\mathrm{d}\Pi(\theta) = \infty$$

even super reasonable generalized Bayes rules can be inadmissible; (See the following example). Unfortunately, the use of improper priors quite often lead to $r(\pi, \delta^{\pi}) = \infty$, hence they are evil....

Example 36. Consider the Bayesian model

$$\begin{cases} y & \sim \text{Ga}(a, 1/b) \\ b & \sim \Pi(b) \end{cases}$$

where the non informative prior such that $\pi(b) \propto \frac{1}{b}$ is used, and $a > 0$ is known.

1. Check if prior $\pi(b)$ is proper.
2. Given the loss function $\ell(b, \delta) = (b - \delta)^2$, find the Bayes rule (estimate).
3. Is that Bayes rule admissible? Check it with the decision rule $\delta_c(y) = cy$ for $c \in \mathbb{R}$.

Hint: Gamma distr.: $x \sim \text{Ga}(a, b)$ has pdf $f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) 1_{(0,+\infty)}(x)$, $\text{E}(x) = \frac{a}{b}$, and $\text{Var}(x) = \frac{a}{b^2}$.

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

Solution.

1. It is $\int_0^{\infty} \text{d}\Pi(b) = \int_0^{\infty} \pi(b)\text{d}b = \int_0^{\infty} \frac{1}{b}\text{d}b = \infty$ so it is an improper prior.
2. By using Bayes theorem, I can compute the posterior distribution and recognize it as $b|y \sim \text{IG}(a, y)$. So the Bayes rule is

$$\begin{aligned} 0 = & \frac{\text{d}}{\text{d}\delta} \varrho(\pi, \delta|y) \Big|_{\delta=\delta^{\pi}} = \frac{\text{d}}{\text{d}\delta} \text{E}_{\text{IG}(b|a,y)}(\ell(b, \delta)) \Big|_{\delta=\delta^{\pi}} \\ = & \frac{\text{d}}{\text{d}\delta} \int_{\Theta} (b - \delta)^2 \text{IG}(b|a, y)\text{d}b \Big|_{\delta=\delta^{\pi}} \int_{\Theta} \frac{\text{d}}{\text{d}\delta} (b - \delta)^2 \text{IG}(b|a, y)\text{d}b \Big|_{\delta=\delta^{\pi}} \\ = & -2 \int_{\Theta} b\text{IG}(b|a, y)\text{d}b + 2\delta^{\pi} = -2\text{E}_{\text{IG}(a,y)}(b) + 2\delta^{\pi} = -2\frac{y}{a-1} + 2\delta^{\pi} \implies \\ \delta^{\pi}(x) = & \frac{y}{a-1} \end{aligned}$$

3. First, I will try to check first if $\delta^{\pi}(y)$ is inadmissible. In order to show that $\delta^{\pi}(y)$ is inadmissible, I need to find a decision rule that dominates $\delta^{\pi}(y)$.

   Consider the decision rule $\delta_c(y) = cy$, where c is some constant. I will try to find a value for c, let's say c_* such that $\delta_{c_*}(y)$ can dominate δ^{π}.

   Decision rule $\delta_c(y)$ has risk

$$\begin{aligned} R(b, \delta_c) = & \text{E}_{\text{Ga}(a,1/b)}(cy - b)^2 = \text{E}_{\text{Ga}(a,1/b)}(cy - c\text{E}_{\text{Ga}(a,1/b)}(y) + c\text{E}_{\text{Ga}(a,1/b)}(y) - b)^2 \\ = & c^2\text{E}_{\text{Ga}(a,1/b)}(y - \text{E}_{\text{Ga}(a,1/b)}(y))^2 + (c\text{E}_{\text{Ga}(a,1/b)}(y) - b)^2 \\ = & c^2\text{Var}_{\text{Ga}(a,1/b)}(x) + (cab - b)^2 = b^2(c^2 a + (ca - 1)^2) \end{aligned}$$

   To specify δ_{c_*}, a nice value c_* that I can use is the one that minimizes $R(b, \delta_c)$. It can be shown that $R(b, \delta_c)$ has minimum at $c_* = 1/(a + 1)$. In fact, it is

$$\frac{\text{d}}{\text{d}c} R(b, \delta_c) \Big|_{c=c_*} = 0 \implies 2b^2 a(c + (ca - 1))\big|_{c=c_*} = 0 \implies c_* = \frac{1}{a+1}$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}c^2}R(b,\delta_c)\bigg|_{c=c_*} = 2b^2a(a+1) > 0$$

Now let's check if δ_{c_*} dominates δ^π (and hence δ^π is inadmissible)

$$\frac{R(b,\delta^\pi)}{R(b,\delta_{c_*})} = \frac{R(b,\delta_{\frac{1}{a-1}})}{R(b,\delta_{\frac{1}{a+1}})} = \overset{\text{calc.}}{\cdots} = \frac{a(a-1)^{-2}+(a/(a-1)-1)^2}{a(a+1)^{-2}+(a/(a+1)-1)^2} = \frac{(a+1)^2}{(a-1)^2} > 1 \tag{9}$$

Hence, δ_{c_*} dominates δ^π, for all b. In fact, here, $R(b,\delta_{c_*}) < R(b,\delta^\pi)$ for all b. This shows that the generalised Bayesian rule $\delta^\pi(y) = \frac{y}{a-1}$ is inadmissible ...

- From (9), we see that the risk of $\delta^\pi(y)$ significantly worsens compared to that of $\delta_{c_*}(y)$ when a decreases.

Long run analysis of the behavior of generalised Bayesian estimator δ^π Consider the case that an objective Bayesian uses non-informative priors (improper in this case) automatically on a routine bases because he/she does not want to contaminate his/her statistical analysis with subjective elements. Due to this repeated use, the Bayesian enters into the frequentist domain; hence it is reasonable to investigate how repeated use of this prior actually performs.
Consider

- a sequence of independent problems $((\theta^{(1)}, y^{(1)}), (\theta^{(2)}, y^{(2)}), ...)$
- a loss function $\ell(\theta,\delta)$ that measures the performance of the procedure in each problem
- a quantity to compare δ_1 and δ_2, in the limit $N\rightarrow\infty$ is

  $$S_N = \sum_{i=1}^{N}(\ell(\theta^{(i)},\delta_1(y^{(i)})) - \ell(\theta^{(i)},\delta_2(y^{(i)})))^2$$

  the limiting behavior of S_N is related to the risk functions $R(\theta,\delta_1)$, $R(\theta,\delta_2)$ as we will see.

Theorem 37. *Consider $\theta = (\theta^{(1)},\theta^{(2)},...)$ to be any fixed sequence of parameters $\theta^{(i)}\in\Theta$, and suppose random variables $y^{(i)}\in\mathcal{Y}$ are independently generated from the parametric models $f(y^{(i)}|\theta^{(i)})$ (here f is the same for the entire sequence). Define the random variables*

$$Z_i = \ell(\theta^{(i)},\delta_1(x^{(i)})) - \ell(\theta^{(i)},\delta_2(x^{(i)}))$$

and assume that $\mathrm{Var}_F(Z_i|\theta^{(i)}) < \infty$ for all i. If $R(\theta,\delta_1) - R(\theta,\delta_2) > \epsilon > 0$ for all $\theta\in\Theta$, then

$$\lim_{N\rightarrow\infty}\inf\frac{1}{N}S_N > \epsilon, \qquad \textit{w.p. } 1 \tag{10}$$

for any sequence of θ.

Proof. It is

$$\mathrm{E}_F(Z_i|\theta^{(i)}) = R(\theta,\delta_1) - R(\theta,\delta_2)$$

Since $\mathrm{Var}_F(Z_i|\theta^{(i)}) < \infty$, by the SLLN, as $N\rightarrow\infty$, it is

$$\frac{1}{N}\sum_{i=1}^{N}(Z_i - \mathrm{E}_F(Z_i|\theta^{(i)})) \rightarrow 0, \qquad \text{w.p. } 1$$

because $\mathrm{E}_F(Z_i|\theta^{(i)}) > \epsilon$, (10) is implied. □

Remark 38. Theorem 37 implies that if our generalized Bayes rule (let's say δ_1) is inadmissible, δ_1 will be always inferior to δ_2 in actual practical use.

Remark 39. If ϵ is too small in (10), we may tolerate to use inadmissible δ_1 in routine use, however, we should proceed consciously. Once again, in Theorem 37, inadmissibility applies only to automated use of δ_1; eg, a computer package.

Question 40. *Practice with Exercise 58 from the Exercise sheet, or the crossed out proof of Theorem 32.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 11: Bayesian point estimation[a]

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

Aim: To explain and produce point estimators in the Bayesian framework.

References:

- Raiffa, H., & Schlaifer, R. (1961; Chapter 6). Applied statistical decision theory
- Berger, J. O. (2013; Section 4.3.1). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- DeGroot, M. H. (2005, Sections 11.1-11.4). Optimal statistical decisions (Vol. 82). John Wiley & Sons.
- Robert, C. (2007; Chapter 4.1). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

Web applets:
`https://georgios-stats-1.shinyapps.io/demo_pointestimation/`

[a] Author: Georgios P. Karagiannis.

1 Set-up and aim

Notation 1. Consider a Bayesian model

$$\begin{cases} y|\theta & \sim \mathrm{d}F(y|\theta) \\ \theta & \sim \mathrm{d}\Pi(\cdot) \end{cases}$$

where $y := (y_1, ..., y_n) \in \mathcal{Y}$ is a sequence of observables, assumed to be generated from the parametric sampling distribution $F(y|\theta)$ with pdf/pmf $f(y|\theta)$ and labeled by an unknown parameter $\theta \in \Theta$ follwoing a prior distribution $\Pi(\theta)$ with pdf/pmf $\pi(\theta)$.

The AIM, in parametric (or predictive) point estimation, is to derive a parametric (or predictive) point estimator $\hat{\delta}(y)$, a quantity summarizing Your believes about the unknown parameter θ (or unknown future outcome sequence $z := (y_{n+1}, ..., y_{n+m})$) or any function of it in an appropriate manner.

It is addressed in the statistical decision theory framework, where the decision rule is the estimator $\hat{\delta}(y)$, (the decision space results consequently) and the loss function $\ell(\cdot, \cdot)$ is set in a subjective manner as a penalty and according to what You may loss or how You may suffer if you use the produced estimator.

Note 2. We define the parametric and predictive point estimation as two distinct concepts, however as it will become clear they are treated in a similar manner.

2 Parametric point estimation[1]

Note 3. Posterior degree of believe about uncertain parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ is quantified via the posterior distribution

$$\mathrm{d}\Pi(\theta|y) = \pi(\theta|y)\mathrm{d}\theta$$

with cdf $\Pi(\theta|y)$ and pdf/pmf $\pi(\theta|y)$.

Definition 4. Bayes point estimator $\hat{\delta} = \hat{\delta}(y)$ of θ under the loss function $\ell(\theta, \delta)$ and the posterior distribution $\mathrm{d}\Pi(\theta|y)$ is an Bayes rule (which minimizes the posterior expected loss $\varrho(\pi, d|y) = \mathrm{E}_\Pi(\ell(\theta, \delta)|y)$); i.e.

$$\hat{\delta} = \arg\min_{\forall \delta \in \mathcal{D}} \mathrm{E}_\Pi(\ell(\theta, \delta)|y) = \arg\min_{\forall \delta \in \mathcal{D}} \underbrace{\int_\Theta \ell(\theta, \delta)\mathrm{d}\Pi(\theta|y)}_{=\varrho(\pi,\delta|y)}$$

Note 5. Traditionally the accuracy of a statistical estimator (called standard error) is described by the squared mean square error of the estimator.

Definition 6. Let $\hat{\delta} = (\hat{\delta}_1, ..., \hat{\delta}_d)$ be the Bayes point estimator of $\theta \in \Theta \subseteq \mathbb{R}^d$ with posterior distribution $\Pi(\theta|y)$. The standard error of the j-th dimension of $\hat{\delta}$ is defined as

$$\mathrm{se}_\Pi\left(\hat{\delta}_j|y\right) = \sqrt{\left[\mathrm{mse}_\Pi(\hat{\delta}|y)\right]_{j,j}}$$

where

$$\mathrm{mse}_\Pi\left(\hat{\delta}|y\right) = \mathrm{E}_\Pi\left((\theta - \hat{\delta})(\theta - \hat{\delta})^\top|y\right)$$

is the mean squared error of $\hat{\delta}$.

Remark 7. MSE of (any) estimator $\hat{\delta}$ of θ following posterior distribution $\Pi(\theta|y)$ can be decomposed as

$$\mathrm{E}_\Pi\left((\theta - \hat{\delta})(\theta - \hat{\delta})^\top|y\right) = \mathrm{Var}_\Pi(\theta|y) + \left(\mathrm{E}_\Pi(\theta|y) - \hat{\delta}\right)\left(\mathrm{E}_\Pi(\theta|y) - \hat{\delta}\right)^\top$$

Remark 8. By Definition **??**, the mse of 1-dim Bayes point estimator δ of θ with posterior distribution $\Pi(\theta|y)$ is

$$\mathrm{se}_\Pi\left(\hat{\delta}|y\right) = \sqrt{\mathrm{mse}_\Pi(\delta|y)}$$

where

$$\mathrm{mse}_\Pi\left(\hat{\delta}|y\right) = \mathrm{E}_\Pi\left((\theta - \hat{\delta})^2|y\right) = \mathrm{Var}_\Pi(\theta|y) + \left(\mathrm{E}_\Pi(\theta|y) - \hat{\delta}\right)^2$$

3 Predictive point estimation

Note 9. The Bayesian point predictive estimator and its standard error are defined similar to the parametric ones.

Note 10. Degree of believe about a future sequence of outcomes $z = (y_{n+1}, ..., y_{n+m}) \in \mathcal{Z}$ is quantified via the predictive distribution

$$\mathrm{d}G(z|y) = g(z|y)\mathrm{d}z$$

with cdf $G(z|y)$ and pdf/pmf $g(z|y)$.

Definition 11. Bayes predictive point estimator of $z = (y_{n+1}, ..., y_{n+m}) \in \mathcal{Z}$ under the loss function $\ell(z, \delta)$ and predictive distribution $G(z|y)$ is the decision rule $\delta \in \mathcal{D} = \mathcal{Z}$ which minimizes $\mathrm{E}_G(\ell(z, \delta)|y)$; i.e.

$$\hat{\delta} = \arg\min_{\forall \delta} \mathrm{E}_G(\ell(z, \delta)|y) = \arg\min_{\forall \delta \in \mathcal{D}} \int_\mathcal{Z} \ell(y, \delta)\mathrm{d}G(z|y)$$

[1] Web applet: `https://georgios-stats-1.shinyapps.io/demo_pointestimation/`

Note 12. The accuracy of the predictive point estimator is traditionally presented by using the squared mean (predictive) square error.

Definition 13. Let $\hat{\delta} = (\hat{\delta}_1, ..., \hat{\delta}_m)$ be the Bayes point predictive estimator of $z \in \mathcal{Z} \subseteq \mathbb{R}^m$ with predictive distribution $G(z|y)$. Then the standard error of the j-th dimension of $\hat{\delta}$ is defined as

$$\mathrm{se}_G(\hat{\delta}_j|y) = \sqrt{\left[\mathrm{mse}_\Pi(\hat{\delta}|y)\right]_{j,j}}$$

where

$$\mathrm{mse}_G(\hat{\delta}|y) = \mathrm{E}_G\left((z - \hat{\delta})(z - \hat{\delta})^\top |y\right)$$

is the mean squared error of $\hat{\delta}$.

4 Popular point estimators[2]

Note 14. A number of standard Bayesian point estimators, along with the corresponding loss functions are examined below. These point estimators correspond to summary statistics of the posterior/predictive distribution (mean, median, mode, quantiles, etc.).

Note 15. Bayesian point estimators are applied to both parametric and predictive inference likewise and hence presented together.

Notation 16. Consider unknown random quantity $x \in \mathcal{X} \subseteq \mathbb{R}^k$ following a distribution

$$\mathrm{d}Q(x|y) = q(x|y)\mathrm{d}x$$

with cdf $Q(x|y)$ and pdf/pmf $q(x|y)$. These are dummies for the following:

- In parametric inference, we have $x \equiv \theta$, $Q \equiv \Pi$, $q \equiv \pi$, and $k = d$.
- In predictive inference, we have $x \equiv z$, $Q \equiv G$, $q \equiv g$, and $k = m$.
- In more extreme cases, we have $x \equiv (z, \theta)$, $Q \equiv P$, $q \equiv p$, and $k = d + m$.
- Note that x can also be any function of θ or z, or (z, θ)...

Quadratic loss function

Proposition 17. *The Bayes point estimate $\hat{\delta}$ of x with respect to the quadratic loss function $\ell(x, \delta) = (x-\delta)^\top H(x-\delta)$, where $H > 0$, is*

$$\hat{\delta} = E_Q(x|y) \tag{1}$$

Proof. It is

$$\begin{aligned} 0 = \frac{\mathrm{d}}{\mathrm{d}\delta}\int \ell(x,\delta)\mathrm{d}Q(x|y)\Big|_{\delta=\hat{\delta}} &= \frac{\mathrm{d}}{\mathrm{d}\delta}\int (x-\delta)H(x-\delta)^\top \mathrm{d}Q(x|y)\Big|_{\delta=\hat{\delta}} \\ = -2H\int (x-\hat{\delta})\mathrm{d}Q(x|y) \quad &= -2H\int \theta \mathrm{d}Q(x|y) + 2H\hat{\delta} \quad = -2H\mathrm{E}_Q(x|y) + 2H\hat{\delta}. \end{aligned}$$

□

Remark 18. If $k = 1$, then $\ell(x, \delta) = H(x - \delta)^2$, with $H > 0$, and the Bayes point estimate is $\hat{\delta} = \mathrm{E}_Q(x|y)$.

Remark 19. The loss in Proposition 17 has the same effect as $\ell(x, \delta) = \|x - \delta\|_2^2$ and hence H has no effect.

[2] Web applet: `https://georgios-stats-1.shinyapps.io/demo_pointestimation/`

Remark 20. The point estimator in Proposition 17 minimizes the standard error, as

$$\mathrm{se}(\hat{\delta}|y)=\sqrt{\mathrm{mse}_Q(\hat{\delta}|y)}=\sqrt{\mathrm{Var}_Q(x|y)+\cancelto{0}{\left(\mathrm{E}_Q(x|y)-\hat{\delta}\right)^2}}=\sqrt{\mathrm{Var}_Q(x|y)}$$

Weighted quadratic loss function

Proposition 21. *The Bayes estimate $\hat{\delta}$ of x under the weighted quadratic loss function $\ell(x,\delta)=w(x)(x-\delta)^2$, where $w(x)$ as a non negative function with $E_Q(w(x)|y)>0$, is*

$$\delta^{\pi}(y)=\frac{E_Q(w(x)x|y)}{E_Q(w(x)|y)}. \tag{2}$$

Proof. It is

$$\begin{aligned}
0&=\frac{\mathrm{d}}{\mathrm{d}\delta}\int_{\mathcal{X}}\ell(x,\delta)\mathrm{d}Q(x|y)\bigg|_{\delta=\hat{\delta}} &&=\frac{\mathrm{d}}{\mathrm{d}\delta}\int_{\mathcal{X}}w(x)(x-\hat{\delta})^2\mathrm{d}Q(x|y)\bigg|_{\delta=\hat{\delta}}\\
&=2\int_{\mathcal{X}}w(x)(x-\hat{\delta})(-1)\mathrm{d}Q(x|y) &&=-2\left[\int_{\mathcal{X}}w(x)x\mathrm{d}Q(x|y)\right]+2\left[\int_{\mathcal{X}}w(\theta)\mathrm{d}Q(x|y)\right]\hat{\delta}.\\
&=-2\mathrm{E}_Q(w(x)x|y)+2\mathrm{E}_Q(w(x)|y)\hat{\delta} &&=2(\mathrm{E}_Q(w(x)x|y)-\mathrm{E}_Q(w(x)|y)\hat{\delta})
\end{aligned}$$

Also, $\frac{\mathrm{d}^2}{\mathrm{d}\delta^2}\int_{\mathcal{X}}\ell(x,\hat{\delta})\mathrm{d}Q(x|y)=-2\mathrm{E}_Q(w(x)|y)<0$. This completes the proof. □

Remark 22. Weighted quadratic loss allows the discrepancy $(x-\delta)^2$ to vary with x. It is appropriate in cases where a given discrepancy in estimation can vary in harm according to what x happens to be.

Example 23. Consider there is interest in performing parametric inference for θ. Show that Proposition 21 exhibits a duality between loss and prior distribution, in the sense that it is equivalent to estimate θ under loss $\ell(\theta,\delta)=w(\theta)(\theta-\delta)^2$ with prior pdf $\pi(\theta)$ (under (2)), or under loss $\tilde{\ell}(\theta,\delta)=(\theta-\delta)^2$ with prior pdf $\tilde{\pi}(\theta)\propto\pi(\theta)$ (under 1).

