Bayesian Analysis (2014) **9**, Number 4, pp. 923–938

Hellinger Distance and Non-informative Priors

Arkady Shemyakin [*]

Abstract. This paper introduces an extension of the Jeffreys' rule to the construction of objective priors for non-regular parametric families. A new class of priors based on Hellinger information is introduced as Hellinger priors. The main results establish the relationship of Hellinger priors to the Jeffreys' rule priors in the regular case, and to the reference and probability matching priors for the non-regular class introduced by Ghosal and Samanta. These priors are also studied for some non-regular examples outside of this class. Their behavior proves to be similar to that of the reference priors considered by Berger, Bernardo, and Sun, however some differences are observed. For the multi-parameter case, a combination of Hellinger priors and reference priors is suggested and some examples are considered.

Keywords: non-informative prior, Jeffreys' rule, reference prior, matching probability prior, Hellinger distance, Hellinger information.

1 Introduction

Non-informative priors play a crucial role in objective Bayesian analysis. The most popular ways to construct non-informative priors include the Jeffreys' rule (Jeffreys, 1961) based on the concept of Fisher's information; matching probability principle proposed by many authors including Tibshirani (1989), Datta and Mukerjee (2004), and Ghosal (1999); and reference priors introduced by Bernardo (1979) and developed in such works as Berger and Bernardo (1989, 1992), Berger, Bernardo and Sun (2009). For a more comprehensive review of these and other approaches, see, e.g., Kass and Wassermann (1996). The main purpose of the present paper is to introduce an extension of the Jeffreys' rule to the construction of objective priors in such cases when Fisher's information might not be defined. We introduce a class of non-informative priors based on Hellinger information. The priors of this class (dubbed Hellinger priors) are attractive due to relative technical simplicity of their derivation. It is also interesting that in many (but not all!) staple examples of Ghosal (1997, 1999), Ghosal and Samanta (1997), and Berger, Bernardo and Sun (2009), the Hellinger prior directly coincides with reference and probability matching priors, which often require much more mathematical sophistication to derive.

In the next section we first remind the reader of the conventional definition of Hellinger distance and mention some of its useful properties. Then we define Hellinger information, which was introduced in Shemyakin (1992) in a much different context of information inequalities for the Bayes risk. Then we define Hellinger priors as suggested in Shemyakin (2011, 2012).

[*] University of St. Thomas a9shemyakin@stthomas.edu

DOI:10.1214/14-BA881

It is still an open question whether the Hellinger prior can be derived as the solution of some information-optimization problems similar to the way the reference priors are obtained in Berger, Bernardo and Sun (2009). One possible motivation to consider Hellinger priors is related to information-theoretic principles. The single-parameter family of distributions endowed with the Hellinger information metric defines a one-dimensional Riemann manifold. The Hellinger prior is natural from an information-geometric viewpoint, following Kass and Wasserman (1996), since it can be treated as the volume element on this manifold. However, the Riemann geometric framework is less appropriate for multi-dimensional non-regular parametric families, which rather exhibit Finsler manifold structure.

Section 3 is dedicated to the formulation of the main results. Here we establish the relationship of Hellinger priors to the Jeffreys' rule priors in the regular case, and to the reference and probability matching priors for the non-regular class introduced in Ghosal and Samanta (1997). Some related results were formulated in Shemyakin (2012) without specifying regularity conditions. The complete original proofs are relegated to Appendix A. In order to illustrate both the similarity and differences between Hellinger and reference priors, in Section 4 we consider examples used in Ghosal (1997, 1999), Ghosal and Samanta (1997), and Berger, Bernardo and Sun (2009). We begin with examples from the Ghosal-Samanta class, for which Hellinger priors coincide with reference and probability matching priors. Then we proceed with non-regular models outside of the Ghosal-Samanta class, for which some discrepancies may be observed. Some detailed derivations of Hellinger priors are placed in Appendix B mostly in order to demonstrate their technical simplicity. Section 5 is dedicated to the multi-parameter case, for which a combination of Hellinger and reference priors is suggested. Section 6 contains conclusions and proposes some directions for future development.

2 Notations and Definitions

2.1 Hellinger Distance

Hellinger distance, which should be more accurately referred to as Bhattacharyya-Hellinger distance, since it was first defined in its modern version in Bhattacharyya (1943), may be used to quantify the distance between two points of a parametric family. Under certain regularity conditions, its limit behavior as the difference in the parameter values goes down to 0 is closely related to Fisher information. Hellinger distance can also be used to study information properties of the parametric set in non-regular situations (e.g., when Fisher information does not exist) as suggested in Birge (1985), Ibragimov and Has'minskii (1981), and Le Cam (1986). Satisfying properties of a distance (symmetry, triangle inequality), it promises certain advantages relative to such alternative information measures as Kullback-Leibler divergence. This explains a natural way in which Hellinger information (the limit of Hellinger distance between two adjacent points of the parametric set) serves to provide the lower bounds for Bayes risk in non-regular situations. Its role is similar to the role played by Fisher information in the regular case. It is also natural to apply Hellinger information to derive priors similar to Jeffreys' rule priors which would be the least informative for non-regular cases. We will derive Hellinger priors based on the concept of Hellinger information and compare their behavior with the Jeffreys' rule prior and reference prior as defined by Berger, Bernardo, and Sun (2009). It should be mentioned that in the regular case Hellinger priors generalize the constructions already known, while under the loss of regularity they can reveal somewhat different behavior.

