Bayesian Analysis (2019) **14**, Number 4, pp. 1013–1036

On the Geometry of Bayesian Inference

Miguel de Carvalho*, Garritt L. Page†, and Bradley J. Barney‡

Abstract. We provide a geometric interpretation to Bayesian inference that allows us to introduce a natural measure of the level of agreement between priors, likelihoods, and posteriors. The starting point for the construction of our geometry is the observation that the marginal likelihood can be regarded as an inner product between the prior and the likelihood. A key concept in our geometry is that of compatibility, a measure which is based on the same construction principles as Pearson correlation, but which can be used to assess how much the prior agrees with the likelihood, to gauge the sensitivity of the posterior to the prior, and to quantify the coherency of the opinions of two experts. Estimators for all the quantities involved in our geometric setup are discussed, which can be directly computed from the posterior simulation output. Some examples are used to illustrate our methods, including data related to on-the-job drug usage, midge wing length, and prostate cancer.

Keywords: Bayesian inference, geometry, Hellinger affinity, Hilbert space, marginal likelihood.

1 Introduction

Assessing the influence that prior distributions and/or likelihoods have on posterior inference has been a topic of research for some time. One commonly used ad-hoc method suggests fitting a Bayes model using a few competing priors, then visually (or numerically) assessing changes in the posterior as a whole or using some pre-specified posterior summary. More rigorous approaches have also been developed. Lavine (1991) developed a framework to assess sensitivity of posterior inference to sampling distribution (likelihood) and the priors. Berger (1991) introduced the concept of Bayesian robustness which includes perturbation models (see also Berger and Berliner 1986). More recently, Evans and Jang (2011) have compared information available in two competing priors. Related to this work, Gelman et al. (2008) advocate the use of so-called weakly informative priors that purposely incorporate less information than available as a means of regularizing. Work has also been dedicated to the so-called prior–data conflict (Evans and Moshonov, 2006; Walter and Augustin, 2009; Al Labadi and Evans, 2016). Such conflict can be of interest in a wealth of situations, such as for evaluating how much prior and likelihood information are at odds at the node level in a hierarchical model (see Scheel, Green and Rougier, 2011, and references therein). Regarding sensitivity of the posterior distribution to prior specifications, Lopes and Tobias (2011) provide a fairly accessible overview.

*School of Mathematics, The University of Edinburgh, UK, miguel.decarvalho@ed.ac.uk
†Department of Statistics, Brigham Young University, Provo, US, page@stat.byu.edu
‡Department of Statistics, Brigham Young University, Provo, US, barney@stat.byu.edu

 https://doi.org/10.1214/18-BA1112

We argue that a geometric representation of the prior, likelihood, and posterior distribution encourages understanding of their interplay. Considering Bayes methodologies from a geometric perspective is not new, but none of the existing geometric perspectives has been designed with the goal of providing a summary on the agreement or impact that each component of Bayes theorem has on inference and predictions. Aitchison (1971) used a geometric perspective to build intuition behind each component of Bayes theorem, Shortle and Mendel (1996) used a geometric approach to draw conditional distributions in arbitrary coordinate systems, and Agarawal and Daumé (2010) argued that conjugate priors of posterior distributions belong to the same geometry giving an appealing interpretation of hyperparameters. Zhu, Ibrahim and Tang (2011) defined a manifold on which a Bayesian perturbation analysis can be carried out by perturbing data, prior and likelihood simultaneously, and Kurtek and Bharath (2015) provide an elegant geometric construction which allows for Bayesian sensitivity analysis based on the so-called ϵ-compatibility class and on comparison of posterior inferences using the Fisher–Rao metric.

In this paper, we develop a geometric setup along with a set of metrics that can be used to provide an informative preliminary 'snap-shot' regarding comparisons between prior and likelihood (to assess the level of agreement between prior and data), prior and posterior (to determine the influence that prior has on inference), and prior versus prior (to compare 'informativeness'—i.e., a density's peakedness—and/or congruence of two competing priors). To this end, we treat each component of Bayes theorem as an element of a geometry formally constructed using concepts from Hilbert spaces and tools from abstract geometry. Because of this, it is possible to calculate norms, inner products, and angles between vectors. Not only do each of these numeric summaries have intuitively appealing individual interpretations, but they may also be combined to construct a unitless measure of compatibility, which can be used to assess how much the prior agrees with the likelihood, to gauge the sensitivity of the posterior to the prior, and to quantify the coherency of the opinions of two experts. Estimating our measures of level of agreement is straightforward and can actually be carried out within an MCMC (Markov Chain Monte Carlo) algorithm. An important advantage of our setting is that it offers a direct link to Bayes theorem, and a unified treatment that can be used to assess the level of agreement between priors, likelihoods, and posteriors—or functionals of these. To streamline the illustration of ideas, concepts, and methods we reference the following example (Christensen et al., 2011, pp. 26–27) throughout the article.

ON-THE-JOB DRUG USAGE TOY EXAMPLE

Suppose interest lies in estimating the proportion $\theta \in [0,1]$ of US transportation industry workers that use drugs on the job. Suppose $n = 10$ workers were selected and tested with the 2nd and 7th testing positive. Let $y = (Y_1, \ldots, Y_n)$ with $Y_i = 1$ denoting that the ith worker tested positive and $Y_i = 0$ otherwise. Let $Y_i \mid \theta \sim \text{Bern}(\theta)$, be independent and identically distributed (iid), for $i = 1, \ldots, n$, and $\theta \sim \text{Beta}(a, b)$, for $a, b > 0$. Then, $\theta \mid y \sim \text{Beta}(a^\star, b^\star)$ with $a^\star = n_1 + a$ and $b^\star = n - n_1 + b$, where $n_1 = \sum_{i=1}^{n} Y_i$.

Some natural questions our treatment of Bayes theorem will answer are: How compatible is the likelihood with this prior choice? How similar are the posterior and prior distributions? How does the choice of $\text{Beta}(a, b)$ compare to other possible prior distribu-

tions? While the drug usage example provides a recurring backdrop that we consistently call upon, additional examples are used throughout the paper to illustrate our methods.

In Section 2 we introduce the geometric framework in which we work and provide definitions and interpretations along with examples. Section 3 considers extensions of the proposed setup, Section 4 contains computational details, and Section 5 provides a regression example illustrating utility of our metric. Section 6 conveys some concluding remarks. Proofs are given in the supplementary materials (de Carvalho et al., 2018).

