III - From prior information to prior distribution

The choice of the prior distribution is probably the most delicate part of Bayesian analysis because

1. The choice of $\pi(\theta)$ can have a large impact on the posterior distribution (and thus on all the inference derived from it).
2. There is no (decision) theory to choose $\pi(\theta)$.

Three approaches can be distinguished:

- Subjective priors. Prior information is available (expert knowledge, previous experiments, etc.). Typically, $\pi(\theta)$ is obtained by using this information to (i) select a parametric family of densities and (ii) determine the corresponding parameters.
- Conjugate priors. Limited prior information is available. The choice of the parametric form of $\pi(\theta)$ is made for ease of computations while the corresponding parameters are determined from some prior information.
- Objective priors. Prior information is not available, or too sparse to be taken into account. One must find a way to express this lack of information.

Conjugate prior distributions

Definition 3.1 *A parametric family $\mathcal{F}$ of distributions is said to be conjugate for the model $\{f(\cdot|\theta), \theta \in \Theta\}$ if every prior distribution in $\mathcal{F}$ yields a posterior distribution for this model that is still in $\mathcal{F}$; that is,*

$$\pi \in \mathcal{F} \Rightarrow \pi(\cdot|x) \in \mathcal{F}, \quad \forall x \in \mathcal{X}.$$

The main advantage of using a conjugate prior distribution is therefore that the posterior distribution has a known expression and, usually, classical Bayesian estimators (notably the posterior mean) have a closed form expression.

However, the existence of a conjugate prior distribution is (mostly) limited to the case where $\{f(\cdot|\theta), \theta \in \Theta\}$ is an exponential family of distributions.

Remark: When the model of interest $\{f(\cdot|\theta), \theta \in \Theta\}$ does not admit a conjugate prior, conjugate prior distributions may still be useful to obtain a posterior distribution $\pi(\theta|x)$ which is "easy" to approximate using Markov Chain Monte Carlo methods. This is in particular the case for hierarchical models (see Chapter 9).

Example: The univariate Gaussian model

Let $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times \mathbb{R}_{>0}$.

Proposition 3.1 *Let $(\alpha_0, \beta_0, \kappa_0) \in \mathbb{R}^3_{>0}$, $\mu_0 \in \mathbb{R}$, $n \in \mathbb{N}_{>0}$ and consider the Bayesian statistical model where the prior distribution $\pi(\theta)$ is such that*

$$\mu|\sigma^2 \sim \mathcal{N}_1(\mu_0, \kappa_0^{-1}\sigma^2), \quad \sigma^2 \sim \textit{Inv-Gamma}(\alpha_0, \beta_0)$$

and where, with $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$f(x|\theta) = \prod_{k=1}^{n} \frac{1}{\sigma}\phi\left(\frac{x_k - \mu}{\sigma}\right), \quad \theta \in \Theta.$$

Then, the posterior distribution $\pi(\theta|x)$ is such that

$$\mu|(\sigma^2, x) \sim \mathcal{N}_1(\mu_n, \kappa_n^{-1}\sigma^2), \quad \sigma^2|x \sim \textit{Inv-Gamma}(\alpha_n, \beta_n)$$

with

$$\alpha_n = \alpha_0 + \frac{n}{2}, \quad \beta_n = \beta_0 + \sum_{k=1}^{n} \frac{(x_k - \bar{x}_n)^2}{2} + \frac{n\kappa_0}{n + \kappa_0}\frac{(\bar{x}_n - \mu_0)^2}{2}$$

$$\kappa_n = \kappa_0 + n, \quad \mu_n = \frac{\kappa_0\mu_0 + n\bar{x}_n}{\kappa_0 + n}.$$

Proof: See Problem Sheet 1, Problem 1.

The non-informative approach

When no prior information is available it may still be worth employing Bayesian approaches because they enjoy some classical optimality criteria (e.g. admissibility of Bayes estimators) and because sometimes computing Bayesian estimators is simpler than computing frequentist estimators.

