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

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

Lecturer & author: 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.
- O'Hagan, A., & Forster, J. J. (2004; Paragraphs 6.55-6.60). Kendall's advanced theory of statistics, volume 2B: Bayesian inference (Vol. 2). Arnold.
- O'Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. The American Statistician, 73(sup1), 69-81.

1 Prior elicitation[1]

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 $\Pi(\theta)$ 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...

Note 3. An informative summary presenting 'Biases in Elicitation', 'Elicitation Protocols', and real examples is given in the following not examinable document:

- O'Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. The American Statistician, 73, 69-81.

[1] O'Hagan, A., & Forster, J. J. (2004; Paragraphs 6.55-6.60).

2 Maximum entropy priors[2]

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...

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

$$m_j = \mathrm{E}_\Pi(h_j(\theta)), \qquad j = 1, ..., k \tag{1}$$

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 $\pi(\cdot)$ 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.[3]
- *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*

$$KL(\pi\|u) = H(\pi, u) - H(\pi)$$
$$\Longleftrightarrow$$
$$\overbrace{\underbrace{\int_\Theta \log\left(\frac{\pi(\theta)}{u(\theta)}\right) d\Pi(\theta)}_{=E_\Pi\left(\log\left(\frac{\pi(\theta)}{u(\theta)}\right)\right)}}^{KL(\pi\|u)=} = \overbrace{\underbrace{-\int_\Theta \log(u(\theta)) d\Pi(\theta)}_{=E_\Pi(\log(u(\theta)))}}^{H(\pi,u)=} - \overbrace{\underbrace{\left[-\int_\Theta \log(\pi(\theta)) d\Pi(\theta)\right]}_{=-E_\Pi(\log(\pi(\theta)))}}^{H(\pi)=} \tag{3}$$

[2] I do not actually demonstrate the concept as it is originally developed, but I give a similar explanation.

[3] can be proved by log-sum and Jensen's inequalities

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.

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:

$$\mathrm{H}(\pi) = \log(|\Theta|) - \int_{\Theta} \left(\log(\pi(\theta)) - \log\left(\frac{1}{|\Theta|}\right)\right) \mathrm{d}\Pi(\theta) = \underbrace{\log(|\Theta|)}_{=\text{constant}} - \mathrm{KL}(\pi\|u)$$

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

$$\begin{aligned} &\text{minimise: } \mathrm{KL}(\pi\|u) && \text{approximate the reference measure } u(\theta) && (4) \\ &\text{subject to: } \mathrm{E}_{\Pi}(h_1(\theta)) = m_1 && \text{satisfy partial prior info} \\ &\qquad\vdots && \vdots \\ &\qquad \mathrm{E}_{\Pi}(h_k(\theta)) = m_k && \text{satisfy partial prior info} && (5) \\ &\qquad \int_{\Theta} \mathrm{d}\Pi(\theta) = 1 && \text{hopefully redive a ptoper prior} && (6) \end{aligned}$$

Based on the method of Lagrange multipliers, solving (4)-(6) is equivalent to minimizing,

$$Q(\pi, \lambda) = \int_{\Theta} \log\left(\frac{\pi(\theta)}{u(\theta)}\right) \mathrm{d}\Pi(\theta) + \sum_{j=1}^{k} \lambda_j \left[\int_{\Theta} h_j(\theta)\mathrm{d}\Pi(\theta) - m_j\right] + \lambda_0 \left[\int_{\Theta} \mathrm{d}\Pi(\theta) - 1\right] \tag{7}$$

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*

$$\pi(\theta) \qquad = g(\lambda)u(\theta)\exp\left(\sum_{j=1}^{k} \lambda_j h_j(\theta)\right) \qquad \propto u(\theta)\exp\left(\sum_{j=1}^{k} \lambda_j h_j(\theta)\right). \tag{8}$$

where,

$$g(\lambda)^{-1} = \int_{\Theta} u(\theta)\exp\left(\sum_{j=1}^{k} \lambda_j h(\theta)\right) d\theta < +\infty$$

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 A, 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

$$\pi(\theta) \propto \pi^{(\mathrm{J})}(\theta) \exp\left(\sum_{j=1}^{k} \lambda_j h_j(\theta)\right) \tag{9}$$

