Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On a prior based on the Wasserstein information matrix

W. Li [a], F.J. Rubio [b,*]

[a] Department of Mathematics, University of South Carolina, SC, USA
[b] Department of Statistical Science, University College London, London, UK

ARTICLE INFO

Article history:
Received 11 February 2022
Received in revised form 6 July 2022
Accepted 27 July 2022
Available online 5 August 2022

Keywords:
Fisher information matrix
Jeffreys prior
Wasserstein-2 distance
Wasserstein information matrix
Wasserstein prior

ABSTRACT

We introduce a prior for the parameters of univariate continuous distributions, based on the Wasserstein information matrix, which is invariant under reparameterisations. We discuss the links between the proposed prior with information geometry. We present sufficient conditions for the propriety of the posterior distribution for general classes of models. We present a simulation study that shows that the induced posteriors have good frequentist properties.

1. Introduction

In Bayesian parametric inference, the choice of the prior plays a fundamental role. In scenarios where the prior information about the model parameters is vague or unreliable, it is desirable to use priors which do not require the user to specify their parameters (hyperparameters). The main aim of *Objective Bayes* (Berger, 2006; Consonni et al., 2018) is indeed to produce priors via formal rules (Kass and Wasserman, 1996), which typically depend only on the statistical model. Such rules usually aim at producing a prior that has little effect on the inference on the parameters, or that is invariant under reparameterisations, or that penalises the model complexity. Priors obtained with formal rules are usually referred to as *Objective priors* or *Non-informative priors*. We refer the reader to Leisen et al. (2020) for a recent review of methods for constructing priors based on formal rules. A pioneering contribution in this area is the *Jeffreys prior* (Jeffreys, 1946), which is obtained by calculating the square root of the determinant of the Fisher information matrix (FIM) (Robert et al., 2009). The aim behind the construction of the Jeffreys prior is to produce a prior that is invariant under reparameterisations.

Another direction for constructing a prior based on a formal rule consists of looking at the genesis of the Jeffreys prior (Kass and Wasserman, 1996). The Jeffreys prior is typically motivated by its invariance under reparameterisations, however, it can also be motivated using concepts from information geometry (Amari, 2016; Nielsen, 2020; Amari, 2021; Amari and Matsuda, 2022). Briefly, the Kullback–Leibler divergence behaves locally as a function of a distance function determined by the Riemannian metric. The Jeffreys prior can be seen as the natural volume associated to such metric, and natural volume elements generate uniform measures on manifolds (Kass and Wasserman, 1996). Moreover, natural volumes of Riemannian metrics are invariant under reparameterisations (Kass and Wasserman, 1996). Intuitively, this suggests that other distances could be used to construct alternative priors. In this line, a natural alternative consists of using the optimal transport induced information matrix (Li and Zhao, 2019), referred to as the Wasserstein information matrix (WIM). The construction of the WIM can be justified using ideas from "transport information geometry", which is

[*] Corresponding author.
E-mail address: f.j.rubio@ucl.ac.uk (F.J. Rubio).

https://doi.org/10.1016/j.spl.2022.109645

the intersection between optimal transport (Villani, 2003) and information geometry (Amari, 2016, 2021). We refer the reader to Li (2021b,a) and Amari (2021) for a more extensive treatment of this area. The idea behind the construction of the WIM consists of using tools from optimal transport, where a distance between distributions is used to construct an information matrix. Li and Zhao (2019) focused on the particular choice of the Wasserstein-2 distance. The Wasserstein-2 distance can be associated to a metric operator, namely the WIM, which is different in nature from the Fisher information matrix (Amari, 2021). Although a vast amount of literature has been devoted to the study of the Fisher information matrix and the Jeffreys prior, there is a void in the study of priors associated to the Wasserstein information matrix.