Solution. The estimator of θ under

- loss $\ell(\theta,\delta)=w(\theta)(\theta-\delta)^2$ and prior pdf $\pi(\theta)$ is $\hat{\delta}(y)=\frac{\mathrm{E}_{\Pi}(w(\theta)\theta|y)}{\mathrm{E}_{\Pi}(w(\theta)|y)}$
- loss $\tilde{\ell}(\theta,\delta)=(\theta-\delta)^2$ the prior pdf $\tilde{\pi}(\theta)\propto\pi(\theta)$ is $\tilde{\delta}(y)=\mathrm{E}_{\tilde{\Pi}}(\theta|y)$

But

$$\begin{aligned}
\tilde{\delta}(y)=\mathrm{E}_{\tilde{\Pi}}(\theta|y)&=\int_{\Theta}\theta\frac{f(y|\theta)\tilde{\pi}(\theta)}{\int_{\Theta}f(y|\theta)\tilde{\pi}(\theta)\mathrm{d}\theta}\mathrm{d}\theta=\int_{\Theta}\theta\frac{f(y|\theta)w(\theta)\pi(\theta)}{\int_{\Theta}f(y|\theta)w(\theta)\pi(\theta)\mathrm{d}\theta}\mathrm{d}\theta=\frac{\int_{\Theta}\theta f(y|\theta)w(\theta)\pi(\theta)\mathrm{d}\theta}{\int_{\Theta}f(y|\theta)w(\theta)\pi(\theta)\mathrm{d}\theta}\\
&=\frac{\int_{\Theta}\theta w(\theta)\frac{f(y|\theta)\pi(\theta)}{\int_{\Theta}f(y|\theta)\pi(\theta)\mathrm{d}\theta}\mathrm{d}\theta}{\int_{\Theta}w(\theta)\frac{f(y|\theta)\pi(\theta)}{\int_{\Theta}f(y|\theta)\pi(\theta)\mathrm{d}\theta}\mathrm{d}\theta}=\frac{\int_{\Theta}\theta w(\theta)\pi(\theta|y)\mathrm{d}\theta}{\int_{\Theta}w(\theta)\pi(\theta|y)\mathrm{d}\theta}=\frac{\mathrm{E}_{\Pi}(w(\theta)\theta|y)}{\mathrm{E}_{\Pi}(w(\theta)|y)}=\hat{\delta}(y)
\end{aligned}$$

- This is an example supporting that loss and prior are difficult to separate and may be specified/analysed simultaneously.

Linear loss function

Proposition 24. *The Bayes estimate $\hat{\delta}$ of x under the linear loss function*

$$\ell(x,\delta)=c_1(\delta-x)1_{\{x\le\delta\}}(\delta)+c_2(x-\delta)1_{\{x\le\delta\}^{\complement}}(\delta)$$

is the $\frac{c_2}{c_1+c_2}$-th quantile of distribution Q, namely

$$\hat{\delta} \text{ is such that } \mathsf{P}_Q\left(x \le \hat{\delta}|y\right) = \frac{c_2}{c_1+c_2}.$$

Proof. It is

$$\begin{aligned}\int \ell(x,\delta)\mathrm{d}Q(x|y) &= \int c_1(\delta-\theta)1_{\{x\le\delta\}}(\delta)\mathrm{d}Q(x|y) + \int c_2(\theta-\delta)1_{\{x\le\delta\}^{\complement}}(\delta)\mathrm{d}Q(x|y)\\ &= c_1\int_{\{x\le\delta\}}(\delta-\theta)\mathrm{d}Q(x|y) + c_2\int_{\{x\le\delta\}^{\complement}}(\theta-\delta)\mathrm{d}Q(x|y)\end{aligned}$$

Then

$$\begin{aligned}&0 = \frac{\mathrm{d}}{\mathrm{d}\delta}\int \ell(x,\delta)\mathrm{d}Q(x|y)\Big|_{\delta=\hat{\delta}} = c_1\int_{\{x\le\hat{\delta}\}}\mathrm{d}Q(x|y) - c_2\int_{\{x\le\hat{\delta}\}^{\complement}}\mathrm{d}Q(x|y) = c_1\mathsf{P}_Q\left(\{x\le\hat{\delta}\}|y\right) - c_2\mathsf{P}_Q\left(\{x\le\hat{\delta}\}^{\complement}|y\right)\\ &\iff -c_2\mathsf{P}_Q\left(\{x\le\hat{\delta}\}|y\right) = c_1\mathsf{P}_Q\left(\{x\le\hat{\delta}\}|y\right) - c_2\mathsf{P}_Q\left(\{x\le\hat{\delta}\}^{\complement}|y\right) - c_2\mathsf{P}_Q\left(\{x\le\hat{\delta}\}|y\right)\\ &\iff \mathsf{P}_Q\left(x\le\hat{\delta}|y\right) = \frac{c_2}{c_1+c_2}.\end{aligned}$$

□

Note 25. Below, we introduce the absolute loss as a special case of the linear loss which leads to median summaries.

Proposition 26. *The Bayes estimate $\hat{\delta}$ of x under the absolute loss function $\ell(x,\delta) = \|x-\delta\|_1$ is the median of distribution $Q(x|y)$*

$$\hat{\delta} = \textit{median}_Q(x|y). \tag{3}$$

Proof. This is straightforward by setting $c_1 = c_2$ in Proposition 24, where we get $\mathsf{P}_Q\left(x\le\hat{\delta}|y\right) = \frac{c_2}{c_1+c_2} = 0.5$. □

Remark 27. The linear loss function in Proposition 24, allows the adjustment of the penalty between over-estimating and under-estimating x, by adjusting $c_1$and c_2. Hence, the absolute loss in Proposition 26, is more appropriate when over-estimation and under-estimation are of the same concern (as penalized the same).

Remark 28. Compared to the linear loss functions $\ell(x,\delta) = \|x-\delta\|_1$, the quadratic loss $\ell(x,\delta) = \|x-\delta\|_2^2$ aims at over-penalizing large but unlikely errors. The linear loss functions increase much slower than the quadratic loss, and hence, while remaining convex, they do not penalize so much the large but unlikely errors.

Zero-one loss functions

Proposition 29. *The Bayes estimate $\hat{\delta}$ of x under the zero-one loss function*

$$\ell(x,\delta) = 1 - 1_{B_\epsilon(\delta)}(x); \text{ where } B_\epsilon(\delta) = (x\in\mathcal{X} \mid \|x-\delta\|\le\epsilon)$$

is

$$\hat{\delta} = \arg\max_{\forall\delta}\mathsf{P}_Q\left(x\in B_\epsilon(\delta)|y\right). \tag{4}$$

Proof. It is

$$\int \ell(x,\delta)\mathrm{d}Q(x|y) = 1 - \int 1_{B_\epsilon(\delta)}(x)\mathrm{d}Q(x|y) \quad = 1 - \mathsf{P}_Q\left(x\in B_\epsilon(\delta)|y\right)$$

which is minimized where $\mathsf{P}_Q\left(x\in B_\epsilon(\delta)|y\right)$ is maximized. □

Proposition 30. *The Bayes estimate $\hat{\delta}$ of x with respect to the zero-one loss $\ell(x,\delta) = 1 - 1_{\delta}(x)$ is the mode of distribution Q, and it is called Maximum A posteriori (MAP) estimator, i.e.*

$$\hat{\delta} = mode_{Q}(x|y). \tag{5}$$

Proof. Straightforward from Proposition 29, by taking any small $\epsilon \to 0$. □

Note 31. Consider that parametric inference about θ is of interest. In the frequentist sense, MAP estimator can be seen as a penalized maximum likelihood estimator in the sense that it essentially maximizes

$$\log(\pi(\theta|y)) = \log f(y|\theta) + \log \pi(\theta) - [\text{normalising const.}] \ \propto \log f(y|\theta) + \log \pi(\theta) \tag{6}$$

which is the log-likelihood penalized by the log-prior (!!!)

Example: Consider the Bayesian model

$$\begin{cases} y|\mu & \sim \mathrm{N}(\mu, \sigma^2) \\ \mu & \sim \mathrm{N}(\mu_0, \sigma_0^2) \end{cases}$$

where $\sigma^2, \mu_0, \sigma_0^2$ are known. Then (6) becomes

$$\log(\pi(\mu|y)) \propto \log(\mathrm{N}(y|\mu, \sigma^2)) - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2$$

Here maximization of the posterior pdf is a compromise between the maximization of the log-likelihood and minimization of the distance $(\mu - \mu_0)^2$ regulated by σ_0^2. Hence the prior shrinks μ towards μ_0 and this shrinkage is adjusted via σ_0^2.

Example: Consider the Bayesian regression model

$$\begin{cases} y|\beta & \sim \mathrm{N}(\Phi\beta, I\sigma^2) \\ \beta & \sim \mathrm{N}(0, \mathrm{diag}(v_1, ..., v_d)) \end{cases}$$

where $\sigma^2, v_1, ..., v_d$ are known. Then (6) becomes

$$\log(\pi(\beta|y)) \propto \log(\mathrm{N}(y|\Phi\beta, I\sigma^2)) - \frac{1}{2}\sum_{j=1}^{d}\frac{1}{v_j}(\beta_j - 0)^2$$

Likewise, the prior shrinks all β_j towards 0 and this shrinkage is adjusted via v_j. This idea has applications to dimension reduction/variable selection in specific high-dimensional regression problems.

Remark 32. Zero-one loss imposes a quite forceful penalization; because the penalty is equal to 0 if δ is the correct answer, and 1 if it is wrong.

Example 33. Consider a Bayesian model

$$\begin{cases} y_i|\theta & \overset{\text{iid}}{\sim} \mathrm{Br}(\theta), \qquad i = 1, ..., n \\ \theta & \sim \mathrm{Be}(a, b) \end{cases}$$

where $a > 0$, $b > 0$, and $n > 2$. Find the Bayesian estimator of parameter θ for the zero-one loss function.

Hint: The posterior is $\theta|y \sim \mathrm{Be}(a + n\bar{y}, b + n - n\bar{y})$.

Hint: Consider PDF/PMF, for Bernoulli, and Beta distributions

$$f_{\mathrm{Br}(\theta)}(y) = \theta^y(1-\theta)^{1-y}1(y \in \{0, 1\})\,; \qquad \pi_{\mathrm{Be}(a,b)}(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}1(\theta \in [0, 1])$$

Solution. Given the absolute loss function, the point estimator is the Maximum Aposteriori Estimator (the posterior mode). Then

$$\log(\pi(\theta|y)) \propto (n\bar{y}+a-1)\log(\theta)+(n-n\bar{y}+b-1)\log(1-\theta)$$

So, for $a>0$, $b>0$

$$0=\frac{\mathrm{d}}{\mathrm{d}\theta}\log(\pi(\theta|y))\Big|_{\theta=\hat{\delta}}=\frac{n\bar{y}+a-1}{\theta}-\frac{n-n\bar{y}+b-1}{1-\theta}\Big|_{\theta=\hat{\delta}}\implies\hat{\delta}=\frac{n\bar{y}+a-1}{n+a+b-2}.$$

It is good to notice (although not asked by the exercise) that

- If $a\rightarrow 1$, $b\rightarrow 1$ (aka $\pi(\theta)\propto 1$), then , like the Frequentists.
- If $a\rightarrow 0$, $b\rightarrow 0$ (aka $\pi(\theta)\propto\theta^{-1}(1-\theta)^{-1}$), then $\hat{\delta}=\frac{n\bar{y}-1}{n-2}$.
- If $a\rightarrow 1/2$, $b\rightarrow 1/2$ (aka $\pi(\theta)\propto\theta^{-1/2}(1-\theta)^{-1/2}$), then $\hat{\delta}=\frac{n\bar{y}-1/2}{n-1}$.
- If $n\rightarrow\infty$, $a>0$, $b>0$, then $\hat{\delta}=\bar{y}$. Like the Frequentists.

Question 34. *Practice with Exercise 59 from the Exercise sheet.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 12: Credible sets[a]

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

Aim: To explain and produce credible regions in the Bayesian framework.

References:

- Berger, J. O. (2013; Section 4.3.2). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- Robert, C. (2007; Section 5.5). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- The Matrix Cookbook (Section 7) http://matrixcookbook.com

Web applets:

- https://georgios-stats-1.shinyapps.io/demo_CredibleSets/

[a] Author: Georgios P. Karagiannis.

1 Set-up and aim

Notation 1. Consider a Bayesian model

$$\begin{cases} y|\theta & \sim \mathrm{d}F(y|\theta) \\ \theta & \sim \mathrm{d}\Pi(\cdot) \end{cases}$$

where $y := (y_1, ..., y_n) \in \mathcal{Y}$ is a sequence of observables, assumed to be generated from the parametric sampling distribution $F(y|\theta)$ with pdf/pmf $f(y|\theta)$ and labeled by an unknown parameter $\theta \in \Theta$ with a prior distribution $\Pi(\theta)$ with pdf/pmf $\pi(\theta)$.

AIM: Instead of just reporting a point value for θ (or z) and the associated standard error, it is often desirable and clearer to report sets of values $C_a \subseteq \Theta$ (or $C_a \subseteq \mathcal{Z}$) with a specified probability a reflecting Your believe that $\theta \in C_a$ (or $z \in C_a$).

Note 2. Recall that

- Posterior degree of believe about uncertain parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ is quantified via the posterior distribution

  $$\mathrm{d}\Pi(\theta|y) = \pi(\theta|y)\mathrm{d}\theta$$

  with cdf $\Pi(\theta|y)$ and pdf/pmf $\pi(\theta|y)$.

- Degree of believe about a future sequence of outcomes $z = (y_{n+1}, ..., y_{n+m}) \in \mathcal{Z}$ is quantified via the predictive distribution

  $$\mathrm{d}G(z|y) = g(z|y)\mathrm{d}z$$

  with cdf $G(z|y)$ and pdf/pmf $g(z|y)$.

Notation 3. We present the parametric and predictive credible intervals in a unified framework. Consider unknown random quantity $x \in \mathcal{X} \subseteq \mathbb{R}^k$ following a distribution

$$\mathrm{d}Q(x|y) = q(x|y)\mathrm{d}x$$

with cdf $Q(x|y)$ and pdf/pmf $q(x|y)$. These are dummies for the following:

- In parametric inference, we have $x \equiv \theta$, $Q \equiv \Pi$, $q \equiv \pi$, and $k = d$.
- In predictive inference, we have $x \equiv z$, $Q \equiv G$, $q \equiv g$, and $k = m$.
- Note that x can also be any function of θ or z.

2 Credible intervals

Definition 4. A set $C_a \subseteq \mathcal{X}$ is called '$100(1-a)\%$' posterior credible set for x, with respect to the posterior distribution $Q(x|y)$ if

$$1 - a \leq \mathsf{P}_Q(x \in C_a|y) = \int 1\left(x \in C_a\right)\mathrm{d}Q(x|y)$$

Note 5. In Bayesian stats (unlike frequetist stats) we can speak correctly and meaningfully say that the $(1-a)100\%$ credible set C_a of unknown parameter θ implies that the probability that θ in in C_a is $(1-a)100\%$. This is theoretically correct as everything unknown/uncertain is a random quantity following a distribution reflecting Your degree of believe.

Note 6. Note that different sets may satisfy Definition 4 and hence we are interested in using the most useful credible set for our application. This is addressed by imposing additional restrictions.

3 Highest probability density Credible intervals[1]

Note 7. Often it is useful to consider credible sets C_a which contain values of x that correspond to the highest pdf/pmf $g(x|y)$ (aka the most likely values of x). Then Definition 4 can be restricted by essentially imposing an extra restriction that: $g(x|y) \geq g(x'|y)$ for all $x \in C_a, x' \in C_a^{\complement}$. This leads to the definition of the highest probability density (HPD) set.

Definition 8. The $100(1-a)\%$ highest probability density (HPD) set for $x \in \mathcal{X}$ with respect to the posterior distribution $Q(x|y)$ is the subset C_a of Θ of the form

$$C_a = \{x \in \mathcal{X} : g(x|y) \geq k_a\}$$

where k_a is the largest constant such that

$$1 - a \leq \mathsf{P}_Q(x \in C_a|y)$$

Note 9. Credible sets are considered as 'set estimators', and hence, they can be produced as Bayes decision rules under a specified loss function.

Note 10. HPD credible sets are credible sets with the minimum size (length, volume, area, etc...). In the decision theory framework, the HPD set is the Bayes estimator (Bayes rule) of the credible set under the loss

$$\ell(x,\delta) = c\,\|\delta\| - 1(x \in \delta), \quad \forall \delta \in \mathcal{D},\ \forall x \in \mathcal{X},\ \forall c > 0. \tag{1}$$

[1] Web applet: `https://georgios-stats-1.shinyapps.io/demo_CredibleSets/`

4 General discussions

Remark 11. HPD credible sets are not, in general, invariant to transformations. If one has computed the HPD set for $x \sim Q(x|y)$, the HPD set for $\varphi = g(x)$ does not necessarily result by converting HPD set for x. To compute the HPD set for φ, one has to compute the posterior distribution

$$\mathrm{d}Q(\varphi|y) = \underbrace{q(g^{-1}(\varphi)|y)|\frac{\mathrm{d}}{\mathrm{d}\varphi}g^{-1}(\varphi)|}_{=\pi(\varphi|y)}\mathrm{d}\varphi,$$

and then compute the HPD set by implementing Definition 8.

Note 12. [2]A (not-that-efficient) algorithm to compute HPD credible sets with a computer is as follows:

1. Create a routine which computes all solutions x^* to the equation $q(x|y) = k_a$, for a given k_a. Typically, $C_a = \{x \in \mathcal{X} : q(x|y) \geq k_a\}$ can be constructed from those solutions.

2. Create a routine which computes

$$\mathsf{P}_Q(x \in C_a|y) = \int 1\left(x \in C_a\right) \mathsf{d}Q(x|y) \tag{2}$$

3. Sequentially solve the equation

$$\mathsf{P}_Q(x \in C_a|y) = 1 - a$$

by increasing incrementally k_a from zero to larger, and stop just before (2) drops below $1 - a$.

Note 13. For the simple 1D case, $x \in \mathcal{X}$ with $\dim(\mathcal{X}) = 1$, the following theorem can be used to compute HPD credible sets.

Theorem 14. *Let $x \in \mathbb{R}$ be a continuous random variable following distribution $Q(x|y)$ with unimodal density $q(x|y)$. If the interval $C_a = [L, U]$ satisfies*

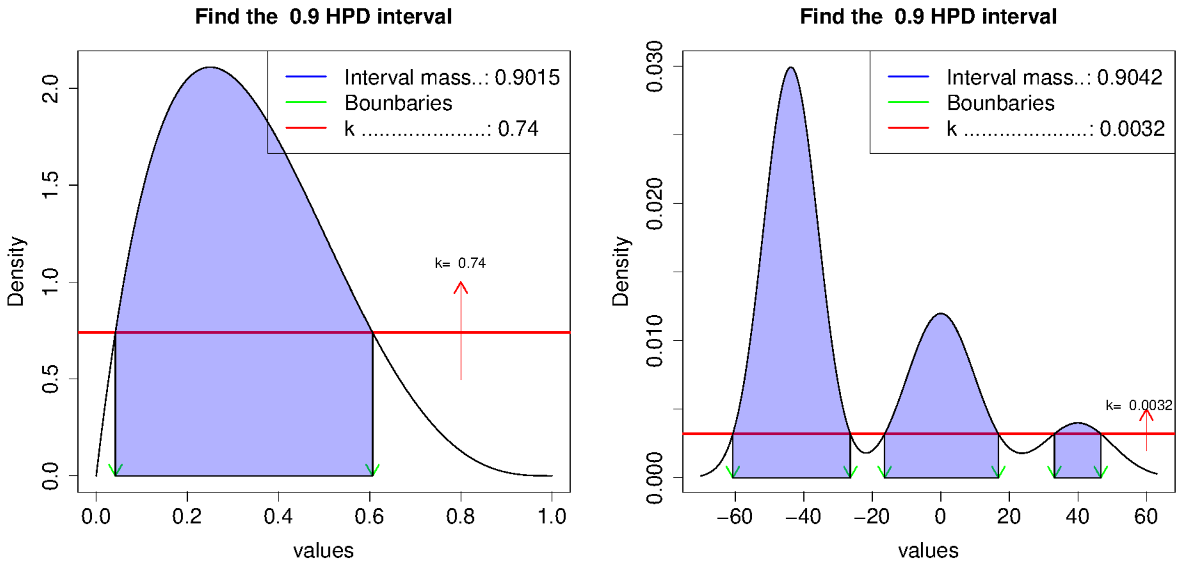
1. $\int_L^U q(x|y)dx = 1 - a$,
2. $q(U) = q(L) > 0$, *and*
3. $x_{mode} \in (L, U)$, *where* x_{mode} is the mode of $q(x|y)$,

then it is the HPD interval of x with respect to $Q(x|y)$.

Proof. Out of scope. Use of the mean values theorem to prove. See, Casella, G., & Berger, R. L. (2002; pp. 441-443). Statistical inference (Vol. 2). Pacific Grove, CA: Duxbury. □

Remark 15. Theorem 14 suggests a procedure to find the boundaries of C_a in 1D cases. As is Figure 1a, we can imagine a horizontal bar which moves from the maximum of the density to zero, and intersects the density at locations which are the potential boundaries of C_a. The limits of the credible set are where the density above the two points the intersection take place (shaded area) is equal to $1 - a$. This trick can also be used in multimodal densities (Figure 1b).

[2] https://georgios-stats-1.shinyapps.io/demo_CredibleSets/



(a) Unimodal case; $C_a = [L, U]$

(b) Multimodal ; $C_a = [L_1, U_1] \cup [L_2, U_2]$

Figure 1: Schematic of Theorem 14 (in Fig. 1(1a)) and Note 12 (in Fig. 1(1a) & Fig. 1(1b))

5 Examples

Example 16. Consider a Bayesian model

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

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

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

Hint-2: It is

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

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

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

Solution. I will use the Definition 8.