Suppose that a parametric family of probability measures $\{P_\theta, \theta \in \Theta\}$ is defined on a measurable space $(\mathcal{X}, \mathcal{B})$ so that all the measures from the family are absolutely continuous with respect to some σ-finite measure λ on $\mathcal{B}$. Throughout Sections 2-4 we will consider the single real parameter case $\Theta \subset \mathbf{R}$, which in Section 5 will be extended to $\Theta \subset \mathbf{R}^m, m = 1, 2, \ldots$. Then Hellinger distance between any two parameter values can be defined in terms of densities $p(x;\theta_i) = \frac{dP_{\theta_i}}{d\lambda}$ as

$$d_H(\theta_1, \theta_2) = \Big(\int_{\mathcal{X}} \big(\sqrt{p(x;\theta_1)} - \sqrt{p(x;\theta_2)}\big)^2 d\lambda \Big)^{1/2}.$$

Thus one can use Hellinger distance to quantify the distance between measures from the same family indexed by different parameters. It does not depend on the choice of the dominating measure λ and is defined for all points of the parametric family. It is a true metric, satisfying the symmetry property and triangle inequality. Further properties of Hellinger distance were reviewed in multiple studies, e.g., Gibbs and Su (2002). Hellinger distance is closely related to other quantities which serve to measure divergence between the points of parametric space, such as Kullback-Leibler divergence

$$d_{KL}(\theta_1, \theta_2) = \int_{supp(P_{\theta_1})} p(x;\theta_2) \log\Big(\frac{p(x;\theta_2)}{p(x;\theta_1)}\Big) d\lambda,$$

where integration is carried out over the support of the distribution, chi-square distance

$$d_{\chi^2}(\theta_1, \theta_2) = \int_{supp(P_{\theta_1}) \bigcup supp(P_{\theta_2})} \frac{\big(p(x;\theta_1) - p(x;\theta_2)\big)^2}{p(x;\theta_1)} d\lambda,$$

or total variation distance

$$d_{TV}(\theta_1, \theta_2) = \sup_{A \in \mathcal{B}} \big| P_{\theta_1}(A) - P_{\theta_2}(A) \big| .$$

Kullback-Leibler divergence plays a very important role in information theory and finds many natural applications in Bayesian parametric estimation. However, neither Kullback-Leibler nor chi-square divergence measures are symmetric. These two divergence measures also cannot be defined for all possible pairs of values of parameter in the case of parameter-dependent density support.

One attractive feature of Hellinger distance used in the sequel is provided by the following formula. For the product measures $\mu = \mu_1 \times \mu_2$ and $\nu = \nu_1 \times \nu_2$, it is true that

$$1 - \frac{1}{2} d_H^2(\mu, \nu) = \big(1 - \frac{1}{2} d_H^2(\mu_1, \nu_1)\big) \cdot \big(1 - \frac{1}{2} d_H^2(\mu_2, \nu_2)\big),$$

from where we can first derive a useful expression for the case of i.i.d. finite samples $x^{(n)} = (x_1, ..., x_n)$and $p^{(n)}(x;\theta) = \frac{dP_\theta^{(n)}}{d\lambda^n}$:

$$d_H^2\big(p(x^{(n)};\theta_1), p(x^{(n)};\theta_2)\big) = 2\Big[1 - \big(1 - \frac{1}{2} d_H^2\big(p(x;\theta_1), p(x;\theta_2)\big)\big)^n\Big],$$

which helps to reduce many considerations to the case of sample size 1.

2.2 Hellinger Information and Lower Bounds for Bayes Risk

From this point on until further notice we assume $\Theta \subset \mathbf{R}$. Extensions for vector parameter cases are deferred to Section 5 for the sake of simplicity. If for almost all θ from Θ there exists such $\alpha \geq 0$ that

$$\lim_{\varepsilon \to 0} \big| \varepsilon \big|^{-\alpha} d_H^2(\theta, \theta + \varepsilon) = j(\theta),$$

we define Hellinger information as

$$I_H(\theta) = j^{2/\alpha}(\theta).$$

The role of the exponent α is that of the proper normalizing order of magnitude in the limiting behavior of the Hellinger distance. The following side-trip will serve to reveal the reasons why Hellinger information is introduced in this slightly unnatural way.

Let us define the least-square Bayes risk for an independent identically distributed sample $X^{(n)} = (X_1, ..., X_n)$ of size n and prior $\pi(\theta)$ as

$$R\big(\hat{\theta}(X^{(n)})\big) = \int_{\mathcal{X}^{(n)} \times \Theta} \big(\hat{\theta}(X^{(n)}) - \theta\big)^2 p(x^{(n)};\theta)\pi(\theta) dx^{(n)} d\theta.$$

We will consider an integral version of the classical Cramer-Frechet-Rao inequality (Cramer, 1946), which under certain standard regularity conditions leads to the following asymptotic lower bound for the Bayes risk in terms of Fisher information $I(\theta) = E\big(\frac{\partial}{\partial\theta} \log p(X;\theta)\big)^2$:

$$R\big(\hat{\theta}(X^{(n)})\big) \geq n^{-1} \int_\Theta I^{-1}(\theta)\pi(\theta) d\theta + o(n^{-1}).$$

This lower bound was obtained by Borovkov and Sakhanienko (1980). See also Bobrovsky et al. (1987) for a different form of the lower bound, and Brown and Gajek (1990) for an alternative method of proof. The order of the last term on the right hand side can be improved to $O(n^{-2})$ under additional regularity conditions. However, we will rather take the opposite direction: release the regularity conditions following the general setup of Ibragimov and Has'minskii (1981). This bound can be extended to the case when Fisher information may not exist. One such extension gives the following asymptotic lower bound for Bayes risk.

Theorem 1 *(Hellinger information inequality). Let Hellinger information I_H be strictly positive for some $\alpha \in (0,2]$, almost surely continuous and bounded on any compact subset of Θ , where Θ is an open subset of real numbers, and $\int_\Theta I_H^{-1}(\theta)\pi(\theta)d\theta < \infty$. Then*

$$\inf_{\hat\theta(X^{(n)})} R\big(\hat\theta(X^{(n)})\big) \geq C(\alpha)n^{-2/\alpha}\int_\Theta I_H^{-1}(\theta)\pi(\theta)d\theta + o(n^{-2/\alpha}),$$

where $C(\alpha)$ is a constant depending on α only.