2 Bayes geometry

2.1 A geometric view of Bayes theorem

Suppose the inference of interest is over a parameter θ which takes values on $\Theta \subseteq \mathbb{R}^p$. We consider the space of square integrable functions $L_2(\Theta)$, and use the geometry of the Hilbert space $\mathcal{H} = (L_2(\Theta), \langle \cdot, \cdot \rangle)$, with inner-product

$$\langle g, h \rangle = \int_\Theta g(\theta) h(\theta) \, \mathrm{d}\theta, \quad g, h \in L_2(\Theta). \tag{1}$$

The fact that $\mathcal{H}$ is a Hilbert space is often known in mathematical parlance as the Riesz–Fischer theorem; for a proof see Cheney (2001, p. 411). Borrowing geometric terminology from linear spaces, we refer to the elements of $L_2(\Theta)$ as vectors, and assess their 'magnitudes' through the use of the norm induced by the inner product in (1), i.e., $\| \cdot \| = (\langle \cdot, \cdot \rangle)^{1/2}$.

The starting point for constructing our geometry is the observation that Bayes theorem can be written using the inner-product in (1) as follows

$$p(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{\int_\Theta \pi(\theta) f(y \mid \theta) \, \mathrm{d}\theta} = \frac{\pi(\theta) \ell(\theta)}{\langle \pi, \ell \rangle}, \tag{2}$$

where $\ell(\theta) = f(y \mid \theta)$ denotes the likelihood, $\pi(\theta)$ is a prior density, $p(\theta \mid y)$ is the posterior density and $\langle \pi, \ell \rangle = \int_\Theta f(y \mid \theta) \pi(\theta) \, \mathrm{d}\theta$ is the marginal likelihood or integrated likelihood. The inner product in (1) naturally leads to considering π and ℓ that are in $L_2(\Theta)$, which is compatible with a wealth of parametric models and proper priors. By considering p, π, and ℓ as vectors with different magnitudes and directions, Bayes theorem simply indicates how one might recast the prior vector so as to obtain the posterior vector. The likelihood vector is used to enlarge/reduce the magnitude and suitably tilt the direction of the prior vector in a sense that will be made precise below.

The marginal likelihood $\langle \pi, \ell \rangle$ is simply the inner product between the likelihood and the prior, and hence can be understood as a measure of agreement between the prior and the likelihood. To make this more concrete, define the *angle measure* between the prior and the likelihood as

$$\pi \angle \ell = \arccos \frac{\langle \pi, \ell \rangle}{\|\pi\| \|\ell\|}. \tag{3}$$

Since π and ℓ are nonnegative, the angle between the prior and the likelihood can only be acute or right, i.e., $\pi \angle \ell \in [0, 90°]$. The closer $\pi \angle \ell$ is to $0°$, the greater the agreement between the prior and the likelihood. Conversely, the closer $\pi \angle \ell$ is to $90°$, the greater the disagreement between prior and likelihood. In the pathological case where $\pi \angle \ell = 90°$ (which requires the prior and the likelihood to have all of their mass on disjoint sets), we say that the prior is orthogonal to the likelihood. Bayes theorem is incompatible with a prior being orthogonal to the likelihood as $\pi \angle \ell = 90°$ indicates that $\langle \pi, \ell \rangle = 0$, thus leading to a division by zero in (2). Similar to the correlation coefficient for random variables in $L_2(\Omega, \mathbb{B}_\Omega, P)$—with $\mathbb{B}_\Omega$ denoting the Borel sigma-algebra over the sample space Ω—, our target object of interest is given by a standardized inner product

$$\kappa_{\pi,\ell} = \frac{\langle \pi, \ell \rangle}{\|\pi\| \|\ell\|}. \tag{4}$$

The quantity $\kappa_{\pi,\ell}$ quantifies how much an expert's opinion agrees with the data, thus providing a natural measure of the level of agreement between prior and data.

Before exploring (4) more fully by providing interpretations and properties we concretely define how the term 'geometry' will be used throughout the paper. The following definition of abstract geometry can be found in Millman and Parker (1991, p. 17).

Definition 1 (Abstract geometry)**.** *An abstract geometry $\mathcal{A}$ consists of a pair $\{\mathcal{P}, \mathcal{L}\}$, where the elements of set $\mathcal{P}$ are designed as points, and the elements of the collection $\mathcal{L}$ are designed as lines, such that:*

1. *For every two points $A, B \in \mathcal{P}$, there is a line $l \in \mathcal{L}$.*
2. *Every line has at least two points.*

Our abstract geometry of interest is $\mathcal{A} = \{\mathcal{P}, \mathcal{L}\}$, where $\mathcal{P} = L_2(\Theta)$ and the set of all lines is

$$\mathcal{L} = \{g + kh : g, h \in L_2(\Theta), k \in \mathbb{R}\}. \tag{5}$$

Hence, in our setting points can be, for example, prior densities, posterior densities, or likelihoods, as long as they are in $L_2(\Theta)$. Lines are elements of $\mathcal{L}$, as defined in (5), so that for example if g and h are densities, line segments in our geometry consist of all possible mixture distributions which can be obtained from g and h, i.e.,

$$\{\lambda g + (1 - \lambda)h : \lambda \in [0, 1]\}. \tag{6}$$

A related interpretation of two-component mixtures as straight lines can be found in Marriott (2002, p. 82).

Vectors in $\mathcal{A} = \{\mathcal{P}, \mathcal{L}\}$ are defined through the difference of elements in $\mathcal{P} = L_2(\Theta)$. For example, let $g \in L_2(\Theta)$ and let $0 \in L_2(\Theta)$. Then $g = g - 0 \in L_2(\Theta)$, and hence g can be regarded both as a point and as a vector. If $g, h \in L_2(\Theta)$ are vectors then we say that g and h are collinear if there exists $k \in \mathbb{R}$, such that $g(\theta) = kh(\theta)$. Put differently, we say g and h are collinear if $g(\theta) \propto h(\theta)$, for all $\theta \in \Theta$.

For any two points in the geometry under consideration, we define their compatibility as a standardized inner product (with (4) being a particular case).

