Bayesian Statistics III/IV (MATH3341/4031) Michaelmas term, 2021

Handout 5: Sufficiency and Exponential family of distributions

Lecturer & author: 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.

1 Sufficiency

It is often of interest to simplify the Bayesian model by reducing the complexity of the observables (data-set) $y = (y_1, ..., y_n)$, where $y \in \mathcal{Y}$. This can be for computational purposes: the data-set may involve large number of observations (large n) or high-dimensional observables (high d); or this can be for theoretical purposes: study how the information for the experiment affects the Bayesian model.

Key concepts to address data-set complexity are

- Summary statistic (or statistic) aims to summarize part or all the relevant information supplied by the sample.
- Sufficient statistic aims to summarize the whole of relevant information supplied by the sample.

1.1 Summary statistics

A statistic (or summary statistic) is a function 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 $t : \mathcal{Y} \to \mathbb{R}^k$) 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 $F(t|\theta)$ (of course it depends on the experiment $e \in \mathcal{E}$as well but conditioning is omitted here) labeled by unknown parameter $\theta \in \Theta$, which is determined by sampling distribution $F(y|\theta)$. Given a prior distribution $\Pi(\theta)$ on θ, the posterior distribution $\Pi(\theta|t)$ of θ given $t = t(y)$ can be calculated from the Bayesian theorem. Hence:

$$\begin{cases} t|\theta & \sim F(t|\theta) \\ \theta & \sim \Pi(\theta) \end{cases}$$

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 $t : \mathcal{Y} \rightarrow \mathcal{T}$ is efficient statistic for θ (in the Frequentist sense) if the conditional distribution $F(y|t,\theta)$ 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

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

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 $t : \mathcal{Y} \rightarrow \mathcal{T}$ is (parametric) sufficient for θ if and only if for any prior distribution $\Pi(\theta)$ with pdf/pmf $\pi(\theta)$ we get

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

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 implies that the derived posterior can be determined by the kernel of the likelihood while it is invariant to the residue.

Theorem 12. *Let $t : \mathcal{Y} \rightarrow \mathcal{T}$ be a statistic. Then t is a parametric sufficient statistic for θ 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 45 in Exercise sheet □

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

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

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 15 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 $(y = (y_1, ..., y_n),\ y \sim F(y|\theta),\ \theta \sim \Pi(\theta))$ if and only if the predictive distribution $G(y_f|t = t')$ of a future outcome y_f given t and the predictive distribution $G(y_f|t = t')$ 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 sufficient 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. See Exercise 44 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 $F(y|\theta)$ 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

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

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 $\{c_j\}$,

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

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

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

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*

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

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

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

$$= \left(\prod_{i=1}^{n} u(y_i)\right)(g(\theta))^n \exp\left(\sum_{j=1}^{k} c_j\phi_j(\theta)\sum_{i=1}^{n} h_j(y_i)\right)$$

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 $y_i|\theta \sim \text{Br}(\theta)$ iid for $i = 1, ..., n$. Then,

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

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

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

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

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

$$= (\frac{1}{2\pi})^{\frac{1}{2}}(\frac{1}{\sigma^2})^{\frac{1}{2}}\exp(-\frac{1}{2}\frac{1}{\sigma^2}y_i^2 + \frac{\mu}{\sigma^2}y_i - \frac{1}{2}\frac{\mu^2}{\sigma^2}) = (\frac{1}{2\pi})^{\frac{1}{2}}(\frac{1}{\sigma^2})^{\frac{1}{2}}\exp(-\frac{1}{2}\frac{\mu^2}{\sigma^2})\exp(-\frac{1}{2}\frac{1}{\sigma^2}y_i^2 + \frac{\mu}{\sigma^2}y_i)$$

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

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

Example 27. (Uniform distribution)Let $y_i|\theta \sim \text{U}(\theta_1, \theta_2)$. Then

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

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

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 28. *Let $y_i \sim \text{Ef}_k(u, g, h, \phi, \theta, c)$ 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*

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

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 $L(y|\theta)$ for the observed y. 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 29. 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 $L(y|\theta)$; but just the experimental information. Other information relevant to the statistical analysis, such as prior information may exist.

Remark 30. 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 31. *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 32. (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 e 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 e_1 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 e_2 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 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 $= \text{Bn}(r \le 3|n = 12, \theta = 1/2) = \sum_{i=0}^{3} \text{Bn}(r|n = 12, \theta = 1/2) = 0.073 > 5\%$ (I do not reject H_0) (ii.) while in case M_2: we get p-value $= \text{Nb}(n \ge 12|\theta = 1/2) = 1 - \sum_{n=0}^{11} \text{Nb}(n|r = 9, \theta = 1/2) = 0.032 < 5\%$ (I reject H_0) !!! The two models lead to quite different conclusions (in Frequentist stats), and are in contradiction to the Likelihood Principle.

Question 33. *For practice try the Exercises 41, 42 from the Exercise Sheet.*