The statement of Theorem 1 directly follows from more general results obtained in Shemyakin (1991, 1992) for the multi-parameter case.

This inequality suggests that Hellinger information can be used as a substitute for Fisher information when the latter does not exist. Notice the role of the exponent $2/\alpha$, which appears in the definition of Hellinger information and also determines the asymptotic order of magnitude of the risk with respect to the sample size. The regular case (asymptotic normality) corresponds to $\alpha = 2$. The non-regular case of densities with jumps (density support depending on parameter) corresponds to $\alpha = 1$. Intermediate cases (smooth transition from finite discontinuity of the density to differentiability) correspond to $\alpha \in (1,2)$. Similar bounds can be obtained in terms of chi-square distance or Kullback-Leibler distance, but they require additional assumptions. While the exact value of $C(\alpha)$ is related to the technical details of the proof and is not necessarily tight, the lower bound in Theorem 1 has the correct order of magnitude with respect to the sample size. It is very tempting to try to justify Hellinger priors as the least favorable under the least-square risk following the arguments of Clarke and Barron (1994) for the entropy risk, but it is just a suggestion for future studies. For further discussion of specific examples, see Shemyakin (2012).

2.3 Hellinger Priors

We will define the Hellinger prior for the parametric set Θ as

$$\pi_H(\theta) \propto \sqrt{I_H(\theta)} = j^{1/\alpha}(\theta).$$

Our goal is to compare Hellinger priors with other objective priors. The following theorems demonstrate that in many cases the Hellinger prior coincides with priors obtained by other approaches: namely, Jeffreys' rule prior, reference priors, and probability matching priors. A special role might be played by the Hellinger prior in the case when Fisher information is not defined, therefore it cannot be used in construction of objective priors. We will define the Jeffreys' rule prior in the one-parameter case being equal (to within a constant multiple) to the square root of Fisher information. A probability matching prior is a prior distribution under which the posterior probability of certain regions coincides with their coverage probabilities, either exactly or approximately. For the exact definition that we will assume in the sequel see Ghosal (1999).

The reference prior according to Berger, Bernardo, and Sun (2009) will be understood as a permissible prior maximizing the missing information (expected Kullback-Leibler distance between the prior and the posterior) with respect to the prior.

3 Main Results

The following relationship between Fisher information and Hellinger distance was established in Borovkov (1998) for non-specified regularity conditions and without mentioning Hellinger information explicitly. We suggest our version of sufficient (probably, not necessary) conditions.

Theorem 2. *If $p(x;\theta)$ is twice continuously differentiable w.r.t. θ for almost all $x \in \mathcal{X}$ w.r.t. λ, $E\big(\frac{\partial^2}{\partial\theta^2}p(X;\theta)\big)^2 < \infty$, Fisher information $I(\theta) = E\big(\frac{\partial}{\partial\theta}\log p(X;\theta)\big)^2$ is continuous, strictly positive and finite for almost all θ from Θ, then*:

$$\alpha = 2,\ I_H(\theta) = \frac{1}{4}I(\theta).$$

Corollary 1. *In the assumptions of Theorem 2, $\pi_H(\theta)$ is the Jeffreys' rule prior and is, therefore, invariant to re-parameterization.*

The statement of the corollary follows directly from the definitions of the Jeffreys' rule and Hellinger priors. The next result deals with the models generalizing the class of non-regular distributions described by Ghosal and Samanta (1997).

Theorem 3. *If (1) probability density $p(x;\theta) = \frac{dP_\theta}{d\theta}$ with support $S(\theta) = \big[a_1(\theta), a_2(\theta)\big]$ is strictly positive on $S(\theta)$, bounded and continuous on any compact set in Θ, (2) both functions $a_1(\theta)$ and $a_2(\theta)$ are continuously differentiable and $\big|a_k'(\theta)\big| > 0,\ k = 1,2$, (3) right limit $q_1(\theta) = \lim_{x\searrow a_1(\theta)} p(x;\theta)$ and left limit $q_2(\theta) = \lim_{x\nearrow a_2(\theta)} p(x;\theta)$ are finite, uniform on compact subsets of Θ, and q_k grow at most polynomially, (4) $\log p(x;\theta)$ is twice continuously differentiable w.r.t. θ on the interior of the support* $\mathrm{int}S(\theta)$ *and for any θ for small enough ε*

$$\sup_{\theta-\varepsilon<u<\theta+\varepsilon}\Big|\frac{\partial^2}{\partial u^2}\log p(x;u)\Big| \le H_\theta(x),$$

where $EH_\theta(X) < \infty$ is continuous, then:

$$a = 1,\ I_H(\theta) = \ \big|\, a_1'(\theta)\,\big| \cdot q_1(\theta) + \ \big|\, a_2'(\theta)\,\big| \cdot q_2(\theta).$$

Corollary 2. *For the non-regular class discussed in Ghosal, Ghosh, and Samanta (1995) and defined by Ghosal and Samanta (1997),$\pi_H(\theta)$ coincides with the probability matching prior (Ghosal, 1999) and also with the reference prior derived in Berger, Bernardo, and Sun (2009).*

Proof of Corollary 2: Corollary 2 follows from the fact that the non-regular class defined in Ghosal and Samanta (1997) is a particular case of the class of probability measures defined by the conditions of Theorem 3. Conditions (1)-(4) of Theorem 3 are weaker than conditions (A1)-(A5) of Ghosal and Samanta (1997) and Ghosal (1999). Additional monotonicity of the support suggested for Ghosal-Samanta class is related to the use of Kullback-Leibler divergence and is not required by Theorem 3.