Definition 2 (Compatibility)**.** *The compatibility between points in the geometry under consideration is defined as*

$$\kappa_{g,h} = \frac{\langle g, h\rangle}{\|g\|\|h\|}, \quad g, h \in L_2(\Theta). \tag{7}$$

The concept of compatibility in Definition 2 is based on the same construction principles as the Pearson correlation coefficient, which would be based however on the inner product

$$\langle X, Y\rangle = \int_\Omega XY \,\mathrm{d}P, \quad X, Y \in L_2(\Omega, \mathbb{B}_\Omega, P), \tag{8}$$

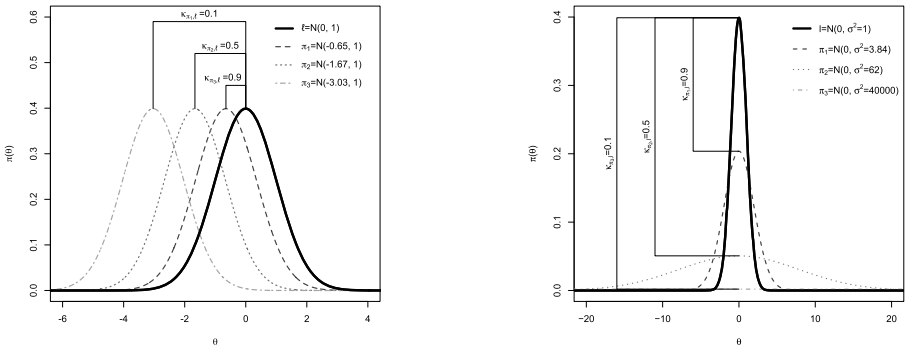
instead of the inner product in (1). However, compatibility is defined for priors, posteriors, and likelihoods in $L_2(\Theta)$ equipped with the inner product (1), whereas Pearson correlation works with random variables in $L_2(\Omega, \mathbb{B}_\Omega, P)$ equipped with the inner product (8). Our concept of compatibility can be used to evaluate how much the prior agrees with the likelihood, to measure the sensitivity of the posterior to the prior, and to quantify the level of agreement of elicited priors. As an illustration consider the following example.

Example 1. Consider the following densities $\pi_0(\theta) = I_{(0,1)}(\theta)$, $\pi_1(\theta) = 1/2I_{(0,2)}(\theta)$, $\pi_2(\theta) = I_{(1,2)}(\theta)$, and $\pi_3(\theta) = 1/2I_{(1,3)}(\theta)$. Note that $\|\pi_0\| = \|\pi_2\| = 1$, $\|\pi_1\| = \|\pi_3\| = \sqrt{2}/2$, and; further, $\kappa_{\pi_0,\pi_1} = \kappa_{\pi_2,\pi_3} = \sqrt{2}/2$, thus implying that $\pi_0 \angle \pi_1 = \pi_2 \angle \pi_3 = 45°$. Also, $\kappa_{\pi_0,\pi_2} = 0$ and hence $\pi_0 \perp \pi_2$.

As can be observed in Example 1, $(\pi_a \angle \pi_b)/90°$ is a natural measure of distinctiveness of two densities. In addition, Example 1 shows us how different distributions can be associated to the same norm and angle. Hence, as expected, any Cartesian representation $(x, y) \mapsto (\|\cdot\| \cos(\cdot\angle\cdot), \|\cdot\| \sin(\cdot\angle\cdot))$, will only allow us to represent some features of the corresponding distributions, but will not allow us to identify the distributions themselves.

To build intuition regarding $\kappa_{\pi,\ell}$, we provide Figure 1, where ℓ is set to $\mathrm{N}(0, 1)$ while $\pi = \mathrm{N}(m, \sigma^2)$ varies according to m and σ^2. Figure 1 (i) corresponds to fixing $\sigma^2 = 1$ and varying m while in the right plot $m = 0$ is fixed and σ^2 varies. Notice that in plot (i) $\kappa_{\pi,\ell} = 0.1$ corresponds to distributions whose means are approximately 3 standard deviations apart while a $\kappa_{\pi,\ell} = 0.9$ corresponds to distributions whose means are approximately 0.65 standard deviations apart. Connecting specific values of κ to specific standard deviation distances between means seems like a natural way to quickly get a rough idea of relative differences between two distributions. In Figure 1 (ii) it appears that if both distributions are centered at the same value, then one distribution must be very disperse relative to the other to produce κ values that are small (e.g., ≤ 0.1). This makes sense as there always exists some mass intersection between the two distributions considered. Thus, $\kappa_{\pi,\ell}$—to which we refer as *compatibility*—can be regarded as a measure of the level of agreement between prior and data. Some further comments regarding our geometry are in order:

- Two different densities π_1 and π_2 cannot be collinear: If $\pi_1 = k\pi_2$, then $k = 1$, otherwise $\int \pi_2(\theta)\,\mathrm{d}\theta \neq 1$.



(i) (ii)

Figure 1: Values of $\kappa_{\pi,\ell}$ when π and ℓ are both Gaussian distributions. (i) Gaussian distributions whose means become more separated. (ii) Gaussian distributions that become progressively more diffuse.

- A density can be collinear to a likelihood: If the prior is Uniform then $p(\theta \mid y) \propto \ell(\theta)$, and hence the posterior is collinear to the likelihood, i.e., in such a case the posterior simply consists of a renormalization of the likelihood.
- Two likelihoods can be collinear: Let ℓ and ℓ^* be the likelihoods based on observing y and y^*, respectively. The strong likelihood principle states that if $\ell(\theta) = f(\theta \mid y) \propto f(\theta \mid y^*) = \ell^*(\theta)$, then the *same* inference should be drawn from both samples (Berger and Wolpert, 1988). According to our geometry, this would mean that likelihoods with the same direction yield the same inference.

As a final comment on reparametrizations of the model, interpretations of compatibility should keep a fixed parametrization in mind. That is, we do not recommend comparing prior–likelihood compatibility for models with different parametrizations. Further comments on reparametrizations will be given below in Sections 2.3, 2.4, and 3.2.

2.2 Norms and their interpretation

As $\kappa_{\pi,\ell}$ is comprised of function norms, we dedicate some exposition to how one might interpret these quantities. We start by noting that in some cases the norm of a density is linked to the variance, as can be seen in the following example.