We propose a formal rule for constructing a prior, which is invariant under reparameterisations, based on the Wasserstein information matrix. The construction of this prior (referred to as the Wasserstein prior hereafter) is analogous to that of the Jeffreys prior. However, as shown later, we find that the Wasserstein prior has a different functional form for several models and, appealingly, requires a lower order of differentiability. This helps overcome some challenges with the Jeffreys prior, where the required higher order of differentiability precludes its construction for some non-regular models (Shemyakin, 2014; Li and Zhao, 2019). Moreover, as we will show later in the simulation study, the Wasserstein prior induces a posterior with good frequentist properties in the models studied here.

2. Wasserstein information matrix

Let X be a continuous random variable with finite second moment, and $F(x \mid \boldsymbol{\theta})$ be the corresponding cumulative distribution function (cdf) with support $\mathcal{D} \subset \mathbb{R}$, and parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, with $d \geq 1$. Let us assume that $F(x \mid \boldsymbol{\theta})$ is absolutely continuous, and let $f(x \mid \boldsymbol{\theta})$ be the corresponding probability density function (pdf).

Consider the Wasserstein information matrix (WIM) proposed in Li and Zhao (2019)

$$W_{ij}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\dfrac{\partial}{\partial \boldsymbol{\theta}_i}F(X \mid \boldsymbol{\theta})\dfrac{\partial}{\partial \boldsymbol{\theta}_j}F(X \mid \boldsymbol{\theta})}{f(X \mid \boldsymbol{\theta})^2}\right], \tag{1}$$

where the expectation is taken with respect to $F(x \mid \boldsymbol{\theta})$. A clear difference between the WIM and the FIM is that the former is based on derivatives of the cdf (with respect to the parameters), while the latter is based on derivatives of the pdf. This is an appealing property as it reduces the conditions for the existence of the WIM (Li and Zhao, 2019), allowing its construction for non-regular models. Next, we present a brief description of the motivation behind the construction of the WIM. The details are somewhat technical, but we refer the reader to Li and Zhao (2019) for a detailed derivation of the WIM.

As discussed in Section 1, a distance between probability distributions can be used to define an information matrix. In our case, we focus on the analysis of the information matrix (WIM) implied by the Wasserstein-2 distance. Given two parameter values $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \Theta$, the Wasserstein-2 distance between two probability distributions with support on $\mathcal{D} \subset \mathbb{R}$, $F(\cdot \mid \boldsymbol{\theta}_0)$ and $F(\cdot \mid \boldsymbol{\theta}_1)$, satisfies the following relationship with the corresponding quantile functions (Villani, 2003)

$$\text{Dist}_W(F(\cdot \mid \boldsymbol{\theta}_0), F(\cdot \mid \boldsymbol{\theta}_1)) = \sqrt{\int_0^1 \left|F^{-1}(u \mid \boldsymbol{\theta}_0) - F^{-1}(u \mid \boldsymbol{\theta}_1)\right|^2 du},$$

where F^{-1} is the quantile function associated to the cdf F. It can be shown that the Wasserstein-2 distance Dist_W defines a Riemannian metric among probability distributions (Villani, 2003), which can be used to establish the connection between such metric with an information matrix. More specifically, the infinitesimal expansion of the squared Wasserstein-2 distance establishes a link of this metric and the Wasserstein information matrix. That is, let $\Delta\boldsymbol{\theta} = (\Delta\boldsymbol{\theta}_1, \ldots, \Delta\boldsymbol{\theta}_d)^\top \in \mathbb{R}^d$ such that $\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta} \in \Theta$, one can show that (Li and Zhao, 2019)

$$\text{Dist}_W(F(\cdot \mid \boldsymbol{\theta}_0), F(\cdot \mid \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}))^2 = \sum_{i,j=1}^{d} W_{ij}(\boldsymbol{\theta}_0)\Delta\boldsymbol{\theta}_i\Delta\boldsymbol{\theta}_j + o(\|\Delta\boldsymbol{\theta}\|^2).$$

This shows a link between the Wasserstein-2 distance and the WIM, which is discussed in detail in section 7.7 of Amari (2021). We can also see from this result that the WIM shares a similar derivation to that of the Fisher information matrix (see Chapter 5 of Ghosh et al. (2006) for an extensive discussion). We remark that one can also define the WIM in higher dimensional sample spaces (that is, for random vectors). However, this requires solving an elliptic partial differential equation (Li and Zhao, 2019).