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

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

where

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

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

- Now let's implement Definition 8. So,

$$\begin{aligned}C_a &= \left\{\mu\in\mathbb{R}^d : \pi(\mu|y)\geq k_a\right\}\\ &= \left\{\mu\in\mathbb{R}^d : \mathrm{N}_q(\mu|\hat{\mu}_n,\hat{\Sigma}_n)\geq k_a\right\}\\ &= \left\{\mu\in\mathbb{R}^d : (\mu-\hat{\mu}_n)^\top\hat{\Sigma}_n^{-1}(\mu-\hat{\mu}_n)\leq\underbrace{-\log(2\pi\det(\hat{\Sigma}_n)))k_a}_{=\tilde{k}_a}\right\}\end{aligned} \tag{3}$$

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

$$\mathsf{P}_\Pi\left(\mu\in C_a|y\right)\geq 1-a \iff$$

$$\mathsf{P}_\Pi\left(\underbrace{(\mu-\hat{\mu}_n)^\top\hat{\Sigma}_n^{-1}(\mu-\hat{\mu}_n)}_{=\xi}\leq\tilde{k}_a\right)\geq 1-a \tag{4}$$

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

$$\xi = (\mu-\hat{\mu}_n)^\top\hat{\Sigma}_n^{-1}(\mu-\hat{\mu}_n)\sim\chi_d^2 \tag{5}$$

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

Hence Eq. 4, (due to Eqs. 3, 5) becomes

$$\mathsf{P}_{\chi_d^2}((\mu-\hat{\mu}_n)^\top\hat{\Sigma}_n^{-1}(\mu-\hat{\mu}_n)\leq\tilde{k}_a) = 1-a \tag{6}$$

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

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

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

Example 17. Consider an exchangeable sequence of observables $y := (y_1, ...y_n)\in\mathbb{R}^n$ from model

$$\begin{cases} y_i|\theta & \overset{\text{iid}}{\sim}\mathrm{Br}(\theta), \qquad i = 1,...,n\\ \theta & \sim\mathrm{Be}(a,b)\end{cases}$$

where $a = b = 2$, $n = 30$, and $\sum_{i=1}^{30} y_i = 15$. Find the 2-sides C_a parametric HPD credible interval for θ. Consider $a = 0.95$.

Solution.

- The posterior distribution of θ is $\text{Be}(a + n\bar{y}, b + n - n\bar{y})$, because

$$\pi(\theta|y) \propto \prod_{i=1}^{n} \text{Br}(y_i|\theta)\text{Be}(\theta|a,b) \propto \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{y_i}\theta^{a-1}(1-\theta)^{b-1} \propto \theta^{n\bar{y}+a-1}(1-\theta)^{n-n\bar{y}+b-1}$$

  After substituting the values of the fixed parameters, I get $\pi(\theta|y) = \text{Be}(\theta|a_n = 17, b_n = 17)$.

- To find the 2-sides C_a parametric HPD credible interval for θ, I use Theorem 14.

$$1 - a = \int_{L}^{U} \text{Be}(\theta|17,17)\text{d}\theta = \mathsf{P}_{\text{Be}(17,17)}(\theta < U) - \mathsf{P}_{\text{Be}(17,17)}(\theta < L)$$

  I note that the posterior is symmetric around 0.5 because $a_n = b_n$. Then,

$$1 - a = \mathsf{P}_{\text{Be}(17,17)}(\theta < U) - \big(1 - \mathsf{P}_{\text{Be}(17,17)}(\theta < U)\big) = 2\mathsf{P}_{\text{Be}(17,17)}(\theta < U) - 1$$

  so $\mathsf{P}_{\text{Be}(17,17)}(\theta < U) = 1 - a/2$ and $L = 1 - U$. For $a = 0.95$, the 95% posterior credible interval for θ is $[L, U] = [0.36, 0.64]$.

Note. Note that, if we follow the same procedure, the compute the 95% prior credible interval for θ is $[L, U] = [0.14, 0.85]$. As expected, the posterior 95 credible interval is narrower than the corresponding posterior one. (Try to check it in R).

```
> install.packages('HDInterval')
> library('HDInterval')
> hdi(qbeta, 0.95, shape1=17, shape2=17)
lower upper
0.3354445 0.6645555
```

Practice

Question 18. *To practice try to work on the Exercise 66 from the Exercise sheet.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 13: Hypothesis tests [a]

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

Aim: To explain, design, and use hypothesis tests in the Bayesian framework

References:

- Berger, J. O. (2013; Section 4.3.3). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- DeGroot, M. H. (2005, Sections 11.5-11.13). Optimal statistical decisions (Vol. 82). John Wiley & Sons
- Robert, C. (2007; Section 5.2(exclude 5.2.6)). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

[a] Author: Georgios P. Karagiannis.

1 Set-up of a hypothesis test

Aim: Let $y = (y_1, ..., y_n)$ generated from the real unknown data-generating process $y \sim \mathrm{d}R(y)$. Statistician approximates/parametrizes $\mathrm{d}R(\cdot)$ by a statistical model $\mathrm{d}F(y|\theta)$ with unknown $\theta \in \Theta$. Then you wish to find useful statements about θ: E.g. is there a smaller $\Theta_\star \subseteq \Theta$ where You can restrict the possible values of unknown θ?

Notation 1. Let $y = (y_1, ..., y_n)$ be a sequence of observables modeled to have been generated from the sampling distribution $\mathrm{d}F(y|\theta)$ labeled by an unknown parameter $\theta \in \Theta$ following a priori distribution $\mathrm{d}\Pi(\theta)$; namely

$$\begin{cases} y|\theta & \sim \mathrm{d}F(y|\theta) \\ \theta & \sim \mathrm{d}\Pi(\theta) \end{cases} \tag{1}$$

Assume there is interest to test/compare the hypotheses/statements

$$\mathrm{H}_0 : \theta \in \Theta_0; \text{ vs. } \mathrm{H}_1 : \theta \in \Theta_1 \tag{2}$$

where $\Theta = \Theta_0 \cup \Theta_1$, under the Bayesian model (1).

Note 2. The pair of hypotheses (2) partitions the overall prior $\mathrm{d}\Pi(\theta)$ (representing overall prior believes about θ) as

$$\mathrm{d}\Pi(\theta) = \pi_0 \times \mathrm{d}\Pi_0(\theta) + \pi_1 \times \mathrm{d}\Pi_1(\theta) \tag{3}$$

where π_0, and π_1 describe the prior probabilities of hypotheses H_0 and H_1

$$\pi_0 = \underbrace{\mathsf{P}_\Pi(\theta \in \Theta_0)}_{=\mathsf{P}_\Pi(\mathrm{H}_0)} = \int 1(\theta \in \Theta_0)\,\mathrm{d}\Pi(\theta), \qquad \pi_1 = \underbrace{\mathsf{P}_\Pi(\theta \in \Theta_1)}_{=\mathsf{P}_\Pi(\mathrm{H}_1)} = \int 1(\theta \in \Theta_1)\,\mathrm{d}\Pi(\theta),$$

respectively while $\mathrm{d}\Pi_0(\theta) := \mathrm{d}\Pi(\theta|\theta \in \Theta_0)$ and $\mathrm{d}\Pi_1(\theta) := \mathrm{d}\Pi(\theta|\theta \in \Theta_1)$ are prior distributions with pdf/pmf

$$\pi_0(\theta) := \underbrace{\pi(\theta|\theta \in \Theta_0)}_{=\pi(\theta|\mathrm{H}_0)} = \frac{\pi(\theta)1(\theta \in \Theta_0)}{\int 1(\theta \in \Theta_0)\,\mathrm{d}\Pi_0(\theta)}; \quad \text{and} \qquad \pi_1(\theta) := \underbrace{\pi(\theta|\theta \in \Theta_1)}_{=\pi(\theta|\mathrm{H}_1)} = \frac{\pi(\theta)1(\theta \in \Theta_1)}{\int 1(\theta \in \Theta_1)\,\mathrm{d}\Pi_1(\theta)},$$

describing how the prior mass of θ is spread out over the hypotheses H_0 and H_1 respectively. Then the Bayesian hypothesis test is can also be expressed as

$$\text{H}_0 : \begin{cases} y|\theta & \sim \mathrm{d}F(y|\theta) \\ \theta & \sim \mathrm{d}\Pi_0(\theta),\ \theta \in \Theta_0 \end{cases} \qquad \text{vs} \qquad \text{H}_1 : \begin{cases} y|\theta & \sim \mathrm{d}F(y|\theta) \\ \theta & \sim \mathrm{d}\Pi_1(\theta),\ \theta \in \Theta_1 \end{cases} \tag{4}$$

with prior $\pi_0 = \text{P}_\Pi\left(\theta \in \Theta_0\right)$ and $\pi_1 = \text{P}_\Pi\left(\theta \in \Theta_1\right)$.

Question 3. *Which Bayesian model (H_0 or H_1) describes 'better' the real data generating process?*

Note 4. In Bayesian framework, hypothesis testing is rather straightforward. All You need to do is to calculate the corresponding posterior probabilities $\text{P}_\Pi(\theta \in \Theta_0|y)$, and $\text{P}_\Pi(\theta \in \Theta_1|y)$, and decide between H_0 and H_1.

2 Decision theory prespective

Note 5. Bayes hypothesis test (4) can be addressed as a Bayesian statistical decision problem with decision space $\mathcal{D} = \{\text{accept H}_0,\ \text{accept H}_1\}$ or simpler $\mathcal{D} = \{0, 1\}$, and under Bayesian model (1). It can be seen as a parametric point inference about the indicator function

$$1_{\Theta_1}(\theta) = \begin{cases} 0 & ,\ \theta \in \Theta_0 \\ 1 & ,\ \theta \in \Theta_1 \end{cases} \tag{5}$$

under Bayesian model (1), prior (3), and a loss function $\ell(\theta, \delta)$, with $\theta \in \Theta$, $\delta \in \mathcal{D}$; E.g., the $0-1$ loss function.

Theorem 6. *The Bayes estimator of $1_{\Theta_1}(\theta)$ in (5), under the prior $d\Pi(\theta)$ in (3) and the $c_I - c_{II}$ loss function*

$$\ell(\theta, \delta) = \begin{cases} 0 & ,\ \textit{if}\ \theta \in \Theta_0,\ \delta = 0 \\ 0 & ,\ \textit{if}\ \theta \notin \Theta_0,\ \delta = 1 \\ c_{II} & ,\ \textit{if}\ \theta \notin \Theta_0,\ \delta = 0 \\ c_I & ,\ \textit{if}\ \theta \in \Theta_0,\ \delta = 1 \end{cases} \tag{6}$$

where $c_I > 0$ and $c_{II} > 0$ are specified by the researcher is

$$\delta(y) = \begin{cases} 0 & ,\ \text{P}_\Pi\left(\theta \in \Theta_0|y\right) > \frac{c_{II}}{c_{II}+c_I} \\ 1 & ,\ \textit{otherwise} \end{cases} \tag{7}$$

where $\{\Theta_0, \Theta_1\}$ constitute a partition for Θ, and $\text{P}_\Pi\left(\theta \in \Theta_0|y\right) = \int 1\left(\theta \in \Theta_0\right) d\Pi(\theta|y)$.

Proof. The posterior expected loss is[1]

$$\begin{aligned} \varrho(\pi, \delta|y) &= \text{E}_\Pi(\ell(\theta,\delta)|y) = \int \ell(\theta,\delta)\mathrm{d}\Pi(\theta|y) = \int_{\Theta_0} \ell(\theta,\delta)\mathrm{d}\Pi(\theta|y) + \int_{\Theta_1} \ell(\theta,\delta)\mathrm{d}\Pi(\theta|y) \\ &= \begin{cases} \cancelto{=0}{\int_{\Theta_0} 0\mathrm{d}\Pi(\theta|y)} + \int_{\Theta_1} c_{\text{II}}\mathrm{d}\Pi(\theta|y) & , \quad \text{if } \delta = 0 \\ \int_{\Theta_0} c_{\text{I}}\mathrm{d}\Pi(\theta|y) + \cancelto{=0}{\int_{\Theta_1} 0\mathrm{d}\Pi(\theta|y)} & , \quad \text{if } \delta = 1 \end{cases} = \begin{cases} c_{\text{II}} \int 1\left(\theta \in \Theta_1\right)\mathrm{d}\Pi(\theta|y) & , \quad \text{if } \delta = 0 \\ c_{\text{I}} \int 1\left(\theta \in \Theta_0\right)\mathrm{d}\Pi(\theta|y) & , \quad \text{if } \delta = 1 \end{cases} \\ &= c_{\text{II}}\text{P}_\Pi\left(\theta \notin \Theta_0|y\right) 1\left(\delta \in \{0\}\right) + c_{\text{I}}\text{P}_\Pi\left(\theta \in \Theta_0|y\right) \end{aligned}$$

[1] Notation: $\int_{\Theta_j} \mathrm{d}\Pi(\theta|y) = \int 1\left(\theta \in \Theta_j\right)\mathrm{d}\Pi(\theta|y)$

The Bayes rule (estimator) of (5) is $\delta(y)=0$ when

$$\varrho(\pi,\delta=0|y)<\varrho(\pi,\delta=1|y) \iff c_{\mathrm{II}}\mathrm{P}_{\Pi}\left(\theta\notin\Theta_0|y\right)<c_{\mathrm{I}}\mathrm{P}_{\Pi}\left(\theta\in\Theta_0|y\right) \iff \mathrm{P}_{\Pi}\left(\theta\in\Theta_0|y\right)>\frac{c_{\mathrm{II}}}{c_{\mathrm{II}}+c_{\mathrm{I}}}$$

The Bayes rule (estimator) of (5) is $\delta(y)=1$ when

$$\varrho(\pi,\delta=0|y)>\varrho(\pi,\delta=1|y) \iff c_{\mathrm{II}}\mathrm{P}_{\Pi}\left(\theta\notin\Theta_0|y\right)>c_{\mathrm{I}}\mathrm{P}_{\Pi}\left(\theta\in\Theta_0|y\right) \iff \mathrm{P}_{\Pi}\left(\theta\in\Theta_0|y\right)<\frac{c_{\mathrm{II}}}{c_{\mathrm{II}}+c_{\mathrm{I}}}$$

So $\varrho(\pi,\delta|y)$ is minimised for (7). □

3 Bayes factors perspective

Note 7. Hypothesis tests in Bayesian statistics can be addressed by using Bayes factors.

Definition 8. The Bayes factor $\mathrm{B}_{01}(y)$ is 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.

$$\mathrm{B}_{01}(y)=\frac{\mathrm{P}_{\Pi}\left(\theta\in\Theta_0|y\right)/\mathrm{P}_{\Pi}\left(\theta\in\Theta_0\right)}{\mathrm{P}_{\Pi}\left(\theta\in\Theta_1|y\right)/\mathrm{P}_{\Pi}\left(\theta\in\Theta_1\right)} \tag{8}$$

where

$$\mathrm{P}_{\Pi}\left(\theta\in\Theta_j\right)=\int 1\left(\theta\in\Theta_j\right)\mathrm{d}\Pi(\theta);\quad \text{and}\quad \mathrm{P}_{\Pi}\left(\theta\in\Theta_j|y\right)=\int 1\left(\theta\in\Theta_j\right)\mathrm{d}\Pi(\theta|y);\quad \text{for } j=0,1.$$

Proposition 9. *For Hypothesis pair (2) and Bayes model (1), where the prior is formed as in (3), the Bayes factor in (8) can be written as*

$$B_{01}(y)=\frac{\int_{\Theta_0}f(y|\theta)d\Pi_0(\theta)}{\int_{\Theta_1}f(y|\theta)d\Pi_1(\theta)}=\frac{f_0(y)}{f_1(y)}$$

where $f_j(y)=\int_{\Theta_j}f(y|\theta)d\Pi_j(\theta)$ is the conditional marginal likelihood (or prior predictive pdf/pmf) given H_j, for $j=0,1$.

Proof. It results by showing that for $j=0,1$, it is

$$\mathrm{P}_{\Pi}(\theta\in\Theta_j|y)=\int_{\Theta_j}\mathrm{d}\Pi(\theta|y)=\int_{\Theta_j}\frac{f(y|\theta)\mathrm{d}\Pi(\theta)}{\int_{\Theta}f(y|\theta)\mathrm{d}\Pi(\theta)}=\int_{\Theta_j}\frac{f(y|\theta)\left(\pi_0\times\mathrm{d}\Pi_0(\theta)+\pi_1\times\mathrm{d}\Pi_1(\theta)\right)}{\underbrace{\int_{\Theta}f(y|\theta)\mathrm{d}\Pi(\theta)}_{=f(y)}}$$

$$=\frac{\pi_0}{f(y)}\int_{\Theta_j}f(y|\theta)\mathrm{d}\Pi_0(\theta)+\frac{\pi_0}{f(y)}\int_{\Theta_j}\pi_1 f(y|\theta)\mathrm{d}\Pi_1(\theta)=\begin{cases}\frac{\pi_0}{f(y)}\int_{\Theta_0}f(y|\theta)\mathrm{d}\Pi_0(\theta) & \text{,if } j=0\\ \frac{\pi_1}{f(y)}\int_{\Theta_1}f(y|\theta)\mathrm{d}\Pi_1(\theta) & \text{,if } j=1\end{cases}$$

□

Remark 10. Obviously, $\mathrm{B}_{10}(y)=1/\mathrm{B}_{01}(y)$

Remark 11. Bayes factor $\mathrm{B}_{01}(y)$:

- ...is the 'odds in favour of H_0 against H_1 that are given by the data' y.
- ...evaluate the modification of the odds of Θ_0 against Θ_1 due to the observations y.
- ...is the ratio of the likelihoods, weighted by the conditional priors $\mathrm{d}\Pi_0(\theta)$ and $\mathrm{d}\Pi_1(\theta)$.

Proposition 12. *One can write*

$$\mathsf{P}_{\Pi}(\theta\in\Theta_0|y) \quad =\left[1+\frac{1-\mathsf{P}_{\Pi}(\theta\in\Theta_0)}{\mathsf{P}_{\Pi}(\theta\in\Theta_0)}B_{01}(y)^{-1}\right]^{-1} \quad =\left[1+\frac{\pi_1}{\pi_0}B_{01}(y)^{-1}\right]^{-1}$$

where $\pi_j=\mathsf{P}_{\Pi}\left(\theta\in\Theta_j\right)$, for $j=0,1$, by rearranging (8) (please check).

Criterion 13. *Consider a hypothesis test $H_0:\theta\in\Theta_0$ versus $H_1:\theta\in\Theta_1$ as described in (4) with loss function (6), and given a Bayesian model (1). The hypothesis H_0 is accepted when*

$$B_{01}(y)>\frac{c_{II}}{c_I}\frac{\pi_1}{\pi_0} \tag{9}$$

where $\pi_j=\mathsf{P}_{\Pi}\left(\theta\in\Theta_j\right)$, for $j=0,1$.

Proof. Straightforward result from Definition 8 and Theorem 6. □

Remark 14. Eq. 9 shows the duality between loss function and the prior distribution. Different combinations of priors and loss functions may lead to the same result. For instance, for $c'_{\text{II}}=c'_{\text{I}}=1$, $\pi'_0=\frac{c_{\text{I}}\pi_0}{c_{\text{I}}\pi_0+c_{\text{II}}\pi_1}$, and $\pi'_1=\frac{c_{\text{II}}\pi_1}{c_{\text{I}}\pi_0+c_{\text{II}}\pi_1}$, we get again (9) !!!

Criterion 15. *Jeffreys developed a scale to judge the strength of evidence in favor of H_0 or against H_0 brought by the data, outside a true decision-theoretic setting (aka; without the need to specify c_I and c_{II} in (9)).*

B_{01}	$\log_{10}(B_{01})$	**Strength of evidence**
$(1,+\infty)$	$(0,+\infty)$	**$\mathbf{H}_0$ is supported**
$(10^{-1/2},1)$	$(-1/2,0)$	**Evidence against $\mathbf{H}_0$: not worth more than a bare**
$(10^{-1},10^{-1/2})$	$(-1,-1/2)$	**Evidence against $\mathbf{H}_0$: substantial**
$(10^{-3/2},10^{-1})$	$(-3/2,-1)$	**Evidence against $\mathbf{H}_0$: strong**
$(10^{-2},10^{-3/2})$	$(-2,-3/2)$	**Evidence against $\mathbf{H}_0$: very strong**
$(0,10^{-2})$	$(-\infty,-2)$	**Evidence against $\mathbf{H}_0$: decisive**

The precise bounds separating one strength from another are a matter of convention. Note that similar criticism exists in frequentist hypothesis tests with the choice of the significance level $a=\{0.01,0.05,0.1,...\}$.

4 Special cases in hypotheses tests

Definition 16. Traditionally, hypotheses, H_j, are categorized as:

- **Single (or point) hypothesis** for θ is called the hypothesis $\text{H}_j:\theta\in\Theta_j$ where $\Theta_j=\{\theta_j\}$ contains a single element, namely when $\Pi_j(\theta)$ assigns probability one to a specific value for θ.
- **Composite hypothesis** for θ is called the hypothesis $\text{H}_j:\theta\in\Theta_j$ where $\Theta_j\subseteq\Theta$ contains many elements. Namely when $\Pi_j(\theta)$ defines a non-degenerate density $\pi_j(\theta)$ over $\Theta_j\subseteq\Theta$.
- **General alternative hypothesis** for θ is called the composite hypothesis $\text{H}_1:\theta\in\Theta_1$ where $\Theta_1=\Theta-\{\theta_0\}$ and θ_0 a single value. It is often denoted as $\text{H}_1:\theta\neq\theta_0$ and compared against a single null hypothesis $\text{H}_0:\theta=\theta_0$.