Example 2. Let $U \sim \text{Unif}(a, b)$ and let $\pi(u) = (b-a)^{-1} I_{(a,b)}(u)$ denote its corresponding density. Then, it holds that $\|\pi\| = 1/(12\sigma_U^2)^{1/4}$, where the variance of U is $\sigma_U^2 = 1/12(b-a)^2$. Next, consider a Normal model $X \sim \text{N}(\mu, \sigma_X^2)$ with known variance σ_X^2 and let ϕ denote its corresponding density. It can be shown that $\|\phi\| = \{\int_{\mathbb{R}} \phi^2(x; \mu, \sigma_X^2)\, \text{d}\mu\}^{1/2} = 1/(4\pi\sigma_X^2)^{1/4}$ which is a function of σ_X^2.

The following proposition explores how the norm of a general prior density, π, relates with that of a Uniform density, π_0.

Proposition 1. *Let $\Theta \subset \mathbb{R}^p$ with $\lambda(\Theta) < \infty$ where λ denotes the Lebesgue measure. Consider π: $\Theta \to [0, \infty)$ a probability density with $\pi \in L_2(\Theta)$ and let π_0 denote a Uniform density on Θ, then*

$$\|\pi\|^2 = \|\pi - \pi_0\|^2 + \|\pi_0\|^2. \tag{9}$$

Since $\|\pi_0\|^2$ is constant, $\|\pi\|^2$ increases as π's mass becomes more concentrated (or less Uniform). Thus, as can be seen from (9), $\|\pi\|$ is a measure of how much π differs from a Uniform distribution over Θ. This interpretation cannot be applied to Θ's that do not have finite Lebesgue measure as there is no corresponding proper Uniform distribution. Nonetheless, the notion that the norm of a density is a measure of its peakedness may be applied whether or not Θ has finite Lebesgue measure. To see this, evaluate $\pi(\theta)$ on a grid $\theta_1 < \cdots < \theta_D$ and consider the vector $p = (\pi_1, \ldots, \pi_D)$, with $\pi_d = \pi(\theta_d)$ for $d = 1, \ldots, D$. The larger the norm of the vector p, the higher the indication that certain components would be far from the origin—that is, $\pi(\theta)$ would be peaking for certain θ in the grid. Now, think of a density as a vector with infinitely many components (its value at each point of the support) and replace summation by integration to get the L_2 norm. Therefore, $\|\cdot\|$ can be used to compare the 'informativeness' of two competing priors with $\|\pi_1\| < \|\pi_2\|$ indicating that π_1 is less informative.

Further reinforcing the idea that the norm is related to the peakedness of a distribution, there is an interesting connection between $\|\pi\|$ and the (differential) entropy (denoted by H_π) which is described in the following proposition.

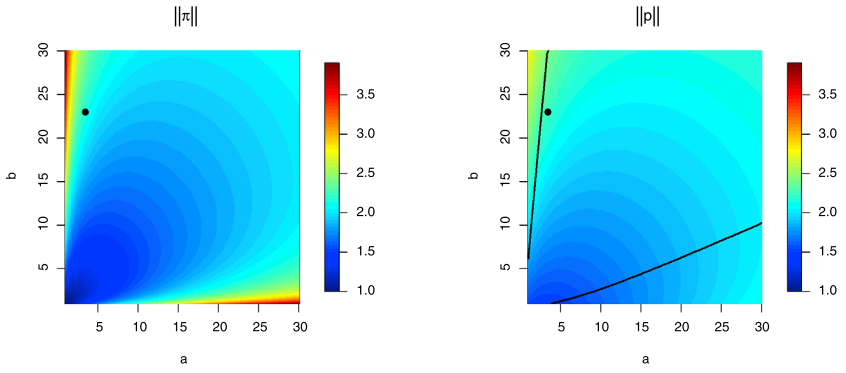
Proposition 2. *Suppose $\pi \in L_2(\Theta)$ is a continuous density on a compact $\Theta \subset \mathbb{R}^p$, and that $\pi(\theta)$ is differentiable on $\operatorname{int}(\Theta)$. Let $H_\pi = -\int_\Theta \pi(\theta) \log \pi(\theta)\, \mathrm{d}\theta$. Then, it holds that*

$$\|\pi\|^2 = 1 - H_\pi + o\{\pi(\theta^*) - 1\}, \tag{10}$$

for some $\theta^* \in \operatorname{int}(\Theta)$.

The expansion in (10) hints that the norm of a density and the entropy should be negatively related, and hence as the norm of a density increases, its mass becomes more concentrated. In terms of priors, this suggests that priors with a large norm should be more 'peaked' relative to priors with a smaller norm. Therefore, the magnitude of a prior appears to be linked to its peakedness (as is demonstrated in (9) and in Example 2). While this might also be viewed as 'informativeness,' the Beta(a, b) density has a higher norm if $(a, b) \in (1/2, 1)^2$ than if $a = b = 1$, possibly placing this interpretation at odds with the notion that a and b represent 'prior successes' and 'prior failures' in the Beta–Binomial setting. As will be further discussed in Section 2.5, a reviewer recognized that this seeming paradox is a consequence of the parameterization employed and is avoided when using the log-odds as the parameter.

As can be seen from (10), the connection between entropy and $\|\pi\|$ is an approximation at best. Just as a first-order Taylor expansion provides a poor polynomial approximation for points that are far from the point under which the expansion is made, the expansion in (10) will provide a poor entropy approximation when π is not similar to a



(i) (ii)

Figure 2: Prior and posterior norms for on-the-job drug usage toy example. Contour plots depicting the $\|\cdot\|$ associated with a Beta(a, b) prior (i) and the corresponding Beta$(a^\star, b^\star)$ posterior (ii), with $a^\star = a + 2$ and $b^\star = b + 8$. Solid lines in (ii) indicate boundaries delimiting the region of values of a and b for which $\|\pi\| > \|p\|$. The solid dot (•) corresponds to $(a, b) = (3.44, 22.99)$ (values employed by Christensen et al. 2011, pp. 26–27).

standard Uniform-like distribution π_0. However, since $\|\pi_0\|^2 = 1 - H_{\pi_0}$, the approximation is exact for a standard Uniform-like distribution. We end this discussion by noting that integrals related to $\|\pi\|^2$ also appear in physical models on L_2-spaces and they are usually interpreted as the total energy of a physical system (Hunter and Nachtergaele, 2005, p. 142), and there is considerable frequentist literature on the estimation of the integrated square of a density (see Giné and Nickl, 2008, and references therein). Now, to illustrate the information that $\|\cdot\|$ and κ provide, we consider the example described in Section 1.

