Computational Statistics and Data Analysis 174 (2022) 107352

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

The Wasserstein Impact Measure (WIM): A practical tool for quantifying prior impact in Bayesian statistics

Fatemeh Ghaderinezhad [a], Christophe Ley [a,*], Ben Serrien [b]

[a] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium
[b] Experimental Anatomy Research Group, Vrije Universiteit Brussel, Brussels, Belgium

ARTICLE INFO

Article history:
Received 5 January 2021
Received in revised form 14 September 2021
Accepted 14 September 2021
Available online 4 October 2021

Keywords:
Effective sample size
Neutrality
Prior distribution
Vallender formula
Wasserstein distance

ABSTRACT

The prior distribution is a crucial building block in Bayesian analysis, and its choice will impact the subsequent inference. It is therefore important to have a convenient way to quantify this impact, as such a measure of prior impact will help to choose between two or more priors in a given situation. To this end a new approach, the Wasserstein Impact Measure (WIM), is introduced. In three simulated scenarios, the WIM is compared to two competitor prior impact measures from the literature, and its versatility is illustrated via two real datasets.

1. Introduction

With the increase in computational power readily available in personal computers and the increase in complexity in statistical models, more and more researchers are shifting towards Bayesian statistics. The Bayesian framework allows more flexibility and more intuitive interpretations compared to frequentist methods, but this same flexibility comes with the cost of argumentation for certain choices and the assessment of their impact on any conclusions drawn from the data. In Bayesian statistics, particularly the choice of the prior can have a profound impact on the inferences drawn from data, especially at small sample sizes. The prior elicitation thus can be perceived as dual: an opportunity to introduce important prior knowledge into the problem, but at the same time a challenge when it comes to choosing a suitable prior. Of course, a natural choice of prior can be obtained if one resorts to general principles such as mathematical tractability (bringing conjugate priors to the forefront) or the invariance property (favouring the Jeffreys prior). However, in absence of such considerations, the prior choice induces an essential research question: how can one quantify the impact of the choice of the prior on the posterior distribution, and hence on subsequent inference? Diaconis and Freedman (1986a,b) showed that under certain regularity conditions the effect of the prior wanes as the sample size increases (think also of the Bernstein-von Mises theorem stating that, under certain regularity conditions, the posterior distribution converges to the Gaussian distribution), however in practice the sample size of course is always finite and, very often, rather small, underlining the relevance of said question. At first sight, a pragmatic answer could be a sensitivity analysis, but such an analysis often depends on how it is carried out and it only covers certain aspects of the inference, hence cannot provide a satisfying general answer.

* Corresponding author.
E-mail addresses: Fatemeh.Ghaderinezhad@ugent.be (F. Ghaderinezhad), Christophe.Ley@ugent.be (C. Ley), ben.serrien@gmail.com (B. Serrien).

https://doi.org/10.1016/j.csda.2021.107352
0167-9473/

There does not exist a formal definition of prior impact in the literature. According to Reimherr et al. (2014), an all-encompassing approach to this problem is seemingly philosophically and mathematically impossible. The mathematical problem lies in the lack of proper definition, while the philosophical problem concerns different schools of Bayesian inference: for a subjective Bayesian, the impact of the prior may be of less concern than for an objective Bayesian or frequentist statistician. Consequently, different measures of prior impact have been proposed over the years. A popular approach is the so-called *effective sample size*, defined as the approximate number of observations equivalent to the information conveyed by the prior, see for instance Lin et al. (2007); Morita et al. (2008); Reimherr et al. (2014); Wiesenfarth and Calderazzo (2019); Jones et al. (2021) and the references therein. Kerman (2011) introduced the Neutrality, corresponding to the posterior's tail probability to the left of the frequentist maximum likelihood estimate. Yet another approach has been taken by Ley et al. (2017) who measure the Wasserstein distance between two posteriors with finite first absolute moments, of which one results from the prior of interest and the other is the no-prior data-only posterior. Since it is mostly impossible to calculate this distance explicitly, the authors have provided sharp lower and upper bounds on the Wasserstein distance and their approach relies on a variant of the famous Stein Method. In order to compare any two priors directly, Ghaderinezhad and Ley (2019) recently extended their approach to any two priors for one-dimensional parameters, provided that the posteriors are nested and have finite first moments; see also Ghaderinezhad and Ley (2020).

For practical purposes, the power of the Wasserstein distance idea has not been exploited so far. The obtained bounds and rare explicit results are obtained for tractable posteriors; indeed, most examples considered in Ley et al. (2017) and Ghaderinezhad and Ley (2019) are (related to) conjugate priors. Moreover, their results are confined to the priors whose supports are nested and to the one-dimensional setting, meaning that prior impact can only be assessed for one scalar parameter at once. Even if the bounds are well computable, they may not give us accurate information on the actual distance if they are spread far apart. Thus, nice as they are, these mostly theoretical results are not yet broadly usable in practice. The aim of the present paper is to precisely fill this important gap and make this theoretically successful method widely usable in practice. More concretely, we will provide in Section 2 the Wasserstein Impact Measure, abbreviated WIM, for assessing prior impact for any type of priors (under the assumption that the posteriors possess finite absolute moments of order 1) and any dimensions. The WIM relies on a numerical computation of Wasserstein distances and will allow us to compare any two priors, thus making the WIM a fully usable alternative to the proposals from the literature. It is important to note that the Wasserstein distance does not enjoy the scale invariance property, meaning that the impact of the prior is not comparable across different parametrizations. By means of Monte Carlo simulations, we will compare the WIM to other prior impact measures from the literature in Section 3, namely the effective sample size (in what follows we will use its most recent version called MOPESS) and the Neutrality mentioned above. We illustrate the practical aspects of the WIM on two data examples in Section 4. Finally, we conclude the paper with a discussion in Section 5.