We present some special cases of hypothesis tests.

Case 1. **Composite vs Composite** is the hypothesis test:

$$\text{H}_0:\theta\in\Theta_0 \qquad \text{vs} \qquad \text{H}_1:\theta\in\Theta_1$$

where both $\Theta_0 \subseteq \Theta$ and $\Theta_1 \subseteq \Theta$ contain more than one elements. Overall prior can be partitioned as

$$\mathrm{d}\Pi(\theta) = \pi_0 \times \mathrm{d}\Pi_0(\theta) + \pi_1 \times \mathrm{d}\Pi_1(\theta)$$

Then the conditional marginal likelihoods are

$$f_0(y) = \int_{\Theta_0} f(y|\theta)\mathrm{d}\Pi_0(\theta); \qquad\qquad f_1(y) = \int_{\Theta_1} f(y|\theta)\mathrm{d}\Pi_1(\theta).$$

Example. A composite vs. composite hypothesis is:

$$\mathrm{H}_0 : \begin{cases} y_i|\mu,\sigma^2 & \overset{\text{IID}}{\sim} \mathrm{N}(\mu,\sigma^2),\ i=1,...,n \\ \mu|\sigma^2 \sim; & \sim \mathrm{N}(\mu_0, \sigma^2\frac{1}{\lambda_0}) \\ \sigma^2 & \sim \mathrm{Ga}(a_0,b_0) \end{cases} \qquad \text{vs} \qquad \mathrm{H}_1 : \begin{cases} y_i|\mu & \overset{\text{IID}}{\sim} \mathrm{T}(\mu,1,k_0),\ i=1,...,n \\ \mu & \sim \mathrm{N}(\xi_0,v_0) \\ & \end{cases}$$

In H_0: I consider a sampling model $y_i \overset{\text{IID}}{\sim} \mathrm{N}(\mu,\sigma^2)$ with prior $(\mu,\sigma^2) \sim \mathrm{N}(\mu_0,\sigma^2/\lambda_0)\mathrm{IG}(a_0,b_0)$, and $\Theta_0 = \{\mathrm{N}\} \cup \mathbb{R} \cup (0,\infty)$. Here $\mu_0, \lambda_0, a_0, b_0$ are fixed.

In H_1: I consider a sampling model $y_i \overset{\text{IID}}{\sim} \mathrm{T}(\mu,1,k_0)$ with prior $(\mu,\sigma^2) \sim \mathrm{N}(\mu_0,\sigma^2/\lambda_0)\mathrm{IG}(a_0,b_0)$, and $\Theta_1 = \{\mathrm{T}\} \cup \mathbb{R}$. Here μ_0, k_0, ξ_0, v_0 are fixed.

Case 2. **Single vs. General alternative** is the pair of hypotheses

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

If θ is continuous, the difficulty is that we cannot use a continuous prior for $\mathrm{d}\Pi_0(\theta)$ to conduct a test with point null hypothesis $\mathrm{H}_0 : \theta = \theta_0$ because it would give a prior probability zero for $\theta = \theta_0$. To overcome this, we specify the conditional distribution $\mathrm{d}\Pi_0(\theta)$ as a Dirac prior distribution with concentration point at θ_0 ; namely $\mathrm{d}\Pi_0(\theta) = 1(\theta \in \{\theta_0\})\,\mathrm{d}\theta$. The conditional distribution $\mathrm{d}\Pi_1(\theta)$ can be any reasonable distribution $\mathrm{d}\Pi_1(\theta) = \mathrm{d}\Pi_1(\theta|\theta \in \Theta_1)$. Then the overall prior is

$$\mathrm{d}\Pi(\theta) = \pi_0 \times 1(\theta \in \{\theta_0\})\,\mathrm{d}\theta + \pi_1 \times \mathrm{d}\Pi_1(\theta|\theta \in \Theta_1) \tag{10}$$

and it is called spike-and-slab. Then the conditional marginal likelihoods are

$$\begin{aligned} f_0(y) &= \int_{\Theta_0} f(y|\theta)\mathrm{d}\Pi_0(\theta) = \int_{\{\theta=\theta_0\}} f(y|\theta)1(\theta \in \{\theta_0\})\,\mathrm{d}\theta = f(y|\theta_0) \\ f_1(y) &= \int_{\Theta_1} f(y|\theta)\mathrm{d}\Pi_1(\theta) = \int_{\{\theta\neq\theta_0\}} f(y|\theta)\mathrm{d}\Pi_1(\theta) \end{aligned}$$

Example. The standard two side test $\mathrm{H}_0 : \mu = \theta_0$ vs. $\mathrm{H}_1 : \mu \neq \theta_0$, where the sampling distribution is assumed to be$y_i \overset{\text{IID}}{\sim} \mathrm{N}(\mu,\sigma^2)$ with known variance σ^2 for $i = 1,...,n$, is a simple vs. general alternative hypothesis test and can also be formulated as:

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

Here it is $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta \in \mathbb{R} : \theta \neq \theta_0\}$, while $\theta_0, \mu_0, \sigma_0^2$ are fixed values.

Case 3. **Single vs. Single** is the pair of hypothesis

$$\mathrm{H}_0 : \theta = \theta_0 \qquad \text{vs} \qquad \mathrm{H}_1 : \theta = \theta_1$$

where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ for some values of θ_0 and θ_1. The hypotheses H_0 and H_1 are single, and hence the corresponding priors can be considered as having a point mass around θ_0 and θ_1. Mathematically, we assign Dirac prior distributions $\mathrm{d}\Pi_0(\theta) = 1\left(\theta \in \{\theta_0\}\right)\mathrm{d}\theta$ and $\mathrm{d}\Pi_1(\theta) = 1(\theta \in \{\theta_1\})\mathrm{d}\theta$, which imply

$$\mathrm{d}\Pi(\theta) = \Bigg(\quad \pi_0 \times 1\left(\theta \in \{\theta_0\}\right) + \pi_1 \times 1(\theta \in \{\theta_1\})\Bigg)\,\mathrm{d}\theta$$

Then the conditional marginal likelihoods are

$$f_0(y) = \int_{\Theta_0} f(y|\theta)\mathrm{d}\Pi_0(\theta) = \int_{\{\theta=\theta_0\}} f(y|\theta)1\left(\theta \in \{\theta_0\}\right)\mathrm{d}\theta = f(y|\theta_0)$$
$$f_1(y) = \int_{\Theta_1} f(y|\theta)\mathrm{d}\Pi_1(\theta) = \int_{\{\theta=\theta_1\}} f(y|\theta)1\left(\theta \in \{\theta_1\}\right)\mathrm{d}\theta = f(y|\theta_1)$$

Note 17. In this case, Bayes factor is the likelihood ratio of H_0 against H_1 which most statisticians (whether Bayesian or not) view as the odds in favor of H_0 against H_1 that are given by the data.

Example. Given the statistical model $y_i \overset{\text{IID}}{\sim} \mathrm{N}(\mu, \sigma^2)$ the comparison $\mathrm{H}_0 : \mu = \theta_0$ vs. $\mathrm{H}_1 : \mu = \theta_1$, is a simple vs. simple hypothesis, where $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$ are sets with a single elements $\theta_0 \neq \theta_1$.

Example. The model comparison

$$\mathrm{H}_0 : y_i|\phi_0 \overset{\text{IID}}{\sim} \mathrm{Nb}(\phi_0, 1) \text{ vs.} \quad \mathrm{H}_1 : y_i|\lambda_0 \overset{\text{IID}}{\sim} \mathrm{Pn}(\lambda_0)$$

where $\phi_0 > 0, \lambda_0 > 0$ are known, is a simple vs. simple hypothesis. Here it is $\Theta_0 = \{\mathrm{Nb}\}$, $\Theta_1 = \{\mathrm{Pn}\}$.

Example 18. Let $y = (y_1, ..., y_n)$ a sequence of observables, and assume that $n = 5$, and $y_* = \sum_{i=1}^{5} y_i = 3$. Assume a sampling distribution $y_i|\theta \overset{\text{iid}}{\sim} \mathrm{Br}(\theta)$, with unknown parameter $\theta \in [0, 1]$, a priori following a uniform distribution.

1. By using Jeffreys' scaling rule, perform the following hypothesis test for $\theta_0 = 1/2$

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

2. Compute the posterior probability of the NULL hypothesis.

Solution. This is a simple vs. general alternative hypothesis. I specify the overall prior with pdf

$$\pi(\theta) = \pi_0 1(\theta = \theta_0) + (1 - \pi_0)\mathrm{U}(\theta|0, 1)$$

for some $\pi_0 > 0$. I leave π_0 abstract, however the usual choice (but maybe not the best) is $\pi_0 = 1/2$.

1. The Bayes factor is

$$\mathrm{B}_{01}(y) = \frac{\prod_{i=1}^{n} \mathrm{Br}(y_i|\theta_0)}{\int_{(0,1)} \prod_{i=1}^{n} \mathrm{Br}(y_i|\theta)\mathrm{U}(\theta|0, 1)\mathrm{d}\theta} = \frac{\theta_0^{y_*}(1 - \theta_0)^{n-y_*}}{\int_{(0,1)} \theta^{y_*}(1 - \theta)^{n-y_*}\mathrm{d}\theta} = \frac{\theta_0^{y_*}(1 - \theta_0)^{n-y_*}}{\mathrm{B}(y_* + 1, n - y_* + 1)} = \frac{(1/2)^5}{\mathrm{B}(4, 3)} = \frac{15}{8}$$

Then $B_{01}(y) = \frac{15}{8} \approx 2$, and $\log_{10}(\mathrm{B}_{01}(y)) \approx 0.27$. According to Jeffreys' scaling rule, H_0 is supported. We can accept the null hypothesis.

2. The posterior probability of H_0 is

$$\mathrm{P}_{\Pi}(\theta = \theta_0|y) \ = \mathrm{P}_{\Pi}(\theta \in \Theta_0|y) \ = [1 + \frac{1 - \pi_0}{\pi_0}\mathrm{B}_{01}(y)^{-1}]^{-1} = \left[1 + \frac{1/2}{1 - 1/2}(\frac{15}{8})^{-1}\right]^{-1} = \frac{15}{23} \approx 0.65$$

and hence the posterior distribution tends to support H_0.

Example 19. Let $y = (y_1, ..., y_n)$ a sequence of observables. There is interest in performing the following hypothesis test

$$\mathrm{H}_0 : \begin{cases} y_i|\phi \sim \mathrm{Nb}(1,\phi); & \phi > 0 \\ \phi \sim \mathrm{Be}(a_0, b_0); & a_0 = 2,\ b_0 = 2 \end{cases} \qquad \text{vs} \qquad \mathrm{H}_1 : \begin{cases} y_i|\lambda \sim \mathrm{Pn}(\lambda); & \lambda > 0 \\ \lambda \sim \mathrm{Ga}(a_1, b_1); & a_1 = 2,\ b_1 = 1 \end{cases}$$

1. Perform the test for $n = 2$, and $y_1 = y_2 = 0$, by using Jeffrey's scaling.

2. Perform the test for $n = 2$, and $y_1 = y_2 = 2$, by using Jeffrey's scaling.

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 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 $\mathbb{N} = \{0, 1, 2, ...\}$.

Hint-3 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-4 Beta distribution $x \sim \mathrm{Be}(a, b)$ has PDF: $\mathrm{Be}(x|a, b) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}1_{(0,1)}(x)$, with $a > 0$ and $b > 0$.

Solution. This is a Composite vs composite hypotheses. The overall a priori distribution $\mathrm{d}\Pi(\theta)$ with $\theta \in \Theta$ and $\Theta = \{\mathrm{Nb}\} \times (0, 1) \cup \{\mathrm{Pn}\} \times (0, \infty)$ has density

$$\pi(\theta) = \pi_0 \mathrm{Be}(\phi|a_0, b_0) + \pi_1 \mathrm{Ga}(\lambda|a_1, b_1);$$

where $\pi_0 = \pi_1 = 0.5$. Let's compute the Bayes factor

$$f_0(y) = \int \prod_{i=1}^{n} \mathrm{Nb}(y_i|\phi, 1)\mathrm{Be}(\phi|a_0, b_0)\mathrm{d}\phi = \frac{1}{B(a_0, b_0)}\int_0^1 \phi^{n+a_0-1}(1-\phi)^{n\bar{y}+b_0-1}\mathrm{d}\phi = \frac{B(n + a_0, n\bar{y} + b_0)}{B(a_0, b_0)}$$

$$f_1(y) = \int \prod_{i=1}^{n} \mathrm{Pn}(y_i|\lambda)\mathrm{Ga}(\lambda|a_1, b_1)\mathrm{d}\lambda = \frac{1}{\prod_{i=1}^{n} y_i!}\frac{b_1^{a_1}}{\Gamma(a_1)}\int_0^\infty \lambda^{n\bar{y}+a_1-1}\exp(-(n + b_1)\lambda)\mathrm{d}\lambda$$
$$= \frac{\Gamma(n\bar{y} + a_1)}{\Gamma(a_1)}\frac{b_1^{a_1}}{(n + b_1)^{n\bar{y}+a_1}}\frac{1}{\prod_{i=1}^{n} y_i!}$$

So the Bayes Factor is

$$B_{01}(y) = \frac{B(n + a_0, n\bar{y} + b_0)}{B(a_0, b_0)}\frac{\Gamma(a_1)}{\Gamma(n\bar{y} + a_1)}\frac{(n + b_1)^{n\bar{y}+a_1}}{b_1^{a_1}}\prod_{i=1}^{n} y_i!$$

1. Then $B_{01}(y) = 2.70$, and $\log_{10}(\mathrm{B}_{01}(y)) \approx 0.43$. According to Jeffrey's scaling rule, H_0 is supported.

2. Then $B_{01}(y) = 0.29$, and $\log_{10}(\mathrm{B}_{01}(y)) \approx -0.53$. According to Jeffrey's scaling rule, the evidence against H_0 is substantial.

Example 20. Let $y = (y_1, ..., y_n)$ a sequence of observables. There is interest in performing the following hypothesis test

$$\mathrm{H}_0 : y_i|\phi \sim \mathrm{Nb}(\phi, 1);\ \text{with}\ \phi = 1/3 \qquad \text{vs} \qquad \mathrm{H}_1 : y_i|\lambda \sim \mathrm{Pn}(\lambda);\ \text{with}\ \lambda = 2$$

1. Perform the test for $n = 2$, and $y_1 = y_2 = 0$, by using Jeffreys' scaling.

2. Perform the test for $n = 2$, and $y_1 = y_2 = 2$, by using Jeffreys' scaling.

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 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 $\mathbb{N} = \{0, 1, 2, ...\}$.

Solution. This is a simple vs simple hypothesis test. I specify priors $\pi(\mathrm{Nb}) = \pi(\mathrm{Pn}) = 1/2$, due to the a priori ignorance about the parametric statistical model, however, we do not really need it now .... The Bayes factor is

$$B_{01}(y) = \frac{f_0(y)}{f_1(y)} = \frac{\prod_{i=1}^n \mathrm{Nb}(y_i|\phi, 1)}{\prod_{i=1}^n \mathrm{Pn}(y_i|\lambda)} = \frac{\phi^n(1-\phi)^{n\bar{y}}}{\lambda^{n\bar{y}}\exp(-n\lambda)/\prod_{i=1}^n y_i!}$$

1. Then $B_{01}(y) = \exp(4)/9 \approx 6.07$, and $\log_{10}(\mathrm{B}_{01}(y)) \approx 0.78$. According to Jeffrey's scaling rule, H_0 is supported.

2. Then $B_{01}(y) = 4\exp(4)/729 \approx 0.30$, and $\log_{10}(\mathrm{B}_{01}(y)) \approx -0.54$. According to Jeffrey's scaling rule, the evidence against H_0 is substantial.

Practice

Question 21. *To practice try to work on the Exercise 67 from the Exercise sheet.*

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 14: Jeffreys-Lindley ‘Paradox’ (?) [a]

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

Aim: To explain, and theorize Jeffreys-Lindley Paradox

References:

- Berger, J. O. (2013; Section 4.3.3). Statistical decision theory and Bayesian analysis. Springer Science & Business Media.
- Robert, C. (2007; Section 5.2(exclude 5.2.6)). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.

[a] Author: Georgios P. Karagiannis.

1 Improper priors situations

Jeffreys-Lindley Paradox describes the overwhelming statistical evidence in favor a hypothesis when the priors of the alternative hypotheses diverge faster. It occurs in Bayesian hypothesis test and model selection with improper priors.

Example 1. [Single-vs-General-alternative]

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

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

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

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

1. The Jeffreys’ (and Laplace) prior of μ is $\pi^{(\mathrm{J})}(\mu) \propto 1(\mu \in \mathbb{R})$ (see Handout 7). Find values μ_*, σ_*^2 such that

$$\tilde{\pi}(\mu|\mathrm{H}_1) \to \tilde{\pi}^{(\mathrm{J})}(\mu)\ ; \text{ as } \quad (\mu_0, \sigma_0^2) \to (\mu_*, \sigma_*^2) \tag{2}$$

   where $\tilde{\pi}(\mu|\mathrm{H}_1)$ denotes the kernel of the pdf of the conditional prior $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$ of μ under H_1in (1), and $\tilde{\pi}^{(\mathrm{J})}(\mu)$ denotes that of the the Jeffreys prior $\pi^{(\mathrm{J})}(\mu) \propto 1(\mu \in \mathbb{R})$.

2. [Lindley’s Paradox] Let $\sigma_0^2 \to \infty$. Investigate, how, and why
   - the conditional prior $\pi(\mu|\mathrm{H}_1)$ given the hypothesis H_1 behaves?
   - the Bayes Factor $\mathrm{B}_{01}(y)$ and the posterior $\mathsf{P}_\Pi(\mathrm{H}_0|y)$ behave?

Solution. The calculation of $\mathrm{B}_{01}(y)$ and of $\mathsf{P}_\Pi(\mathrm{H}_0|y)$ is presented in the Appendix A.

1. The kernel $\tilde{\pi}(\mu|\mathrm{H}_1)$ of the prior $\pi(\mu|\mathrm{H}_1)$ is

$$\pi(\mu|\mathrm{H}_1) = \mathrm{N}(\mu|\mu_0, \sigma_0^2) \propto \exp\left(-\frac{1}{2}\frac{1}{\sigma_0^2}(\mu-\mu_0)^2\right) 1(\mu \in \mathbb{R}) = \tilde{\pi}(\mu|\mathrm{H}_1)$$

   So for $\sigma_*^2 = \infty$ and any $|\mu_*| < \infty$, I get the (2).

2. If $\sigma_0^2 \to \infty$, the prior $\mathrm{N}(\mu_0, \sigma_0^2)$ of H_1 becomes flat (the mass spreads uniformly around $\mathbb{R}$) and improper:

$$\pi(\mu|\mathrm{H}_1) = \mathrm{N}\left(\mu|\mu_0, \sigma_0^2\right) \propto \exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right) 1(\mu \in \mathbb{R}) \xrightarrow{\sigma_0^2\to\infty} 1(\mu \in \mathbb{R})$$

   In fact $\pi(\mu|\mathrm{H}_1)$ meets Jeffreys' (or Laplace) prior of μ in the limit $\sigma_0^2 \to \infty$.

   Then in the limit the Bayes factor $\mathrm{B}_{01}(y)$ and the posterior $\mathsf{P}_{\Pi}(\mathrm{H}_0|y)$ approach the values

$$\mathrm{B}_{01}(y) \to \infty, \text{ and } \mathsf{P}_{\Pi}(\mathrm{H}_0|y) \to 1 \quad \text{as} \qquad \sigma_0^2 \to \infty$$

   We observe that regardless the number of the observables n, when the prior variance σ_0^2 of μ given H_1 becomes huge $\left(\sigma_0^2 \to \infty\right)$, the prior of μ in H_1 becomes more diverge

$$\pi(\mu|\mathrm{H}_1) \xrightarrow{\sigma_0^2\to\infty} 1(\mu \in \mathbb{R}), \quad \text{as} \qquad \sigma_0^2 \to \infty$$

   spreading the prior mass uniformly in $\mathbb{R}$ while the evidence in favor $\mathrm{H}_0 : \mu = \theta_0$ that μ is equal to single value θ_0 becomes overwhelming.

$$\mathrm{B}_{01}(y) \xrightarrow{\sigma_0^2\to\infty} \infty \quad \text{and} \quad \mathsf{P}_{\Pi}(\mathrm{H}_0|y) \xrightarrow{\sigma_0^2\to\infty} 1$$

   Paradoxical behavior: One would not expect for $\sigma_0^2 \to \infty$ favor $\mathrm{H}_0 : \mu = \theta_0$ because $\sigma_0^2 \to \infty$ gives increases ignorance and give uniformly positive mass to more values. Also, we wouldn't expect increasing the sample size n to have no effect.

   Essentially, the above paradoxical behavior says that: When $\sigma_0^2 \to \infty$, it means that a priori, I know $\mu = \theta_0$ with probability π_0, but I know nothing about μ probability $1 - \pi_0$; however after I get any observation y I am a posteriori certain that $\mu = \theta_0$.

