Chapter

On the Impact of the Choice of the Prior in Bayesian Statistics

Fatemeh Ghaderinezhad and Christophe Ley

Abstract

A key question in Bayesian analysis is the effect of the prior on the posterior, and how we can measure this effect. Will the posterior distributions derived with distinct priors become very similar if more and more data are gathered? It has been proved formally that, under certain regularity conditions, the impact of the prior is waning as the sample size increases. From a practical viewpoint it is more important to know what happens at finite sample size n. In this chapter, we shall explain how we tackle this crucial question from an innovative approach. To this end, we shall review some notions from probability theory such as the Wasserstein distance and the popular Stein's method, and explain how we use these a priori unrelated concepts in order to measure the impact of priors. Examples will illustrate our findings, including conjugate priors and the Jeffreys prior.

Keywords: conjugate prior, Jeffreys prior, prior distribution, posterior distribution, Stein's method, Wasserstein distance

1. Introduction

A key question in Bayesian analysis is the choice of the prior in a given situation. Numerous proposals and divergent opinions exist on this matter, but our aim is not to delve into a review or discussion, rather we want to provide the reader with a description of a useful new tool allowing him/her to make a decision. More precisely, we explain how to effectively measure the effect of the choice of a given prior on the resulting posterior. How much do two posteriors, derived from two distinct priors, differ? Providing a quantitative answer to this question is important as it also informs us about the ensuing inferential procedures. It has been proved formally in [1, 2] that, under certain regularity conditions, the impact of the prior is waning as the sample size increases. From a practical viewpoint it is however more interesting to know what happens at finite sample size n, and this is precisely the situation we are considering in this chapter.

Recently, [3, 4] have devised a novel tool to answer this question. They measure the Wasserstein distance between the posterior distributions based on two distinct priors at fixed sample size n. The Wasserstein (more precisely, Wasserstein-1) distance is defined as

$$d_{\mathcal{W}}(P_1, P_2) = \sup_{h \in \mathcal{H}} |\mathrm{E}[h(X_1)] - \mathrm{E}[h(X_2)]|$$

for X_1 and X_2 random variables with respective distribution functions P_1 and P_2, and where $\mathcal{H}$ stands for the class of Lipschitz-1 functions. It is a popular distance

between two distributions, related to optimal transport and therefore also known as *earth mover distance* in computer science, see [5] for more information. The resulting distance thus gives us the desired measure of the difference between two posteriors. If one of the two priors is the flat uniform prior (leading to the posterior coinciding with the data likelihood), then this measure quantifies how much the other chosen prior has impacted on the outcome as compared to a data-only posterior. Now, the Wasserstein distance being mostly impossible to calculate exactly, it is necessary to obtain sharp upper and lower bounds, which will partially be achieved by using techniques from the so-called Stein method, a famous tool in probabilistic approximation theory. We opt for the Wasserstein metric instead of, e.g., the Kullback-Leibler divergence because of precisely its nice link with the Stein method, see [3].

The chapter is organized as follows. In Section 2 we provide the notations and terminology used throughout the paper, provide the reader with the minimal necessary background knowledge on the Stein method, and state the main result regarding the measure of the impact of priors. Then in Section 3 we illustrate how this new measure works in practice, by first working out a completely new example, namely priors for the scale parameter of the inverse gamma distribution, and second giving new insights into an example first treated in both [3, 4], namely priors for the success parameter in the binomial distribution.

2. The measure in its most general form

In this section we provide the reader with the general form of the new measure of the impact of the choice of prior distributions. Before doing so, we however first give a very brief overview on Stein's method that is of independent interest.

2.1 Stein's method in a nutshell

Stein's method is a popular tool in applied and theoretical probability, typically used for Gaussian and Poisson approximation problems. The principal goal of the method is to provide quantitative assessments in distributional comparison statements of the form $W \approx Z$ where Z follows a known and well-understood probability distribution (typically normal or Poisson) and W is the object of interest. Charles Stein [6] in 1972 laid the foundation of what is now called "Stein's method" by aiming at normal approximations.

Stein's method consists of two distinct components, namely

Part A: a framework allowing to convert the problem of bounding the error in the approximation of W by Z into a problem of bounding the expectation of a certain functional of W.

Part B: a collection of techniques to bound the expectation appearing in Part A; the details of these techniques are strongly dependent on the properties of W as well as on the form of the functional.