2. The WIM for prior impact

We now present our new practice-oriented measure of the impact of the choice of the prior, the WIM. For a given dataset and, hence, a given likelihood, it quantifies the Wasserstein or Wasserstein-1 distance (Vaserstein, 1969) between two posteriors resulting from two distinct priors (still under the assumption that the posteriors possess finite absolute moments of order 1). In particular, if one prior is the (improper) uniform/flat prior, then this distance reveals how much information the other prior adds to the posterior in comparison to the likelihood alone. We stress that the WIM is not an absolute measure, but rather a relative measure allowing to compare two priors. If they are quite close for instance, then one can opt for the simpler prior. To quantify the uncertainty of the WIM for a given dataset, we suggest having recourse to bootstrapping as we shall illustrate in Section 4. This measure of prior impact is highly intuitive, simple and can be used in univariate as well as multivariate, one-parameter as well as multi-parameter settings.

One way to efficiently compute the WIM is by having recourse to the Vallender formula (Vallender, 1974) which allows one to calculate the Wasserstein distance by using the cumulative distribution functions $F_i(\theta; x)$ of two posterior distributions P_1 and P_2, where $\theta \in \mathbb{R}$ is the parameter of interest and x represents the data. The Wasserstein distance $d_W(P_1, P_2)$ is obtained as

$$d_W(P_1, P_2) = \int_a^b |F_1(\theta; x) - F_2(\theta; x)| d\theta, \tag{1}$$

where a and b are the bounds of the support of the parameter of interest (support that obviously can be $\mathbb{R}$). This formula can also readily be extended to $m > 1$ parameters $\theta_1, \ldots, \theta_m$ as follows:

$$d_W(P_1, P_2) = \int_{a_1}^{b_1} \ldots \int_{a_m}^{b_m} |F_1(\theta_1, \ldots, \theta_m; x) - F_2(\theta_1, \ldots, \theta_m; x)| d\theta_1 \cdots d\theta_m,$$

where a_j, b_j are the bounds of the support of θ_j, $j = 1, \ldots, m$. The Vallender formula gives an exact expression, and numerical integration techniques permit to calculate this quantity irrespective of any assumption of nested supports between P_1

and P_2. Now, it is not uncommon to encounter complicated posterior densities for which the cdfs are not computable. In such cases, we suggest to have recourse to computational techniques such as Markov Chain Monte Carlo (MCMC) methods to generate random samples from each distribution P_1 and P_2 and then use Monte Carlo integration. For instance, the *"transport"* package in R offers functions for estimating the Wasserstein distance of any order between two sets of samples from different distributions, see Schuhmacher et al. (2019). Most of the functions in the package have been designed for data with two or higher dimensions.

One may wonder why we opted for the Wasserstein distance and not other distances such as the Total Variation distance. The reasons are manifold. First, we started off from the theoretically successful proposals of Ley et al. (2017) and Ghaderinezhad and Ley (2019), which rely on the Wasserstein distance. Second, this distance can be numerically calculated. Third, it bares a clear and intuitive interpretation as the amount of "work" required to turn one probability distribution into another (therefore it has some different names in other fields of studies such as optimal transport distance in optimization or earth mover's distance in computer science). Finally, it incorporates a ground distance on the space in question that causes adequacy compared to competing metrics such as the Total Variation or χ^2-metrics which do not consider any similarity structure on the ground space (Sommerfeld and Munk, 2017). Finally we note that a good alternative choice would be the Kolmogorov metric which is also computable and, contrary to the Wasserstein distance, is invariant under scale transformations.

3. Comparison with prior impact measures from the literature

In order to judge its quality as new measure of prior impact, we now compare the WIM to other competitors from the literature, namely the Neutrality and the MOPESS (abbreviation to be defined below). We start by describing both approaches in Sections 3.1 and 3.2, respectively, before proceeding to the comparison with our WIM in Section 3.3.

3.1. Neutrality

Kerman (2011) introduced the concept of *Neutrality* N for a given prior with corresponding posterior Θ, say. The Neutrality of Θ is defined as the probability of Θ to lie on the left of the frequentist maximum likelihood estimate $\hat{\theta}_{MLE}$, which in mathematical terms means

$$N = P(\Theta < \hat{\theta}_{MLE}) = \int_a^{\hat{\theta}_{MLE}} p(\theta; x) d\theta,$$

where $p(\theta; x)$ is the posterior and a the lower bound of its support. The closer this tail probability is to 1/2, the less informative or the more neutral the prior is. Kerman (2011) showed that for the binomial and Poisson likelihoods, the conjugate $Beta(1/3, 1/3)$ and $Gamma(1/3, 0)$ are respectively the most neutral priors over the entire parameter space. The advantage of the Neutrality N as a metric of prior impact is that it is an *absolute* metric for each prior, in contrast to the WIM and the MOPESS (see next section) which are relative measures. Another advantage is the ease of calculation in conjugate models, since it can be calculated analytically and, when MCMC has been used, it can be readily calculated based on the posterior samples. In addition, the scale of N is the same for all models no matter how complex they are since $N \in [0, 1]$. However, it cannot be used properly in cases where the frequentist MLE is at the boundaries of the parameter space which is a notable disadvantage in Bayesian analysis. For example, in a binomial model when the MLE is 0 or 1, then N will be the same for any prior. Moreover, Kerman (2011) has not mentioned how to extend this concept to multivariate or multiparameter situations, so we do not discuss this issue here.

3.2. Mean Observed Prior Effective Sample Size (MOPESS)

In certain studies one distinguishes between the effect of the prior and the amount of information the prior does contain. For conjugate models, the nominal amount of information in the prior is known and is often expressed in terms of the number of pseudo-observations (prior sample size (PSS)). For instance, in a beta-binomial model, the parameters of the $Beta(\alpha, \beta)$ prior can be interpreted as the number of successes (α) and failures (β) in the available prior information. When the prior and the likelihood function differ substantially or are very comparable, the impact will be different although the prior has the same amount of information in each situation. Therefore the concept of effective prior sample size (EPSS) has been introduced.