Example 3 (On-the-job drug usage toy example, cont. 1)**.** From the example in the Introduction we have $\theta \mid y \sim \text{Beta}(a^\star, b^\star)$ with $a^\star = n_1 + a = 2 + a$ and $b^\star = n - n_1 + b = 8 + b$. The norm of the prior, posterior, and likelihood are respectively given by

$$\|\pi(a, b)\| = \frac{\{B(2a - 1, 2b - 1)\}^{1/2}}{B(a, b)}, \tag{11}$$

and $\|p(a, b)\| = \|\pi(a^\star, b^\star)\|$, with $a, b > 1/2$, and

$$\|\ell\| = \binom{n}{n_1} \{B\left(2n_1 + 1, 2\left(n - n_1\right) + 1\right)\}^{1/2},$$

where $B(a, b) = \int_0^1 u^{a-1}(1 - u)^{b-1}\, \mathrm{d}u$.

Figure 2 (i) plots $\|\pi(a, b)\|$ and Figure 2 (ii) plots $\|p(a, b)\|$ as functions of a and b. We highlight the prior values $(a_0, b_0) = (3.44, 22.99)$ which were employed by Christensen et al. (2011). Because prior densities with large norms will be more peaked relative to

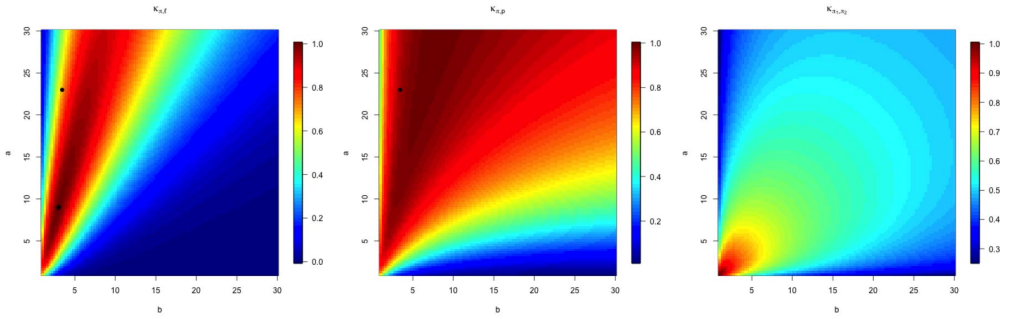
priors with small norms, $\|\pi(a_0, b_0)\| = 2.17$ is more peaked than $\|\pi(1,1)\| = 1$ (Uniform prior) indicating that $\|\pi(a_0, b_0)\|$ is more 'informative' than $\|\pi(1,1)\|$. The norm of the posterior for these same pairs is $\|p(a_0, b_0)\| = 2.24$ and $\|p(1,1)\| = 1.55$, meaning that the posteriors will have mass more concentrated than the corresponding priors. The lines found in Figure 2 (ii) represent boundary lines such that all (a, b) pairs that fall outside of the boundary produce $\|\pi(a,b)\| > \|p(a,b)\|$ which indicates that the prior is more peaked than the posterior (typically an undesirable result). If we used an extremely peaked prior, say $(a_1, b_1) = (40, 300)$, then we would get $\|\pi(a_1, b_1)\| = 4.03$ and $\|p(40, 300)\| = 4.04$ indicating that the peakedness of the prior and posterior densities is essentially the same.

Considering $\kappa_{\pi,\ell}$, it follows that

$$\kappa_{\pi,\ell}(a,b) = \frac{B(a^\star, b^\star)}{\{B(2a-1, 2b-1)B(2n_1+1, 2(n-n_1)+1)\}^{1/2}}, \tag{12}$$

with $a^\star = n_1 + a$ and $b^\star = n - n_1 + b$. Figure 3 (i) plots values of κ as a function of prior parameters a and b with $\kappa_{\pi,\ell}(a_0, b_0) \approx 0.69$ being highlighted indicating a great deal of agreement with the likelihood. In this example a lack of prior–data compatibility would occur (e.g., $\kappa_{\pi,\ell} \leq 0.1$) for priors that are very peaked at $\theta > 0.95$ or for priors that place substantial mass at $\theta < 0.05$.

The values of the hyperparameters (a, b) which, according to $\kappa_{\pi,\ell}$, are more compatible with the data (i.e., those that maximise κ) are given by $(a^*, b^*) = (3, 9)$ and are highlighted with a star $(*)$ in Figure 3 (i). In Section 2.4 we provide some connections between this prior and maximum likelihood estimators.



(i) (ii) (iii)

Figure 3: Compatibility (κ) for on-the-job drug usage toy illustration as found in (12) and Example 4. (i) Prior–likelihood compatibility, $\kappa_{\pi,\ell}(a,b)$; the black star $(*)$ corresponds to (a^*, b^*) which maximise $\kappa_{\pi,\ell}(a,b)$. (ii) Prior–posterior compatibility, $\kappa_{\pi,p}(a,b)$. (iii) Prior–prior compatibility, $\kappa_{\pi_1,\pi_2}(1,1,a,b)$, where $\pi_1 \sim \text{Beta}(1,1)$ and $\pi_2 \sim \text{Beta}(a,b)$. In (i) and (ii) the solid dot $(\bullet)$ corresponds to $(a,b) = (3.44, 22.99)$ (values employed by Christensen et al. 2011, pp. 26–27).

2.3 Angles between other vectors

As mentioned, we are not restricted to use κ only to compare π and ℓ. Angles between densities, and between likelihoods and densities or even between two likelihoods are available. We explore these options further using the example provided in the Introduction.