We refer the interested reader to [7, 8] for detailed recent accounts on this powerful method. The reader will understand in the next sections why Stein's method has been of use for quantifying the desired measure, even without formal proofs or mathematical details.

2.2 Notation and formulation of the main goal

We start by fixing our notations. We consider independent and identically distributed (discrete or absolutely continuous) observations $X_1, ..., X_n$ from a parametric model with parameter of interest $\theta \in \Theta \subseteq \mathbb{R}$. We denote the likelihood of $X_1, ..., X_n$ by $\ell(x; \theta)$ where $x = (x_1, ..., x_n)$ are the observed values. Take two different (possibly improper) prior densities $p_1(\theta)$ and $p_2(\theta)$ for our parameter θ; the famous Bayes' theorem then readily yields the respective posterior densities

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

where $\kappa_1(x)$, $\kappa_2(x)$ are normalizing constants that depend only on the observed values. We denote by (Θ_1, P_1) and (Θ_2, P_2) the couples of random variables and cumulative distribution functions associated with the densities $p_1(\theta; x)$ and $p_2(\theta; x)$.

These notations allow us to formulate the main goal: measure the Wasserstein distance between $p_1(\theta; x)$ and $p_2(\theta; x)$, as this will exactly correspond to the difference between the posteriors resulting from the two priors p_1 and p_2. Sharp upper and lower bounds have been provided for this Wasserstein distance, first in [3] for the special case of one prior being flat uniform, then in all generality in [4]. The determination of the upper bound has been achieved by means of the Stein Method: first a relevant Stein operator has been found (Part A), and then a new technique designed in [3] has been put to use for Part B. The reader is referred to these two papers for details about the calculations; since this chapter is part of a book on Bayesian inference, we prefer to keep out those rather probabilistic manipulations.

2.3 The general result

The key element in the mathematical developments underlying the present problem is that the densities $p_1(\theta; x)$ and $p_2(\theta; x)$ are *nested*, meaning that one support is included in the other. Without loss of generality we here suppose that $I_2 \subseteq I_1$, allowing us to express $p_2(\theta; x)$ as $\frac{\kappa_2(x)}{\kappa_1(x)} \rho(\theta) p_1(\theta; x)$ with

$$\rho(\theta) = \frac{p_2(\theta)}{p_1(\theta)}.$$

The following general result has been obtained in [4], where we refer the reader to for a proof.

Theorem 1.1 Consider $\mathcal{H}$ the set of Lipschitz-1 functions on $\mathbb{R}$ and define

$$\tau_i(\theta; x) = \frac{1}{p_i(\theta; x)} \int_{a_i}^{\theta} (\mu_i - y) p_i(y; x) dy, \quad i = 1, 2, \tag{1}$$

where a_i is the lower bound of the support $I_i = (a_i, b_i)$ of p_i. Suppose that both posterior distributions have finite means μ_1 and μ_2, respectively. 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

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

and, if the variance of Θ_1 exists,

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

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

This result quantifies in all generality the measure of the difference between two priors p_1 and p_2, and comprises of course the special case where one prior is flat uniform. Quite nicely, if ρ is a monotone increasing or decreasing function, the bounds do coincide, leading to

$$d_{\mathcal{W}}(P_1; P_2) = \frac{\mathrm{E}[\tau_1(\Theta_1; x)|\rho'(\Theta_1)|]}{\mathrm{E}[\rho(\Theta_1)]}, \tag{2}$$

hence an exact result. The reader notices the sharpness of these bounds given that they contain the same quantities in both the upper and lower bounds; this fact is further underpinned by the equality Eq. (2). Finally we wish to stress that the functions $\tau_i(\theta; x)$, $i = 1, 2,$ from Eq. (1) are called Stein kernel in the Stein method literature and that these functions are always positive and vanish at the boundaries of the support.

3. Applications and illustrations

Numerous examples have been treated in [3, 4], such as priors for the location parameter of a normal distribution, the scale parameter of a normal distribution, the success parameter of a binomial or the event-enumerating parameter of the Poisson distribution, to cite but these. In this section we will, on the one hand, investigate a new example, namely the scale parameter of an inverse gamma distribution, and, on the other hand, revisit the binomial case. Besides providing the bounds, we will also for the first time plot numerical values for the bounds and hence shed new intuitive light on this measure of the impact of the choice of the prior.