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

$$\pi(\theta) \propto \exp\left(\sum_{j=1}^{k} \lambda_j h_j(\theta)\right)$$

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[4] 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

$$\pi(\theta) \propto \exp(\lambda_2 \theta^2 + \lambda_1 \theta) \propto \exp\left(-\frac{1}{2} \frac{\left(\theta - \left(-\frac{\lambda_1}{2\lambda_2}\right)\right)^2}{-\frac{1}{2\lambda_2}}\right) \propto \mathrm{N}\left(\theta | \, \mu = -\frac{\lambda_1}{2\lambda_2}, \sigma^2 = -\frac{1}{2\lambda_2}\right)$$

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.

[4] This is the actual 'Maximum entropy prior' , however we will call the same those in (8) and (9).

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.

A Appendix

Proof. Sketch of the proof of Theorem 15.

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 $\tau : \Theta \to \mathbb{R}$ of sufficiently small norm. It is

$$\begin{aligned} Q(\pi(\theta) + a\tau(\theta)) = \int_\Theta \log(\frac{\pi(\theta) + a\tau(\theta)}{u(\theta)})(\pi(\theta) + a\tau(\theta))\mathrm{d}\theta + \sum_{j=1}^k \lambda_j [\int_\Theta h_j(\theta)(\pi(\theta) + a\tau(\theta))\mathrm{d}\theta - m_j] \\ + \lambda_0 [\int_\Theta (\pi(\theta) + a\tau(\theta))\mathrm{d}\theta - 1] \qquad \Longrightarrow \\ \frac{\partial}{\partial a} Q(\pi(\theta) + a\tau(\theta)) = \int_\Theta \left(\tau(\theta) + \log(\frac{\pi(\theta) + a\tau(\theta)}{u(\theta)})\tau(\theta) \right) \mathrm{d}\theta + \sum_{j=1}^k \lambda_j \int_\Theta h_j(\theta)\tau(\theta)\mathrm{d}\theta + \lambda_0 \int_\Theta \tau(\theta)\mathrm{d}\theta \\ = \int_\Theta \left(1 + \log(\frac{\pi(\theta) + a\tau(\theta)}{u(\theta)}) + \sum_{j=1}^k \lambda_j h_j(\theta) + \lambda_0 \right) \tau(\theta)\, \mathrm{d}\theta \end{aligned}$$

Hence, the condition

$$\frac{\partial}{\partial a} Q(\pi(\theta) + a\tau(\theta))|_{a=0} = 0$$

reduces to

$$\int_\Theta (\log(\frac{\pi(\theta)}{u(\theta)}) + \sum_{j=1}^k \lambda_j h_j(x) + (\lambda_0 + 1))\tau(\theta)\mathrm{d}\theta = 0$$

which implies

$$\pi(\theta) = \underbrace{\left(\frac{1}{\exp(\lambda_0 + 1)} \right)}_{=g(\lambda)} u(\theta) \exp(\sum_{j=1}^k \lambda_j h_j(\theta)), \tag{10}$$

by setting $\lambda_j \leftarrow -\lambda_j$ for $j = 1, ..., k$. By applying the normalizing constraints, we get,

$$g(\lambda)^{-1} = \int_\Theta u(\theta) \exp(\sum_{j=1}^k \lambda_j h_j(\theta))\mathrm{d}\theta.$$

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

$$\begin{aligned} \mathrm{E}(y|\psi) = \frac{\mathrm{d}}{\mathrm{d}\psi} b(\psi) \Longleftrightarrow \\ [\mathrm{E}(y|\psi)]_j = \frac{\partial}{\partial \psi_j} b(\psi), \text{ for all, } j = 1, ..., k \Longleftrightarrow \\ \mathrm{E}(h_j(\theta)|\lambda) = \frac{\partial \lambda_j}{\partial \psi_j} \frac{\partial}{\partial \lambda_j} (-\log(g(\lambda))), \text{ for all, } j = 1, ..., k \Longleftrightarrow \\ m_j = -\frac{\partial}{\partial \lambda_j} \log(g(\lambda)), \text{ for all, } j = 1, ..., k \end{aligned}$$

□