Below we will see two types of non-informative prior distributions:

1. Laplace's prior
2. Jeffreys prior.

Definition 3.2 *A prior density $\pi(\theta)$ is said to be improper if it does not integrate to a finite value; that is, if*

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta = +\infty.$$

Often, non-informative prior densities are improper but the resulting Bayesian model is considered as valid provided that the corresponding posterior density (as defined by Bayes formula) is proper; i.e.

$$\int_\Theta \pi(\theta) f(x|\theta)\,\mathrm{d}\theta < +\infty.$$

In this case the estimator δ^π is called generalized Bayes estimator because typically the Bayes risk is infinite when an improper prior is used (i.e. $r(\pi) = +\infty$) and thus δ^π is not a Bayes estimator as defined in Chapter 2.

Laplace's prior

Laplace, who was the first to use non-informative techniques (recall the last example of Chapter 1), defines a non-informative prior density as a prior density $\pi(\theta)$ that puts equal value on every $\theta \in \Theta$; that is

$$\pi(\theta) \propto 1, \quad \forall \theta \in \Theta.$$

Note that if Θ is unbounded, the prior density $\pi(\theta)$ is improper.

Example: *Consider the multivariate Gaussian model*

$$X \sim \mathcal{N}_d(\theta, I_d), \quad \theta \in \mathbb{R}^d.$$

Then, the Laplace's prior yields a proper posterior density since in this case

$$\theta|(X = x) \sim \mathcal{N}_d(x, I_d), \quad \forall x \in \mathbb{R}^d.$$

Under a quadratic loss function, δ^π is unique and defined by

$$\delta^\pi(x) = \mathbb{E}_\pi[\theta|x] = x, \quad x \in \mathbb{R}^d.$$

However, $r(\pi) = +\infty$ and thus δ^π is not a Bayes estimator. In addition, δ^π is not admissible when $d > 2$ (see Theorem 2.7).

Remark: In the above example, the inference derived under the Laplace's prior corresponds to the "limiting inference" where the prior variance goes to infinity in a conjugate analysis. This observation often holds true when a conjugate prior distribution is available.

Laplace's prior: re-parametrization issues

The most important problem with the Laplace's prior is the problem of invariance under re-parametrization.

Namely, if $\pi(\theta) = 1$ for all $\theta \in \Theta$ and $g : \Theta \to \Theta$ is a one-to-one and continuously differentiable mapping, then the parameter $\eta = g^{-1}(\theta)$ has prior density

$$\pi^*(\eta) = \left| \det \left[\frac{\partial}{\partial \eta} g(\eta) \right] \right| \pi(g(\eta)) = \left| \det \left[\frac{\partial}{\partial \eta} g(\eta) \right] \right|$$

by the Jacobian formula.

Therefore, the prior density $\pi^*(\eta)$ is, in general, not constant (and hence not non-informative in the sense of Laplace) although the information on η is the same as on θ!

The Jeffreys prior

The Jeffreys prior is based on the idea that a non-informative prior density is a prior density derived entirely from the model $\{f(\cdot|\theta), \theta \in \Theta\}$ since this is the only available information.

Definition 3.3 *For a given parametric model $\{f(\cdot|\theta), \theta \in \Theta\}$, the Jeffreys prior is defined by the density*

$$\pi_J(\theta) \propto \{\det[I(\theta)]\}^{1/2}, \quad \theta \in \Theta$$

with $I(\theta)$ the Fisher Information matrix; that is,

$$I(\theta) = \mathbb{E}_\theta \left[\frac{\partial \log f(X|\theta)}{\partial \theta} \frac{\partial \log f(X|\theta)}{\partial \theta^T} \right] = -\mathbb{E}_\theta \left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta \partial \theta^T} \right]$$

where the second inequality holds only under some conditions and where $\mathbb{E}_\theta$ denotes the expectations under $f(x|\theta)\mathrm{d}x$.

Compared to the Laplace's prior, the Jeffreys prior has the advantage to be invariant by re-parametrisation, as shown in the next result.