3.1 Priors for the scale parameter of the inverse gamma (IG) distribution

The inverse gamma (IG) distribution has the probability density function

$$x \to \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left\{-\frac{\beta}{x}\right\}, \; x>0,$$

where α and β are the positive shape and scale parameters, respectively. This distribution corresponds to the reciprocal of a gamma distribution (if $X \sim Gamma(\alpha, \beta)$ then $\frac{1}{X} \sim IG(\alpha, \beta)$) and is frequently encountered in domains such as machine learning, survival analysis and reliability theory. Within Bayesian Inference, it is a popular choice as prior for the scale parameter of a normal distribution. In the present setting, we consider $\theta = \beta$ as the parameter of interest and α is fixed. The observations sampled from this distribution are written $x_1, \ldots, x_n$.

The first prior is the popular noninformative Jeffreys prior. It is invariant under reparameterization and is proportional to the square root of the Fisher information quantity associated with the parameter of interest. In the present setting simple calculations show that it is proportional to $\frac{1}{\beta}$. The resulting posterior P_1 then has a density of the form

$$p_1(\beta|x) \propto \frac{1}{\beta}\beta^{n\alpha}\exp\left\{-\beta\sum_{i=1}^{n}\frac{1}{x_i}\right\} = \beta^{n\alpha-1}\exp\left\{-\beta\sum_{i=1}^{n}\frac{1}{x_i}\right\}$$

which is none other than a gamma distribution with parameters $\left(n\alpha, \sum_{i=1}^{n}\frac{1}{x_i}\right)$.

Now, the gamma distribution happens to be the conjugate prior for the scale parameter of an IG distribution. We consider thus as second prior a general gamma distribution with density $\beta \mapsto \frac{\kappa^\eta}{\Gamma(\eta)}\beta^{\eta-1}\exp\{-\kappa\beta\}$, where the shape and scale parameters η and κ are strictly positive. The ensuing posterior distribution P_2 has then the density

$$p_2(\beta|x) \propto \beta^{\eta-1}\exp\{-\kappa\beta\} \times \beta^{n\alpha}\exp\left\{-\beta\sum_{i=1}^{n}\frac{1}{x_i}\right\} = \beta^{n\alpha+\eta-1}\exp\left\{-\beta\left(\sum_{i=1}^{n}\frac{1}{x_i}+\kappa\right)\right\}$$

which is a gamma distribution with updated parameters $\left(n\alpha+\eta, \sum_{i=1}^{n}\frac{1}{x_i}+\kappa\right)$.

Considering Jeffreys prior as p_1 and the gamma prior as p_2 leads to the ratio

$$\rho(\beta) = \frac{p_2(\beta)}{p_1(\beta)} \propto \frac{\frac{\kappa^\eta}{\Gamma(\eta)}\beta^{\eta-1}\exp\{-\kappa\beta\}}{\frac{1}{\beta}} = \frac{\kappa^\eta}{\Gamma(\eta)}\beta^{\eta}\exp\{-\kappa\beta\}.$$

One can easily check that all conditions of Theorem 1.1 are fulfilled, hence we can calculate the bounds. The lower bound is directly obtained as follows:

$$d_{\mathcal{W}}(P_1,P_2) \geq |\mu_1-\mu_2| = \left|\frac{n\alpha}{\sum_{i=1}^{n}\frac{1}{x_i}} - \frac{n\alpha+\eta}{\sum_{i=1}^{n}\frac{1}{x_i}+\kappa}\right| \tag{3}$$

$$= \left|\frac{n\alpha\sum_{i=1}^{n}\frac{1}{x_i} + n\alpha\kappa - n\alpha\sum_{i=1}^{n}\frac{1}{x_i} - \eta\sum_{i=1}^{n}\frac{1}{x_i}}{\sum_{i=1}^{n}\frac{1}{x_i}\left(\sum_{i=1}^{n}\frac{1}{x_i}+\kappa\right)}\right| \tag{4}$$

$$= \left|\frac{n\alpha\kappa - \eta\sum_{i=1}^{n}\frac{1}{x_i}}{\sum_{i=1}^{n}\frac{1}{x_i}\left(\sum_{i=1}^{n}\frac{1}{x_i}+\kappa\right)}\right|. \tag{5}$$