We will follow here the definition of Reimherr et al. (2014) of a new class of effective prior sample size measures based on prior-likelihood discordance, which is also referred to as (the degree of) prior-likelihood conflict in Jones et al. (2021). Reimherr et al. (2014) published the first algorithm on how to adjust for prior-likelihood conflict in the calculation of EPSS to answer the question: how many extra observations are needed to transform a posterior based on a baseline prior into the posterior based on the prior of interest? The EPSS can be lower or higher than the nominal PSS (information) depending on the actual data that are observed. When the prior mean is arbitrarily far from the maximum likelihood estimate, then the impact will become larger. The EPSS can also be negative, indicating that the baseline prior is in fact more impactful than

the prior of interest. Jones et al. (2021) have extended the method of Reimherr et al. (2014) to a more general setting and their method also works for lower sample sizes (where the impact of the prior is particularly important). They calculate the mean observed prior effective sample size (MOPESS) according to the following steps:

1. Derive the posteriors based on the prior of interest and the baseline prior.
2. Use the posterior predictive distribution based on the prior of interest to sample two sets of m additional observations ($m = 1, \ldots, L$ where L is the maximum feasible value for EPSS).
3. For each m, calculate
   (i) the Wasserstein-2 distance (see their paper for a formal definition of this Wasserstein distance) between posteriors based on (1) original data combined with prior of interest versus (2) original + additional data combined with baseline prior ($W1$).
   (ii) the Wasserstein-2 distance between posteriors based on (3) original data combined with baseline prior versus (4) original + additional data combined with prior of interest ($W2$).
4. The m for which this distance is smallest is the OPESS. The lowest value of the set of $W1$ and $W2$ determines the sign: when $W1$ contains the lowest distance, then the prior of interest is more impactful than the baseline prior (OPESS > 0) and vice versa (OPESS < 0).
5. This process is repeated several times and the mean of the OPESS values is the MOPESS.

An advantage of the MOPESS algorithm is that it indicates which of the two priors has the most impact thanks to the added sign. Similarly to our Wasserstein approach, the aim of the MOPESS is to compare any two posteriors and measure the relative impact of the priors. Hence, it is necessary to label one of the priors as the baseline. A practical disadvantage of this method is the computation time for non-conjugate models where advanced sampling algorithms are necessary and need to be repeated several times. A further disadvantage of the MOPESS appears in models with covariates such as in regression settings, because it requires assumptions to be made on the distribution of each covariate. Jones et al. (2021) mention this issue, and also show through a simple linear regression example that the MOPESS in principle should work in higher dimensions, though this extension has not been touched upon outside of this one example.

3.3. Comparison with the WIM

We will study in how far our WIM relates to the Neutrality and the MOPESS via Monte Carlo simulations. In order to see whether they provide similar information or we can learn different aspects from them, we plot the WIM versus both the Neutrality N and the MOPESS for various priors corresponding to the Poisson, binomial and normal settings. The simulation strategy is the same for each setting. We draw 1000 random samples for each combination of sample size ($n = 10, 50, 100$ and 200) and parameter value (this depends of course on the setting). We calculate the WIM for each pair of posteriors by drawing 10000 random samples from them and calculating the Wasserstein distance between them with the *transport* package.

3.3.1. Poisson case

The most famous count distribution is the Poisson with probability mass function

$$x \mapsto \frac{\exp(-\theta)\theta^x}{x!},$$

where $\theta \in \mathbb{R}^+$ is the parameter of interest indicating the average number of events in a given time interval, and $x \in N$ is the number of occurrences. The Gamma distribution is one of the most popular choices of priors for the Poisson model, in large parts due to the fact that it is a conjugate prior. For example, it was used to study asthma mortality rates by Gelman et al. (2004). The Gamma probability density function is

$$\theta \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta),$$

where $\alpha, \beta > 0$. The Gamma prior contains as special cases the exponential prior ($\alpha = 1$), the uniform prior ($\alpha = 1, \beta = 0$), and the Jeffreys prior which is proportional to $\theta^{-1/2}$ ($\alpha = 1/2, \beta = 0$). In addition, Kerman (2011) showed that the $Gamma(1/3, 0)$ prior is optimal in terms of Neutrality for a large range of sample sizes and values of θ.

Fig. 1 shows the WIM (x-axis) versus the Neutrality and the MOPESS (y-axis), respectively, for different Gamma priors and sample sizes. For the WIM and the MOPESS, the baseline prior is the uniform, which is not required for the Neutrality N. We observe that all prior impact measures decrease with sample size, as one could expect.

The first clear conclusion for Neutrality against WIM is that these measures are very alike for the two considered Gamma priors, except for $n = 10$. This is remarkable because the shapes of these priors diverge considerably with increasing β. Moreover, in some panels the Neutrality quickly reaches its maximum value meaning the Neutrality is no longer sensitive to distinguish between priors. In such cases, the WIM is a much better choice to quantify the impacts of the priors as it is unbounded. Also in some cases such as $\theta = 1$ the WIM is not much affected by β.