Proposition 3.2 *Assume that θ has prior density $\pi_J(\theta)$ and let $g : \Theta \to \Theta$ be a one-to-one and continuously differentiable mapping. Then, $\eta = g^{-1}(\theta)$ has prior density*

$$\pi^*(\eta) \propto \left| \det \left[\frac{\partial}{\partial \eta} g(\eta) \right] \right| \pi_J(g(\eta)) = \left\{ \det[\tilde{I}(\eta)] \right\}^{1/2}$$

where

$$\tilde{I}(\eta) = \mathbb{E}_{g(\eta)} \left[\frac{\partial \log f(X|g(\eta))}{\partial \eta} \frac{\partial \log f(X|g(\eta))}{\partial \eta^T} \right].$$

Proof: Use the fact that $\det(AB) = \det(A)\det(B)$.

Example: The Jeffreys prior for the Binomial model is the Beta$(1/2, 1/2)$ distribution.

Some remarks concerning the Jeffreys prior

- The Fisher information matrix $I(\theta)$ is related to the curvature of the Kullback-Leiber divergence and hence provides a measure of the ability of the model to discriminate between θ and $\theta + \mathrm{d}\theta$ (see the Appendix for more details on this).

  Therefore, the Jeffreys prior puts
  - a small prior probability on the values of θ for which the observations provide little information.
  - a large prior probability on the values of θ for which the observations provide a "significant" amount of information.

- The Jeffreys prior is often improper.

- The Jeffreys prior is not always computable.

Jeffreys prior as a bias reduction technique

Let $X^{(n)} = (X_1, \ldots, X_n)$, where the X_i's are i.i.d. with common distribution $\tilde{f}(\cdot|\theta_0)$ for some $\theta_0 \in \Theta \subset \mathbb{R}^d$ and where $\{\tilde{f}(\cdot|\theta), \theta \in \Theta\}$ is an exponential family model with canonical parameter θ.

Let $f(X^{(n)}|\theta) = \prod_{i=1}^n \tilde{f}(X_i|\theta)$, $\pi(\theta)$ be a prior distribution for θ such that $\pi(\theta) > 0$ for all $\theta \in \Theta$ and $\hat{\theta}_n$ be the maximum likelihood estimator (MLE) of θ_0.

Then, it is well-known that

- $\|\mathbb{E}_{\theta_0}[\hat{\theta}_n] - \theta_0\| = \mathcal{O}(n^{-1})$
- $\|\mathbb{E}_{\theta_0}[\delta^{\pi}_{\text{MAP}}(X^{(n)})] - \theta_0\| = \mathcal{O}(n^{-1})$ (with $\delta^{\pi}_{\text{MAP}}$ as in Chapter 2).

However, it can be shown that[a]

$$\|\mathbb{E}_{\theta_0}[\delta^{\pi_J}_{\text{MAP}}(X^{(n)})] - \theta_0\| = o(n^{-1}) \tag{1}$$

so that, under the Jeffreys prior, the MAP outperforms the MLE in term of bias!

[a] see Firth, D. (1993). "Bias reduction of maximum likelihood estimates". *Biometrika*, 80(1), 27-38.

General remarks on the objective approach

- A more basic way of obtaining a non-informative prior density is to take a vague prior; that is, a prior distribution with a 'large' variance.
- Automated use of non-informative prior distributions does not make sense in any settings. In particular, this is 'sub-optimal' if some prior information is available.
- Careless use of improper prior densities is dangerous. In particular, it may be that the posterior density does not define a proper probability distribution.
- Generalized Bayes estimators do not share the same optimality properties than Bayes estimators. In particular, the formers may not be admissible and may have very poor properties (see e.g. the next example).
- Improper prior densities should be avoided for hypothesis testing (see Chapter 4).

Generalized Bayes estimators and admissibility

The goal of this example is to show that, although Bayes estimators are (under weak conditions) admissible, generalized Bayes estimators can have very poor properties.