In order to acquire the upper bound we need to calculate

$$\rho'(\beta) = \frac{\kappa^\eta}{\Gamma(\eta)}\beta^{\eta-1}\exp(-\kappa\beta)[\eta-\kappa\beta]$$

and, writing Θ_1 the random variable associated with $Gamma\left(n\alpha, \sum_{i=1}^{n}\frac{1}{x_i}\right)$ and $f_{Gamma\left(n\alpha, \sum_{i=1}^{n}\frac{1}{x_i}\right)}(\beta)$ the related density, we get

$$\mathbb{E}[\rho(\Theta_1)] = \int_0^\infty \frac{\kappa^\eta}{\Gamma(\eta)}\beta^\eta\exp\{-\kappa\beta\} \times f_{Gamma\left(n\alpha, \sum_{i=1}^{n}\frac{1}{x_i}\right)}(\beta)d\beta \tag{6}$$

$$= \frac{\kappa^\eta}{\Gamma(\eta)}\frac{\left(\sum_{i=1}^{n}\frac{1}{x_i}\right)^{n\alpha}}{\Gamma(n\alpha)}\int_0^\infty \beta^\eta\exp\{-\kappa\beta\}\beta^{n\alpha-1}\exp\left\{-\beta\sum_{i=1}^{n}\frac{1}{x_i}\right\}d\beta \tag{7}$$

$$= \frac{\kappa^{\eta}}{\Gamma(\eta)} \frac{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{n\alpha}}{\Gamma(n\alpha)} \int_0^{\infty} \beta^{n\alpha+\eta-1} \exp\left\{-\beta\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)\right\} d\beta \tag{8}$$

$$= \frac{\kappa^{\eta}}{\Gamma(\eta)} \frac{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{n\alpha}}{\Gamma(n\alpha)} \frac{\Gamma(n\alpha+\eta)}{\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta}} \tag{9}$$

$$= \frac{\kappa^{\eta}}{Beta(n\alpha, \eta)} \frac{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{n\alpha}}{\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta}}. \tag{10}$$

From the Stein literature we know that the Stein kernel for the gamma distribution with parameters $\left(n\alpha, \sum_{i=1}^{n} \frac{1}{x_i}\right)$ corresponds to $\tau(\beta;x) = \frac{\beta}{\sum_{i=1^{x_i}}^{n} \frac{1}{x_i}}$. Employing the triangular inequality we have thus

$$\mathbb{E}[\tau(\Theta_1; x)|\rho'(\Theta_1)|] = \mathbb{E}\left[\frac{\Theta_1}{\sum_{i=1}^{n} \frac{1}{x_i}} \frac{\kappa^{\eta}}{\Gamma(\eta)} \Theta_1^{\eta-1} \exp\{-\kappa\Theta_1\}|\eta - \kappa\Theta_1|\right] \tag{11}$$

$$\leq \frac{\kappa^{\eta}}{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)\Gamma(\eta)} \mathbb{E}\left[\Theta_1^{\eta} \exp\{-\kappa\Theta_1\}(\eta + \kappa\Theta_1)\right]. \tag{12}$$

Now we need to calculate the expectation

$$\mathbb{E}\left[\Theta_1^{\eta} \exp\{-\kappa\Theta_1\}(\eta + \kappa\Theta_1)\right] \tag{13}$$

$$= \int_0^{\infty} \beta^{\eta} \exp\{-\kappa\beta\}(\eta + \kappa\beta) \times f_{Gamma\left(n\alpha, \sum_{i=1^{x_i}}^{n} \frac{1}{x_i}\right)}(\beta) d\beta \tag{14}$$

$$= \frac{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{n\alpha}}{\Gamma(n\alpha)} \int_0^{\infty} \eta\beta^{n\alpha+\eta-1} \exp\left\{-\beta\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)\right\} d\beta \tag{15}$$

$$+ \frac{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{n\alpha}}{\Gamma(n\alpha)} \int_0^{\infty} \kappa\beta^{n\alpha+\eta} \exp\left\{-\beta\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)\right\} d\beta \tag{16}$$

$$= \frac{\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{n\alpha}}{\Gamma(n\alpha)} \left(\eta \frac{\Gamma(n\alpha+\eta)}{\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta}} + \kappa \frac{\Gamma(n\alpha+\eta+1)}{\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta+1}}\right). \tag{17}$$