Example 4 (On-the-job drug usage toy example, cont. 2)**.** Extending Example 3 and (12) we calculate

$$\kappa_{\pi,p}(a,b) = \frac{B(a+a^\star-1, b+b^\star-1)}{\{B(2a-1,2b-1)B(2a^\star-1,2b^\star-1)\}^{1/2}},$$

with $a^\star = n_1 + a$ and $b^\star = n - n_1 + b$; for $\pi_1 \sim \text{Beta}(a_1, b_1)$ and $\pi_2 \sim \text{Beta}(a_2, b_2)$,

$$\kappa_{\pi_1,\pi_2}(a_1,b_1,a_2,b_2) = \frac{B(a_1+a_2-1, b_1+b_2-1)}{\{B(2a_1-1,2b_1-1)B(2a_2-1,2b_2-1)\}^{1/2}}.$$

To visualize how the hyperparameters influence $\kappa_{\pi,p}$ and κ_{π_1,π_2} we provide Figures 3 (ii) and (iii). Figure 3 (ii) again highlights the prior used in Christensen et al. (2011) with $\kappa_{\pi,p}(a_0,b_0) \approx 0.95$; see solid dot (•). This value of $\kappa_{\pi,p}$ implies that both prior and posterior are concentrated on essentially the same subset of $[0,1]$, indicating a large amount of agreement between them. Disagreement between prior and posterior takes place with priors concentrated on high probabilities of θ being greater than 0.8. In Figure 3 (iii), κ_{π_1,π_2} is largest when π_2 is close to $\text{Unif}(0,1)$ (the distribution of π_1) and gradually drops off as π_2 becomes more peaked and/or less symmetric.

In the next example, we use another data illustration to demonstrate the application of κ to a two-parameter model.

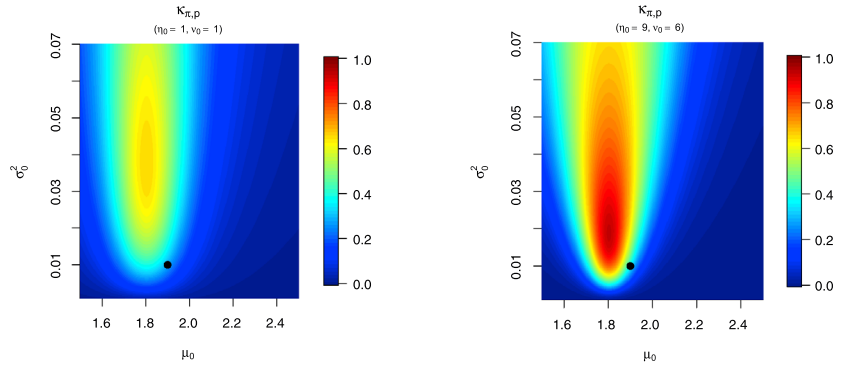
Example 5 (Midge wing length data)**.** Let $Y_1, \ldots, Y_n \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$, and $\mu \mid \sigma^2 \sim \text{N}(\mu_0, \sigma^2/\eta_0)$ and $\sigma^2 \sim \text{IG}(\nu_0/2, \sigma_0^2\nu_0/2)$; we refer to this conjugate prior distribution as $\text{NIG}(\mu_0, \eta_0, \nu_0, \sigma_0^2)$. In comparing $\pi_1 = \text{NIG}(\mu_1, \eta_1, \nu_1, \sigma_1^2)$ and $\pi_2 = \text{NIG}(\mu_2, \eta_2, \nu_2, \sigma_2^2)$, κ_{π_1,π_2} may be expressed as,

$$\kappa_{\pi_1,\pi_2} = \frac{(\pi_A\pi_B)^{1/2}}{\pi_C}\Big|_{\mu=0,\sigma^2=1}, \tag{13}$$

with

$$\begin{aligned}
\pi_A &= \text{NIG}(\mu_1, 2\eta_1, 2\nu_1+3, \nu_1\sigma_1^2/(\nu_1+3/2)),\\
\pi_B &= \text{NIG}(\mu_2, 2\eta_2, 2\nu_2+3, \nu_2\sigma_2^2/(\nu_2+3/2)),\\
\pi_C &= \text{NIG}((\eta_1\mu_1+\eta_2\mu_2)/(\eta_1+\eta_2), \eta_1+\eta_2, \nu_1+\nu_2+3,\\
&\qquad \{\nu_1\sigma_1^2+\nu_2\sigma_2^2+\eta_1\eta_2(\mu_1-\mu_2)^2/(\eta_1+\eta_2)\}/(\nu_1+\nu_2+3)).
\end{aligned}$$

Note that (13) (whose derivation can be found in Section 5.1 of the Supplementary Materials) may also be used to compute $\kappa_{\pi_1,p}$, since $p = \text{NIG}(\mu^\star, \eta^\star, \nu^\star, \sigma^{2\star})$, with



(i) (ii)

Figure 4: Prior–posterior compatibility, $\kappa_{\pi,p}(\mu_0, \eta_0, \nu_0, \sigma_0^2)$, for midge wing lengths data from Example 5. In (i) η_0 and ν_0 are fixed at one, whereas in (ii) η_0 is fixed at nine and ν_0 is fixed at six. The solid dot (•) corresponds to $(\mu_0, \sigma_0^2) = (1.9, 0.01)$ which is here used as a baseline given that hyperparameters employed by Hoff (2009, pp. 72–76) are $\mu_0 = 1.9, \eta_0 = 1, \nu_0 = 1$, and $\sigma_0^2 = 0.01$.

$$\begin{cases} \mu^\star = (n\bar{Y} + \eta_0\mu_0)/(n + \eta_0), \quad \eta^\star = \eta_0 + n, \quad \nu^\star = \nu_0 + n, \\ \sigma^{2\star} = \left\{ \nu_0\sigma_0^2 + \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \eta_0 n (\eta^\star)^{-1} (\mu_0 - \bar{Y})^2 \right\} / \nu^\star \end{cases}$$

(see Hoff, 2009, pp. 73–75). Computation of $\kappa_{\pi_1,\ell}$ also adheres to (13) if $n > 3$ and $\pi_2 = \text{NIG}(\bar{Y}, n, n-3, \sum_{i=1}^{n} (Y_i - \bar{Y})^2/(n-3))$ because then ℓ is collinear to π_2. Hoff (2009, pp. 72–76) applied this model to a dataset of nine midge wing lengths, where he set $\mu_0 = 1.9$, $\eta_0 = 1$, $\nu_0 = 1$, and $\sigma_0^2 = 0.01$, while $\bar{Y} = 1.804$ and $\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \approx 0.135$. This yields $\kappa_{\pi,p} \approx 0.28$, and thus the agreement between the prior and posterior is not particularly strong. Figure 4 (i) displays $\kappa_{\pi,p}$, as a function of μ_0 and σ_0^2 while fixing $\nu_0 = 1$ and $\eta_0 = 1$. To evaluate how $\kappa_{\pi,p}$ is affected by ν_0 and η_0, the analogous plot is displayed as Figure 4 (ii) when these values are fixed at $\nu_0 = 6$ and $\eta_0 = 9$; these alternative values for ν_0 and η_0 are those which allow the compatibility between the prior and likelihood to be maximised. It is apparent from Figure 4 that a larger σ_0^2 increases $\kappa_{\pi,p}$ substantially, and a simultaneous increase of ν_0 and η_0 would further propel this increase.