4 Single-parameter Examples

The following examples illustrate in what ways Hellinger priors are similar (or, possibly, different) to reference and matching probability priors. We will begin with several single-parameter examples. Apparently, due to Theorem 2 and Corollary 1, regular cases (continuous Fisher information) do not present anything new: we end up with the Jeffreys' rule prior. Therefore we will concentrate on the non-regular class of Ghosal and Samanta (1997) and some non-regular models outside of this class. One important advantage of the Hellinger information approach (as discussed in Section 2) is that it allows for the analysis of a single observation $x \in \mathcal{X}$ so that for i.i.d. samples the results are invariant of the sample size.

Example 1. Uniform with parameter-dependent support $\mathrm{Unif}(0,\theta),\ \theta \in (0,\infty)$. We will assume $\varepsilon > 0$, which due to the symmetry of Hellinger distance does not lead to loss of generality, and perform integration

$$d_H^2(\theta, \theta+\varepsilon) = \int \left(\sqrt{p(x;\theta)} - \sqrt{p(x;\theta+\varepsilon)}\right)^2 dx$$

$$= 2 - 2\int_0^\theta \frac{1}{\sqrt{\theta(\theta+\varepsilon)}}dx = 2\biggl(1 - \frac{1}{\sqrt{1+\varepsilon/\theta}}\biggr) = \frac{\varepsilon}{\theta} + o(\varepsilon),$$

which follows from the property $\sqrt{1+\delta} = 1 + \frac{1}{2}\delta + o(\delta)$ as $\delta \to 0$.

Therefore $\alpha = 1,\ \pi_H(\theta) \propto j(\theta) = \theta^{-1}$, which is consistent with well-known results for reference priors. We can also use Theorem 3.

Example 2. Uniform from Ghosal-Samanta class $\mathrm{Unif}(\theta^{-1}, \theta),\ \theta \in (1,\infty)$. In this case,

$$\alpha = 1,\ \pi_H(\theta) \propto j(\theta) = \frac{(\theta^2+1)}{\theta(\theta^2-1)}.$$

This result can be obtained directly from Theorem 3 with $a_1(\theta) = \theta^{-1},\ a_2(\theta) = \theta,\ p(x;\theta) = \theta/(\theta^2-1)$. The same prior can be constructed as a probability matching prior (Ghosal, 1999) or reference prior (Berger, Bernardo, and Sun, 2009).

Example 3. Uniform not belonging to Ghosal-Samanta class $\mathrm{Unif}(\theta,\theta^2),\ \theta \in (1,\infty)$. Due to the lack of monotonicity of support (neither $S(\theta) \subseteq S(\theta+\varepsilon),\ \varepsilon > 0$ nor $S(\theta) \subseteq S(\theta+\varepsilon),\ \varepsilon < 0$), this model lies outside of the non-regular class defined in

Ghosal and Samanta (1997). The reference prior

$$\pi_H(\theta) \propto \frac{2\theta-1}{\theta(\theta-1)} \exp\Big\{\psi\Big(\frac{2\theta}{2\theta-1}\Big)\Big\},$$

where $\psi(z)$ is the digamma function (poly-gamma function of order 0) defined as $\psi(z) = \frac{d}{dz}\log\Gamma(z),\ z > 0$ was obtained by Berger, Bernardo, and Sun (2009). The Hellinger prior obtained directly from Theorem 3 with $a_1(\theta) = \theta,\ a_2(\theta) = \theta^2,\ p(x;\theta) = 1/(\theta(\theta-1))$ is different:

$$\pi_H(\theta) \propto \frac{2\theta+1}{\theta(\theta-1)}.$$

However, we can see that both priors are relatively close numerically, because the digamma function for $1 \le \theta < \infty$ changes monotonically and

$$\frac{1}{3}\exp\big\{\psi(2)\big\} \le \frac{\pi_R(\theta)}{\pi_H(\theta)} = \frac{2\theta-1}{2\theta+1}\exp\Big\{\psi\Big(\frac{2\theta}{2\theta-1}\Big)\Big\} \le \exp\big\{\psi(1)\big\},$$

so that the ratio of two priors increases monotonically with respect to θ and is bounded both from above and below:

$$\frac{1}{3}\exp\big\{1-\gamma\big\} \le \frac{\pi_R(\theta)}{\pi_H(\theta)} \le \exp\big\{-\gamma\big\},$$

where $\gamma \approx .5772$ is the Euler's constant, or numerically as $.5087 \le \frac{\pi_R(\theta)}{\pi_H(\theta)} \le .5615$.

Example 4. Non-symmetric standard triangular distribution.

For this distribution with density

$$p(x;\theta) = \begin{cases} \frac{2x}{\theta}, & 0 \le x \le \theta \\ \frac{2(1-x)}{\theta}, & \theta \le x \le 1 \end{cases}$$

the Fisher information function cannot be properly defined, and Jeffreys' rule prior does not exist. The beta prior $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ is the reference prior which was first suggested by numerical approximations and then proved analytically (Berger, Bernardo, and Sun, 2009). This is also the Hellinger prior as demonstrated in Appendix B.

Example 5. Shifted Gamma distribution.