The final expression for the upper bound then corresponds to

$$d_{\mathcal{W}}(P_1, P_2) \leq \frac{\frac{\kappa^{\eta}}{\left(\sum_{i=1^{x_i}}^{n} \frac{1}{x_i}\right)\Gamma(\eta)} \times \frac{\left(\sum_{i=1^{x_i}}^{n} \frac{1}{x_i}\right)^{n\alpha}}{\Gamma(n\alpha)} \left[\eta \frac{\Gamma(n\alpha+\eta)}{\left(\sum_{i=1^{x_i}}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta}} + \kappa \frac{\Gamma(n\alpha+\eta+1)}{\left(\sum_{i=1^{x_i}}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta+1}}\right]}{\frac{\kappa^{\eta}}{Beta(n\alpha, \eta)} \times \frac{\left(\sum_{i=1^{x_i}}^{n} \frac{1}{x_i}\right)^{n\alpha}}{\left(\sum_{i=1^{x_i}}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta}}} \tag{18}$$

$$= \frac{Beta(n\alpha, \eta)\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta}}{\frac{\Gamma(n\alpha)\Gamma(\eta)}{\Gamma(n\alpha+\eta)}\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)} \times \frac{1}{\left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa\right)^{n\alpha+\eta}} \left(\eta + \kappa \frac{n\alpha+\eta}{\sum_{i=1}^{n} \frac{1}{x_i} + \kappa}\right) \tag{19}$$

$$= \frac{1}{\sum_{i=1}^{n} \frac{1}{x_i}} \left(\eta + \kappa \frac{n\alpha + \eta}{\sum_{i=1}^{n} \frac{1}{x_i} + \kappa} \right). \tag{20}$$

The Wasserstein distance between the posteriors based on the Jeffreys prior and conjugate gamma prior for the scale parameter β of the IG distribution is thus bounded as

$$\left| \frac{n\alpha\kappa - \eta \sum_{i=1}^{n} \frac{1}{x_i}}{\left(\sum_{i=1}^{n} \frac{1}{x_i} \right) \left(\sum_{i=1}^{n} \frac{1}{x_i} + \kappa \right)} \right| \leq d_{\mathcal{W}(P_1, P_2)} \leq \frac{1}{\sum_{i=1}^{n} \frac{1}{x_i}} \left(\eta + \kappa \frac{n\alpha + \eta}{\kappa + \sum_{i=1}^{n} \frac{1}{x_i}} \right).$$

It can be seen that both the lower and upper bound are of the order of $O(n^{-1})$. In addition, it is noticeable that for the larger observations, the rate of convergence is getting slower.

In order to show the performance of the methodology which leads to have the lower and upper bounds, we have conducted a simulation study including two parts. First we simulate $N = 100$ samples for each sample size $n = 10, 11, \cdots, 100$ from the inverse gamma distribution with parameters $(\alpha, \beta) = (0.5, 1)$ in each iteration. For each of these samples we calculate the lower and upper bounds of the Wasserstein distance and calculate the average over all N replications, together with the difference between the bounds. Finally we plot these values for each sample size in **Figure 1**. We repeat the same process for $N = 1000$ samples with the same sizes. The hyperparameters from the prior gamma distribution are $(\kappa, \eta) = (0.2, 2)$. We clearly observe how fast these values decrease with the sample size. Of course, augmenting the number of replications does not increase the speed of convergence, however the curves become noticeably smoother.

This methodology not only can help the practitioners to make a decision between existing priors in theory, but also helps them to know from what sample size on the effect of choosing one prior becomes less important, especially in situations when the cost and time matter. This can be particularly useful when the