3. The Wasserstein prior

In this section, we propose the Wasserstein prior, whose main motivation is to obtain an invariant prior. We also describe the precise meaning of the invariance property and its connection with the Jeffreys prior.

3.1. One parameter case

Consider the case where $d = 1$, that is, we focus on the case where $F(x \mid \theta)$ contains only one parameter. Then, the WIM (1) becomes

$$W(\theta) = \mathbb{E}\left[\frac{\left\{\frac{\partial}{\partial\theta}F(X \mid \theta)\right\}^2}{f(X \mid \theta)^2}\right].$$

Let $\varphi = h(\theta)$ be a reparameterisation of $F(x \mid \theta)$. Let us denote the WIM associated to $F(x \mid \theta)$ by $W(\theta)$, and the WIM associated to $F(x \mid \varphi)$ by $\tilde{W}(\varphi)$. From the above expression, we can see that

$$\tilde{W}(\varphi) = W(\theta)\left(\frac{d\theta}{d\varphi}\right)^2.$$

Indeed, the FIM also satisfies this relationship (Robert et al., 2009). This suggests the construction of an invariant prior, based on the WIM, in a similar fashion as the Jeffreys prior (which is based on the FIM). Define the prior (up to a positive proportionality constant)

$$\pi_W(\theta) \propto \sqrt{W(\theta)}.$$

It follows that this prior is invariant under reparameterisations in the sense that

$$\pi_{\tilde{W}}(\varphi) = \pi_W(\theta)\left|\frac{d\theta}{d\varphi}\right|,$$

where $\pi_{\tilde{W}}(\varphi) \propto \sqrt{\tilde{W}(\varphi)}$. That is, the priors $\pi_W(\theta)$ and $\pi_{\tilde{W}}(\varphi)$ are related by the corresponding change of variable. Therefore, this represents a strategy for constructing a prior based on a formal rule (Kass and Wasserman, 1996) which is invariant under reparameterisations, in the same spirit as the invariance property of the Jeffreys prior (Jeffreys, 1946). We formalise this idea next.

3.2. Multi-parameter case

Consider now the general case $d \geq 1$ and let $\boldsymbol{\varphi} = h(\boldsymbol{\theta})$ be a reparameterisation of $F(x \mid \boldsymbol{\theta})$. Note first that the WIM of $\boldsymbol{\varphi}$, $\tilde{W}(\boldsymbol{\varphi})$, can be written after a change of variable as:

$$\tilde{W}(\boldsymbol{\varphi}) = \mathbf{J}^\top W(\boldsymbol{\theta})\mathbf{J},$$

where $\mathbf{J}$ is the Jacobian matrix with entries

$$\mathbf{J}_{ij} = \frac{\partial\boldsymbol{\theta}_i}{\partial\boldsymbol{\varphi}_j}.$$

The proof of this result is analogous to the proof of the invariance property of the FIM, which can be found in Lehmann and Casella (2006). Consequently, we have that

$$\det\tilde{W}(\boldsymbol{\varphi}) = \det W(\boldsymbol{\theta})(\det\mathbf{J})^2.$$

This result suggests the construction of an invariant prior, based on the WIM, in a similar fashion as the Jeffreys prior is obtained from the FIM. The construction of this prior is formalised in the following definition.

Definition 1. The Wasserstein prior is defined, up to a positive proportionality constant, as

$$\pi_W(\boldsymbol{\theta}) \propto \sqrt{\det W(\boldsymbol{\theta})}, \tag{2}$$

where $W(\boldsymbol{\theta})$ denotes the Wasserstein information matrix (1).

