# To choose or not to choose a prior. That's the question!

Fatemeh Ghaderinezhad
Universiteit Gent

Joint work with Christophe Ley (Universiteit Gent)

26th Annual Meeting of the RSSB, Ovifat, Belgium

October 17-19 2018

# Outline

1. The effect of the priors
2. Quantification of the effects of two distinct priors
3. Comparison of various priors for various distributions

## Outline

1. The effect of the priors
2. Quantification of the effects of two distinct priors
3. Comparison of various priors for various distributions

# Bayesian paradigm

We wish to perform inference on parameters $\theta$ from a statistical model.

The Bayesian treats the unobservable parameters $\theta$ as random quantities $\Theta$.

When no data are available, a prior distribution is used to quantify our knowledge about the parameter : the prior reflects the statistician's understanding of $\theta$.

When data are available, we update our prior knowledge using the conditional distribution of parameters, given the data ; combining prior and data we get a posterior distribution.

The transition from the prior to the posterior is possible via the Bayes formula

$$\pi(\theta|x) = \frac{\pi(x|\theta)\pi(\theta)}{\int \pi(x|\theta)\pi(\theta)d\theta}$$

according to Gelman et al. (2013, 3th edition), the essential characteristic of Bayesian methods is the explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis, and it can be expressed in three steps as follows

1. Setting up a full probability model : choosing a joint probability distribution for observable and unobservable quantities consistent with knowledge about the underlying scientific problem and the data collection process
2. Conditioning on observed data
3. Evaluating the fit of the model and the implication of the resulting posterior distribution

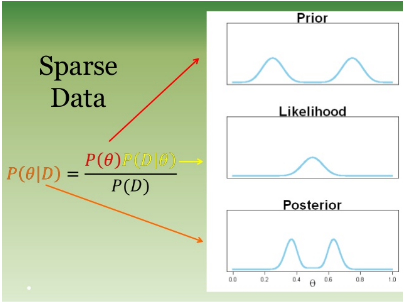
The second step is involving computational methodology and the third step is a delicate balance of techniques and judgements.

- If we consider different prior choices, how can we assess the difference between them on the basis of the final outcome ?

## Choosing the priors

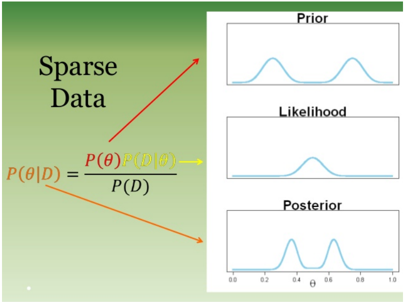
There are different ways of choosing the priors :

- Conjugate priors : these are priors such that the posterior belongs to the same family, and the effect of multiplication by the likelihood is only to "update" the prior parameters
- Jeffrey's prior : which is proportional to $\mid I(\theta) \mid^{\frac{1}{2}}$ the determinant of the Fisher information matrix $I(\theta)$
- Invariant priors, i.e. priors which are invariant under some group action
- Maximum entropy priors



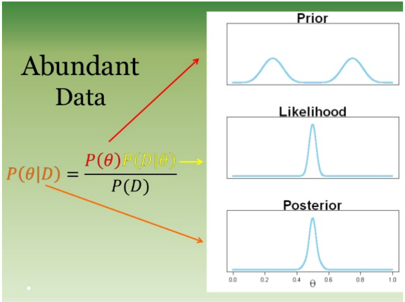
Sparse Data

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$$



Sparse Data

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$$



Abundant Data

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$$

There are some cases in which choice of prior does not have so much importance. Some quotations :

- with well-identified parameters and large sample sizes, reasonable choices of prior distributions will have minor effects on posterior inferences
- when we have enough data, the effect of the prior we choose will be small compared to the data. In that case we find that we can get very similar posteriors despite starting from quite different priors
- For the binomial case it is advised to choose the prior a member of the beta family ; it doesn't matter very much which one you chose

Diaconis and Freedman (1986, Ann. Statist.) have provided formal conditions under which the effect of the prior vanishes asymptotically.

# Goals

**Question** : How can the effect of priors be assessed in Bayesian analysis ? As more and more data are collected, will the posterior distributions derived with different priors be very similar ?

**Aim** : detecting the effect of the prior, at *fixed* sample size, by means of explicit bounds on the Wasserstein distance between posteriors based on two distinct priors.

Recently Ley, Reinert and Swan (2017, Ann. Appl. Probab.) have considered this question about the effect of priors between nested densities by considering the Uniform distribution as a non-informative prior.

We extend their methodology to the setting of any two desired priors.

# Outline

1. The effect of the priors
2. Quantification of the effects of two distinct priors
3. Comparison of various priors for various distributions

# Ingredients

Consider iid observations $(X_1, X_2, \ldots, X_n)$ with parameter of interest $\theta \in \Theta$ and the sampling density (likelihood) $\ell(x;\theta)$.