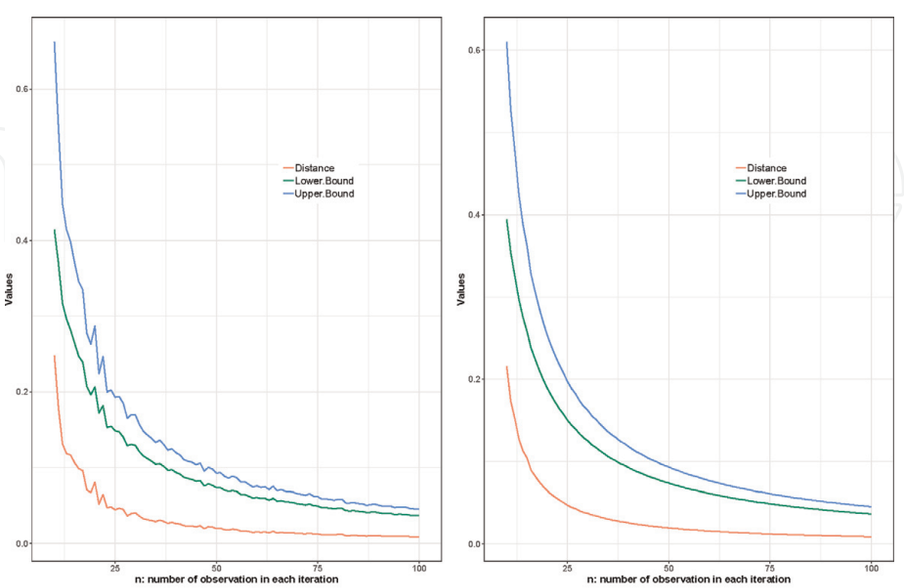


Figure 1.
(a) Shows the bounds and the distances between the bounds for $N = 100$ iterations for each sample size 10–100 by steps of 1, and (b) illustrates the same situation for $N = 1000$. The hyperparameters are $\kappa = 0.2$ and $\eta = 2$, while the fixed parameter α equals 0.5.

hesitation is between a simple, closed-form prior and a more complicated one. It is advisable to use the simpler one when there is no considerable difference between the effect of the two priors.

3.2 The impact of priors for the success parameter of the binomial model

The probability mass function of a binomial distribution is given by

$$x \mapsto \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

where $x \in \{0, 1, \cdots, n\}$ is the number of observed successes, the natural number n indicates the number of binary trials and $\theta \in (0, 1)$ stands for the success parameter. In this setting we suppose n is fixed and the underlying parameter of interest is θ.

A comprehensive comparison of various priors for the binomial distribution including a beta prior, the Haldane prior and Jeffreys prior, has been done in [9], based on the methodology described above. Therefore, since there is a complete reference for the reader in this case, we use the binomial distribution as a second example to show numerical results.

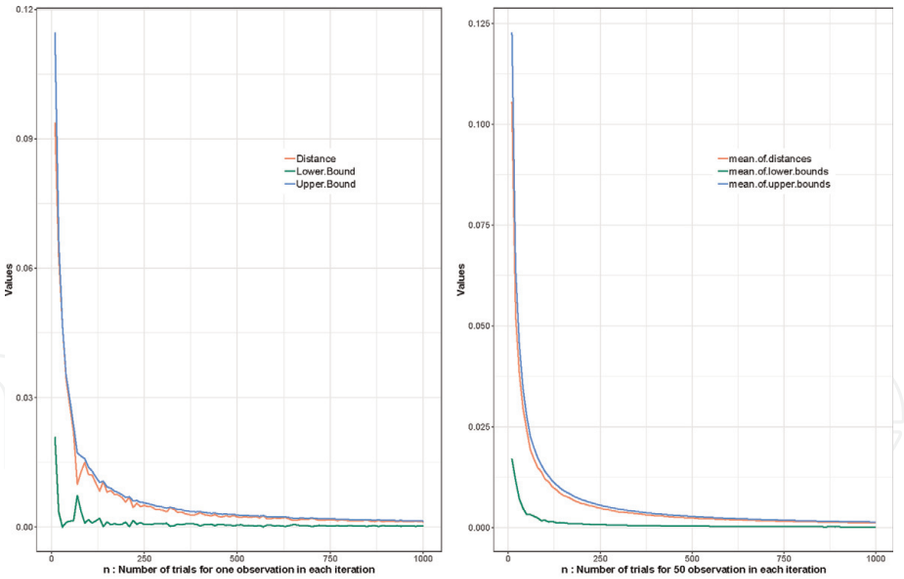
The theoretical lower and upper bounds between a $Beta(\alpha, \beta)$ prior and the flat uniform prior are given by

$$\left| \frac{x+1}{n+2} \left(\frac{\alpha+\beta-2}{n+\alpha+\beta} \right) - \frac{\alpha-1}{n+\alpha+\beta} \right| \leq d_W(P_1, P_2) \leq \frac{1}{n+2} \left\{ |\alpha-1| + \frac{x+\alpha}{n+\alpha+\beta} (|\beta-1| - |\alpha-1|) \right\},$$

where x is the observed number of successes. We see that both lower and upper bounds are of the order of $O(n^{-1})$. This rate of convergence remains even in the extreme cases $x = 0$ and $x = n$. We invite the reader to see [3, 9] for more details.