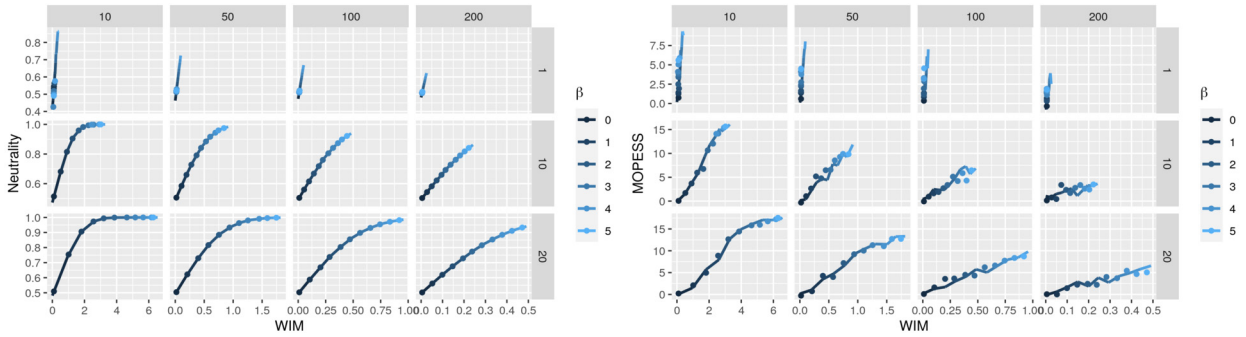

Fig. 1. WIM vs. Neutrality (left panel) and WIM vs. MOPESS (right panel) for the Poisson model. The columns correspond to different sample sizes and the rows to different values of θ. The β parameter of the priors is shown on the colour scale. The full line represents the impact measures for the $Gamma(1, \beta)$ vs. uniform prior and the dots for the $Gamma(\beta, \beta)$ vs. uniform prior. (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

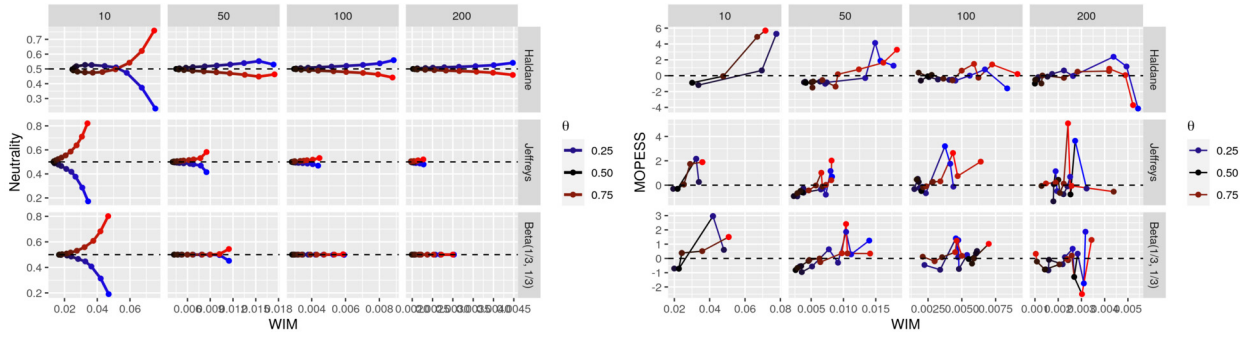

Fig. 2. WIM vs. Neutrality (left panel) and WIM vs. MOPESS (right panel) for the binomial model for different priors (rows) and sample sizes (columns). The colour scale presents the binomial success rate $\theta = 0.05, 0.10, \cdots, 0.95$. The baseline prior is the uniform. Note the different axis scales.

The relation between the MOPESS and the WIM is not monotone in some panels due to the known high variability of the MOPESS (Jones et al., 2021), which speaks in favour of the WIM. One can notice the higher MOPESS values compared to the WIM, in particular for $\theta = 1$ there is more variability in the MOPESS. Its positive sign indicates that the Gamma priors are more impactful than the baseline uniform prior, an information the WIM does not yield.

3.3.2. Binomial case

Let us now consider the binomial distribution $Bin(n, \theta)$ with probability mass function

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

where $x \in \{0, \ldots, n\}$ is the number of observed successes, the natural number n indicates the number of binary experiments and $\theta \in (0, 1)$ stands for the success parameter. The conjugate prior for the binomial distribution is the beta model $Beta(\alpha, \beta)$ with parameters $\alpha, \beta > 0$. The $Beta(\alpha, \beta)$ prior contains some specific cases based on the definition of α and β, such as the uniform prior ($\alpha = \beta = 1$), the Jeffreys prior ($\alpha = \beta = 1/2$), Haldane's prior ($\alpha = \beta = 0$) which gives a complete uncertainty, and the neutral prior ($\alpha = \beta = 1/3$) (Kerman, 2011).

Fig. 2 portrays the WIM (x-axis) versus the Neutrality and the MOPESS (y-axis), respectively, for different priors and sample sizes. For the WIM and the MOPESS, the baseline prior is the uniform, which is not required for the Neutrality N.

The optimal Neutrality of 1/2 is obtained for values of θ near 1/2, just at the point where the WIM also reports the smallest impact. The further we move towards the edges of the parameter space, the higher the WIM between priors and the less Neutrality the priors have. For Jeffreys' prior, this relation is monotone for all sample sizes. For Haldane's prior the relation is monotone only for $n = 100, 200$; for $n = 50$ we see a small non-monotonicity near the outer edge and for $n = 10$, we see that the Neutrality reverses in sign with respect to 0.50 around $\theta = 0.20$ and 0.80. The WIM is here a better interpretable, because monotone prior measure when moving away in either sense from $\theta = 0.50$. For the neutral $Beta(1/3, 1/3)$ prior, the graph shows that it optimally preserves Neutrality near 1/2 for $n = 50, 100, 200$, where one thus resorts to the WIM to detect differences. Again both prior impact measures decrease with the sample size.