4. Examples

In this section, we present three examples where we illustrate the calculation of the WIM and the Wasserstein prior. In all cases, we provide sufficient conditions for the propriety of the posterior distribution.

The location-scale family

Let f_0 be a symmetric and unimodal pdf with mode at 0 and support on $\mathbb{R}$, and F_0 be the corresponding cdf. Let

$$F(x \mid \mu, \sigma) = F_0\left(\frac{x-\mu}{\sigma}\right), \quad f(x \mid \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}, \tag{3}$$

denote the cdf and pdf of the class of symmetric and unimodal location-scale family of distributions, with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma \in \mathbb{R}_+$.

Theorem 1. *Suppose that $\int_{-\infty}^{\infty} t^2 f_0(t)dt < \infty$. The WIM and the Wasserstein prior of (μ, σ) in the location-scale family (3) are:*

$$W(\mu, \sigma) = \begin{pmatrix} 1 & 0 \\ 0 & \int_{-\infty}^{\infty} t^2 f_0(t)dt. \end{pmatrix}, \quad \pi_W(\mu, \sigma) \propto 1. \tag{4}$$

An important class of location-scale models is the family of scale mixtures of normal distributions. A pdf f_0 is said to belong to family of scale mixtures of normal distributions if it can be represented as:

$$f_0(x) = \int_0^{\infty} \lambda^{\frac{1}{2}} \exp\left\{-\frac{\lambda x^2}{2}\right\} dH(\lambda), \tag{5}$$

where H is a cumulative distribution function with support on $\mathbb{R}_+$. The family of scale mixtures of normal distributions contains important distributions such as the Normal, Logistic, Laplace, Student-t, among other distributions (see Rubio and Steel, 2014 for a discussion). The next result provides sufficient conditions for the propriety of the posterior distribution of (μ, σ) under the Wasserstein prior (4) for the case when f_0 belongs to the family of scale mixtures of normal distributions.

Theorem 2. *Let $\mathbf{x} = (x_1, \dots, x_n)^\top$ be an i.i.d. sample from (3) with f_0 given by (5). Suppose that $\int_{-\infty}^{\infty} t^2 f_0(t)dt < \infty$. Then, the posterior distribution of (μ, σ) associated to the Wasserstein prior (4) is proper if $n > 2$ and*

$$\int_0^{\infty} \lambda^{1/2} dH(\lambda) < \infty.$$

The skew-normal distribution

We now present a result in a one-parameter model, where we obtain the Wasserstein prior for the skewness parameter of the skew-normal distribution (Azzalini, 1985). Let $\phi(x)$ and $\Phi(x)$ be the pdf and cdf of the standard normal distribution. The skew-normal pdf is defined as (Azzalini, 1985):

$$s(x \mid \alpha) = 2\phi(x)\Phi(\alpha x), \quad x \in \mathbb{R}, \tag{6}$$

where $\alpha \in \mathbb{R}$ is a skewness parameter. The following result characterises the WIM and Wasserstein prior of α.

Theorem 3. *Consider the skew-normal distribution (6). Then,*

(i) The WIM of α is given by

$$W(\alpha) = \int_{-\infty}^{\infty} \frac{\sqrt{2} e^{-\frac{1}{2}\left(2\alpha^2+1\right)x^2}}{\pi^{3/2}\left(\alpha^2+1\right)^2 \left(erf\left(\frac{\alpha x}{\sqrt{2}}\right)+1\right)} dx.$$

(ii) The Wasserstein prior

$$\pi_W(\alpha) \propto \sqrt{\int_{-\infty}^{\infty} \frac{\sqrt{2} e^{-\frac{1}{2}(2\alpha^2+1)x^2}}{\pi^{3/2}\left(\alpha^2+1\right)^2 \left(erf\left(\frac{\alpha x}{\sqrt{2}}\right)+1\right)} dx}, \tag{7}$$

is symmetric about 0.