   Observe that the overall posterior is

$$\pi(\mu) = \pi_0 1(\mu \in \{\theta_0\}) + (1-\pi_0)\left[(\frac{1}{2\pi\sigma_0^2})^{\frac{1}{2}} \exp(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}) 1(\mu \in \mathbb{R})\right]$$

   When $\sigma_0^2 \to \infty$, it involves one conditional component that concentrates the mass $\pi(\mu|\mathrm{H}_0) = 1_{\{\theta_0\}}(\mu)$ and one component with infinite normalizing constant. So we cannot apply the Bayes theorem by using the kernels of the priors and in the sense of canceling normalizing constants as $\pi(\mu|y) \propto f(y|\mu)\tilde{\pi}(\mu)$ where $\pi(\mu) \propto \tilde{\pi}(\mu)$.

Lindley's paradox also appears in composite-vs-composite tests with improper priors, when the one prior diverges faster than the other in the limit. See the following realistic example in Bayesian regression Variable selection problem.

Example 2. [Composite-vs-Composite]

Consider a Normal linear regression model with dependent variable y and a set of regressors $\{\Phi_j\}_{j\in\mathscr{M}}$ where $\mathscr{M}$ is the set of size d that includes the labels of the available regressors; e.g.

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

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

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

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

The Bayes factor $B_{01}(y)$ and the posterior marginal probability $\mathrm{P}_\Pi(\mathrm{H}_0|y)$ are (See the Appendix B)

$$B_{01}(y) = \sqrt{\frac{|V_1|}{|V_0|}}\sqrt{\frac{|V_0^*|}{|V_1^*|}}\left(\frac{k_0^*}{k_1^*}\right)^{-\frac{n}{2}-a}; \qquad \mathrm{P}_\Pi(\mathrm{H}_0|y) = \left(1+\frac{1-\pi_0}{\pi_0}B_{01}^{-1}(y)\right)^{-1}$$

where for $j = 0, 1$

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

Assume that $V_{\mathscr{M}_0} = vI_{d_0}$ and $V_{\mathscr{M}_1} = vI_{d_1}$. Let $v\to\infty$, how $B_{01}(y)$ and $\mathrm{P}_\Pi(\mathrm{H}_0|y)$ behave when (1.) $d_0 < d_1$, (2.) $d_0 > d_1$, and (3.) $d_0 = d_1$

Solution. For $V_0 = vI_{d_0}$ and $V_1 = vI_{d_1}$ it is

$$\lim_{v\to\infty} B_{01}(y) = \lim_{v\to\infty}(v)^{\frac{d_1-d_0}{2}} \times \sqrt{\frac{|\Phi_0^\top\Phi_0|}{|\Phi_1^\top\Phi_1|}}\left(\frac{k-\frac{1}{2}y^\top\Phi_0\left(\Phi_0^\top\Phi_0\right)^{-1}\Phi_0^\top y_0 + y^\top y}{k-\frac{1}{2}y^\top\Phi_1\left(\Phi_1^\top\Phi_1\right)^{-1}\Phi_1^\top y_1 + y^\top y}\right)^{-\frac{n}{2}-a}$$

$$\text{So } B_{01}(y)\xrightarrow{v\to\infty}\begin{cases} +\infty, & d_0<d_1 \\ 0, & d_0>d_1 \\ <\infty, & d_0=d_1 \end{cases} \quad\text{and}\quad \mathrm{P}_\Pi(\mathrm{H}_0|y)\xrightarrow{v\to\infty}\begin{cases} 1, & d_0<d_1 \\ 0, & d_0>d_1 \\ \in(0,1), & d_0=d_1 \end{cases}$$

As $v\to\infty$, both conditional priors $\beta_0|\sigma^2\sim\mathrm{N}(\mu_0,\sigma^2 Iv)$ and $\beta_1|\sigma^2\sim\mathrm{N}(\mu_1,\sigma^2 Iv)$ under hypothesis H_0 and H_1 become more and more diverge, while the evidence becomes more overwhelming in favor of the hypothesis that the prior diverges slower. In our example, when $d_0 < d_1$ the conditional prior $\pi(\beta|\sigma^2,\mathrm{H}_0)$ given H_0 diverges slower.

2 Informal but intuitive investigation of the phenomenon

Consider

$$\mathrm{H}_0:\theta\in\Theta_0,\ \text{vs}\ \mathrm{H}_1:\theta\in\Theta_1$$

where $\{\Theta_0,\Theta_1\}$ unbounded sets partitioning $\Theta\subseteq\mathbb{R}^d$. Consider overall prior with pdf

$$\pi(\theta) = \pi_0\ \pi_0(\theta) + \pi_1\ \pi_1(\theta)$$

Let conditional priors $\pi_0(\theta) = \frac{1}{C_0}\tilde{\pi}_0(\theta)$ and $\pi_1(\theta) = \frac{1}{C_1}\tilde{\pi}_1(\theta)$ where $\tilde{\pi}_0(\theta)$ and $\tilde{\pi}_1(\theta)$ are pdf kernels.

The posterior probability $\mathrm{P}_\Pi(\mathrm{H}_0|y)$ of the hypothesis H_0 is

$$\mathrm{P}_\Pi(\mu\in\Theta_0|y) = \frac{\pi_0 C_0^{-1}\int_{\Theta_0}\tilde{\pi}_0(\theta)f(y|\theta)\mathrm{d}\theta}{\pi_0 C_0^{-1}\int_{\Theta_0}\tilde{\pi}_0(\theta)f(y|\theta)\mathrm{d}\theta + \pi_1 C_1^{-1}\tilde{\pi}_1(\theta)\int_{\Theta_1}f(y|\theta)\mathrm{d}\theta} = \left[1+\frac{C_0}{C_1}\frac{\int_{\Theta_1}\tilde{\pi}_1(\theta)f(y|\theta)\mathrm{d}\theta}{\int_{\Theta_0}\tilde{\pi}_0(\theta)f(y|\theta)\mathrm{d}\theta}\right]^{-1}$$

The Bayes factor is

$$\mathrm{B}_{01}(y) = \frac{C_1}{C_0} \frac{\int_{\Theta_0} \tilde{\pi}_0(\theta) f(y|\theta) \mathrm{d}\theta}{\int_{\Theta_1} \tilde{\pi}_1(\theta) f(y|\theta) \mathrm{d}\theta}.$$

When $\tilde{\pi}_0(\theta) \rightarrow$ const and $\tilde{\pi}_1(\theta) \rightarrow$ const, the normalizing constants of $\pi_0(\theta)$ and $\pi_1(\theta)$ become infinite $C_0 \rightarrow \infty$ and $C_1 \rightarrow \infty$. Then whether $\mathrm{B}_{01}(y) \rightarrow 0$ or $\mathrm{B}_{01}(y) \rightarrow \infty$ depends on which prior $\pi_0(\theta)$ or $\pi_1(\theta)$ becomes improper faster than the other, namely

$$\frac{C_1}{C_0} \xrightarrow[\tilde{\pi}_1(\theta)\rightarrow\text{const}]{\tilde{\pi}_0(\theta)\rightarrow\text{const}} \begin{cases} \infty & \text{, if } C_0 << C_1 \\ 0 & \text{, if } C_1 << C_0 \end{cases}$$

In Example 2,

$$C_j = (2\pi)^{\frac{n}{2}} |V_j|^{\frac{1}{2}} \overset{V_j=vI}{=} (2\pi)^{\frac{n}{2}} v^{\frac{d_j}{2}}; \Longrightarrow \qquad \frac{C_1}{C_0} = (v)^{\frac{d_1-d_0}{2}} \xrightarrow{v\rightarrow\infty} \begin{cases} \infty & \text{, if } d_0 < d_1 \\ 0 & \text{, if } d_0 > d_1 \end{cases} \text{.as}$$

Summary 3. Improper priors do not really work in hypothesis tests, or model comparison. Be cautious when You use improper priors.

3 Large samples situation

It is not a secrete[1] that in Frequentist hypothesis tests you can reject H_0 if you get a large enough sample. P-values is a misleading description of the evidence against H_0. What is the case in Bayesian hypothesis tests?

Example 4. (Cont. of Example 1) Set $\mu_0 = \theta_0$ and let $z_n = \frac{\bar{y}-\mu_0}{\sigma}\sqrt{n}$. Then, it is

$$\mathrm{B}_{01}(y) \overset{\text{calc.}}{=} \frac{\left(1 + n\sigma_0^2/\sigma^2\right)^{\frac{1}{2}}}{\exp\left(\frac{1}{2} z_n^2 \left(1 + \sigma^2/(\sigma_0^2 n)\right)^{-1}\right)} \geq \frac{\left(1 + n\sigma_0^2/\sigma^2\right)^{\frac{1}{2}}}{\exp\left(\frac{1}{2} z_n^2\right)},$$

Assume that $|z_n| \geq z^*_{1-\frac{a}{2}}$; compare the result of the Bayesian hypothesis test against the a-sig. level Frequentist test.

Solution. Consider z_n as fixed. For the Bayesian hypothesis test, I get

$$\mathrm{B}_{01}(y) \geq \frac{\left(1 + n\sigma_0^2/\sigma^2\right)^{\frac{1}{2}}}{\exp\left(\frac{1}{2} z_n^2\right)}, \text{ and } \mathsf{P}_{\Pi}(\mathrm{H}_0|y) \geq \left(1 + \frac{1-\pi_0}{\pi_0} \frac{\exp\left(\frac{1}{2} z_n^2\right)}{\left(1 + n\sigma_0^2/\sigma^2\right)^{\frac{1}{2}}}\right)^{-1}$$

As $n \rightarrow \infty$ $|z_n| < \infty$ is bounded due to the CLT and hence

$$\mathrm{B}_{01}(y) \rightarrow \infty, \text{ and } \mathsf{P}_{\Pi}(\mathrm{H}_0|y) \rightarrow 1$$

so H_0 is accepted for sure. In the frequentist hypothesis test, I reject H_0 at a-sig. level, because $|z_n| > z^*_{1-\frac{a}{2}}$.

The Bayesian behavior is the reasonable one: It has to be $\bar{y} \approx \mu_0$ for the observed $z_n = \frac{\bar{y}-\mu_0}{\sigma}\sqrt{n}$ to be a fixed finite number and not an infinite as $n \rightarrow \infty$

Note 5. It is not difficult to see that for Single-vs-General-alternative tests $\mathrm{H}_0 : \mu = \theta_0 \quad \text{vs} \quad \mathrm{H}_1 : \mu \neq \theta_0$

$$\mathrm{B}_{01}(y) = \frac{f_0(y)}{f_1(y)} \geq \frac{f(y|\theta_0)}{\sup_{\mu\neq\theta_0} f(y|\mu)} = \frac{f(y|\theta_0)}{f(y|\hat{\mu}_{\text{MLE}})} \equiv \text{Max. Likl. Ratio}$$

implying that the $\mathrm{B}_{01}(y)$ has the maximum likelihood ratio as a lower bound.

Question. *Practice with Exercises 68 and 69 in the Exercise sheet.*

[1] ...maybe it is kept as a secrete from students attending only frequentist courses in any university ...

Appendix

The following calculations are given for completeness. They are part of Exercises 68 and 69 in the Exercise sheet.

A Calculations

Hint-1: It is

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

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

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

Solution. The overall prior is

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

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

The Bayes factor is

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

So

$$\begin{aligned}
f_0(y)=f(y|\theta_0)=\prod_{i=1}^{n}\mathrm{N}(y_i|\theta_0,\sigma^2)&=(\frac{1}{2\pi\sigma^2})^{\frac{n}{2}}\exp(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\theta_0)^2}{\sigma^2})\\
&=(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2}\frac{1}{\sigma^2}(\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{=ns^2}+n(\bar{y}-\theta_0)^2))\\
&=(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2}\frac{1}{\sigma^2/n}s^2)\exp(-\frac{1}{2}\frac{1}{\sigma^2/n}(\bar{y}-\theta_0)^2)
\end{aligned}$$

$$\begin{aligned}
f_1(y)=\int_{\mathbb{R}}f(y|\theta)\mathrm{d}\Pi_1(\theta)&=\int_{\mathbb{R}}\prod_{i=1}^{n}\mathrm{N}(y_i|\mu,\sigma^2)\mathrm{N}(\mu|\mu_0,\sigma_0^2)\mathrm{d}\mu\\
&=\int_{\mathbb{R}}(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}\exp(-\frac{1}{2}\frac{1}{\sigma^2}(\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{=ns^2}+n(\bar{y}-\mu)^2))\times\\
&\qquad\times(2\pi)^{-\frac{1}{2}}(\sigma_0^2)^{-\frac{1}{2}}\exp(-\frac{1}{2}\frac{1}{\sigma_0^2}(\mu-\mu_0)^2)\mathrm{d}\mu\\
&=(2\pi)^{-\frac{n}{2}-\frac{1}{2}}(\sigma^2)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{1}{2}}\exp(-\frac{1}{2}\frac{1}{\sigma^2/n}s^2)\\
&\qquad\times\int_{\mathbb{R}}\underbrace{\exp(-\frac{1}{2}\frac{1}{\sigma^2/n}(\bar{x}-\mu)^2-\frac{1}{2}\frac{1}{\sigma_0^2}(\mu-\mu_0)^2)}_{=A(\mu)}\mathrm{d}\mu
\end{aligned}$$

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

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

where

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

and

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

So

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

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

It is

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

It is

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

B Calculations

You may use the following identity:

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

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

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

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

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

$$\times\frac{k^a}{\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\left(-\frac{k}{\sigma^2}\right)\mathrm{d}\beta\mathrm{d}\sigma^2 \; = \; ...$$

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

Where

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

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

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

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

and

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

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 15: Inference under model uncertainty [a]

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

Aim: To explain, design, and apply model selection and model determination Bayesian procedures. The calculations in the Example are challenging but feasible given the Hints provided.

References:

- Robert, C. (2007; Section 7.1 & 7.2). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- O'Hagan, A., & Forster, J. J. (2004; Section 7 Model comparison). Kendall's advanced theory of statistics, volume 2B: Bayesian inference (Vol. 2). Arnold.

[a] Author: Georgios P. Karagiannis.

1 Model uncertainty problem

Note 1. All (statistical/mathematical) models are approximations/simplifications of the reality. They are arguable. Their purpose is to allow You draw conclusions about reality and take decisions.

- "All models are wrong, but some are useful". –George Box's quote

Note 2. Let $y = (y_1, ..., y_n)$ be observables generated from the real but unknown data-generating process $y \sim R(y)$. Let $z = (y_{n+1}, ..., y_{n+m})$ be future outcomes from the same process $R(\cdot)$.

Note 3. Statistician may be unsure of the proper way to formulate the statistical model $\mathscr{M} = \{F(y|\theta);\, \theta \in \Theta\}$ in order to acceptably represent the data-generating process $R(y)$. We call this situation 'model uncertainty'.

Note 4. To address problems under model uncertainty, often

- statistician initially considers a restricted set $\{\mathscr{M}_k;\, k \in \mathcal{K}\}$ of available - reasonable - potential - competing - alternative statistical models $\mathscr{M}_k = \{F_k(y|\theta_k);\, \theta_k \in \Theta_k\}$. Then,
- interest lies in learning a model $\mathscr{M}$ which 'best' represents $R(y)$, performing inference (predictive or parametric), and drawing conclusions under the pretense of model uncertainty.

Example 5. Consider the Normal linear regression problem. Let $\{\phi_0, \phi_1, ..., \phi_{d-1}\}$ be a set of d possible regressor variables, which may affect the values of the response variable $y \in \mathcal{Y} \subseteq \mathbb{R}$. We are interested in learning the mapping

$$\{\phi_0, ..., \phi_{d-1}\} \overset{?}{\longmapsto} y$$

Assume there are available n pairs of observations: (Φ_1, y_1), (Φ_2, y_2), ..., (Φ_n, y_n).

Consider the following approximations of the real mapping $\{\phi_0, ..., \phi_{d-1}\} \longmapsto y$: Assume that observables y_i are Normally distributed and independent for $i = 1, ..., n$, with unknown variance σ^2, and unknown mean which can be parametrised as a linear combination of some of the regressors $\{\phi_0, ..., \phi_{d-1}\}$ with unknown coefficients β. Finally, consider, we are uncertain which set of $\{\phi_0, ..., \phi_{d-1}\}$ actually affects values of y.

2 Standard framework

Note 6. Consider a collection of statistical models $\mathcal{M} = \{\mathscr{M}_k;\ k \in \mathcal{K}\}$, where each model

$$\mathscr{M}_k = \{F_k(y|\theta_k);\ \theta_k \in \Theta_k\},$$

is labeled by an index $k \in \mathcal{K}$, and associated with the sampling distribution $F_k(y|\theta_k)$ with pdf/pmf $f_k(y|\theta_k)$, unknown parameter $\theta_k \in \Theta_k$, and parametric space Θ_k.

Definition 7. The encompassing model $\mathcal{M}$ is defined as the collection of statistical models

$$\begin{aligned}\mathcal{M} &= \{\mathscr{M}_k;\ k \in \mathcal{K}\}\\ &= \{\{F_k(y|\theta_k);\ \theta_k \in \Theta_k\};\ k \in \mathcal{K}\}\end{aligned}$$

labeled by parameter $\vartheta = (k, \theta_k) \in \varTheta$ which is defined on the joint parameter space

$$\varTheta = \cup_{k \in \mathcal{K}}\{k\} \times \Theta_k$$

Example 8. (Cont...) We set-up the encompassing model $\mathcal{M} = \{\mathscr{M}_k;\ k \in \mathcal{K}\}$ as the collection of statistical models

$$\mathscr{M}_k = \left\{y|\beta_k, \sigma^2, k \sim \mathrm{N}\left(\Phi_k\beta_k, I\sigma^2\right)\ ;\ \beta_k \in \mathbb{R}^{d_k},\ \sigma^2 \in \mathbb{R}_+\right\}, \tag{1}$$

where $k \subseteq \{0, 1, ..., d-1\}$ is the label of the model indicating the sub-set of the regressors included in the statistical model $\mathscr{M}_k$, and $\mathcal{K}$ is the collection of all sub-sets of $\{0, 1, ..., d-1\}$ so that $k \in \mathcal{K}$. Given $\mathscr{M}_k$, Φ_k is the design matrix consisting of the regressors with indexes in k, $\beta_k \in \mathbb{R}^{d_k}$ is the associated vector of regression coefficients, and d_k is the number of regressors. Notice that σ^2 is kept to be common among all the models, however one could have set it to be different between different models (e.g.; σ^2_k). The within model parameters are $\theta_k = \left(\beta_k, \sigma^2\right)$ defined on space $\Theta_k = \mathbb{R}^{d_k} \times \mathbb{R}$. The parameter of $\mathcal{M}$ is $\vartheta = (k, \beta_k, \sigma^2)$ and defined on space $\varTheta = \cup_{k \in \mathcal{K}}\{k\} \times \mathbb{R}^{d_k} \times \mathbb{R}$.

Note 9. To complete the Bayesian model we specify prior distributions on the joint parametric space $\varTheta$. Consider a collection of conditional priors on θ_k given model $\mathscr{M}_k$ as

$$\theta_k|\mathscr{M}_k \sim \Pi(\theta_k|\mathscr{M}_k)$$

with pdf/pmf $\pi(\theta_k|\mathscr{M}_k)$; and consider a marginal prior distribution on $\mathscr{M}_k$

$$\mathscr{M}_k \sim \Pi(\mathscr{M}_k)$$

with pdf/pmf $\pi(\mathscr{M}_k)$, for all $k \in \mathcal{K}$. The encompassing prior model has distribution $\Pi(\mathscr{M}_k, \theta_k)$ with pdf/pmf

$$\pi(\mathscr{M}_k, \theta_k) = \pi(\theta_k|\mathscr{M}_k)\pi(\mathscr{M}_k),\ \ \forall(k, \theta_k) \in \varTheta$$

Note 10. Specifying the prior distribution $\Pi(\mathscr{M}_k, \theta_k)$ is a delicate matter.

1. Within model priors $\theta_k|\mathscr{M}_k \sim \Pi(\theta_k|\mathscr{M}_k)$ should not be improper because the normalizing constants are not canceled in the joint posterior (7), and hence the Lindley-Jeffreys padadox may kick in.

2. If some models are embedded (nested) into others, e.g. $\mathscr{M}_1 \subseteq \mathscr{M}_2$, possibly the marginal model prior should be $\pi(\mathscr{M}_1) \leq \pi(\mathscr{M}_2)$, in order to be coherent –this is not a panacea.

3. Common parameters between models (such as σ^2 in Example) may be treated as separate parameters. However, it is common to be regarded as the same parameter and be assigned the same prior; this is an additional approximation/simplification that modeler does for computational convenience –less parameters to learn ;-).

   - In the Example, σ^2 is common to all models. If we wanted to be in accordance with this note, we should have specified σ^2 separately for each model $\mathscr{M}_k$, as σ^2_k, but we do not do this now (naughty).

Note 11. The full Bayesian model can be summarized as

$$\begin{cases} y|\theta_k, \mathscr{M}_k \sim F_k(y|\theta) & \text{, data generation fromt he sampling distribution} \\ \\ \theta_k|\mathscr{M}_k \sim \Pi(\theta_k|\mathscr{M}_k) & \text{, within model parameter generation from the conditional prior} \\ \\ \mathscr{M}_k \sim \Pi(\mathscr{M}_k) & \text{, model generation fromt he marginal model prior} \end{cases} \tag{2}$$

hence the joint distribution is such that $\mathrm{d}P(y, \theta_k, \mathscr{M}_k) = \mathrm{d}F_k(y|\theta)\mathrm{d}\Pi(\theta_k|\mathscr{M}_k)\mathrm{d}\Pi(\mathscr{M}_k)$.