Since the MOPESS is a highly variable metric, the relationship with the WIM is not nicely monotone as with the Neutrality (even after averaging over 1000 replicates). For Haldane's prior, the MOPESS lies close to zero except for $n = 10$ and $\theta = 0.95$. In addition, for the other combinations of sample size and success rate we do observe an increase in the MOPESS

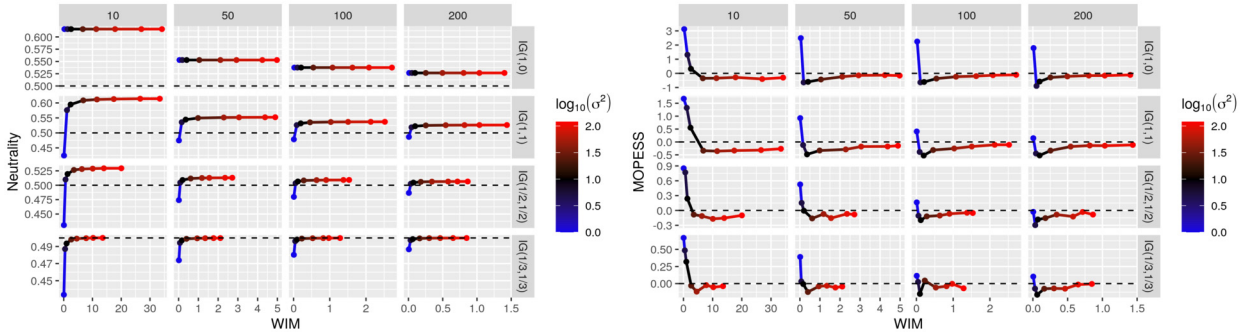

Fig. 3. WIM vs. Neutrality (left panel) and WIM vs. MOPESS (right panel) for the normal model for different priors (rows) and sample sizes (columns). The colour scale presents the $\log_{10}$(variance). The baseline prior is the Jeffreys prior. Note the different axis scales.

towards the edges of the parameter space which makes sense as the Haldane prior inflates towards 0 and 1. A similar behaviour exists for the neutral prior and Jeffreys' prior for the sample size $n = 10$ (increasing of the MOPESS near the edge of the parameter space), however for larger sample sizes, the relationship between the WIM, the MOPESS and the success rate is very erratic due to the MOPESS' variability. We also observe that the MOPESS does not decrease with sample size; the clearer structure of the WIM makes it a better understandable measure for the binomial case.

3.3.3. Normal case

Finally we choose as continuous model the normal distribution with probability density function

$$x \mapsto \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R},$$

with location parameter $\mu \in \mathbb{R}$ and dispersion parameter $\sigma > 0$. We consider σ^2 to be the parameter of interest and μ is fixed (known a priori). We wish to compare here the improper Jeffreys prior $p(\sigma^2) \propto 1/\sigma^2$, invariant under reparameterization, and the Inverse Gamma (IG) prior with positive real parameters α and β which happens to be the conjugate prior for the normal distribution.

Fig. 3 presents the WIM (x-axis) versus the Neutrality and the MOPESS (y-axis), respectively, for different priors and sample sizes. For the WIM and the MOPESS, the baseline prior is the Jeffreys prior, which is not required for the Neutrality N. All prior impact measures decrease with the sample size.

We first notice that the uniform $IG(1, 0)$ prior's Neutrality does not depend on the variance and converges slowly to 1/2 as the sample size is increasing. Thus, for the comparison of the $IG(1, 0)$ prior with Jeffreys' prior, the WIM is a more informative choice. For the other priors, the relationship is monotone at first until it reaches an asymptote for the Neutrality which is dependent on the sample size. For the $IG(1/3, 1/3)$ prior, we observe that this asymptote is at 1/2 and reached rather quickly, indicating that Neutrality of this prior is optimal for a large range of values and the $IG(1/3, 1/3)$ thus can be considered a neutral prior as defined by Kerman (2011). It is notable that while the Neutrality converges to an asymptote and is thus no longer sensitive to distinguish between priors, the unbounded WIM is still able to distinguish further between the two priors.

For smaller values of σ^2, the MOPESS values are positive for all priors and sample sizes, which was expected because the priors under consideration have the dominant part of their probability mass near the limit of zero. So when the prior and likelihood are strongly similar, this corresponds to additional observations. However, only for the $IG(1, 0)$ and $IG(1, 1)$ priors is the MOPESS larger than 1. The MOPESS shrinks to zero or slightly below with increasing values of σ^2, indicating more impact from the Jeffreys prior. Although the relationship between the WIM and the MOPESS is less erratic than for the binomial model, the relationship is still not monotone. A possible reason for this behaviour is that the Wasserstein distance has no sign to indicate which of the two priors is more impactful. A second reason is that finite sampling variability limits the display of a monotone relationship (Jones et al., 2021). Both the MOPESS and WIM report insightful information and it seems hard to pick a better choice in this setting.

4. Illustration of the WIM on two data sets

4.1. Priors for the skewness parameter of the skew-normal model

Many distributions have been built to capture skewness in data. Arguably the most famous instance is the skew-normal distribution of Azzalini (1985) with probability density function

$$x \mapsto \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\alpha\frac{(x-\mu)}{\sigma}\right), \quad x \in \mathbb{R}, \tag{2}$$

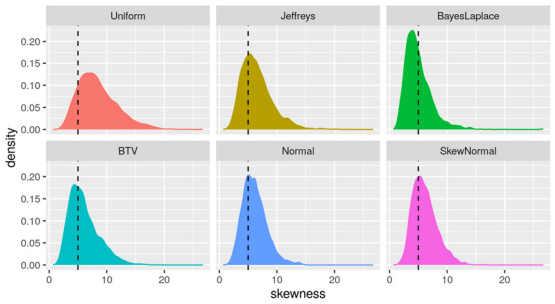

Fig. 4. Posterior distributions for the skewness parameter of the skew-normal model for the frontier dataset based on different priors.