Some comments on reparametrizations are in order. We focus on the case of compatibility between two priors with a single parameter, but the rationale below also applies to compatibility between a prior and posterior, and in multiparameter settings. Let $\theta_1 \sim \pi_1$ and $\theta_2 \sim \pi_2$; further, let $g(\theta) = \lambda$ be a monotone increasing function, with range Λ, and let

$$\pi_1^g(\lambda) = \frac{\pi_1(g^{-1}(\lambda))}{g'(g^{-1}(\lambda))}, \quad \pi_2^g(\lambda) = \frac{\pi_2(g^{-1}(\lambda))}{g'(g^{-1}(\lambda))},$$

be prior densities of the transformed parameters, $g(\theta_1)$ and $g(\theta_2)$. It thus follows that

$$\frac{\int_\Lambda \pi_1^g(\lambda)\pi_2^g(\lambda)\,\mathrm{d}\lambda}{[\int_\Lambda\{\pi_1^g(\lambda)\}^2\mathrm{d}\lambda\int_\Lambda\{\pi_2^g(\lambda)\}^2\mathrm{d}\lambda]^{1/2}} = \frac{\int_\Theta \pi_1(\theta)\pi_2(\theta)/g'(\theta)\,\mathrm{d}\theta}{[\int_\Theta\{\pi_1(\theta)\}^2/g'(\theta)\,\mathrm{d}\theta\int_\Theta\{\pi_2(\theta)\}^2/g'(\theta)\,\mathrm{d}\theta]^{1/2}}.$$

The version of compatibility discussed in this section is thus invariant to linear transformations of the parameter. A variant to be discussed in Section 3.2 is more generally invariant to monotone increasing transformations.

2.4 Max-compatible priors and maximum likelihood estimators

In Example 3, we briefly alluded to a connection between priors maximising prior–likelihood compatibility $\kappa_{\pi,\ell}$ (to be termed as max-compatible priors) and maximum likelihood (ML) estimators, on which we now elaborate. Below, we use the notation $\pi(\theta \mid \alpha)$ to denote a prior on $\theta \in \Theta$, with $\alpha \in A$ are hyperparameters, and where $\dim(A) = q$ and $\dim(\Theta) = p$. (Think of the Beta–Binomial model, where $\theta \in \Theta = (0,1)$, and $\alpha = (a,b) \in A = (0,\infty)^2$.)

Definition 3 (Max-compatible prior)**.** *Let $y \sim f(\cdot \mid \theta)$, and let $\mathcal{P} = \{\pi(\theta \mid \alpha) : \alpha \in A\}$ be a family of priors for θ. If there exists $\alpha_y^* \in A$, such that $\kappa_{\pi,\ell}(\alpha_y^*) = 1$, the prior $\pi(\theta \mid \alpha_y^*) \in \mathcal{P}$ is said to be max-compatible, and α_y^* is said to be a max-compatible hyperparameter.*

The max-compatible hyperparameter, α_y^*, is by definition a random vector, and thus a max-compatible prior density is a random function. Geometrically, a prior is max-compatible if and only if it is collinear to the likelihood in the sense that $\kappa_{\pi,\ell}(\alpha_y^*) = 1$ if and only if $\pi(\theta \mid \alpha_y^*) \propto f(y \mid \theta)$, for all $\theta \in \Theta$.

The following example suggests there could be a connection between the ML estimator of θ and the max-compatibility parameter α_y^*.

Example 6 (Beta–Binomial)**.** Let $n_1 \mid \theta \sim \mathrm{Bin}(n,\theta)$, and suppose $\theta \sim \mathrm{Beta}(a,b)$. Here, $\mathcal{P} = \{\beta(\theta \mid a,b) : (a,b) \in (1/2,\infty)^2\}$, with $\beta(\theta \mid a,b) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b)$. It can be shown that the max-compatible prior is $\pi(\theta \mid a^*,b^*) = \beta(\theta \mid a^*,b^*)$, where $a^* = 1+n_1$, and $b^* = 1+n-n_1$, so that

$$\widehat{\theta} = \arg\max_{\theta\in(0,1)} f(n_1 \mid \theta) = \frac{n_1}{n} = \frac{a^*-1}{a^*+b^*-2} =: m(a^*,b^*), \tag{14}$$

with $f(n_1 \mid \theta) = \binom{n}{n_1}\theta^{n_1}(1-\theta)^{n-n_1}$.

A natural question is whether there always exists a function $m : A \to \Theta$, as in (14), linking the max-compatible parameter with the ML estimator? The following theorem addresses this.

Proposition 3. *Let $y \sim f(\cdot \mid \theta)$, and let $\widehat{\theta}$ be the ML estimator of θ. In addition, let $\mathcal{P} = \{\pi(\theta \mid \alpha) : \alpha \in A\}$ be a family of priors for θ. If there exists a unimodal max-compatible prior, then*

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} f(y \mid \theta) = m_\pi(\alpha_y^*) := \arg\max_{\theta\in\Theta} \pi(\theta \mid \alpha_y^*).$$

Proposition 3 states that the mode of the max-compatible prior coincides with the ML estimator, and in Example 6, $m(a^*, b^*) = (a^* - 1)/(a^* + b^* - 2)$ is indeed the mode of a Beta prior. A comment on parametrizations is in order. A corollary to Proposition 3 is that, due to invariance of ML estimators, if $m_\pi(\alpha_y^*)$ is the mode of the max-compatible prior for θ and $g(\theta) = \lambda$ is a function, then $g(m_\pi(\alpha_y^*))$ is the mode of the max-compatible prior of the transformed parameter $\pi^g(\lambda \mid \alpha_y^*)$. Formally,

$$g(\widehat{\theta}) = \widehat{\lambda} = \arg\max_{\lambda \in \Lambda} \sup_{\theta \in \Theta_\lambda} f(y \mid \theta) = g(m_\pi(\alpha_y^*)) = \arg\max_{\lambda \in \Lambda} \pi^g(\lambda \mid \alpha_y^*),$$

with $\Theta_\lambda = \{\theta : g(\theta) = \lambda\}$ and where Λ is the range of g.