Let $f(\cdot|\theta)$ be the p.d.f. of the $\mathcal{N}_d(\theta, I_d)$ distribution, $\theta \in \mathbb{R}^d$ and $\pi(\theta) \propto 1$ so that

$$\theta|(X = x) \sim \mathcal{N}_d(x, I_d), \quad \forall x \in \mathbb{R}^d.$$

Assuming a quadratic loss function, the corresponding generalized Bayes estimator $\delta^\pi(x) := \mathbb{E}_\pi[\theta|x] = x$ is reasonable as mentioned in Chapter 2 (i.e. it is minimax and has good asymptotic properties).

Assume now that the parameter of interest is $\eta = \|\theta\|^2$. Then, under the above prior distribution, the generalized Bayes estimator of η is

$$\mathbb{E}_\pi[\eta|x] = \|x\|^2 + d.$$

However, as shown in the next result, this estimator has very poor properties:

Proposition 3.3 *For $c \in \mathbb{R}$, let $\delta_c : \mathbb{R}^d \to \mathbb{R}_+$ be the estimator of η defined by*

$$\delta_c(\tilde{x}) = \|\tilde{x}\|^2 + c, \quad \tilde{x} \in \mathbb{R}^d.$$

Then, under a quadratic loss function,

$$R(\eta, \delta_{-d}) < R(\eta, \delta_d), \quad \forall \eta \in \mathbb{R}_+$$

and $\sup_{\eta \in \mathbb{R}_+} R(\eta, \delta_{-d}) = \sup_{\eta \in \mathbb{R}_+} R(\eta, \delta_d) = +\infty$.

Proof: See Problem Sheet 2, Problem 3.

Alternative approaches

As mentioned above, the use of improper prior densities as non-informative prior may be problematic.

Below we describe two approaches that aim at reducing the impact of the prior distribution on the inference.

Let $\{\tilde{\pi}(\theta|\lambda),\, \lambda \in \Lambda\}$ be a family of probability density functions on Θ.

- **Hierarchical modelling:** Instead of choosing a fixed $\lambda \in \Lambda$, we can reduce the influence of the prior distribution by assigning a prior distribution $\pi_2(\lambda)$ on Λ and then by considering the prior distribution $\pi(\theta)$ defined by

$$\pi(\theta) = \int_\Lambda \tilde{\pi}(\theta|\lambda)\pi_2(\lambda)\mathrm{d}\lambda, \quad \theta \in \Theta.$$

- **Empirical Bayes:** The idea is to use the observation $x \in \mathcal{X}$ to select a value $\lambda_x \in \Lambda$ and then to use the prior distribution $\pi(\theta)$ defined by

$$\pi(\theta) = \tilde{\pi}(\theta|\lambda_x), \quad x \in \mathcal{X}.$$

  Typically, λ_x is estimated from the marginal distribution

$$m(x|\lambda) = \int_\Theta f(x|\theta)\tilde{\pi}(\theta|\lambda)\mathrm{d}\theta$$

  by taking, for instance, $\lambda_x \in \mathrm{argmax}_{\lambda\in\Lambda}\, m(x|\lambda)$.

  Remark: This approach is not fully Bayesian since the observation x is used twice (i.e. in the prior and in the likelihood).

Appendix: KL divergence and Fisher information matrix

We first recall the definition of the Kullback-Leiber (KL) divergence.

Definition 3.4 *The KL divergence* $KL(\cdot|\cdot) : \Theta \times \Theta \to \mathbb{R}$ *is defined by*

$$KL(\theta'|\theta) = \mathbb{E}_{\theta'}\left[\log \frac{f(X|\theta')}{f(X|\theta)}\right] = \int_{\mathcal{X}} \log \frac{f(x|\theta')}{f(x|\theta)} f(x|\theta')\mathrm{d}x, \quad (\theta', \theta) \in \Theta^2.$$

The next result provides two basic properties of the KL divergence which imply that the quantity $KL(\theta'|\theta)$ can be used to measure the difference between the probability distributions $f(x|\theta')\mathrm{d}x$ and $f(x|\theta)\mathrm{d}x$.