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma \in \mathbb{R}^+$ the scale parameter, both inherited from the standard normal distribution with pdf denoted by ϕ and cumulative distribution function Φ, and $\alpha \in \mathbb{R}$ is called the skewness parameter. The density (2) is an asymmetric model for $\alpha \neq 0$ and reduces to a standard normal when $\alpha = 0$. A well-known problem of the skew-normal model in relation with frequentist inference is that, for some datasets, the maximum likelihood estimate of the skewness parameter becomes infinite (Azzalini and Capitanio, 1999). A famous example of such a situation is the *frontier* dataset, to be found in the *sn* package of R where the MLE for the skewness is 1.4e+06 for 50 draws from a skew-normal with $\mu = 0, \sigma = 1$ and skewness $\alpha = 5$. In these cases a Bayesian estimation procedure seems a meaningful alternative. A review of different priors for the skew-normal model and more general skew-symmetric distributions is given in Ghaderinezhad et al. (2020). We will apply various priors discussed in that paper to the frontier dataset and calculate the WIM between the resulting posteriors. The following priors will be examined (the * indicates that some priors are approximated by a known parametric form for ease of computation):

- Uniform/flat prior: $p(\alpha) \propto 1$.
- Jeffreys' prior (*): a tractable approximation of the Jeffreys prior $p(\alpha)$ is $t_{0,\pi^2/4,0.5}(\alpha)$ where $t_{a,b,c}$ is the density of the Student t-distribution with location $a \in \mathbb{R}$, scale $b > 0$ and $c > 0$ degrees of freedom (Bayes and Branco, 2007). This prior is proper, symmetric around zero, decreasing in $|\alpha|$ and its tails are of the order $O(\alpha^{-3/2})$.
- Bayes-Laplace prior: following the Bayes-Laplace rule, Bayes and Branco (2007) proposed a uniform prior on the interval $[-1, 1]$ for $\frac{\alpha}{\sqrt{1-\alpha^2}}$, which corresponds, for α, to the prior $p(\alpha) = t_{0,0.5,2}(\alpha)$.
- Beta Total Variation (BTV) prior (*): belonging to the class of distance-based priors proposed by Dette et al. (2018), the rationale for this prior on the skewness is that α not only controls the skewness, but shifts the entire distribution and hence a prior should rather be set on the Wasserstein distance between the normal and the skew-normal and, from there, one can derive a prior for α. We choose the so-called $BTV(1, 1)$ prior which can be approximated by $p(\alpha) = t_{0,0.92,1}(\alpha)$.
- Normal prior: an informative prior, chosen such that the mean is zero in order to be comparable to the location of the other priors, and the variance is set to cover a reasonable scope of values, resulting in the prior $\mathcal{N}(0, 5)$ (Canale and Scarpa, 2013).
- Skew-normal prior: another informative prior suggested by Canale and Scarpa (2013). It combines the location and scale of the $\mathcal{N}(0, 5)$ with a skewness value of 2.

All posterior distributions are numerically sampled by Markov Chain Monte Carlo with the T-walk algorithm in R (Christen and Fox, 2010). Based on preliminary analyses, we used chains of length 100k with a 50k burn-in and thinning rate of 5 as this yields stable results for this dataset. Fig. 4 gives an overview of the posteriors obtained from the different aforementioned priors. We clearly observe that the mode of each posterior is very close to the true value $\alpha = 5$ of the frontier data.

Table 1 contains the WIM for each pair of posteriors resulting from different priors. It reflects what can be seen in Fig. 4, namely that the posterior based on the uniform/flat prior is considerably different from all the others. Setting the uniform prior as baseline prior, the Jeffreys prior would have the least impact compared to the other priors. The smallest distance occurs between the informative normal and skew-normal priors, indicating that it makes little difference which of these two we choose. It is interesting to note that the $BTV(1, 1)$ prior, considered to be non-informative (Dette et al., 2018), seems to have slightly bigger impact than these informative priors when compared to the prior-free case (uniform). This requires further investigation, which is why we also quantify the uncertainty behind these distances. To this end we use bootstrap resampling (250 bootstrap samples) to obtain the sampling distribution of the WIM for all pairs of posteriors by calculating for each bootstrap sample of the frontier dataset the distance between the posteriors. The results are summarized in Fig. 5. The WIMs involving the uniform prior are not only the largest, but also have the highest variability. In particular, the WIM between the uniform and the informative priors has the largest variability, clearly larger than with the $BTV(1, 1)$ prior, which is now well in line with the non-informative/informative character of these priors. Not surprisingly, the least variable distance occurs between the normal and the skew-normal priors.

Table 1
WIM between pairs of posteriors resulting from different priors for the skewness parameter α in the skew-normal model for the frontier dataset.

Priors	Jeffreys	Bayes-Laplace	BTV	Normal	SkewNormal
Uniform	1.604	3.138	2.213	2.081	2.067
Jeffreys		1.534	0.609	0.504	0.515
Bayes-Laplace			0.926	1.072	1.077
BTV				0.452	0.453
Normal					0.040

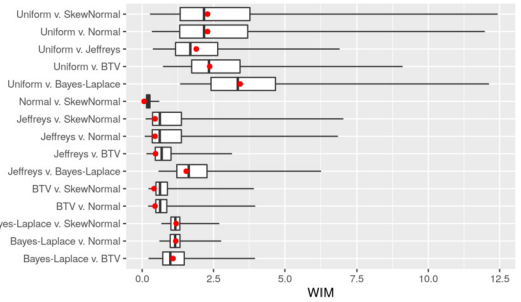

Fig. 5. Boxplot of the WIMs between all pairs of posteriors resulting from different priors for the skewness parameter in the skew-normal model. The boxplots are based on $B = 250$ bootstrap simulations and the red dots indicate the value of the WIM of the original "frontier" dataset.

4.2. Logistic regression and weakly informative priors

In the present section, we compare the uniform versus so-called *weakly informative* priors for the logistic regression model involving a single continuous covariate:

$$logit(\pi_i) = \beta_0 + \beta_1 x_i,$$

where $\beta_0, \beta_1 \in \mathbb{R}$ are the regression parameters, π_i is the probability that observation i is a success and the *logit* function is the logarithm of the odds. An important application of Bayesian inference occurs in dose-response studies where the parameter of interest is the "LD50" (lethal dose), the dose (x) where the probability of death (π) is exactly 50%. Using maximum likelihood estimation, we can get a point estimate for this parameter $LD50 = -\frac{\beta_0}{\beta_1}$, however there is no standard solution for the standard error. Instead Bayesian inference allows deriving the posterior for LD50 from the posterior samples of β_0 and β_1 (Gelman et al., 2004). Here we examine different priors for the data of Racine et al. (1986) that are reported in Gelman et al. (2008). This is a small-scale bioassay experiment with only $n = 4$ binomial observations (each based on 5 replicates) for different doses. With small samples, the choice of the prior is even more important than for larger sample sizes.