In order to illustrate the behavior of the lower and upper bounds and the distances between them, we have conducted a two-part simulation study for the binomial distribution. First, we consider 100 sample sizes (number of trials in the binomial distribution) varying from 10 to 1000 by steps of 10, and generate binomial data exactly once for every sample size (with $\theta = 0.2$). The results of the bounds, obtained for hyperparameters $(\alpha, \beta) = (2, 4)$ from the beta prior, are reported in **Figure 2a** and we can see that, even with only one iteration, when the number of trials (the sample size) increases the lower and upper bound become closer, which is a numerical quantification of the fact that the influence of the choice of the prior wanes asymptotically. This becomes also visible from the distance between the two bounds. Sampling only once for each sample size leads to slightly unpleasant variations in the lower bounds (non-monotone behavior), which however nearly disappear in the second considered scenario. Indeed, in **Figure 2b** we increased the number of iterations to 50 for the same different sample sizes and took averages. A better smoothness is the consequence. This simulation study not only provides the reader with numerical values for the bounds, to which he/she can compare his/her bounds obtained for real data, but also gives a nice visualization of the impact of the choice of the prior at fixed sample size. The main conclusion is that the impact drops fast at small sample sizes, and the bounds start to become very close for medium-to-large sample sizes.

Finally, we investigate the impact of the hyperparameters on the upper and lower bounds. To this end, we varied both α and β in **Table 1**. The situation with α fixed to two and relatively small β corresponds well with $p = 0.2$, which explains why the upper and lower bounds, and hence the Wasserstein distance and thus the impact of the prior, are the smallest. Increasing β more augments the distance.



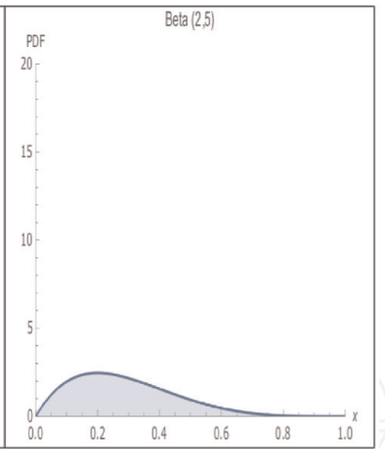
(a) (b)

Figure 2.
(a) Shows the lower and upper bounds and the distances for the number of trials {n = 10,...,1000} for one iteration. (b) Shows the same situation, however this time based on averages obtained for 50 iterations. In both situations the hyperparameters from the beta prior are $\alpha = 2$ and $\beta = 4$.

Hyperparameters (α, β)	Average of the lower bounds	Average of the upper bounds
(0.2, 0.4)	0.002561383	0.003726728
(0.2, 0.8)	0.00296002	0.003344393
(2, 2)	0.002699325	0.00490119
(2, 5)	0.0008115384	0.007984289
(2, 10)	0.004506271	0.01241273
(2, 15)	0.008208887	0.01626326
(2, 30)	0.01750177	0.02581062
(2, 50)	0.02739205	0.0359027
(2, 100)	0.04592235	0.05470826
(2, 200)	0.07071766	0.07976386
(2, 500)	0.1103048	0.1196464
(2, 1000)	0.1399961	0.1495087
(10, 2)	0.02813367	0.03132908
(35, 2)	0.08571115	0.09033568
(50, 2)	0.1127136	0.1178113
(100, 2)	0.1830272	0.189071
(200, 2)	0.2783722	0.2853418
(400, 2)	0.3933338	0.401145
(700, 2)	0.4901209	0.4985089
(1000, 2)	0.5482869	0.5569829

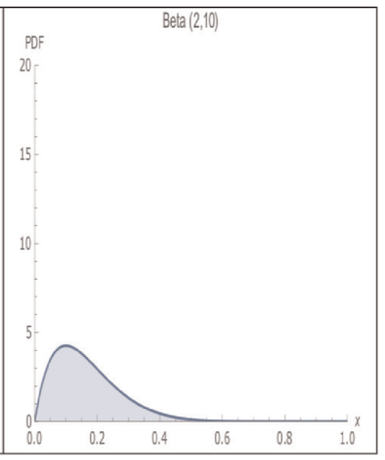
Table 1.
The summary of upper and lower bounds for different hyperparameters, with $p = 0.2$ and for $N = 50$ iterations.