For the case of $x \sim \text{Gamma}(\beta,\varphi,\theta)$ on the semi-interval $\big[\theta,\infty)$

$$p(x;\beta,\varphi,\theta) = \frac{(x-\theta)^{\beta-1}\exp\{-(x-\theta)/\tau\}}{\varphi^{\beta}\cdot\Gamma(\beta)},\ x \in \big[\theta,\infty),$$

if the parameter of interest is the threshold (location) while the shape $\beta \in [1,2]$ and the scale τ can be treated as known,

$$d_H^2(\theta,\theta+\varepsilon) = \int_0^{\theta}\Big(\sqrt{p(x;\beta,\varphi,\theta)} - \sqrt{p(x;\beta,\varphi,\theta+\varepsilon)}\Big)^2 dx$$

$$= \frac{\gamma(\beta, \varepsilon)}{\Gamma(\beta)} \sim_{\varepsilon \to 0} \frac{\varepsilon^{\beta}}{\beta \cdot \Gamma(\beta)},$$

where

$$\gamma(\beta, \varepsilon) = \int_0^{\varepsilon} t^{\beta - 1} \exp\{-t\} dt$$

is the lower incomplete gamma-function, so that $\alpha = \beta$ and $\pi(\theta) \propto j(\theta) = \text{const}$, which is consistent with the general form of reference priors for location parameters. This example demonstrates the range of non-regularities in terms of α.

5 Multi-parameter models

Let us extend the definitions of Section 2 to the multi-parameter case $\Theta \subset \mathbf{R}^m, m = 1, 2, ...$ in such a way that the newly defined Hellinger information matrix may be naturally reduced to the Fisher information matrix in the regular case and Hellinger priors may be reduced to Jeffreys' rule priors, but also making sure that the following non-regular examples are properly addressed. We will introduce the Hellinger distance matrix following Shemyakin (1992) as a matrix with elements

$$D_{ij}(\theta, U) = \int_{\mathcal{X}} \left(\sqrt{p(x; \theta)} - \sqrt{p(x; \theta + \mathbf{u}_i)} \right) \left(\sqrt{p(x; \theta)} - \sqrt{p(x; \theta + \mathbf{u}_j)} \right) d\lambda,$$

where U is an $m \times m$ matrix with columns $\mathbf{u}_i$. Then define vectors $\alpha = (\alpha_1, ..., \alpha_m)$ and $\delta = (\delta_1, ..., \delta_m)$ with components $\delta_i = \varepsilon^{2/\alpha_i}$ such that for all $i = 1, ..., m$ there exist finite non-degenerate limits

$$0 < \lim_{\varepsilon \to 0} \left| \varepsilon \right|^{-2} D_{ii}(\theta, \delta \mathbf{1}) < \infty,$$

where $\mathbf{1}$ is an $m \times m$ unit matrix.

The Hellinger information matrix I_H will be defined by its components

$$(I_H)_{ij}(\theta) = \lim_{\varepsilon \to 0} \left| \varepsilon \right|^{-2} D_{ij}^{(1/\alpha_i + 1/\alpha_j)}(\theta, \delta \mathbf{1}).$$

The following two examples illustrate possible multi-parameter non-regularities. Let us restrict ourselves for simplicity to the two-parameter case. The shifted uniform model in Example 6 addresses non-regularity in two parameter components, while the truncated Weibull model in Example 7 suggests a combination of regular and non-regular parameters. For both examples we develop two approaches. The first requires elicitation of "full" Hellinger priors π_{FH} similar to "full" Jeffreys' rule priors based on the Hellinger information matrix defined above:

$$\pi_{FH}(\theta) \propto \sqrt{\left| \det I_H(\theta) \right|}.$$

Notice that while in most natural examples the matrix I_H is positive definite and $\det I_H(\theta) > 0$, this statement is not proven in the general case, thus the definition of the full Hellinger prior also allows for the possibility $\det I_H(\theta) < 0$.

The second derives the joint priors π_{JH} based on conditional priors using the methodology of Sun and Berger (1998): we begin with identifying the order of parameter components as $\theta = (\theta_1, \theta_2)$, then derive the conditional prior using the conditional Fisher information or conditional Hellinger information if Fisher information cannot be defined:

$$\pi_H(\theta_1 \big| \theta_2) \propto j_1^{1/\alpha_1}(\theta_1 \big| \theta_2).$$

Then we calculate the integrated one-dimensional likelihood using the conditional prior obtained at the first stage:

$$p_2(x; \theta_2) = \int_{\Theta_1} p(x; \theta) \pi_H(\theta_1 \big| \theta_2) d\theta_1,$$

obtain the marginal prior

$$\pi_H(\theta_2) \propto j_2^{1/\alpha_2}(\theta_2)$$

by calculating Fisher or Hellinger information for the integrated one-dimensional likelihood, and finally derive the joint Hellinger prior as

$$\pi_{JH}(\theta_1, \theta_2) \propto \pi_H(\theta_1 \big| \theta_2) \pi_H(\theta_2).$$

The "novelty" of this second approach in our case consists just in the calculation of conditional Hellinger information

$$j_1(\theta_1 \big| \theta_2) = \lim_{\varepsilon \to 0} \big|\varepsilon\big|^{-\alpha_1} d_H^2(\theta_1, \theta_1 + \varepsilon \big| \theta_2)$$

$$= \lim_{\varepsilon \to 0} \big|\varepsilon\big|^{-\alpha_1} \int_{\mathcal{X}} \Big(\sqrt{p(x; \theta_1, \theta_2)} - \sqrt{p(x; \theta_1 + \varepsilon, \theta_2)} \Big) d\lambda,$$

$$j_1(\theta_2) = \lim_{\varepsilon \to 0} \big|\varepsilon\big|^{-\alpha_2} d_{H,2}^2(\theta_2, \theta_2 + \varepsilon) = \lim_{\varepsilon \to 0} \big|\varepsilon\big|^{-\alpha_2} \int_{\mathcal{X}} \Big(\sqrt{p_2(x; \theta_2)} - \sqrt{p_2(x; \theta_2 + \varepsilon)} \Big) d\lambda$$

instead of conditional Fisher information. The main advantage of Hellinger information is revealed in non-regular cases when Fisher information is not available, which sometimes helps to avoid the technicalities of other approaches. Notice that for the second approach we do not need to define Hellinger distance or information for the vector parameter.