The difference of this application with previous models is that we are now in a multivariate situation. There are two regression parameters that are of interest plus a derived parameter, LD50. This means that the measures of prior impact should be inspected both on the joint posterior distribution and on the marginals, as they could potentially yield different information on prior impact. The two priors we will consider are the following:

- Uniform prior: The uniform prior places a uniform distribution on $\mathbb{R}^2 : p(\beta_0, \beta_1) \propto 1$.
- Cauchy prior: Gelman et al. (2008) propose to use the Cauchy prior centered at zero with scale parameters 2.5 and 10 as default priors for the slope and intercept parameters, respectively. Before using these priors, all covariates (in this case only the dose) need to be rescaled to have mean zero and standard deviation 1/2. The authors showed that these priors are weakly informative, not adding excess information on the analysis.

We will complement the already done analysis by computing the WIM between the uniform and Cauchy priors on the bioassay data. Besides the suggested scale parameter of the slope (2.5), we will also take 5 and 10 as scale parameters as in Reimherr et al. (2014) who investigated a similar problem for a distinct dataset. All models were fitted in `R` using the STAN language (Brukner, 2018) with the NUTS sampler (4 chains of length 2000, burn-in 1000, thinning 3). Fig. 6 plots the posteriors for the regression coefficients and the LD50 parameter, respectively.

Table 2 presents the WIM for the bioassay dataset. For all Cauchy-prior based posteriors, as much as the scale parameter of the slope's prior increases, the joint and marginal Wasserstein distances with the posterior based on the uniform prior decrease, which is not surprising since a larger spread of the Cauchy distribution implies more and more uniform-like behaviour. This conclusion of decreasing WIM holds also for the quantity of interest LD50. We further notice the dependency between both marginals since changing only the slope's scale impacts also the intercept's WIM. A look at the marginals

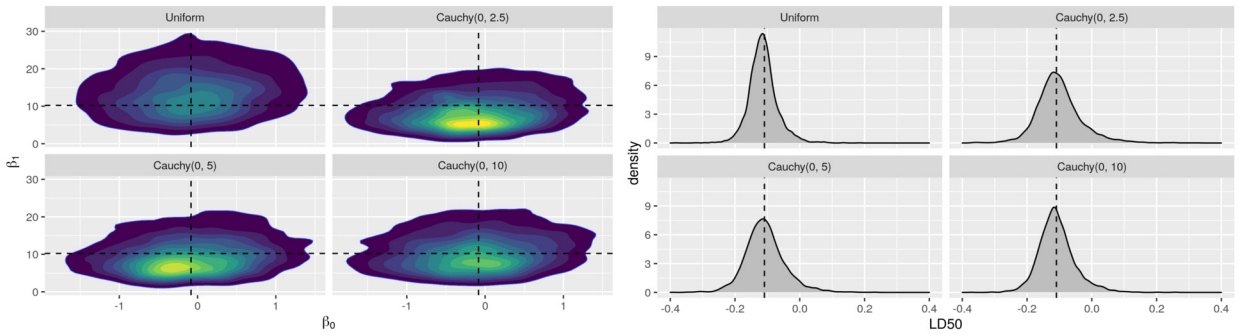

Fig. 6. Left: Joint posterior densities for β_0 and β_1 (the covariate was scaled according to Gelman et al. (2008)). The colour scale reflects posterior density (dark blue = low density, yellow = high density). Right: Posterior densities for the LD50 (lethal dose at which the probability of death is 50%). Only the prior for β_1 is indicated in the Cauchy settings, since for β_0 we always use Cauchy(0,10) (Gelman et al., 2008). The dashed lines present the point estimate based on the MLE of the regression coefficients.

Table 2
WIM between the posteriors (joint, marginal and LD50) based on the uniform prior and the weakly-informative Cauchy prior for the bioassay data. Only the prior for β_1 is indicated in the Cauchy settings, since for β_0 we always use Cauchy(0,10) (Gelman et al., 2008).

	(β_0, β_1)	β_0	β_1	LD50
Uniform vs. Cauchy(0, 2.5)	6.115	0.129	6.113	0.028
Uniform vs. Cauchy(0, 5.0)	5.162	0.090	5.161	0.017
Uniform vs. Cauchy(0, 10.0)	3.851	0.060	3.850	0.013

reveals that the distance between joint posteriors seems mainly guided by the slope whose marginal WIM is quite close to the WIM of the joint distribution. This however is likely due to the larger values of β_1 compared to β_0; though smaller in absolute values, the WIM for β_0 varies more in percentage than the WIM for β_1. We thus recommend that in future studies the judgement of the prior impact should be done on the marginal posteriors to get a full picture.

5. Discussion

In this paper we have introduced the WIM, a practically oriented measure of prior impact that allows comparing any two priors by quantifying the Wasserstein distance between the resulting posteriors (provided they have finite absolute moments of order 1). Our proposal thus retains the appealing intuitive touch of the approach from Ley et al. (2017) and Ghaderinezhad and Ley (2019) while palliating its drawbacks, meaning that the WIM can also deal with multi-parameter and multi-dimensional situations, with non-nested priors and complicated forms of the posteriors (which are becoming more and more routine in modern applications). Through a Monte Carlo simulation study we compared our WIM to two prior impact measures from the literature, namely the concepts of Neutrality and MOPESS. We could see that the WIM is an attractive alternative to these known proposals, since in various cases it provides more information (e.g., when the Neutrality reaches its upper bound from a certain point on) and is better interpretable (thanks to a monotone prior measure, which both the Neutrality and the MOPESS can lack). Moreover, it does not suffer from the high variability that the MOPESS exhibits.