(α, β)	$(2, 5)$
Lower Bound (L)	0.0008115384
Upper Bound (U)	0.007984289
Distance (U-L)	0.007172751



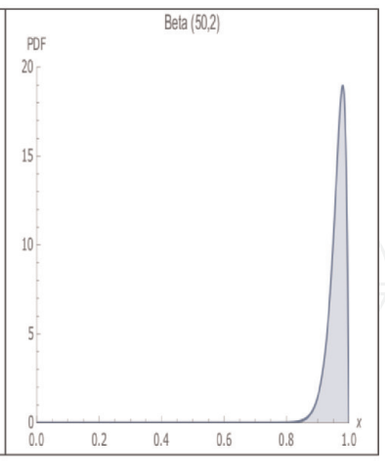
(a)

(α, β)	$(2, 10)$
Lower Bound (L)	0.004506271
Upper Bound (U)	0.01241273
Distance (U-L)	0.007906458



(b)

(α, β)	$(50, 2)$
Lower Bound (L)	0.1127136
Upper Bound (U)	0.1178113
Distance (U-L)	0.00509779



(c)

Figure 3.
Plots of the beta prior densities together with the average lower and upper bounds (and their difference) on the Wasserstein distance between the data-based posterior and the posterior resulting from each beta prior.

On the contrary, fixing $\beta = 2$ yields priors rather centered around large values of p and hence bigger distances. Moreover, the more α is increased, the more the distance augments, as the prior is further away from the data and hence impacts more on the posterior at a fixed sample size. For the sake of illustration, we present three

choices of hyperparameters together with the bounds and the related prior density in **Figure 3**. This will help understanding our conclusions.

4. Conclusions

In this chapter we have presented a recently developed measure for the impact of the choice of the prior distribution in Bayesian statistics. We have presented the general theoretical result, explained how to use it in a particular example and provided some graphics to illustrate it numerically. The practical importance of this study is when practitioners hesitate between two proposed priors in a given situation. For instance, Kavetski et al. [10] considered a storm depth multiplier model to represent rainfall uncertainty where the errors appear under multiplicative form and are assumed to be normal. They fix the mean, but state that "less is understood about the degree of rainfall uncertainty," i.e., the multiplier variance, and therefore studied various priors for the variance. Knowledge of the tools presented in this chapter would have simplified the decision process.

In case of missing data, the present methodology can still be used. Either the data get imputed, in which case nothing changes, or the missing data simply are left out from the calculation of upper and lower bounds, whose expression does of course not alter.

Further developments on this new measure might lead to a more concrete quantification of words such as "informative, weakly informative, noninformative" priors, and we hope to have stimulated interest in this promising new line of research within Bayesian Inference.

Acknowledgements

This research is supported by a BOF Starting Grant of Ghent University.

Author details

Fatemeh Ghaderinezhad and Christophe Ley*
Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

*Address all correspondence to: christophe.ley@ugent.be

References

[1] Diaconis F, Freedman D. On the consistency of Bayes estimates (with discussion and rejoinder by the authors). The Annals of Statistics. 1986; **14**:1-67

[2] Diaconis F, Freedman D. On inconsistent Bayes estimates of location. The Annals of Statistics. 1986;**14**:68-87

[3] Ley C, Reinert G, Swan Y. Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. Annals of Applied Probability. 2017;**27**:216-241

[4] Ghaderinezhad F, Ley C. Quantification of the impact of priors in Bayesian statistics via Stein's method. Statistics & Probability Letters. 2019; **146**:206-212

[5] Rüschendorf L. Wasserstein metric. In: Michiel H, editor. Encyclopedia of Mathematics. Netherlands: Springer Science+Business Media B.V./Kluwer Academic Publishers; 2001

[6] Stein C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Univ. California, Berkeley, CA, 1970/ 1971. 1972. pp. 583-602

[7] Ross N. Fundamentals of Stein's method. Probability Surveys. 2011;**8**: 210-293

[8] Ley C, Reinert G, Swan Y. Stein's method for comparison of univariate distributions. Probability Surveys. 2017; **14**:1-52

[9] Ghaderinezhad F. New insights into the impact of the choice of the prior for the success parameter of binomial distributions. Journal of Mathematics, Statistics and Operations Research, forthcoming

[10] Kavetski D, Kuczera G, Franks SW. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research. 2006;**42**:W03407