(iii) $\pi_W(\alpha)$ is integrable.

(iv) The tails of $\pi_W(\alpha)$ are of order $\mathcal{O}(|\alpha|^{-5/2})$.

The tail behaviour of the Wasserstein prior $\pi_W(\alpha)$ differs from that of the Jeffreys prior of α (Rubio and Liseo, 2014), which has tails of order $\mathcal{O}(|\alpha|^{-3/2})$, and the total variation prior proposed in Dette et al. (2018), which has tails of order $\mathcal{O}(|\alpha|^{-2})$. The characterisation of the propriety and tail behaviour of $\pi_W(\alpha)$ in the previous theorem suggests that one could approximate it using a symmetric distribution with the same tail behaviour. A natural candidate is the Student-t distribution with $\nu = 3/2$ degrees of freedom. We found that a scale parameter $\sigma_t = 0.757$ produces a good approximation in the main body of the distribution, while the tails have the exact same weight (see Fig. 1).

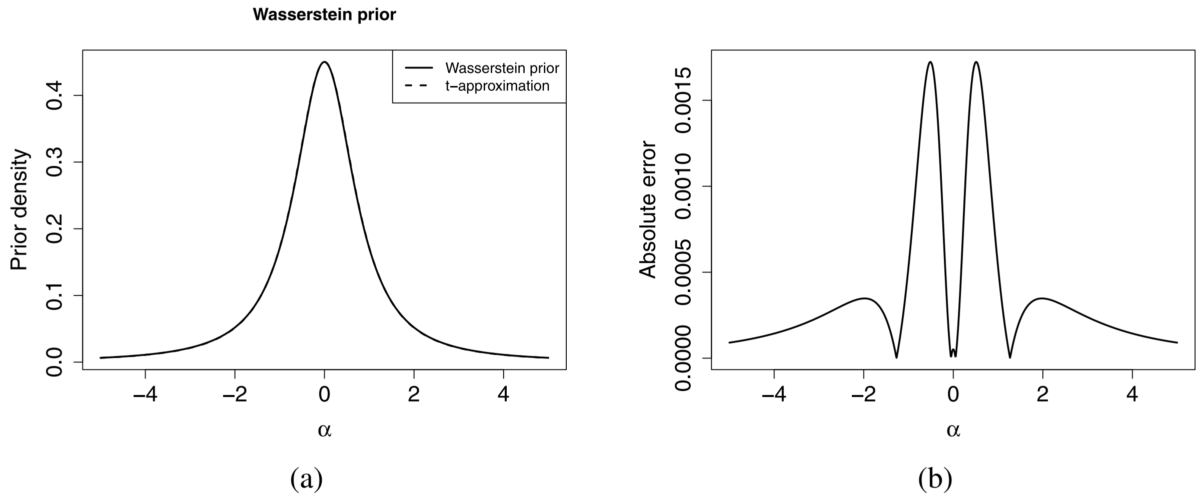

Fig. 1. (a) Wasserstein prior and Student-t approximation. (b) Absolute error of the Student-t approximation.

In the next theorem, we construct a prior for the skew normal distribution with location and scale parameters (μ, σ) and skewness parameter α, based on a product prior structure using the priors (4) and (7). We show that the posterior distribution is proper under mild conditions. This prior can be interpreted as an *Independence Wasserstein prior* (analogous to the *independence Jeffreys prior*, Rubio and Steel, 2014; Rubio and Liseo, 2014), in the sense that it is constructed as the product of the Wasserstein priors for each parameter (or groups of parameters) while considering the other parameters as fixed.