Consider two distinct (possibly improper) prior densities $p_1(\theta)$ and $p_2(\theta)$ for $\theta$ which leads to two resulting posterior distributions for $\theta$ :

$$p_i(\theta;x) = \kappa_i(x)p_i(\theta)\ell(x;\theta), i = 1, 2,$$

where $\kappa_1(x)$ and $\kappa_2(x)$ are normalizing constants.

The densities $p_1(\theta;x)$ and $p_2(\theta;x)$ are nested, meaning that one support is included in the other. It means we are allowed to write $\mathcal{I}_2 \subset \mathcal{I}_1$ and $p_2(\theta;x) = \frac{\kappa_2(x)}{\kappa_1(x)}\rho(\theta)p_1(\theta;x)$ with $\rho(\theta) = \frac{p_2(\theta)}{p_1(\theta)}$.

We write $\Theta_i \sim p_i(\theta;x)$ and its corresponding distribution as $P_i$, $i = 1, 2$.

# Direct comparison between two priors

### Theorem (Ghaderinezhad - Ley)

*Consider* $\mathcal{H}$ *the set of Lipschitz-1 functions on* $\mathbb{R}$*. Assume that* $\theta \mapsto \rho(\theta)$ *is differentiable on* $I_2$ *and satisfies (i)* $\mathrm{E}[(\Theta_1 - \mu_1)\rho(\Theta_1)] < \infty$*, (ii)* $\left(\rho(\theta) \int_{a_1}^{\theta} (h(y) - \mathrm{E}[h(\Theta_1)])p_1(y;x)dy\right)'$ *is integrable for all* $h \in \mathcal{H}$ *and (iii)* $\lim_{\theta \to a_2, b_2} \rho(\theta) \int_{a_1}^{\theta} (h(y) - \mathrm{E}[h(\Theta_1)])p_1(y;x)dy = 0$ *for all* $h \in \mathcal{H}$*. Then*

$$|\mu_1 - \mu_2| = \frac{|\mathrm{E}[\tau_1(\Theta_1;x)\rho'(\Theta_1)]|}{\mathrm{E}[\rho(\Theta_1)]} \leq d_{\mathcal{W}}(P_1, P_2) \leq \frac{\mathrm{E}[\tau_1(\Theta_1;x)|\rho'(\Theta_1)|]}{\mathrm{E}[\rho(\Theta_1)]} \quad (1)$$

*and, if the variance of* $\Theta_1$ *exists,*

$$\mid \mu_1 - \mu_2 \mid \leq d_{\mathcal{W}}(P_1, P_2) \leq ||\rho'||_\infty \frac{Var(\Theta_1)}{\mathrm{E}[\rho(\Theta_1)]} \quad (2)$$

*where* $|| \cdot ||_\infty$ *stands for the infinity norm.*

- If $\rho$ is a monotone increasing or decreasing function, the following Corollary comes out as

## Corollary

*If in addition to the conditions of the theorem we assume that the ratio $\rho$ is monotone increasing or decreasing, then*

$$d_{\mathcal{W}}(P_1, P_2) = \frac{\mathrm{E}[\tau_1(\Theta_1; x)|\rho'(\Theta_1)|]}{\mathrm{E}[\rho(\Theta_1)]}.$$

- The main result by Ley, Reinert and Swan (2017, Ann. Appl. Probab.) is a special case of the aforementioned theorem by fixing one prior to be the Uniform data-only prior.

# The Stein kernel

We still need to define an important concept, namely the *Stein kernel* $\tau_P$ of a distribution $P$ with density $p$. It is defined as follows :

- Let $P$ be a probability distribution with mean $\mu$, and let $X \sim P$. A *Stein kernel* of $P$ is a random variable $\tau_P(X)$ such that

  $$\mathrm{E}[\tau_P(X)\phi'(X)] = \mathrm{E}[(X-\mu)\phi(X)]$$

  for all differentiable $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ for which the expectation $\mathrm{E}[(X-\mu)\phi(X)]$ exists.
- If $P$ has interval support with closure $[a, b]$, we have :

  $$\tau_P(x) = \frac{1}{p(x)} \int_a^x (\mu - y)p(y)dy$$

  is the unique Stein kernel of $P$.
- Moreover the following properties of the Stein kernel are immediate consequences of its definition :

  for all $x \in \mathbb{R}$ we have that $\tau_P(x) \geq 0$ and $\mathrm{E}[\tau_P(X)] = Var[X]$.

# Outline

1. The effect of the priors
2. Quantification of the effects of two distinct priors
3. Comparison of various priors for various distributions

# Normal Model

Recall : The $Normal(\mu, \sigma^2)$ density is

$$f(x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} \right\}$$

Here the data $x = (x_1, x_2, \ldots, x_n)$ have sampling density

$$\ell(x;\sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} \right\}$$

The parameter of interest is $\theta = \sigma^2$ and the mean is supposed fixed.