Example 12. (Cont...) One can specify the following Bayesian linear regression model

$$y|\beta_k, \sigma^2, \mathscr{M}_k \sim \mathrm{N}\left(\Phi_k\beta_k, I\sigma^2\right) \qquad \text{, the sampling distribution} \tag{3}$$

$$\beta_k|\sigma^2, \mathscr{M}_k \sim \mathrm{N}\left(\mu_k, \sigma^2 V_k\right) \qquad \text{, conditional prior of the within model parameter} \tag{4}$$

$$\sigma^2|\mathscr{M}_k \sim \mathrm{IG}\left(a, \lambda\right) \qquad \text{conditional prior of the within model parameter} \tag{5}$$

$$\mathscr{M}_k \sim \pi(k) = \frac{1}{|\mathcal{K}|} = \frac{1}{2^d} \qquad \text{, marginal model prior} \tag{6}$$

Eq. 3 is the sampling distribution $F(y|\beta_k, \sigma^2, \mathscr{M}_k)$ for model $\mathscr{M}_k$. Eq. 4 and 5 are the conditional (or within) model priors $\Pi(\beta_k, \sigma^2|\mathscr{M}_k)$; recall that they are conjugate to $F(y|\beta_k, \sigma^2, \mathscr{M}_k)$ and hence chosen for our computational convenience. Notice that parameter σ^2, which is common to all available models $\{\mathscr{M}_k\}$, is treated as obtaining the same values among all models, and following the same prior. This is against the suggestion in Note 10((3)) but we assume that the computational benefits of reducing the dimensionality of the parametric space dominate the losses from the (possibly) worse approximation of the reality. In Eq. 6, the marginal model prior $\pi(k)$ is chosen to be uniform across models (we have $|\mathcal{K}| = 2^d$ possible combinations or regressors).

Notation 13. To make the notation easier, we will denote $\mathscr{M}_k$ as k and the probability measures as

- $F(y|\theta_k, k) := F_k(y|\theta_k)$ and $f(y|\theta_k, k) := f_k(y|\theta_k)$, as well as
- $\Pi(\theta_k|k) := \Pi(\theta_k|\mathscr{M}_k)$, $\pi(\theta_k|k) := \pi(\theta_k|\mathscr{M}_k)$, $\Pi(k) := \Pi(\mathscr{M}_k)$, and $\Pi(k, \theta_k) := \Pi(\mathscr{M}_k, \theta_k)$, etc...

3 Posterior distributions

According to the Bayesian, the following posteriors can be derived.

Proposition 14. *Given the Bayesian model (2), the joint posterior distribution* $\Pi(k, \theta_k|y)$ *has pdf/pmf*

$$\pi(k, \theta_k|y) = \frac{f(y|\theta_k, k)\pi(\theta_k|k)\pi(k)}{\sum_{k\in\mathcal{K}}\int_{\Theta_k} f(y|\theta_k, k)\pi(\theta_k|k)d\theta_k\pi(k)}, \ \forall(k, \theta_k) \in \mathit{\Theta}; \tag{7}$$

Proposition 15. *The joint posterior density (7) can be factorized as*

$$\pi(k, \theta_k|y) = \pi(\theta_k|y, k)\pi(k|y)$$

where the first part is the pdf/pmf of the conditional posterior $\Pi(\theta_k|y, k)$ *of* θ_k *given model* $\mathscr{M}_k$ *i.e.*

$$\pi(\theta_k|y, k) = \frac{f(y|\theta_k, k)\pi(\theta_k|k)}{\int_{\Theta_k} f(y|\theta_k, k)\pi(\theta_k|k)d\theta_k} = \frac{f(y|\theta_k, k)\pi(\theta_k|k)}{f(y|k)}, \ \forall\theta_k \in \Theta_k\,; \tag{8}$$

and the second term is the pdf/pmf marginal model posterior $\Pi(k|y)$ of model $\mathscr{M}_k$ with pdf/pdf

$$\pi(k|y) = \frac{f(y|k)\pi(k)}{f(y)},\ \forall k \in \mathcal{K}\,.$$

Here, $f(y|k)$ is the conditional marginal likelihood (or the model evidence) of model $\mathscr{M}_k$ with pdf/pdf

$$f(y|k) = \int_{\Theta_k} f(y|\theta_k, k)\pi(\theta_k|k)d\theta_k$$

and $f(y)$ is the marginal likelihood with

$$f(y) = \sum_{k\in\mathcal{K}}\left[\int_{\Theta_k} f(y|\theta_k, k)\pi(\theta_k|k)d\theta_k\right]\pi(k) = \sum_{k\in\mathcal{K}} f(y|k)\pi(k)$$

Example 16. To calculate the conditional posterior pdf $\pi\left(\beta_k, \sigma^2|y\right)$, we use the Bayes Theorem

Appendix A

$$\begin{aligned}\pi\left(\beta_k, \sigma^2|y, k\right) \propto & f(y|\theta_k, k)\pi(\theta_k|k) \;=\; \text{N}\left(y|\Phi\beta_k, I\sigma^2\right)\text{N}\left(\beta_k|\mu_k, V_k\sigma^2\right)\text{IG}\left(\sigma^2|a, \lambda\right)\\ \propto & \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(\beta_k-\mu_k^*\right)^\top\left(V_k^*\right)^{-1}\left(\mu_k-\beta_k^*\right)\right)}_{} \times \underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a+1}\exp\left(-\frac{\lambda}{\sigma^2}\right)}_{}\\ \propto & \text{N}\left(\beta_k|\mu_k^*, V_k^*\sigma^2\right)\text{IG}\left(\sigma^2|a^*, \lambda_k^*\right)\end{aligned}$$

where

$$V_k^* = \left(V_k^{-1}+\Phi_k^\top\Phi_k\right)^{-1};\qquad \mu_k^* = V_k^*\left(V_k^{-1}\mu_k+\Phi_k^\top y\right)\mathscr{M}_k;\quad a^* = \frac{n}{2}+a;\quad \lambda_k^* = \lambda+\frac{1}{2}S_k^*;$$
$$S_k^* = \mu_k^\top V_k^{-1}\mu_k - \left(\mu_k^*\right)^\top\left(V_k^*\right)^{-1}\left(\mu_k^*\right) + y^\top y$$

Hence

$$\Longrightarrow \begin{cases}\beta_k|y\sigma^2, k & \sim \text{N}\left(\mu_k^*, V_k^*\sigma^2\right)\\ \sigma^2|y, k & \sim \text{IG}\left(a^*, \lambda^*\right)\end{cases},\ \forall k\in\mathcal{K}$$

Example 17. To calculate the model evidence / conditional marginal likelihood for $\mathscr{M}_k$, we integrate

Appendix B

$$f(y|k) = \int \overbrace{\text{N}\left(y|\Phi_k\beta_k, I\sigma^2\right)}^{f(y|\theta_k,k)}\overbrace{\text{N}\left(\beta_k|\mu, V_k\sigma^2\right)\text{IG}\left(\sigma^2|a,\lambda\right)}^{\pi(\theta_k|k)}\text{d}\beta_k\text{d}\sigma^2 = \text{...calc...} = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\frac{\Gamma\left(a^*\right)}{\left(\lambda_k^*\right)^{a^*}}$$

Since $\mathcal{K}$ is finite, the marginal likelihood $f(y)$ of y can be computed as

$$f(y) = \sum_{k\in\mathcal{K}} f(y|k)\pi(k) \;=\; \left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\lambda^a\frac{\Gamma\left(a^*\right)}{\Gamma(a)}\frac{1}{2^d}\sum_{k\in\mathcal{K}}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}}$$

So the marginal model posterior probability $\pi(k|y)$ of $\mathscr{M}_k$ is

$$\pi(k|y) = \frac{f(y|k)\pi(k)}{f(y)} = \left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}}\Bigg/\sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_{k'}^*\right)^{a^*}} \tag{9}$$

Proposition 18. *Let $z = (y_{n+1}, ..., y_{n+m})$ be a sequence of future outcomes. The conditional predictive distribution $G(z|y,k)$ of z given observables and model $\mathscr{M}_k$ has pdf/pmf*

$$g(z|y,k) = \int f(z|y,\theta_k,k)d\Pi(\theta_k|y,k) \overset{z\ indep.\ y|(\theta_k,k)}{=} \int f(z|\theta_k,k)d\Pi(\theta_k|y,k) \tag{10}$$

The marginal predictive distribution $G(z|y)$ of z given the observables y has pdf/pmf

$$g(z|y)=\sum_{k\in\mathcal{K}}\int f(z|y,\theta_k,k)d\Pi(\theta_k|y,k)\pi(k|y)=\sum_{k\in\mathcal{K}}g(z|y,k)\pi(k|y)$$

Example 19. The predictive distribution $G\left(z|y,k\right)$ of future observables $z=(y_{n+1},...,y_{n+m})$ at any new Φ_k^{new} given the observables y and the model $\mathscr{M}_k$; i.e. $z=\Phi_k^{\text{new}}\beta_k+\epsilon$ and $\epsilon\sim\text{N}\left(0,I_m\sigma^2\right)$, has pdf

Appendix C

$$\begin{aligned}g(z|y,k)=\int f(z|\theta_k,k)\pi(\theta_k|y,k)\mathrm{d}\theta_k&=\int\text{N}\left(z|\Phi_k^{\text{new}}\beta_k,I\sigma^2\right)\text{N}\left(\beta_k|\mu_k^*,V_k^*\sigma^2\right)\text{IG}\left(\sigma^2|a^*,\lambda_k^*\right)\mathrm{d}\beta_k\mathrm{d}\sigma^2\\&=\text{...calc...}=\text{T}_m\left(z|\Phi^{\text{new}}\mu_k^*,[I+V_k^*]\frac{\lambda_k^*}{a^*},2a^*\right)\end{aligned}$$

Example 20. The marginal predictive distribution $G\left(z|k\right)$ of future observables $z=(y_{n+1},...,y_{n+m})$ at any new Φ_k^{new} given the observables y has pdf

$$\begin{aligned}g(z|y)&=\sum_{k\in\mathcal{K}}g(z|y,k)\pi(k|y)\\&=\sum_{k\in\mathcal{K}}\text{T}_m\left(z|\Phi^{\text{new}}\mu_k^*,V_k^*\frac{\lambda_k^*}{a^*},2a^*\right)\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{(\lambda_k^*)^{a^*}}\Big/\sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\frac{1}{(\lambda_{k'}^*)^{a^*}}\end{aligned}\tag{11}$$

4 Inference based on a single Best model

For parametric or predictive inference, the statistician may proceed as follows:

1. select a single 'Best' model $\mathscr{M}_{k^*}$ (optimal in some sense) from a set of available models $\mathcal{M}=\{\mathscr{M}_1,\mathscr{M}_2,...\}$ and then
2. perform Bayesian inference conditional on the selected model $\mathscr{M}_{k^*}$ only and as if it had generated the data; i.e. based on

$$\begin{cases}y|\theta_{k^*},\mathscr{M}_{k^*}\sim F(y|\theta_{k^*},k^*) & \text{, data generation fromt he sampling distribution}\\ \\ \theta_{k^*}|\mathscr{M}_{k^*}\sim\Pi(\theta_{k^*}|k^*) & \text{, within model parameter generation from the conditional prior}\end{cases}$$

4.1 Bayesian model selection: How to select the single 'Best' model $\mathscr{M}_{k^*}$.

Note 21. Model selection/choice is the selection of the "Best" statistical model from a set of competing statistical models $\mathcal{M}=\{\mathscr{M}_k;\ k\in\mathcal{K}\}$. It can be naturally performed as a parametric point estimation problem, statistical decision problem, or hypothesis test (all equivalent). See Criteria 22 and 23.

Criterion 22. *We produce the Bayes estimator (of Bayes rule) $\hat{\delta}$ of parameter $k\in\mathcal{K}$ under a loss function $\ell(\delta,k,\theta_k)$ with decision space $\mathcal{D}=\mathcal{K}$ and posterior distribution $(k,\theta_k)\sim\Pi(k,\theta_k|y)$. I.e find $\hat{\delta}$:*

$$\hat{\delta}=\arg\min_{\forall\delta\in\mathcal{K}}E_{\Pi}\left(\ell(\delta,k,\theta_k)\,|y\right)=\arg\min_{\forall\delta\in\mathcal{K}}\left(\int_{\Theta_k}\ell(\delta,k,\theta_k)d\Pi(k,\theta_k|y)\right)$$

Criterion 23. *Under the $0-1$ loss $\ell(\delta,k,\theta_k)=1\left(\delta\neq k\right)$ the best model corresponds to the MAP $\hat{\delta}$ that is model with the highest marginal model posterior probability $\pi(k|y)$, I.e $\hat{\delta}$* such as*:*

$$\hat{\delta}=\arg\max_{\forall k\in\mathcal{K}}\{\pi(k|y)\}$$

or equivalently $\hat{\delta}$ such that $\pi(\hat{\delta}|y)\geq\pi(k|y)$, for all $k\in\mathcal{K}$.

Bayes factors can be used for model selection. We re-define the Bayes factor to be suitable to the model selection framework.

Definition 24. Bayes factor $B_{k,j}(y)$ of model $\mathscr{M}_k$ against model $\mathscr{M}_j$ is defined as

$$B_{k,j}(y) = \frac{\pi\left(k|y\right)/\pi\left(k\right)}{\pi\left(j|y\right)/\pi\left(j\right)} = \frac{f(y|k)}{f(y|j)}; \quad \forall k, j \in \mathcal{K}$$

Example 25. (Cont.) The marginal model posterior $\pi(k|y)$ has been calculated in Example 17 in Eq. 9. Under the 0-1 loss $\ell(\delta, k, \theta_k) = 1\,(\delta \neq k)$, the best regression model is that with the largest marginal model posterior $\pi(k|y)$ (9).

Consider data-set airquality {datasets} available from R repository, that involves $n = 153$ pairs of observations on the response variable y 'Ozone (ppb) ' and the possible regressors

ϕ_0: Constant value 1

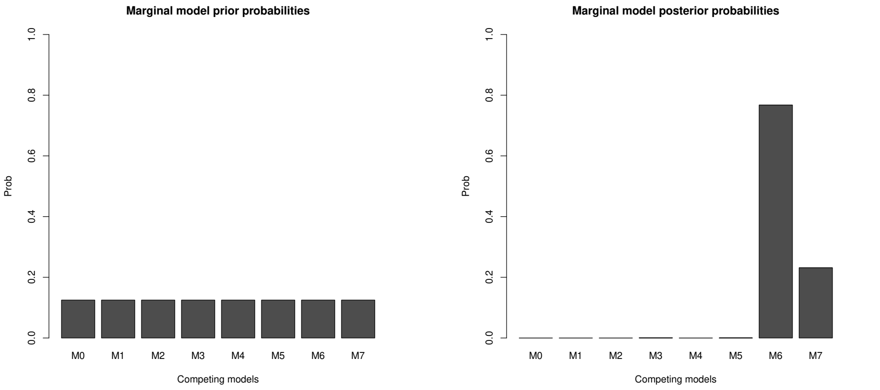
ϕ_1: Solar R (lang)

ϕ_2: Wind (mph)

ϕ_3: Temperature (degrees F)

The regressors have been standardized to adjust their unites in the model. The available models are: $\mathscr{M}_0 = \{1\}$, $\mathscr{M}_1 = \{1, 2\}$, $\mathscr{M}_2 = \{1, 3\}$, $\mathscr{M}_3 = \{1, 4\}$, $\mathscr{M}_4 = \{1, 2, 3\}$, $\mathscr{M}_5 = \{1, 2, 4\}$, $\mathscr{M}_6 = \{1, 3, 4\}$, and $\mathscr{M}_7 = \{1, 2, 3, 4\}$. The prior hyper-parameters are $\mu_k = 0$, $V_k = 100I_{d_k}$, $a_k = 1.0$, and $\lambda_k = 1.0$. The marginal model prior was $\pi(k|y) = 1/8$ for $k = \{0, ..., 7\}$.

In Figure 1, we see the marginal model prior which is uniform, and the marginal model posterior which denotes that the MAP 'best' model is $\mathscr{M}_6$.



(a) Marginal model prior probabilities (b) Marginal model posterior probabilities

Figure 1: Marginal model prior and posterior probabilities produced in Example 25.
`https://github.com/georgios-stats/Bayesian_Statistics/blob/master/HandoutsSupplementary/LinearRegressionModelUncertainty/MarginalModelPosteriroProbability.R`

The Bayes factor is computed as

$$B_{k,j}(y) = \frac{f(y|k)}{f(y|j)} = \left(\frac{|V_j|}{|V_k|}\right)^{\frac{1}{2}} \left(\frac{|V_k^*|}{|V_j^*|}\right)^{\frac{1}{2}} \left(\frac{\lambda_j^*}{\lambda_k^*}\right)^{a^*} \tag{12}$$

4.2 Within model inference

Note 26. Given that $\mathscr{M}_{k^*}$ is the selected “Best” model, interest may lie on learning the parameter $\theta_{k^*} \in \Theta_{k^*}$ (or any function of it), or future outputs $z = (y_{n+1}, ..., y_{n+m})$. Then inference is performed based on the Bayesian model

$$\begin{cases} y|\theta_{k^*}, \mathscr{M}_{k^*} \sim F(y|\theta_{k^*}, k^*) & \text{, data generation fromt he sampling distribution} \\ \\ \theta_{k^*}|\mathscr{M}_{k^*} \sim \Pi(\theta_{k^*}|k^*) & \text{, within model parameter generation from the conditional prior} \end{cases}$$

- Parametric inference about θ_{k^*} (or functions of it) is performed based on the conditional posterior in (8):

  $$\theta_{k^*}|y, k^* \sim \Pi(\theta_{k^*}|y, k^*)$$

- Predictive inference about a sequence of future outcomes $z = (y_{n+1}, ..., y_{n+m})$ is performed based on the conditional predictive distribution :

  $$z|y, k^* \sim G(z|y, k^*)$$

  in (10) with pdf

  $$g(z|y, k^*) = \int f(z|y, \theta_{k^*}, k^*)\mathrm{d}\Pi(\theta_{k^*}|y, k^*)$$

5 Bayesian model averaging (BMA)

Note 27. Selecting a single (best) model from a set of available ones and then proceeding as if the selected model had generated may not be optimal. It ignores the uncertainty due to model selection, which leads to over-confident inferences and decisions. BMA offers a coherent mechanism to account model uncertainty.

Definition 28. Model averaging is called the process where a posterior summary is obtained as a weighted mean of the summaries under each model with weights given by the marginal model posterior probability.

Parametric inference There may be certain model parameters or functions of model parameters which are present in all available models and retain a consistent interpretation across all models. Interesting may be to average out all the models.

Note 29. Let partition $\theta_k = (\psi, \phi_k)$, where $\psi \in \Psi$ is common in all models in $\mathcal{M}$. It is possible to obtain the marginal posterior distribution $\Pi(\psi|y)$ of ψ as a mixture of posterior distributions under each model with weights given by the marginal posterior probability of each model. i.e. $\Pi(\psi|y)$ has pdf/pmf

$$\pi(\psi|y) = \sum_{k\in\mathcal{K}} \pi(\psi|y, k)\pi(k|y); \qquad \text{with} \quad \pi(\psi|y, k) = \int \pi(\psi, \phi_k|y, k)\mathrm{d}\phi_k.$$

Note 30. Then any moment (expectation, variance, probability, etc...) of $\psi \in \Psi$ can be expressed in the aforesaid weighted average form. Let $h(\cdot)$ be a function defined on Ψ. Then

$$\mathrm{E}_\Pi\left(h(\psi)|y\right) = \sum_{k\in\mathcal{K}} \int_{\Theta_k} h(\psi)\mathrm{d}\Pi\left(\psi|y, k\right)\pi(k|y) = \sum_{k\in\mathcal{K}} \mathrm{E}_\Pi\left(h(\psi)|y, k\right)\pi(k|y)$$

Example 31. The sampling distribution variance σ^2 is the common parameter across all models, so

$$\pi\left(\sigma^2|y\right) = \sum_{k\in\mathcal{K}} \overbrace{\mathrm{IG}\left(\sigma^2|a^*, \lambda_k^*\right)}^{\pi(\sigma^2|y,k)}\pi(k|y) = \frac{\sum_{k\in\mathcal{K}}\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a^*+1}\exp\left(-\frac{\lambda_k^*}{\sigma^2}\right)\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\Big/\left(\lambda_k^*\right)^{a^*}}{\sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\Big/\left(\lambda_{k'}^*\right)^{a^*}} 1\left(\sigma^2 > 0\right)$$

and, because $\mathrm{E}_{\sigma^2\sim\mathrm{IG}(a^*,\lambda_k^*)}\left(\sigma^2\right)=\frac{\lambda_k^*}{a^*-1}$ for $a^*>1$, we can get its marginal posterior expectation as

$$\mathrm{E}\left(\sigma^2|y\right)=\sum_{k\in\mathcal{K}}\mathrm{E}_{\Pi}\left(\sigma^2|y,k\right)\pi(k|y)=\frac{\sum_{k\in\mathcal{K}}\frac{\lambda_k^*}{a^*-1}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\Big/\left(\lambda_k^*\right)^{a^*}}{\sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\Big/\left(\lambda_{k'}^*\right)^{a^*}}$$

Predictive inference BMA plays an important role in the predictive inference.