Example 6. Shifted uniform distribution (see also Example 2) Unif$(\mu + \tau^{-1}, \mu + \tau)$, $\tau \in (1, \infty)$, $\mu \in (-\infty, \infty)$, $\theta = (\mu, \tau)$. Full Hellinger prior is

$$\pi_{FH}(\tau, \mu) \propto \sqrt{2\tau^2 + 1} \big(\tau^2 - 1\big)^{-2}.$$

However, for any order of parameterization ($\theta = (\mu, \tau)$ or $\theta = (\tau, \mu)$), the second approach yields $\pi_{JH}(\theta) \propto \frac{\tau^2+1}{\tau(\tau^2-1)}$. It is demonstrated in Appendix B.

Example 7. Truncated Weibull distribution. Let us consider the case of truncated Weibull distribution with density

$$p(x; \beta, \varphi, \tau) = \beta \varphi^\beta x^{\beta - 1} \exp\big\{ -\varphi^\beta (x^\beta - \tau^\beta) \big\}, \; x \in [\tau, \infty),$$

when the parameters of interest are the threshold τ (location) and $\varphi > 0$, while the shape $\beta > 0$ is treated as known.

Full Hellinger prior is $\pi_{FH}(\theta) \propto \varphi^{\beta-1}\tau^{\beta-1}$, which coincides with the reference prior for vector parameter $\theta = (\tau, \varphi)$ derived in Ghosal (1997). However, with the second approach for any order of parameterization we arrive at $\pi_{JH}(\theta) \propto \varphi^{-1}\tau^{\beta-1}$, which coincides with the reference priors obtained in Ghosal (1997) when either τ or φ is the main parameter of interest, and also with the probability matching priors from Ghosal (1999). The details are offered in Appendix B.

6 Conclusions

The examples considered above give the evidence that for many one-parameter non-regular models Hellinger priors bring about the same results as the reference and the probability matching approaches. However, Example 3 reveals an intriguing deviation demonstrating that there is no equivalence between Hellinger and reference priors outside of the Ghosal-Samanta non-regular class. In the multi-parameter case as illustrated by Examples 6 and 7, substituting Hellinger information instead of Fisher information provides a technical means to treat non-regular models in the same fashion as regular ones for the derivation of reference priors. Here the most attractive feature of Hellinger priors is the relative technical simplicity of their derivation.

It is still not clear whether there is a better justification of Hellinger priors than the invariance argument and Corollaries 1 and 2. It is possible that either the information-geometric approach discussed in the introduction or the information-theoretic discussion of the least favorable priors in Subsection 2.2 can bring about more definitive results.

Acknowledgments

The author deeply appreciates the referees' and editors' suggestions which led to substantial revision of the first drafts and hopefully improved the quality of the paper. He is very grateful to Dongchu Sun, Fumiyasu Komaki and Misha Shvartsman for extremely helpful discussions, and also to the participants of O-Bayes-2011 in Shanghai for many valuable comments.

References

Berger, J. O. and Bernardo, J. M. (1989) ”Estimating a product of means: Bayesian analysis with reference priors.” *Journal of the American Statistical Association*, 84: 200-207.

Berger, J. O. and Bernardo, J. M. (1992) ”On the development of reference priors (with discussion).” In: *Bayesian Statistics*, 4: 35-60, Oxford University Press.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). ”The formal definition of reference priors.” *Annals of Statistics*, 37, 1: 905-938.

Bernardo, J. M. (1979). ”Reference posterior distributions for Bayesian inference (with discussions).” *Journal of the Royal Statistical Society, Ser. B*, 41: 113-147.

Bhattacharyya, A. (1943). ”On a measure of divergence between two statistical populations defined by their probability distributions.” *Bulletin of the Calcutta Mathematical Society*, 35: 99-109. MR 0010358

Birge, L. (1985). ”Non-asymptotic minimax risk for Hellinger balls.” *Probability and Mathematical Statistics*, 5: 21-29.

Bobrovsky, B. Z., Mayer-Wolf, E., and Zakai, M. (1987). ”Some classes of global Cramer -Rao bounds.” *Annals of Statistics*, 15, 4: 1421-1438.

Borovkov, A. A. (1998). *Mathematical Statistics.* Nauka, Moscow. In English, translated by A. Moullagaliev, Gordon and Breach (1998).

Borovkov, A. A. and Sakhanienko, A. I. (1980). ”On estimates for the average quadratic risk.” *Probability and Mathematical Statistics*, 1: 185-195.

Brown, L. D. and Gajek, L. (1990). ”Information inequalities for the Bayes risk.” *Annals of Statistics*, 18, 4: 1578-1594.

Clarke, B. and Barron, A. (1994). ”Jeffreys' prior is asymptotically least favorable under entropy risk.” *Journal of Statistical Planning and Inference*, 41: 37-60.

Cramer, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics.* Springer Verlag. New York. MR2053794.

Ghosal, S., Ghosh, J. K., and Samanta, T. (1995). ”On convergence of posterior distributions.” *Annals of Statistics*, 23, 6: 2145-2152.

Ghosal, S. (1997). "Reference priors in multiparameter nonregular cases." *Test*, 6: 159-186.

Ghosal, S. and Samanta, T. (1997). "Expansion of Bayes risk for entropy loss and reference prior in nonregular cases." *Statistical Decisions*, 15: 129-140, MR1475184.

Ghosal, S. (1999). "Probability matching priors for non-regular cases." *Biometrika*, 86: 956-964.

Gibbs, A. L. and Su, E. W. (2002). "On choosing and bounding probability metrics." *International Statistical Review*, 70, 3: 419:435.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer Verlag, New York.

Jeffreys, H. (1946). "An invariant form for the prior probability in estimation problems." *Proceedings of the Royal Statistical Society, London, Ser. A*, 186: 453-461.