We will now wrap up the comparison by discussing further properties. While the MOPESS is a comparative measure like the WIM, the Neutrality is an absolute measure. Unlike the WIM, the MOPESS requires choosing a baseline prior; an advantage of the MOPESS however is its sign which allows finding out which of the two priors is closer w.r.t. the current data set. Our WIM is rather quick to compute (as is Neutrality), unlike the MOPESS; indeed, for non-conjugate models advanced sampling algorithms are necessary for the MOPESS and need to be repeated several times (in particular for higher dimensions), implying a long computation time. A further appealing property of our WIM is its broad usage, of which the two different real data sets testify. They could not have been tackled via the two competitor impact measures. Indeed, the Neutrality cannot be applied on the frontier data since the maximum likelihood estimate lies at the boundary of the parameter space, a situation the Neutrality cannot handle by definition. Since no multivariate extension of the Neutrality exists in the literature, it can also not be used for the LD50 logistic regression problem. The MOPESS cannot be used there, either, because it would require to know the distribution for the covariate which is unrealistic. We did apply the MOPESS on the frontier dataset, however ran into the following issue: the high variability of the MOPESS led to OPESS ranges $[5\%, 95\%]$ that are considerably wide and contain the zero value. This observation was also noted in Jones et al. (2021) who emphasized that care should be taken in interpreting the impact solely on basis of the MOPESS. The aforementioned problems of the Neutrality and the MOPESS are all restrictions that the WIM does not possess.

Acknowledgements

This research is supported by a BOF Starting Grant of Ghent University. The authors would further like to thank David E. Jones, Robert N. Trangucci and Yang Chen for helping them with the R code for the MOPESS. They are very grateful to the Associate Editor and two anonymous reviewers for their helpful comments that allowed a clear improvement of the presentation.

References

Diaconis, P., Freedman, D., 1986a. On the consistency of Bayes estimates (with discussion and rejoinder by the authors). Ann. Stat. 14, 1–67.
Diaconis, P., Freedman, D., 1986b. On inconsistent Bayes estimates of location. Ann. Stat. 14, 68–87.
Reimherr, M., Meng, X., Nicolae, D.L., 2014. Being an informed Bayesian: assessing prior informativeness and prior likelihood conflict. arXiv:1406.5958.
Lin, X., Pittman, J., Clarke, B., 2007. Information conversion, effective samples, and parameter size. IEEE Trans. Inf. Theory 53, 4438–4456.
Morita, S., Thall, P.F., Müller, P., 2008. Determining the effective sample size of a parametric prior. Biometrics 64, 595–602.
Wiesenfarth, M., Calderazzo, S., 2019. Quantification of prior impact in terms of effective current sample size. Biometrics 76, 595–602.
Jones, D.E., Trangucci, R.N., Chen, Y., 2021. Quantifying observed prior impact. Bayesian Anal. 1, 1–28.
Kerman, J., 2011. Neutral noninformative and informative conjugate beta and gamma prior distributions. Electron. J. Stat. 5, 1450–1470.
Ley, C., Reinert, G., Swan, Y., 2017. Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. Ann. Appl. Probab. 27, 216–241.
Ghaderinezhad, F., Ley, C., 2019. Quantification of the impact of priors in Bayesian statistics via Stein's method. Stat. Probab. Lett. 146, 206–212.
Ghaderinezhad, F., Ley, C., 2020. On the impact of the choice of the prior in Bayesian statistics. In: Tang, N. (Ed.), Bayesian Inference for Complicated Data. ISBN 978-1-83880-386-5. IntechOpen.
Vaserstein, L.N., 1969. Markov processes over denumerable products of spaces describing large system of automata. Probl. Pereda. Inf. 5, 64–72.
Vallender, S., 1974. Calculation of the Wasserstein distance between probability distributions on the line. Theory Probab. Appl. 18, 784–786.
Schuhmacher, D., Bhre, B., Gottschlich, C., Hartmann, V., Heinemann, F., Schmitzer, B., 2019. Transport: Computation of optimal transport plans and Wasserstein distances. R package version 0.12-1.
Sommerfeld, M., Munk, A., 2017. Inference for empirical Wasserstein distances on finite spaces. J. R. Stat. Soc. B 80, 219–238.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. Chapman & Hall/CRC, London.
Azzalini, A., 1985. A class of distributions which includes the normal ones. Scand. J. Stat. 12, 171–178.
Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew-normal distribution. J. R. Stat. Soc. B 61, 579–602.
Ghaderinezhad, F., Ley, C., Loperfido, N., 2020. Bayesian inference for skew-symmetric distributions. Symmetry 12, 491.
Bayes, C., Branco, E., 2007. Bayesian inference for the skewness parameter of the scalar skew-normal distribution. J. Stat. Plan. Inference 21, 141–163.
Dette, H., Ley, C., Rubio, F.J., 2018. Natural (non-)informative priors for skew-symmetric distributions. Scand. J. Stat. 45, 405–420.
Canale, A., Scarpa, B., 2013. Informative Bayesian inference for the skew-normal distribution. arXiv:1305.3080.
Christen, J., Fox, C., 2010. A general purpose sampling algorithm for continuous distributions (the t-walk). Bayesian Anal. 5, 263–282.
Racine, A., Grieve, A.P., Fluhler, H., Smith, A.F.M., 1986. Bayesian methods in practice: experiences in the pharmaceutical industry (with discussion). J. Appl. Stat. 35, 93–150.
Gelman, A., Jakulin, A., Pittau, G., Su, Y.S., 2008. A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. 2, 1360–1383.
Brukner, P., 2018. Advanced Bayesian multilevel modeling with the R package brms. R J. 10, 395–411.