Note 32. Let $z=(y_{n+1},...,y_{n+m})$ be a sequence of future outcomes. Then the predictive distribution $\mathrm{d}G(z|y)$ of z given the observables y has pdf/pmf

$$g(z|y)=\sum_{k\in\mathcal{K}}g(z|y,k)\pi(k|y)$$

where

$$g(z|y,k)=\int f(z|y,\theta_k,k)\pi(\theta_k|y,k)\mathrm{d}\theta_k$$

is the predictive pdf of z given model $\mathscr{M}_k$.

Example 33. The BMA predictive distribution for the Normal linear regression model has been calculated in Example 19, Eq. c11.

Note 34. By averaging out all models in this fashion, BMA provides better predictive ability as measured by the logarithmic scoring rule $-\log(\pi(A))$, than using any single model $\mathscr{M}_k$, conditional on $\mathcal{M}$. See the exercise below.

6 Discussion

In model uncertainty framework, we often meet two situations;

$\mathcal{M}$-close perspective: the encompassing Bayesian model includes the real data-generation process $R(\cdot)$. It is argued that this is not a realistic situation. The introduced model uncertainty framework is suitable in this situation.

$\mathcal{M}$-open perspective: the encompassing Bayesian model does not include the real data-generation process $R(\cdot)$. This is a more realistic situation. It is argued that the introduced model uncertainty framework is not suitable in this situation in the sense that $\pi(\mathscr{M}_k)$ does not make sense to exist.

What do we do? This is a rather open problem, and may depends on the application under consideration. Advise; Try to set-up an encompassing Bayesian model rich enough to promise, in some sense, that it acceptably approximate (at least) the real data generating process.

Question 35. *For practice address the Exercises 70, 71, and 72, from the Exercise sheet.*

Exercise 36. Show that for any $k\in\mathcal{K}$

$$-\mathrm{E}\left(\log\left(\sum_{k\in\mathcal{K}}\pi(z|y,k)\pi(k|y)\right)\right)\leq-\mathrm{E}\left(\log\left(\pi(z|y,k)\right)\right)$$

where the expectations are with respect to $\sum_{k\in\mathcal{K}}\pi(z|y,k)\pi(k|y)$.

Solution. It is

$$-\mathrm{E}\left(\log\left(\sum_{k\in\mathcal{K}}\pi(z|y,k)\pi(k|y)\right)\right)\leq-\mathrm{E}\left(\log\left(\pi(z|y,k)\right)\right)\Longleftrightarrow 0\leq\underbrace{-\mathrm{E}\left(\log\left(\frac{\pi(z|y,k)}{\sum_{k\in\mathcal{K}}\pi(z|y,k)\pi(k|y)}\right)\right)}_{=\mathrm{KL}(\pi(z|y)||\pi(z|y,k))}$$

A Solution to Example 16

Use the following identity

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

Solution. For simplicity, we suppress the indexing $\cdot_k$ denoting the sub-set of the regressors. It is

$$
\begin{aligned}
\pi\left(\beta_k,\sigma^2|y,k\right) \propto & f(y|\theta_k,k)\pi(\theta_k|k) \;= \mathrm{N}\left(y|\Phi\beta_k, I\sigma^2\right)\mathrm{N}\left(\beta_k|\mu_k,V_k\sigma^2\right)\mathrm{IG}\left(\sigma^2|a,\lambda\right)\\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(y-\Phi\beta)^{\top}(y-\Phi\beta)\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2\sigma^2}(\beta-\mu)^{\top}V^{-1}(\beta-\mu)\right)\\
&\qquad\times\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\left(-\frac{\lambda}{\sigma^2}\right)\\
&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{2\sigma^2}(y-\Phi\beta)^{\top}(y-\Phi\beta)-\frac{1}{2\sigma^2}(\beta-\mu)^{\top}V^{-1}(\beta-\mu)\right)\exp\left(-\frac{\lambda}{\sigma^2}\right)\\
&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{2\sigma^2}\left[(y-\Phi\beta)^{\top}(y-\Phi\beta)+(\beta-\mu)^{\top}V^{-1}(\beta-\mu)\right]\right)\exp\left(-\frac{\lambda}{\sigma^2}\right)\\
&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\mu^{*}\right)^{\top}\left(V^{*}\right)^{-1}\left(\mu-\mu^{*}\right)-\frac{1}{2\sigma^2}S^{*}\right)\exp\left(-\frac{\lambda}{\sigma^2}\right)\\
&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\mu^{*}\right)^{\top}\left(V^{*}\right)^{-1}\left(\mu-\mu^{*}\right)\right)\exp\left(-\frac{1}{\sigma^2}\left(\lambda+\frac{1}{2}S^{*}\right)\right)\\
&\propto\underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{d}{2}}\exp\left(-\frac{1}{2\sigma^2}\left(\beta-\mu^{*}\right)^{\top}\left(V_k^{*}\right)^{-1}\left(\mu-\mu^{*}\right)\right)}_{\propto\mathrm{N}(\beta_k|\mu^{*},V^{*}\sigma^2)}\times\underbrace{\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+a+1}\exp\left(-\frac{1}{\sigma^2}\left(\lambda+\frac{1}{2}S^{*}\right)\right)}_{\propto\mathrm{IG}(\sigma^2|a^{*},\lambda^{*})}\\
&\propto\mathrm{N}\left(\beta|\mu^{*},V^{*}\sigma^2\right)\mathrm{IG}\left(\sigma^2|a^{*},\lambda^{*}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
S^{*}&=\mu^{\top}V^{-1}\mu-(\mu^{*})^{\top}\left(V^{*}\right)^{-1}(\mu^{*})+y^{\top}y\\
a^{*}=\frac{n}{2}+a;\quad \lambda^{*}=\lambda+\frac{1}{2}S^{*};\qquad V^{*}&=\left(V^{-1}+\Phi^{\top}\Phi\right)^{-1};\qquad \mu^{*}=V^{*}\left(V^{-1}\mu+\Phi^{\top}y\right)
\end{aligned}
$$

The result in Example 16 results by indexing with $\cdot_k$ the quantities that depend on it. I.e.

$$
\begin{aligned}
\pi\left(\beta_k,\sigma^2|y,k\right)&=\mathrm{N}\left(\beta_k|\mu_k^{*},V_k^{*}\sigma^2\right)\mathrm{IG}\left(\sigma^2|a^{*},\lambda_k^{*}\right)\\
S_k^{*}&=\mu_k^{\top}V_k^{-1}\mu_k-(\mu_k^{*})^{\top}\left(V_k^{*}\right)^{-1}(\mu_k^{*})+y^{\top}y\\
a^{*}=\frac{n}{2}+a;\quad \lambda_k^{*}=\lambda+\frac{1}{2}S_k^{*};\qquad V_k^{*}&=\left(V_k^{-1}+\Phi_k^{\top}\Phi_k\right)^{-1};\qquad \mu_k^{*}=V_k^{*}\left(V_k^{-1}\mu_k+\Phi_k^{\top}y\right)
\end{aligned}
$$

Hence

$$
\Longrightarrow\begin{cases}\beta_k|y\sigma^2,k & \sim\mathrm{N}\left(\mu_k^{*},V_k^{*}\sigma^2\right)\\ \sigma^2|y,k & \sim\mathrm{IG}\left(a^{*},\lambda^{*}\right)\end{cases},\ \forall k\in\mathcal{K}
$$

B Solution to Example 17

Use the following identity

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

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

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

The conditional marginal likelihood $f(y|k)$ (or model evidence of $\mathscr{M}_k$) is

$$\begin{aligned}
f(y|k)=&\int_{\Theta_k}f(y|\theta_k,k)\pi(\theta_k|k)\mathrm{d}\theta_k\;=\;\int \mathrm{N}\left(y|\Phi\beta,I\sigma^2\right)\mathrm{N}\left(\beta|\mu,V\sigma^2\right)\mathrm{IG}\left(\sigma^2|a,\lambda\right)\mathrm{d}\beta\mathrm{d}\sigma^2\\
=&\int\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}(y-\Phi\beta)^\top(y-\Phi\beta)\right)\times\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\left(\frac{1}{|V|}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(\beta-\mu)^\top V^{-1}(\beta-\mu)\right)\\
&\qquad\times\frac{\lambda^a}{\Gamma(a)}\exp\left(-\frac{\lambda}{\sigma^2}\right)\mathrm{d}\beta\mathrm{d}\sigma^2\\
=&\left(\frac{1}{2\pi}\right)^{\frac{n}{2}+\frac{d}{2}}\left(\frac{1}{|V|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\int\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{\sigma^2}\left(\frac{(y-\Phi\beta)^\top(y-\Phi\beta)+(\beta-\mu)^\top V^{-1}(\beta-\mu)}{2}+\lambda\right)\right)\\
=&\left(\frac{1}{2\pi}\right)^{\frac{n}{2}+\frac{d}{2}}\left(\frac{1}{|V|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\int\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}S^*+\lambda\right)\right)\\
&\qquad\times\int\left[\exp\left(-\frac{1}{2}\frac{1}{\sigma^2}(\beta-\beta^*)^\top\left(V^*\right)^{-1}(\beta-\beta^*)\right)\mathrm{d}\beta\right]\mathrm{d}\sigma^2\\
=&\left(\frac{1}{2\pi}\right)^{\frac{n}{2}+\frac{d}{2}}\left(\frac{1}{|V|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\int\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+\frac{d}{2}+a+1}\exp\left(-\frac{1}{\sigma^2}\left(\frac{1}{2}S^*+\lambda\right)\right)\times\left[(2\pi)^{\frac{d}{2}}\left(\sigma^2\right)^{\frac{d}{2}}|V^*|^{\frac{1}{2}}\right]\mathrm{d}\sigma^2\\
=&\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{|V^*|}{|V|}\right)^{\frac{1}{2}}\frac{k^a}{\Gamma(a)}\frac{\Gamma\left(\frac{n}{2}+a\right)}{\left(\frac{1}{2}S^*+\lambda\right)^{\frac{n}{2}+a}}\;=\;\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{|V^*|}{|V|}\right)^{\frac{1}{2}}\frac{k^a}{\Gamma(a)}\frac{\Gamma\left(a^*\right)}{\left(\lambda^*\right)^{a^*}}
\end{aligned}$$

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

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

The result in Example 17 results by indexing with $\cdot_k$ the quantities that depend on it. I.e.

$$f(y|k)=\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{\lambda^a}{\Gamma(a)}\frac{\Gamma\left(a^*\right)}{\left(\lambda_k^*\right)^{a^*}}$$

$$S_k^*=\mu_k^\top V_k^{-1}\mu_k-(\mu_k^*)^\top\left(V_k^*\right)^{-1}(\mu_k^*)+y^\top y$$

$$a^*=\frac{n}{2}+a;\quad \lambda_k^*=\lambda+\frac{1}{2}S_k^*;\qquad V_k^*=\left(V_k^{-1}+\Phi_k^\top\Phi_k\right)^{-1};\qquad \mu_k^*=V_k^*\left(V_k^{-1}\mu_k+\Phi_k^\top y\right)$$

The marginal likelihood $f(y)$ is

$$f(y)=\sum_{k\in\mathcal{K}}f(y|k)\pi(k)\;=\;\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\lambda^a\frac{\Gamma\left(a^*\right)}{\Gamma(a)}\frac{1}{2^d}\sum_{k\in\mathcal{K}}\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}}$$

The marginal model posterior probability $\pi(k|y)$ of $\mathscr{M}_k$ is

$$\pi(k|y)=\frac{f(y|k)\pi(k)}{f(y)}=\quad\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}}\Bigg/\sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_{k'}^*\right)^{a^*}}$$

C Solution to Example 19

Use the following identity:

$$\begin{cases} y|\beta \sim \text{N}_m\left(\Phi\beta + c, \Sigma\right) \\ \\ \beta \sim \text{N}_d\left(\mu, V\right) \end{cases} \qquad \Longrightarrow \qquad y \sim \text{N}_m\left(\Phi\mu + c, \Sigma + \Phi V \Phi^\top\right)$$

Use the following property of the Student T distribution

$$\begin{cases} x|\xi \sim \text{N}_m(\mu, \Sigma\xi v) \\ \\ \xi \sim \text{IG}(\frac{v}{2}, \frac{1}{2}) \end{cases} \qquad \Longrightarrow \qquad x \sim \text{T}_m(\mu, \Sigma, v)$$

Solution. For simplicity, we suppress the indexing $\cdot_k$ denoting the sub-set of the regressors. The conditional predictive distribution $G(z|y,k)$ given model $\mathscr{M}_k$ and data y has pdf

$$g(z|y,k) = \int f(z|\theta_k, k)\pi(\theta_k|y,k)\text{d}\theta_k = \int \text{N}_m\left(z|\Phi^{\text{new}}\beta, I\sigma^2\right)\text{N}_d\left(\beta|\mu^*, V^*\sigma^2\right)\text{IG}\left(\sigma^2|a^*, \lambda^*\right)\text{d}\beta\text{d}\sigma^2$$

$$\begin{aligned}
&= \int\left[\int \text{N}_m\left(z|\Phi^{\text{new}}\beta, I\sigma^2\right)\text{N}_d\left(\beta|\mu^*, V^*\sigma^2\right)\text{d}\beta\right]\text{IG}\left(\sigma^2|a^*, \lambda^*\right)\text{d}\sigma^2 \\
&= \int \text{N}_m\left(z|\Phi^{\text{new}}\mu^*, \sigma^2\left[I + V^*\right]\right)\text{IG}\left(\sigma^2|a^*, \lambda^*\right)\text{d}\sigma^2 \\
\\
&= \int \text{N}_m\left(z|\Phi^{\text{new}}\mu^*, \sigma^2\left[I + V^*\right]\right)\text{IG}\left(\sigma^2|a^*, \lambda^*\right)\text{d}\sigma^2 \\
\\
&= \int \text{N}_m\left(z|\Phi^{\text{new}}\mu^*, \sigma^2\frac{\lambda^*}{\lambda^*}\frac{2a^*}{2a^*}\left[I + V^*\right]\right)\text{IG}\left(\sigma^2|a^*, \lambda^*\right)\text{d}\sigma^2 \\
&\overset{\xi=\frac{\sigma^2}{2\lambda^*}\sim\text{IG}\left(a^*,\frac{1}{2}\right)}{=} \int \text{N}_m\left(z|\Phi^{\text{new}}\mu^*, \xi\lambda^*\frac{2a^*}{a^*}\left[I + V^*\right]\right)\text{IG}\left(\xi|a^*, \frac{1}{2}\right)\text{d}\xi \\
&= \int \text{N}_m\left(z|\Phi^{\text{new}}\mu^*, \xi 2a^*\frac{\lambda^*}{a^*}\left[I + V^*\right]\right)\text{IG}\left(\xi|\frac{a^*}{2}, \frac{1}{2}\right)\text{d}\xi \\
&= \text{T}_m\left(z|\Phi^{\text{new}}\mu^*, \left[I + V^*\right]\frac{\lambda^*}{a^*}, 2a^*\right)
\end{aligned}$$

The result in Example 19 results by indexing with $\cdot_k$ the quantities that depend on it. I.e.

$$g(z|y,k) = \text{T}_m\left(z|\Phi^{\text{new}}\mu_k^*, \left[I + V_k^*\right]\frac{\lambda_k^*}{a^*}, 2a^*\right)$$

The marginal predictive distribution $G\left(z|y\right)$ given data y has pdf/pmf

$$\begin{aligned}
g(z|y) &= \sum_{k\in\mathcal{K}} g(z|y,k)\pi(k|y) \\
&= \sum_{k\in\mathcal{K}} \text{T}_m\left(z|\Phi^{\text{new}}\mu_k^*, V_k^*\frac{\lambda_k^*}{a^*}, 2a^*\right)\left(\frac{|V_k^*|}{|V_k|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_k^*\right)^{a^*}} \Bigg/ \sum_{k'\in\mathcal{K}}\left(\frac{|V_{k'}^*|}{|V_{k'}|}\right)^{\frac{1}{2}}\frac{1}{\left(\lambda_{k'}^*\right)^{a^*}}
\end{aligned}$$

Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2019

Handout 16: Hierarchical Bayesian model[a]

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

Aim

To be able to specify and analyze a Hierarchical Bayesian, as well as to extend previously introduces concepts in the Hierarchical Bayes framework.

Basic reading list:

- Robert, C. (2007, Sections 10.1-10.3). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- Robert, C. P., & Reber, A. (1998). Bayesian modelling of a pharmaceutical experiment with heterogeneous responses. Sankhyā: The Indian Journal of Statistics, Series B, 145-160. (`https://www.jstor.org/stable/pdf/25053027.pdf`)

R-scripts:

- `https://github.com/georgios-stats/Bayesian_Statistics/blob/master/HandoutsSupplementary/HierarchicalBayes/HierarchicalBayesPharmaceutical.R`

[a] Author: Georgios P. Karagiannis.

1 Hierarchical Bayesian Model

A Bayesian model can be hierarchical due to the modeling of the observations or due to the decomposition of the prior information. A Bayesian hierarchical model involves several levels / layers of conditional prior distributions.

Definition 1. A hierarchical Bayes model is a Bayesian statistical model with sampling distribution $x \sim f(y|\theta)$ and prior $\theta \sim \pi(\theta)$, where the prior distribution $\pi(\theta)$ is decomposed in conditional distributions. The Bayesian model is

$$\begin{cases} y & \sim f(y|\theta), \text{ is the sampling distribuition} & \begin{cases} y \sim f(y|\theta) \end{cases} \\ \theta & \sim \pi(\theta) \;\; \text{is the marginal prior which is specified as} & \begin{cases} \theta & \sim \pi_1(\theta|\phi_1) & \text{1st level prior} \\ \phi_1|\phi_2 & \sim \pi_2(\phi_1|\phi_2) & \text{2nd level hyper-prior} \\ \vdots & & \\ \phi_j|\phi_{j+1} & \sim \pi_{j+1}(\phi_j|\phi_{j+1}) & j\text{th level hyper-prior} \\ \vdots & & \\ \phi_{m-1}|\phi_m & \sim \pi_m(\phi_{m-1}|\phi_m) & m\text{th level hyper-prior} \end{cases} \end{cases}$$

$$\text{and we write} \quad \begin{cases} y|\theta & \sim f(y|\theta) \\ \theta|\phi_1 & \sim \pi_1(\theta|\phi_1) \\ \phi_1|\phi_2 & \sim \pi_2(\phi_1|\phi_2) \\ \vdots & \\ \phi_j|\phi_{j+1} & \sim \pi_{j+1}(\phi_j|\phi_{j+1}) \\ \vdots & \\ \phi_{m-1} & \sim \pi_m(\phi_{m-1}|\phi_m) \end{cases} \tag{1}$$

The joint distribution $p(y, \theta, \phi_1, ..., \phi_j, ...\phi_{m-1})$ has pdf

$$p(y, \theta, \phi_1, ..., \phi_j, ...\phi_{m-1}) = f(y|\theta)\pi_1(\theta|\phi_1)\pi_2(\phi_1|\phi_2)\pi_3(\phi_2|\phi_3)...\pi(\phi_{m-1}|\phi_m)$$

The marginal prior distribution $\pi(\theta)$ has pdf

$$\pi(\theta) = \int_{\Phi_1\times\Phi_{m-1}} \pi_1(\theta|\phi_1)\pi_2(\phi_1|\phi_2)\mathrm{d}\phi_1\pi_3(\phi_2|\phi_3)\mathrm{d}\phi_2...\pi(\phi_{m-1}|\phi_m)\mathrm{d}\phi_{m-1}.$$

The parameters $\phi_j \in \Phi_j$ are called random hyper-parameters of level j for $1 \leqslant j \leqslant m-1$.

Remark 2. Hierarchical Bayesian model is simply a special type of Bayesian model, where

$$\begin{cases} y|\theta & \sim f(y|\theta) \\ \theta|\phi & \sim \pi(\theta|\phi) \\ \phi|\phi_m & \sim \pi(\phi|\phi_m) \end{cases} \tag{2}$$

for $\phi = (\phi_1, ..., \phi_{m-1})$, and ϕ_m fixed hyper-parameter.

Remark 3. The Bayesian model with sampling distribution $y \sim f(y|\theta)$ and prior $\theta \sim \pi(\theta)$, can be recovered from 2 by marginalizing the prior as

$$\pi(\theta) = \int_{\Phi} \pi(\theta|\phi)\pi(\phi|\phi_m)\mathrm{d}\phi = \int_{\Phi_1\times\Phi_{m-1}} \pi(\theta|\phi_1)\pi(\phi_1|\phi_2)\mathrm{d}\phi_1...\pi(\phi_{m-1}|\phi_m)\mathrm{d}\phi_{m-1}, \tag{3}$$

where ϕ_m is just a fixed hyper-parameter. This reduction shows that hierarchical modelings are indeed included in the Bayesian paradigm.

Note 4. A hierarchical Bayesian model can be used as a mean to specify more diverse priors. This is achieved by setting ϕ to be a random hyper-parameter with $\phi|\phi_m \sim \pi_2(\phi|\phi_m)$ instead of setting ϕ to have a fixed value. See Example 5.

Example 5. Consider the 'Challenger O-ring' example from the Computer practicals. Let y_i denote the presence of a defective O-ring in the ith flight (0 for absence, and 1 for presence).

Assume that y_i can be modeled as observations generated independently from a Bernoulli distribution with with parameter p_i. Here, p_i denotes the relative frequency of defective O-rings at flight i. We study if 'presence of a defective O-ring' (y) depends on the 'temperature' (t), or the 'pressure' (s).