Theorem 4. *Let* $\mathbf{x} = (x_1, \dots, x_n)^\top$ *be an i.i.d. sample from the skew normal distribution with pdf*

$$s(x \mid \mu, \sigma, \alpha) = \frac{2}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)\Phi\left(\alpha\frac{x-\mu}{\sigma}\right).$$

Consider the improper product prior structure using the Wasserstein priors (4) and (7)

$$\pi(\mu, \sigma, \alpha) \propto \sqrt{\int_{-\infty}^{\infty} \frac{\sqrt{2}e^{-\frac{1}{2}(2\alpha^2+1)x^2}}{\pi^{3/2}\left(\alpha^2+1\right)^2\left(erf\left(\frac{\alpha x}{\sqrt{2}}\right)+1\right)}dx}. \tag{8}$$

Then, the posterior distribution of (μ, σ, α) *is proper if* $n > 2$.

Normal linear regression

We now study the WIM and the Wasserstein prior for the normal linear regression model,

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n. \tag{9}$$

where $\mathbf{x}_i^\top \in \mathbb{R}^p$ is a vector of covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients, $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ denote the errors. Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ denote the design matrix and $\mathbf{y} = (y_1, \dots, y_n)^\top$ the vector of response variables.

Theorem 5. *Consider the linear regression model* (9), *and suppose that* $\mathbf{X}$ *has full column rank. Then, the WIM and the Wasserstein prior are given by,*

$$W(\boldsymbol{\beta}, \sigma) = \begin{pmatrix} \mathbf{X}^\top\mathbf{X} & 0 \\ 0 & 1 \end{pmatrix}, \quad \pi_W(\boldsymbol{\beta}, \sigma) \propto 1. \tag{10}$$

The next result presents sufficient conditions for the propriety of the posterior distribution of $(\boldsymbol{\beta}, \sigma)$ under the Wasserstein prior (10).

Theorem 6. *Consider the Normal linear regression model* (9) *together with the Wasserstein prior* (10). *Suppose that* $\mathbf{X}$ *has full column rank and that* $\mathbf{y}$ *is not in the column space of* $\mathbf{X}$. *Then, the posterior distribution of* $(\boldsymbol{\beta}, \sigma)$ *is proper if* $n > p + 1$.

5. Simulation studies

In this section we present two simulation studies to assess the performance of the posterior distributions induced by the Wasserstein prior.

In the first simulation scenario, we evaluate the performance of the independence Wasserstein prior (8) and compare it against the independence Jeffreys prior (Rubio and Liseo, 2014). We simulate $N = 250$ samples of size $n = 50, 250, 500$ from a skew-normal distribution (6) with $\mu = 10$, $\sigma = 1$, and $\lambda = 1, 3, 5$. We emphasise that the value $\lambda = 1$ represents a very challenging scenario as the skew-normal distribution is weakly identifiable for $|\lambda| < 1.25$; in the sense that the skew-normal pdf is virtually symmetric for values of λ in this region (Rubio and Genton, 2016). In the second simulation scenario, we evaluate the performance of the Wasserstein prior (9) in linear regression models. We simulate $N = 250$ samples of size $n = 50, 250, 500$ from the linear regression model (9), with $\boldsymbol{\beta} = (1, 0, 0.5, 1)^\top$ and $\sigma = 0.5$, where the first entry of $\boldsymbol{\beta}$ represents the intercept. The entries of the design matrix are simulated from a multivariate normal distribution with zero mean, unit variance, and pairwise correlations of 0.5. The values of $\boldsymbol{\beta}$ are chosen to reflect different levels of signal-to-noise ratio and the effect of a spurious variable. For each of these samples, we simulate a posterior sample of size 1000 using the R package 'Rtwalk', using a burn-in period of 5000 iterations and a thinning period of 25 iterations (this is, a total of 300,000 posterior samples were obtained for each sample). In all scenarios, we also compare the results against those associated to the maximum likelihood estimators (MLE). We choose the following performance measures to evaluate the different estimation methods and priors: 'mMean' denoting the average of the posterior means across the N simulated samples; 'mSD' denoting the average of the posterior standard deviations; 'mRMSE' denoting the average of the root mean squared errors; 'Coverage' denoting the coverage proportion of the 95% credible intervals; 'mMLE' denoting the average of the maximum likelihood estimators; and 'RMSE-MLE' denoting the root mean squared error of the maximum likelihood estimators across the N samples.