Kass, R. and Wassermann, L. (1996). "Selecting prior distributions by formal rules." *Journal of the American Statistical Association*, 91: 1343-1370.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer Verlag, New York.

Shemyakin, A. (1991). "Rao-Cramer type multidimensional integral inequalities for parametric families with singularities." *Siberian Mathematical Journal*, 32, 4: 706-715.

Shemyakin, A. (1992) "On Information Inequalities in the Parametric Estimation." *Theory of Probability and its Applications*, 37, 1: 89-91.

Shemyakin, A. (2011). "Information inequalities, reference priors and Hellinger distance." *Proceedings of International Conference in Business and Economics*, 7: 231-235, Astrakhan, Russia.

Shemyakin, A. (2012). "A new approach to construction of objective priors: Hellinger information." *Applied Econometrics*, 28, 4:124-137 (in Russian). *J. of Statist. Plann. Inference*, 61, 2: 319-338.

Sun, D. and Berger, J. O. (1998). "Reference priors with partial information." *Biometrika*, 85: 55-71.

Tibshirani, R. (1989). "Noninformative priors for one parameter of many." *Biometrika*, 76: 604-608.

Appendix A

Proof of Theorem 2: Let us denote $L(\varepsilon) = d_H^2(\theta, \theta+\varepsilon)$ and study its behavior in the vicinity of 0. We will use the Taylor expansion

$$L(\varepsilon) = L(0) + \varepsilon \frac{d}{d\varepsilon} L(\varepsilon) + \frac{1}{2}\varepsilon^2 \frac{d^2}{d\varepsilon^2} L(\varepsilon) + o(\varepsilon^2).$$

Evidently,

$$L(0) = 0;$$

$$\frac{dL(\varepsilon)}{d\varepsilon} = \int_{\mathcal{X}} \left(1 - \sqrt{\frac{p(x;\theta)}{p(x;\theta+\varepsilon)}}\right) \frac{\partial p(x;\theta+\varepsilon)}{\partial \varepsilon} d\lambda;$$

$$\frac{d^2L(\varepsilon)}{d\varepsilon^2} = \int_{\mathcal{X}} \left(1 - \sqrt{\frac{p(x;\theta)}{p(x;\theta+\varepsilon)}}\right) \frac{\partial^2 p(x;\theta+\varepsilon)}{\partial \varepsilon^2} d\lambda$$

$$+\frac{1}{2}\int_{\mathcal{X}} \left(\frac{p(x;\theta)}{p(x;\theta+\varepsilon)}\right)^{3/2} \frac{1}{p(x;\theta)} \left(\frac{\partial p(x;\theta+\varepsilon)}{\partial \varepsilon}\right)^2 d\lambda.$$

Taking into account that $L(\varepsilon)$ has a local minimum at 0, $\frac{\partial p(x;\theta+\varepsilon)}{\partial\varepsilon}\Big|_{\varepsilon=0} = \frac{\partial p(x;\theta)}{\partial\theta}$ and $\frac{p(x;\theta)}{p(x;\theta+\varepsilon)} \underset{\varepsilon\to 0}{\to} 1$, we obtain

$$\frac{d}{d\varepsilon}L(\varepsilon) = 0,\ \frac{d^2}{d\varepsilon^2}L(\varepsilon) \to \frac{1}{2}E\Big(\frac{1}{p(x;\theta)}\frac{\partial}{\partial\theta}p(x;\theta)\Big)^2 = \frac{1}{2}I(\theta).$$

Therefore, $L(\varepsilon) = \frac{1}{4}\varepsilon^2 I(\theta) + o(\varepsilon^2)$, Q.E.D.

Proof of Theorem 3: Let us split $d_H^2(\theta, \theta+\varepsilon)$ into two integrals:

$$d_H^2(\theta, \theta+\varepsilon) = I_1(\theta,\varepsilon) + I_2(\theta,\varepsilon),$$

$$I_1(\theta,\varepsilon) = \int_{S(\theta)\cap S^c(\theta+\varepsilon)} p(x;\theta)\ d\lambda + \int_{S^c(\theta)\cap S(\theta+\varepsilon)} p(x;\theta)\ d\lambda;$$

$$I_2(\theta,\varepsilon) = \int_{S(\theta)\cap S(\theta+\varepsilon)} \big(\sqrt{p(x;\theta)} - \sqrt{p(x;\theta+\varepsilon)}\big)^2\ d\lambda.$$

Due to conditions (1)-(3),

$$I_1(\theta,\varepsilon) \le K\Big(\big|a_1'(\theta)\big| \cdot q_1(\theta) + \big|a_2'(\theta)\big| \cdot q_2(\theta)\Big) \cdot \big|\varepsilon\big|,$$

and due to condition (4)

$$I_2(\theta,\varepsilon) \le 4\big|\varepsilon\big| EH_\theta(X).$$

Furthermore,

$$I_1(\theta,\varepsilon) = \big(\big|a_1'(\theta)\big| \cdot q_1(\theta) + \big|a_2'(\theta)\big| \cdot q_2(\theta)\big) \cdot \big|\varepsilon\big| + o(\varepsilon),$$

and due to condition (4)

$$I_2(\theta,\varepsilon)=o(\varepsilon).$$

Combining the last two expressions we obtain the statement of the theorem.

Appendix B

Review of Example 4.