Let t_i denote the temperature (in F) in the platform, and let s_i denote the Leak check pressure (in PSI) before the ith flight. Here are some possible models of interest:

$$\begin{aligned} \mathscr{M}^{I} :\ p(t; \beta_{\mathscr{M}^{I}}, \mathscr{M}^{I}) =& \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \quad &; \mathscr{M}^{IV} :\ p(t; \beta_{\mathscr{M}^{IV}}, \mathscr{M}^{IV}) =& \frac{\exp(\beta_0+\beta_1 t+\beta_2 s)}{1+\exp(\beta_0+\beta_1 t+\beta_2 s)} \\ \mathscr{M}^{II} :\ p(t; \beta_{\mathscr{M}^{II}}, \mathscr{M}^{II}) =& \frac{\exp(\beta_0+\beta_1 t)}{1+\exp(\beta_0+\beta_1 t)} \quad &; \mathscr{M}^{V} :\ p(t; \beta_{\mathscr{M}^{V}}, \mathscr{M}^{V}) =& \frac{\exp(\beta_0+\beta_1 t+\beta_2 s+\beta_3 ts)}{1+\exp(\beta_0+\beta_1 t+\beta_2 s+\beta_3 ts)} \\ \mathscr{M}^{III} :\ p(t; \beta_{\mathscr{M}^{III}}, \mathscr{M}^{III}) =& \frac{\exp(\beta_0+\beta_2 s)}{1+\exp(\beta_0+\beta_2 s)} \quad & \text{etc...} \end{aligned}$$

The Bayesian hierarchical model under consideration is:

$$\begin{cases} y_i|\theta \sim f(y_i|\theta) \; :: & \left\{ y_i|\mathscr{M}, \beta_{\mathscr{M}} \sim \text{Br}\left(y_i | \frac{\exp(x_i^\top \beta_{\mathscr{M}})}{1+\exp(x_i^\top \beta_{\mathscr{M}})} \right), \quad \text{for, } i = 1, ..., n \right. \\ \\ \theta|\phi_1 \sim \pi(\theta|\phi_1) \; :: & \begin{cases} \beta_j|\mathscr{M} \sim (1-\gamma_j) 1_0(\beta_j) + \gamma_j \text{N}(\beta_j|\mu_0, \sigma_0^2) \;\; j = 1, ..., d \\ \\ \begin{cases} \mathscr{M} & = (\gamma_1, ..., \gamma_d) \\ \gamma_j|\varpi & \sim \text{Br}(\varpi), \;\; j = 1, ..., d \end{cases} \\ \\ \end{cases} \\ \\ \phi_1|\phi_2 \sim \pi(\phi_1|\phi_2) :: & \left\{ \varpi \quad \sim \text{Be}(a_0, b_0) \right. \end{cases}$$

where $\theta = (\mathscr{M}, \beta_{\mathscr{M}})$, $\phi_1 = \varpi$, and $\phi_2 = (a_0, b_0)$. Above, in the prior we considered an extra level of uncertainty by considering $\varpi \sim \text{Be}(a_0, b_0)$.

- Here we added an additional level of uncertainty, and set $\varpi \sim \text{Be}(a_0, b_0)$ which creates a more diverse prior model, compared to the computer practical handout example where we had set $\varpi = 0.5$.

Now the joint probability distribution has pdf

$$p(y, \beta_{\mathscr{M}}, \mathscr{M}, \varpi) = \underbrace{\prod_{i=1}^{n} \text{Br}\left(y_i | \frac{\exp(x_i^\top \beta_{\mathscr{M}})}{1+\exp(x_i^\top \beta_{\mathscr{M}})} \right)}_{f(y|\theta)} \underbrace{\prod_{i=1}^{n} \left((1-\gamma_j) 1_0(\beta_j) + \gamma_j \text{N}(\beta_j|\mu_0, \sigma_0^2) \right) \prod_{i=1}^{n} \text{Br}(\gamma_i|\varpi)}_{\pi(\theta|\phi_1)} \underbrace{\text{Be}(\varpi|a_0, b_0)}_{\pi(\phi_1|\phi_2)}$$

Note 6. A hierarchical Bayesian model can be used when the sampling distribution or the prior distributions justify a certain structure. See Example 7.

Example 7. Robert and Reber (1998) considers an experiment under which rats are intoxicated by a substance, then treated by either a placebo or a drug. (Sess: `https://www.jstor.org/stable/pdf/25053027.pdf`)

Statistical model ($f(y|\theta)$): The model associated with this experiment is a linear additive model effect: given x_{ij} , y_{ij} and z_{ij}, jth responses of the ith rat at the control, intoxication and treatment stages, respectively. The statistical model was specified such as that $(1 \leqslant i \leqslant I)$

$$\begin{aligned} x_{i,j} &\sim \text{N}(\theta_i, \sigma_c^2) & , 1 \leqslant j \leqslant J_i^c \\ y_{i,j} &\sim \text{N}(\theta_i + \delta_i, \sigma_a^2) & , 1 \leqslant j \leqslant J_i^a, \\ z_{i,j} &\sim \text{N}(\theta_i + \delta_i + \xi_i, \sigma_t^2) & , 1 \leqslant j \leqslant J_i^t, \end{aligned}$$

where θ_i is the average control measurement, δ_i the average intoxication effect and ξ_i the average treatment effect for the ith rat, the variances of these measurements being constant for the control, the intoxication and the treatment effects. An additional (observed) variable is w_i, which is equal to 1 if the rat is treated with the drug, and 0 otherwise.

Prior model $\pi(\theta|\phi)$: The different individual averages are related through a common (conjugate) prior distribution,

$$\theta_i \sim \text{N}(\mu_\theta, \sigma_\theta^2), \qquad \delta_i \sim \text{N}(\mu_\delta, \sigma_\delta^2), \qquad \xi_i|w_i \sim \begin{cases} \text{N}(\mu_P, \sigma_P^2) & , w_i = 0 \\ \text{N}(\mu_D, \sigma_D^2) & , w_i = 1 \end{cases}$$

$$\sigma_c \sim \pi(\sigma_c) \propto \frac{1}{\sigma_c}, \qquad \sigma_a \sim \pi(\sigma_a) \propto \frac{1}{\sigma_a}, \qquad \sigma_t \sim \pi(\sigma_t) \propto \frac{1}{\sigma_t}, \tag{4}$$

This modeling seems to describe the natural phenomenon realistically enough, in the sense the responses x_{ij} , y_{ij} and z_{ij}

Hyper-priors $\pi(\phi|\phi_m)$: For the higher levels of prior ($\pi(\phi|\phi_m)$ in Eq 2), they considered improper (Jeffrey's) hyper-priors.

$$(\mu_\theta,\sigma_\theta)\sim\pi(\mu_\theta,\sigma_\theta)\propto\frac{1}{\sigma_\theta},\qquad(\mu_\delta,\sigma_\delta)\sim\pi(\mu_\delta,\sigma_\delta)\propto\frac{1}{\sigma_\delta},\qquad(\mu_P,\sigma_P)\sim\pi(\mu_P,\sigma_P)\propto\frac{1}{\sigma_P},\tag{5}$$

$$(\mu_D,\sigma_D)\sim\pi(\mu_D,\sigma_D)\propto\frac{1}{\sigma_D}.\tag{6}$$

The priors in lines (4), (5) and (6) are improper non-informative priors. One could have have specify proper priors, like Normal-Inverse Gamma which are conjugate, however in that case he/she should have to specify the values for the fixed hyper-parameters.

As improper priors are specified, one need to study under what conditions the above improper priors lead to a proper (well defined) posterior –we omit this step here...

Note 8. A particularly appealing aspect of hierarchical models is that they allow for conditioning on all levels, and this easy decomposition of the posterior. Consider the Bayesian hierarchical model (2) a parametric model $f(y|\theta)$ with a hierarchical prior $\theta\sim\pi_1(\theta|\phi)$, and $\phi\sim\pi(\phi)$. The posterior distribution of θ is

$$\pi(\theta|y)=\int_\Phi\pi(\theta|y,\phi)\pi(\phi|y)\mathrm{d}\phi\tag{7}$$

where

$$\begin{aligned}\pi(\theta|y,\phi)&=\frac{f(y|\theta)\pi_1(\theta|\phi)}{f_1(y|\phi)}\\ f_1(y|\phi)&=\int_\Theta f(y|\theta)\pi_1(\theta|\phi)\mathrm{d}\theta\\ \pi(\phi|y)&=\frac{f_1(y|\phi)\pi_2(\phi)}{f(y)}\\ f(y)&=\int_\Theta f_1(y|\phi)\pi_2(\phi)\mathrm{d}\phi\end{aligned}$$

Remark 9. It has important consequences in terms of the computation of Bayes estimators, though, since it shows that $\pi(\theta|y)$ can be simulated by generating, first, ϕ from $\pi(\phi|y)$ and then θ from $\pi(\theta|y,\phi)$, if these two conditional distributions are easier to work with. (Snapshot from Term 2).

Note 10. Hierarchical decomposition (2) may facilitate the computation of intractable posterior moments. Let h be a function $h:\Theta\rightarrow\mathbb{R}$, then

$$\mathrm{E}_\pi(h(\theta)|y)=\mathrm{E}_\pi\left(\mathrm{E}_\pi\left(h(\theta)|y,\phi\right)|y\right).$$

If $\mathrm{E}_\pi(h(\theta)|y)=\int h(\theta)\pi(\theta|y)\mathrm{d}\theta$ is intractable and θ has high dimensionality, one could possibly try to specify the prior decomposition $\pi(\theta)=\int_\Phi\pi_1(\theta|\phi)\pi_2(\phi|\phi_m)\mathrm{d}\phi$ in (3) such that $\mathrm{E}_\pi\left(h(\theta)|y,\phi\right)$ can be computed analytically, and ϕ has low conventionality. In that case one would have to compute the equivalent but lower dimensional (and hence easier) integral $\mathrm{E}_\pi\left(\mathrm{E}_\pi\left(h(\theta)|y,\phi\right)|y\right)=\int\mathrm{E}_\pi\left(h(\theta)|y,\phi\right)\pi(\phi|y)\mathrm{d}\phi$.

Example 11. Regarding the fully hierarchical model (1), the full conditionals distributions of each element of $\vartheta=(\theta,\phi_1,...,\phi_{m-1})\in\Theta\times\Phi$ are given as:

$$\pi(\vartheta_j|y,\vartheta_{-j})=\pi(\vartheta_j|y,\vartheta_{j-1},\vartheta_{j+1})$$

with the convention

$$\vartheta_j = \begin{cases} \theta & , j = 1 \\ \phi_{j-1} & , j = 2, ..., m \\ \phi_m & , j = m \end{cases}$$

and $\vartheta_{-j} = (\vartheta_1, ..., \vartheta_{j-1}, \vartheta_{j+1}, ...\vartheta_m)$.

Proof. Straightforward by using the Bayesian theorem. □

Example 12. (Cont...) You may use

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

The joint posterior pdf of $\vartheta = (\theta_{1:I}, \delta_{1:I}, \xi_{1:I}, \sigma_c^2, \sigma_a^2, \sigma_t^2, \sigma_\theta^2, \sigma_\delta^2, \sigma_P^2, \sigma_D^2, \mu_\theta, \mu_\delta, \mu_P, \mu_D)$ given obs. x, y, z is

$$\begin{aligned}\pi(\vartheta|x,y,z) \propto & \prod_{i=1}^{I}\left[\exp\left(-\frac{(\theta_i-\mu_\theta)^2}{2\sigma_\theta^2} - \frac{(\delta_i-\mu_\delta)^2}{2\sigma_\delta^2}\right)\prod_{j=1}^{J_i^c}\exp\left(-\frac{(x_{i,j}-\theta_i)^2}{2\sigma_c^2}\right) \times \prod_{j=1}^{J_i^a}\exp\left(-\frac{(y_{i,j}-\theta_i-\delta_i)^2}{2\sigma_a^2}\right)\right. \\ & \left. \times \prod_{j=1}^{J_i^t}\exp\left(-\frac{(z_{i,j}-\theta_i-\delta_i-\xi_i)^2}{2\sigma_t^2}\right) \times \prod_{w_i=0}\exp\left(-\frac{(\xi_i-\mu_P)^2}{2\sigma_P^2}\right)\prod_{w_i=0}\exp\left(-\frac{(\xi_i-\mu_D)^2}{2\sigma_D^2}\right)\right] \\ & \times \sigma_c^{-\sum_i J_i^c - 1}\sigma_a^{-\sum_i J_i^a - 1}\sigma_t^{-\sum_i J_i^t - 1}\sigma_\theta^{I-1}\sigma_\delta^{I-1}\sigma_D^{I_D-1}\sigma_P^{I_P-1}.\end{aligned}$$

The joint posterior distributions is not of standard form, and its pdf is intractable. However the full conditionals are of standard form. For instance, the full conditional posterior distribution density

$$\begin{aligned}&\pi(\delta_{1:I}|x_{\text{all}}, y_{\text{all}}, z_{\text{all}}, \theta_{1:I}, \xi_{1:I}, \sigma_c^2, \sigma_a^2, \sigma_t^2, \sigma_\theta^2, \sigma_\delta^2, \sigma_P^2, \sigma_D^2, \mu_\theta, \mu_\delta, \mu_P, \mu_D) \\ &\quad \propto \prod_{i=1}^{I}\left[\exp\left(-\frac{(\delta_i-\mu_\delta)^2}{2\sigma_\delta^2}\right) \times \prod_{j=1}^{J_i^a}\exp\left(-\frac{(y_{i,j}-\theta_i-\delta_i)^2}{2\sigma_a^2}\right) \times \prod_{j=1}^{J_i^t}\exp\left(-\frac{(z_{i,j}-\theta_i-\delta_i-\xi_i)^2}{2\sigma_t^2}\right)\right] \\ &\quad \propto \prod_{i=1}^{I}\left[\exp\left(-\frac{(\delta_i-\mu_\delta)^2}{2\sigma_\delta^2} - \sum_{j=1}^{J_i^a}\frac{(\delta_i-(y_{i,j}-\theta_i))^2}{2\sigma_a^2} - \sum_{j=1}^{J_i^t}\frac{(\delta_i-(z_{i,j}-\theta_i-\xi_i)}{2\sigma_t^2}\right)\right] \\ &\quad \propto \prod_{i=1}^{I}\left[\exp\left(-\frac{(\delta_i-\mu_{\delta,i}^*)^2}{2\left(\sigma_{\delta,i}^*\right)^2} + \text{const...}\right)\right] \propto \prod_{i=1}^{I}\left[\exp\left(-\frac{(\delta_i-\mu_{\delta,i}^*)^2}{2\left(\sigma_{\delta,i}^*\right)^2} + \text{const...}\right)\right] \\ &\quad \propto \prod_{i=1}^{I}\text{N}\left(\delta_i|\mu_{\delta,i}^*, \left(\sigma_{\delta,i}^*\right)^2\right)\end{aligned}$$

with

$$\delta_i|\text{rest}, ... \overset{\text{ind}}{\sim} \text{N}\left(\mu_{\delta,i}^*, \left(\sigma_{\delta,i}^*\right)^2\right), \ \forall i = 1, ..., n$$

where

$$\left(\sigma_{\delta,i}^*\right)^2 = \left(\frac{1}{\sigma_\delta^2} + \frac{1}{\sigma_a^2}J_i^a + \frac{1}{\sigma_t^2}J_i^t\right)^{-1};\quad \mu_{\delta,i}^* = \left(\sigma_{\delta,i}^*\right)^2\left(\frac{\mu_\delta}{\sigma_\delta^2} + \frac{\sum_{j=1}^{J_i^a} y_{i,j} - J_i^a\theta_i}{\sigma_a^2} + \frac{\sum_{j=1}^{J_i^a} y_{i,j} - J_i^t\theta_i - J_i^t\xi_i}{\sigma_t^2}\right)$$

Notice that δ_i are a postriori independent given all the resp unknown parameters $\big(\theta_{1:I}, \xi_{1:I}, \sigma_c^2, \sigma_a^2, \sigma_t^2, \sigma_\theta^2, \sigma_\delta^2, \sigma_P^2, \sigma_D^2, \mu_\theta, \mu_\delta, \mu_P, \mu_D\big)$. Notice that the prior $\delta_i \sim \text{N}(\mu_\delta, \sigma_\delta^2)$ in Example 7 is conditional conjugate prior of δ_i.

Try to compute the rest

$$
\begin{array}{rrc}
\pi(\theta_{1:I}|\text{rest}, ...) \sim ? \,; & \pi(\sigma_t^2|\text{rest}, ...) \sim ? & etc... \\
\pi(\xi_{1:I}|\text{rest}, ...) \sim ? \,; & \pi(\sigma_\theta^2|\text{rest}, ...) \sim ? & \\
\pi(\sigma_c^2|\text{rest}, ...) \sim ? \,; & \pi(\sigma_\delta^2|\text{rest}, ...) \sim ? & \\
\pi(\sigma_a^2|\text{rest}, ...) \sim ? \,; & \pi(\sigma_P^2|\text{rest}, ...) \sim ? &
\end{array}
$$

- See the solutions in: Robert, C. P., & Reber, A. (1998). Bayesian modelling of a pharmaceutical experiment with heterogeneous responses. Sankhy: The Indian Journal of Statistics, Series B, 145-160. from the link (`https://www.jstor.org/stable/pdf/25053027.pdf`).
- I have an R script with a demo in `https://github.com/georgios-stats/Bayesian_Statistics/blob/master/HandoutsSupplementary/HierarchicalBayes/HierarchicalBayesPharmaceutical.R`

2 Non-identifiability issue

A parametric model for which an element of the parametrisation is redundant is said to be non-identified. Let Bayesian model $(f(y|\theta), \pi(\theta))$, where $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$, and assume that the parametric model does not depend on θ_1; i.e. $f(y|\theta_1, \theta_2) = f(y|\theta_2)$. The fact that the likelihood does not depend on θ_1 suggests that y does not provide information about θ_1 directly.

Bayesian analysis of a non-identified model is always possible if a suitable prior $\Pi(\theta_1, \theta_2)$ on all the parameters is specified. For instance, if one specifies a priori that learning the value of θ_2 may change his belief about θ_1, via $\pi(\theta_1|\theta_2) \neq \pi(\theta_1)$.

Factorize the prior distribution as $\pi(\theta_1, \theta_2) = \pi(\theta_1|\theta_2)\pi(\theta_2)$. Then, we have the following PDF/PMF

$$\begin{aligned}
&\pi(\theta_1, \theta_2|y)\propto f(y|\theta_1, \theta_2)\pi(\theta_1, \theta_2) = f(y|\theta_2)\pi(\theta_1|\theta_2)\pi(\theta_2) \implies \\
&\pi(\theta_1, \theta_2|y) = \pi(\theta_2|y)\pi(\theta_1|\theta_2) \implies \\
&\pi(\theta_1|y, \theta_2) = \pi(\theta_1|\theta_2) \qquad (8)\\
&\quad\pi(\theta_2|y) = \frac{f(y|\theta_2)\pi(\theta_2)}{\int_{\Theta_2} f(y|\theta_2)\pi(\theta_2)\mathrm{d}\theta_2} \quad . \\
&\quad\pi(\theta_1|y) = \int_{\Theta_2} \pi(\theta_1|\theta_2)\pi(\theta_2)\mathrm{d}\theta_2 \qquad (9)
\end{aligned}$$

Here, θ_1 is said to be non-identifiable parameter from the data y, because y provides no direct information about θ_1. Inference about θ_1 based on marginal posterior $\pi(\theta_1|y)$ depends on y but the information provided about θ_1 comes indirectly through the marginal posterior of θ_2, see (9). Equivalently, (9) implies that y provides no information about θ_1 given θ_2.

If we a priori specify that learning the value of θ_2 does not change our belief about θ_1 $\pi(\theta_1|\theta_2) = \pi(\theta_1)$, then (9) becomes $\pi(\theta_1|y) = \pi(\theta_1)$ and hence data y provide no information about θ_1 at all.

Example 13. (A simple example) Consider a production process where manufactured items are classified as acceptable, with probability $1-\theta_1-\theta_2$, or defective, with probability $\theta_1+\theta_2$. Assume that there are two exclusive assignable causes of failure that occur with probabilities θ_1 andθ_2, respectively, $\theta_1, \theta_2 > 0$ with $\theta_1 + \theta_2 < 1$.

- For a random sample y, the statistical model for the total number of defective items may be considered as $r_n \sim \text{Bn}(n, \theta_1 + \theta_2)$.
- The data are fully informative for $\theta_1 + \theta_2$, however the individual parameters of interest, (θ_1, θ_2), are non-identifiable.
- The problem may be mitigated if a suitable a priori on θ is assigned, e.g., $\pi(\theta_1, \theta_2) = \text{Di}_2(\theta|a)$.

  Hint: Dirichlet distribution, $\theta \sim \text{Di}_k(a)$ has PDF,

$$\text{Di}_k(\theta|a) = \frac{\Gamma(\sum_{j=1}^{k+1} a_j)}{\prod_{j=1}^{k+1} \Gamma(a_j)} \prod_{j=1}^{k} \theta_j^{a_j-1}(1 - \sum_{j=1}^{k} \theta_j)1(\{\sum_{j=1}^{k} \theta_j \in (0,1)\} \cap \{\theta_j \in (0,1)\})$$

  and $a_j > 0$ for all $j = 1, ..., k + 1$. It is a generalization of Beta distribution in many dimensions.