Tables 1–3 in the Appendix show the results associated to the first simulation scenario. From Table 1 in the Appendix, we observe that (in the case $\lambda = 1$) the estimation of the parameter λ is indeed quite challenging for all sample sizes as the true model is very close to symmetry. Both priors (independence Jeffreys and independence Wasserstein) induce a marked shrinkage of the parameter λ towards zero as the likelihood is relatively flat. This shrinkage naturally induces a bias in the Bayesian point estimators (posterior mean) for both priors. Although, for $n = 50$, the coverage produced by the Jeffreys prior is slightly better than that produced by the Wasserstein prior, the average RMSE and standard deviations of the Bayes estimators associated to the Jeffreys prior are much larger. This is likely a consequence of the very heavy tails of the Jeffreys prior which, together with the flatness of the likelihood, produce a heavy tailed posterior. Indeed, the MLE also exhibits a very large RMSE for $n = 50$. The stronger regularisation induced by the Wasserstein prior also produces a faster concentration of the predictive posterior densities around the true model. The fit of the posterior predictive pdfs is particularly better than that obtained with the fitted pdfs using the MLEs (Figure 1). The cases $\lambda = 3, 5$ (Tables 2–3 and Figures 1–3 in the Appendix) show that the estimation of the parameter λ is much better behaved when the true value of λ is away from $\lambda = 0$, and the density function is clearly asymmetric. The performance of the independence Jeffreys and the independence Wasserstein in terms of all measures is quite similar. Since the true value of the parameter lies in the tails of the prior, the shrinkage effect of the priors is minimal. In those cases, the MLE also exhibits a much larger RMSE for $n = 50$.

Table 4 shows the results associated to the second simulation scenario. We notice that the performance of the Wasserstein prior is good for all sample sizes in terms of the chosen measures. Indeed, given that the prior is flat, the performance of the MLE coincides with that of the maximum a posteriori (MAP).

6. Discussion

We have introduced the Wasserstein prior, a prior based on the Wasserstein information matrix, which is invariant under reparameterisations. We have briefly discussed the link of the construction of this prior with concepts from information geometry. We have also introduced the independence Wasserstein prior, which aims at reducing the functional dependence between the parameters in a similar fashion as the independence Jeffreys prior (and more generally, the reference prior (Yang and Berger, 1997)). The simulation study (results presented in the Appendix) shows that the Wasserstein prior induces a posterior with good frequentist properties (at least for the models studied here), compared to the posteriors induced by the Jeffreys prior and the fitted models using maximum likelihood estimation. Additional numerical examples related to the models presented here can be found at https://github.com/FJRubio67/PIW.

As discussed in the introduction, objective priors are based on formal rules with specific aims. The Wasserstein prior is based on a formal rule aiming at obtaining a prior that is invariant under reparameterisations. Consequently, the construction of such prior does not necessarily penalise model complexity, and thus may produce suboptimal results in sparse scenarios, such as linear regression models with many spurious variables.

Natural extensions of our work include the calculation of the Wasserstein prior for other univariate continuous distributions (with bounded support, with positive support or supported on the entire real line). In this paper, we have taken a conservative position as we do not claim superiority of the Wasserstein prior over the Jeffreys prior in terms of a specific optimality criterion, even though the simulation study illustrates a competitive performance. Our work represents a step forward in the analysis of invariant priors obtained by a formal rule, and shows that it is possible to go beyond those induced by the Kullback–Leibler divergence. We believe it would be interesting to provide a theoretical treatment of the inferential properties of the Wasserstein prior, beyond the propriety of the posterior shown here. This includes the study of the asymptotic normality of the posterior distribution; establishing more formal links of the Wasserstein prior with information geometry (Kass and Wasserman, 1996; Kass, 1989; Nielsen, 2020); and the effect of the parameterisation on the orthogonality of parameters (Cox and Reid, 1987) based on the Wasserstein information matrix.