Proposition 3.4 *For any* $(\theta, \theta') \in \Theta^2$, $KL(\theta'|\theta) \geq 0$ *and, if the model* $\{f(\cdot|\theta),\, \theta \in \Theta\}$ *is well identified,* $KL(\theta'|\theta) = 0$ *if and only of* $\theta' = \theta$.

Proof: Let $(\theta', \theta) \in \Theta^2$. Then,

$$\mathbb{E}_{\theta'}\left[\log \frac{f(X|\theta')}{f(X|\theta)}\right] = \mathbb{E}_{\theta'}\left[-\log \frac{f(X|\theta)}{f(X|\theta')}\right] \geq -\log \mathbb{E}_{\theta'}\left[\frac{f(X|\theta)}{f(X|\theta')}\right] = \log(1)$$

where the inequality uses Jensen's inequality. The second part of the proposition is trivial.

Remark: If $KL(\theta'|\theta)$ is a measure of the difference between the probability distributions $f(x|\theta')\mathrm{d}x$ and $f(x|\theta)\mathrm{d}x$, the KL divergence is not a true metric since it does not satisfy the triangular inequality and is, in general, not symmetric.

Proposition 3.5 below shows that the Fisher information matrix is related to the curvature of the KL divergence and hence provides a measure of the ability of the model to discriminate between θ and $\theta + \mathrm{d}\theta$.

Appendix: KL divergence and Fisher information matrix (end)

Proposition 3.5 *Under some regularity conditions (that notably ensure that we can swap integration and differentiation operations)*

$$KL(\theta'|\theta) = \frac{1}{2}(\theta' - \theta)^T I(\theta)(\theta' - \theta) + o(\|\theta' - \theta\|^2) \quad as\ \|\theta' - \theta\|^2 \to 0.$$

Proof: Using a Taylor expansion of order 2, we have (as $\|\theta' - \theta\|^2 \to 0$)

$$KL(\theta'|\theta) = (\theta' - \theta)^T \frac{\partial KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}}\Big|_{\tilde{\theta}=\theta} + \frac{1}{2}(\theta' - \theta)^T \frac{\partial^2 KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta} \partial \tilde{\theta}^T}\Big|_{\tilde{\theta}=\theta}(\theta' - \theta) + o(\|\theta' - \theta\|^2)$$

where

$$\frac{\partial KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}} = \frac{\partial}{\partial \tilde{\theta}} \int_{\mathcal{X}} f(x|\theta) \log \frac{f(x|\theta)}{f(x|\tilde{\theta})} \mathrm{d}x = \int_{\mathcal{X}} \frac{\partial}{\partial \tilde{\theta}} \Big(\log \frac{f(x|\theta)}{f(x|\tilde{\theta})} \Big) f(x|\theta) \mathrm{d}x$$
$$= -\int_{\mathcal{X}} \frac{\partial \log f(x|\tilde{\theta})}{\partial \tilde{\theta}} f(x|\theta) \mathrm{d}x$$

so that

$$\frac{\partial KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta}}\Big|_{\tilde{\theta}=\theta} = -\int_{\mathcal{X}} \frac{\partial f(x|\theta)}{\partial \theta} \mathrm{d}x = -\frac{\partial}{\partial \theta} \int_{\mathcal{X}} f(x|\theta) \mathrm{d}x = 0.$$

Moreover,

$$\frac{\partial^2 KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta} \partial \tilde{\theta}^T} = -\int_{\mathcal{X}} \frac{\partial^2 \log f(x|\tilde{\theta})}{\partial \tilde{\theta} \partial \tilde{\theta}^T} f(x|\theta) \mathrm{d}x$$

so that

$$\frac{\partial^2 KL(\tilde{\theta}|\theta)}{\partial \tilde{\theta} \partial \tilde{\theta}^T}\Big|_{\tilde{\theta}=\theta} = -\int_{\mathcal{X}} \frac{\partial^2 \log f(x|\theta)}{\partial \tilde{\theta} \partial \tilde{\theta}^T}\Big|_{\tilde{\theta}=\theta} f(x|\theta) \mathrm{d}x = I(\theta).$$

The proof is complete.