- Jeffreys' prior, which is proportional to $\frac{1}{\sigma^2}$, leads to the posterior

$$\begin{aligned} p_1(\sigma^2;x) &\propto \frac{1}{\sigma^2} \times (\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} \right\} \\ &\propto (\sigma^2)^{-\frac{n}{2}-1} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2} \right\} \end{aligned}$$

which we can see to have an $Inverse\ Gamma$ distribution with parameters $(\frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2)$.

- The Inverse Gamma distribution with parameters $(\alpha, \beta)$ as conjugate prior leads to the posterior $IG(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \beta)$.

Comparing those two priors leads to

$$\frac{|(\frac{n}{2} - 1)\beta - \frac{\alpha}{2}\sum_{i=1}^{n}(x_i - \mu)^2|}{(\frac{n}{2} + \alpha - 1)(\frac{n}{2} - 1)} \leq d_{\mathcal{W}}(P_1, P_2) \leq \alpha\frac{\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2}{(\frac{n}{2} + \alpha - 1)(\frac{n}{2} - 1)} + \frac{\beta}{\frac{n}{2} - 1}$$

which is of order $O(1/n)$.

For Normal data, the distance between Jeffreys' and conjugate prior increases when the sum of observations increases.

# Application in the real world

- Our motivation for studying various priors for $\sigma^2$ comes from a study about storm depth multiplier model to represent rainfall uncertainty by Kavetski (2006), where the errors appear under multiplicative form and are assumed to be normal.

- They fix the mean $\mu$ but state that "less is understood about the degree of rainfall uncertainty, i.e., the multiplier variance" and therefore study various priors for $\sigma^2$. The two priors that they investigate are Jeffreys' prior and an Inverse Gamma prior, and we have shown here how our methods allow to directly measure the difference in impacts between these two priors.

# Binomial Model

Recall : Many random phenomena worth studying have binary outcomes and therefore can be modelled using the famous $Binomial$ distribution $Bin(n, \theta)$ with probability mass function :

$$p(x;\theta) = \binom{n}{x} \theta^x (1-\theta)^{1-x}$$

where $x \in \{1, \ldots, n\}$ is the number of observed successes and $\theta$ stands for the success parameter.

- The $Haldane$ prior, a non-informative prior representing complete uncertainty about the probability, is proportional to $\frac{1}{\theta(1-\theta)}$ for $\theta \in (0, 1)$ and leads to the posterior $Beta(x, n-x)$.
- The conjugate prior $Beta(\alpha, \beta)$ leads to the posterior $Beta(x+\alpha, n-x+\beta)$.

# Measuring the effects of the priors

As the conditions of the theorem are satisfied, the resulting bound is

$$\frac{|n\alpha - (\alpha+\beta)x|}{n(n+\alpha+\beta)} \leq d_{\mathcal{W}}(P_1, P_2) \leq \frac{1}{n}\left(\alpha + (\beta-\alpha)\frac{\alpha+x}{\alpha+\beta+n}\right)$$

The bound is again of the order $O(\frac{1}{n})$.

# Poisson Model

Recall : The $Poisson$ distribution is used to model the number of events occurring within a given time interval, and the probability mass function is

$$f(x;\theta) = \frac{\exp(-\theta)\theta^{x}}{x!}$$

The data $x = (x_1, x_2, \dots, x_n)$ have sampling density

$$L(x;\theta) = \frac{\exp(-n\theta)\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

Ley, Reinert and Swan (2017, Ann. Appl. Probab.) have presented a bound for Uniform and exponential prior distributions.

- We opt here to compare in all generality two $Gamma$ priors with non-negative real parameters $(\alpha_1, \beta_1)$ and $(\alpha_2, \beta_2)$, as this contains all the special cases.
- The resulting posteriors are Gamma with updated parameters $(\sum_{i=1}^{n} x_i + \alpha_j, n + \beta_j)$, $j = 1, 2$.
- The parameter combination given by $\alpha_1 < \alpha_2 \cap \beta_1 > \beta_2$ (increasing) and $\alpha_1 > \alpha_2 \cap \beta_1 < \beta_2$ (decreasing) actually satisfy the conditions of the Corollary, and hence we have the equality

$$d_{\mathcal{W}}(P_1, P_2) = \frac{1}{n + \beta_1} \left| \alpha_2 - \alpha_1 - (\beta_2 - \beta_1) \frac{\sum_{i=1}^{n} x_i + \alpha_2}{n + \beta_2} \right|$$

This exact distance is of the order of $O(n^{-1})$.

# Practical benefits

- Why we need to measure this distance ?

It often occurs that practitioners hesitate between two proposed priors in a given situation. Our results then allow them to know how different the two priors actually are, and to decide whether or not it is relevant to consider both priors or just stick to one of them. In particular, when hesitating between a simple, closed-form prior and a much more complicated prior.

- How ?

if the effects of both priors are similar, then it is advisable to use the simpler one. Our quantification provides the theoretical background for making such choices in practice.

# Thank you for your attention !