Acknowledgment

W. Li thanks the start-up funding from the University of South Carolina and NSF RTG: 2038080.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2022.109645.

References

Amari, S., 2016. Information Geometry and Its Applications. Vol. 194. Springer, Oxford.

Amari, S., 2021. Information geometry. Jpn. J. Math. 16 (1), 1–48.

Amari, S., Matsuda, T., 2022. Wasserstein statistics in one-dimensional location scale models. Ann. Inst. Statist. Math. 74 (1), 33–47.

Azzalini, A., 1985. A class of distributions which includes the normal ones. Scand. J. Stat. 12, 171–178.

Berger, J., 2006. The case for objective Bayesian analysis. Bayesian Anal. 1 (3), 385–402.

Consonni, G., Fouskakis, D., Liseo, B., Ntzoufras, I., 2018. Prior distributions for objective Bayesian analysis. Bayesian Anal. 13 (2), 627–679.

Cox, D., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. J. R. Stat. Soc. Ser. B Stat. Methodol. 49 (1), 1–18.

Dette, H., Ley, C., Rubio, F., 2018. Natural (non-) informative priors for skew-symmetric distributions. Scand. J. Stat. 45 (2), 405–420.

Ghosh, J., Delampady, M., Samanta, T., 2006. An Introduction to Bayesian Analysis: Theory and Methods. Vol. 725. Springer, New York.

Jeffreys, H., 1946. An invariant form for the prior probability in estimation problems. Proc. R. Soc. Lond. Ser. A Math. Phys. Sci. 186 (1007), 453–461.

Kass, R., 1989. The geometry of asymptotic inference. Statist. Sci. 188–219.

Kass, R., Wasserman, L., 1996. The selection of prior distributions by formal rules. J. Amer. Statist. Assoc. 91 (435), 1343–1370.

Lehmann, E., Casella, G., 2006. Theory of Point Estimation. Springer, New York.

Leisen, F., Villa, C., Walker, S., 2020. On a class of objective priors from scoring rules (with discussion). Bayesian Anal. 15 (4), 1345–1423.

Li, W., 2021a. Transport information bregman divergences. Inf. Geom. 4 (2), 435–470.

Li, W., 2021b. Transport information geometry: Riemannian calculus on probability simplex. Inf. Geom..

Li, W., Zhao, J., 2019. Wasserstein information matrix. arXiv preprint arXiv:1910.11248.

Nielsen, F., 2020. An elementary introduction to information geometry. Entropy 22 (10), 1100.

Robert, C., Chopin, N., Rousseau, J., 2009. Harold jeffreys's theory of probability revisited (with discussion). Statist. Sci. 24 (2), 141–172.

Rubio, F., Genton, M., 2016. BayesIan linear regression with skew-symmetric error distributions with applications to survival analysis. Stat. Med. 35 (14), 2441–2454.

Rubio, F., Liseo, B., 2014. On the independence jeffreys prior for skew-symmetric models. Statist. Probab. Lett. 85, 91–97.

Rubio, F., Steel, M., 2014. Inference in two-piece location-scale models with jeffreys priors (with discussion). Bayesian Anal. 9 (1), 1–22.

Shemyakin, A., 2014. Hellinger distance and non-informative priors. Bayesian Anal. 9 (4), 923–938.

Villani, C., 2003. Topics in Optimal Transportation. American Mathematical Society, Providence, Rhode Island.

Yang, R., Berger, J., 1997. A Catalogue of Noninformative Priors. Institute of Statistics and Decision Science. Duke University Discussion Papers, pp. 42–97.