The main purpose of this proof is to demonstrate that Hellinger priors can be often obtained by means of elementary calculus. In a standard way, split Hellinger distance into three integrals:

$$d_H^2(\theta,\theta+\varepsilon)=I_1+I_2+I_3,$$

where

$$I_1=\int_0^\theta\left(\sqrt{\frac{2x}{\theta}}-\sqrt{\frac{2x}{\theta+\varepsilon}}\right)^2dx=2\Big(\sqrt{\frac{1}{\theta}}-\sqrt{\frac{1}{\theta+\varepsilon}}\Big)\int_0^\theta xdx$$

$$=\big(\sqrt{\theta+\varepsilon}-\sqrt{\theta}\big)^2\frac{1}{1+\varepsilon/\theta}=\frac{\varepsilon^2}{\big(\sqrt{\theta+\varepsilon}+\sqrt{\theta}\big)^2}\cdot\frac{1}{1+\varepsilon/\theta}\simeq\frac{\varepsilon^2}{4\theta}.$$

Similarly,

$$I_2=\int_{\theta+\varepsilon}^1\left(\sqrt{\frac{2(1-x)}{1-\theta}}-\sqrt{\frac{2(1-x)}{1-\theta-\varepsilon}}\right)^2dx$$

$$=\big(\sqrt{1-\theta}-\sqrt{1-\theta-\varepsilon}\big)^2(1+o(1))=\frac{\varepsilon^2}{\big(\sqrt{1-\theta-\varepsilon}+\sqrt{1-\theta}\big)^2}\cdot(1+o(1))\simeq\frac{\varepsilon^2}{4\big(1-\theta\big)}.$$

The integral over the middle of the interval is negligible

$$I_3=\int_\theta^{\theta+\varepsilon}\left(\sqrt{\frac{2x}{\theta}}-\sqrt{\frac{2(1-x)}{1-\theta-\varepsilon}}\right)^2dx\le K\varepsilon^3,\ \text{as}\ \left|\sqrt{\frac{2x}{\theta}}-\sqrt{\frac{2(1-x)}{1-\theta-\varepsilon}}\right|\le K_1\varepsilon.$$

Combining asymptotical expressions for the first two integrals yields

$$d_H^2(\theta,\theta+\varepsilon)=\frac{\varepsilon^2}{4\theta}+\frac{\varepsilon^2}{4(1-\theta)}+o(\varepsilon^2)=\frac{\varepsilon^2}{4\theta(1-\theta)}+o(\varepsilon^2),$$

and thus $\alpha=2$ and $j(\theta)=1/4\theta(1-\theta)$; $\pi_H(\theta)\propto j^{1/2}(\theta)=1/\sqrt{\theta(1-\theta)}$.

Review of Example 6.

First, we will use the "full Hellinger" approach and calculate the matrix of Hellinger information. Using notation $\theta=(\mu,\tau)$, we take into account that $\mu=\theta_1$ is a location parameter and then follow Example 2 to obtain $\alpha_1 = \alpha_2 = 1$, $\delta_1 = \delta_2 = \varepsilon^2$, and

$$D(\theta, \delta\mathbf{1}) = \varepsilon^2 \left[\begin{array}{cc} \frac{\tau}{\tau^2-1} & \frac{\tau}{\tau^2-1} \\ \frac{\tau}{\tau^2-1} & \frac{\tau^2+1}{\tau(\tau^2-1)} \end{array} \right]$$

so that

$$I_H(\theta) = \left[\begin{array}{cc} \frac{\tau^2}{(\tau^2-1)^2} & \frac{\tau^2}{(\tau^2-1)^2} \\ \frac{\tau^2}{(\tau^2-1)^2} & \frac{(\tau^2+1)^2}{\tau^2(\tau^2-1)^2} \end{array} \right]$$

and $\pi_{FH}(\mu, \tau) \propto \frac{\sqrt{2\tau^2+1}}{\tau(\tau^2-1)^2}$. For the second approach we will also assume the order of parameterization $\mu = \theta_1$ and $\tau = \theta_2$. Here also $\alpha_1 = 1$, $\pi_H(\mu|\tau)$ is a uniform improper prior for the location parameter. Integrating out the location parameter, we obtain $p_2(x; \tau) = \tau/(\tau^2 - 1)$, and finally, the joint Hellinger prior (leading to a proper posterior for sample sizes $n \geq 2$ with not all sample elements identical) can be expressed as

$$\pi_{JH}(\theta) \propto \frac{\tau^2 + 1}{\tau(\tau^2 - 1)}.$$

We can also reverse the order of treatment of parametric components $\tau = \theta_1$ and $\mu = \theta_2$ with the same result.

Review of Example 7.

We will begin with the calculation of the matrix of Hellinger information. Using notation $\theta = (\tau, \varphi)$, we take into account that θ_2 is a regular parameter and we can substitute Fisher information for the second diagonal element of the matrix below. So we obtain $\alpha_1 = 1$, $\alpha_2 = 2$, $\delta_1 = \varepsilon^2$, $\delta_2 = \varepsilon$, and

$$D(\theta, \delta\mathbf{1}) = \left[\begin{array}{ll} \varepsilon^2\varphi^\beta\tau^{\beta-1} & o(\varepsilon^3) \\ o(\varepsilon^3) & \varepsilon^2\varphi^{-2} \end{array} \right]$$

so that

$$I_H(\theta) = \left[\begin{array}{ll} \varphi^{2\beta}\tau^{2\beta-2} & 0 \\ 0 & \varphi^{-2} \end{array} \right]$$

and $\pi_{FH}(\tau, \varphi) \propto \varphi^{\beta-1}\tau^{\beta-1}$. This result coincides with the prior obtained in Ghosal (1997) for the case when both parameters are important. However, the second approach leads to a different result for any order of parameterization. If, for instance, we assume $\tau = \theta_1$ and $\varphi = \theta_2$, $\alpha_1 = 1$, $\pi_H(\tau|\varphi) \propto \tau^{\beta-1}$. Integrating out the threshold parameter τ, we obtain $p_2(x; \varphi) = x^{\beta-1}\exp\{-\varphi^\beta x^\beta\}$, and using Fisher information for parameter φ, we express the joint Hellinger prior as

$$\pi_{JH}(\tau, \varphi) \propto \tau^{\beta-1}\varphi^{-1}.$$