The max-compatible prior is a 'prior' to the extent that it belongs to a family of priors, but it is basically a posterior distribution (it depends on the data). Also, there are some links between the max-compatible prior and Hartigan's maximum likelihood prior (Hartigan, 1998), which will be clarified in Section 2.5.

2.5 Compatibility in the exponential family

We now consider compatibility in the exponential family with density

$$f_\theta(y) = h(y) \exp\{\eta_\theta^{\mathrm{T}} T(y) - A(\eta_\theta)\},$$

for given functions T and h, and with $A(\eta_\theta) = \log[\int h(y) \exp\{\eta_\theta^{\mathrm{T}} T(y)\} \, \mathrm{d}y] < \infty$ denoting the so-called cumulant function. Given a random sample from an exponential family, $Y_1, \ldots, Y_n \mid \theta \sim f_\theta$, it follows that

$$\ell(\theta) = \left[\prod_{i=1}^n h(Y_i)\right] \exp\left\{\eta_\theta^{\mathrm{T}} \sum_{i=1}^n T(Y_i) - nA(\eta_\theta)\right\}.$$

The conjugate prior is known to be

$$\pi(\theta \mid \tau, n_0) = K(\tau, n_0) \exp\{\tau^{\mathrm{T}} \eta_\theta - n_0 A(\eta_\theta)\}, \tag{15}$$

where τ and n_0 are parameters, and

$$K(\tau, n_0) = \left[\int_\Theta \exp\{\tau^{\mathrm{T}} \eta_\theta - n_0 A(\eta_\theta)\} \, \mathrm{d}\theta\right]^{-1}. \tag{16}$$

The posterior density is $\pi(\theta \mid \tau + \sum_{i=1}^n T(Y_i), n_0 + n)$, with $\pi(\theta \mid \tau, n_0)$ defined as in (15); cf Diaconis and Ylvisaker (1979). In this context, compatibility can be expressed using normalizing constants from various members of the conjugate prior family as follows

$$\begin{cases} \kappa_{\pi,\ell}(\tau, n_0) = \dfrac{\{K(2\tau, 2n_0) K(2\sum_{i=1}^n T(Y_i), 2n)\}^{1/2}}{K(\tau + \sum_{i=1}^n T(Y_i), n_0 + n)}, \\ \kappa_{\pi,p}(\tau, n_0) = \dfrac{\{K(2\tau, 2n_0) K(2\{\tau + \sum_{i=1}^n T(Y_i)\}, 2\{n_0 + n\})\}^{1/2}}{K(2\tau + \sum_{i=1}^n T(Y_i), 2n_0 + n)}, \\ \kappa_{p,\ell}(\tau, n_0) = \dfrac{\{K(2\{\tau + \sum_{i=1}^n T(Y_i)\}, 2\{n_0 + n\}) K(2\sum_{i=1}^n T(Y_i), 2n)\}^{1/2}}{K(\tau + 2\sum_{i=1}^n T(Y_i), n_0 + 2n)}, \end{cases} \tag{17}$$

for (τ, n_0) for which the normalizing constants in (17) are defined. The max-compatible prior in the exponential family is given by the following data-dependent prior

$$\pi\left(\theta \mid \sum_{i=1}^{n} T(Y_i), n\right), \tag{18}$$

with $\pi(\theta \mid \tau, n)$ as in (15). Special cases of the results in (17) and (18) were manifest for instance in (12), Example 4, and Example 6.

As pointed out by a reviewer, working with the canonical parametrization brings numerous advantages, especially when measuring compatibility. Since the parametrization of a model is arbitrary (and hence the interpretation of the parameter may be different for each model) it is desirable to work in terms of a parametrization that preserves the same meaning regardless of the model under consideration. For exponential families, a natural choice is the canonical parameter $\eta_\theta = \theta$. For one thing, the conjugate prior on the canonical parameter always exists under very general conditions (Diaconis and Ylvisaker, 1979). In contrast, the conjugate family for an alternative parametrization as defined in (15) can be empty; see Gutiérrez-Peña and Smith (1995, Example 1.2). In what follows, we revisit the Beta–Binomial setting and showcase yet another advantage of working with the canonical parametrization.

Example 7. Let $\eta = \log\{\theta/(1-\theta)\}$ be the natural parameter of $\text{Bin}(n, \theta)$ and consider the prior for θ as $\text{Beta}(a, b)$. The conjugate prior for the natural parameter is

$$\pi(\eta \mid a, b) = \frac{1}{B(a, b)} \exp\{a\eta - (a + b) \log(1 + \mathrm{e}^{\eta})\}.$$

It is readily apparent that

$$\|\pi\| = \frac{\{B(2a, 2b)\}^{1/2}}{B(a, b)}, \quad a, b > 0.$$

More informative priors (i.e. larger values of a and/or b) will always be more 'peaked' than less informative ones, and there is no need to constrain the range of values of the hyperparameters to the set $(1/2, \infty)$, as it was the case in (11). Finally, note that the max-compatible prior under the canonical parametrization is $\pi(\eta \mid n_1, n - n_1)$, whereas the max-compatible prior under the parametrization used earlier in Example 6 was $\beta(\theta \mid 1 + n_1, 1 + n - n_1)$.

There are some links between the max-compatible prior introduced in Section 2.4 and Hartigan's maximum likelihood prior (Hartigan, 1998). In the context of the exponential family, Hartigan's maximum likelihood prior is a uniform distribution on the canonical parameter η. Equation (18) then implies that the max-compatible prior on the canonical parameter $\pi(\eta \mid \sum_{i=1}^{n} T(Y_i), n)$, can be regarded as a posterior derived from Hartigan's maximum likelihood prior.

3 Extensions

3.1 Local prior–likelihood compatibility

In some cases, when assessing the level of agreement between prior and likelihood, integrating over Θ may not be feasible, but one can still assess the level of agreement over priors supported on a subset of the parameter space. Below Θ represents the parameter space and Π denotes the support of the prior. More specifically, let π be a prior supported on $\Pi = \{\theta : \pi(\theta) > 0\} \subseteq \Theta$. We define local prior–likelihood compatibility as

$$\kappa^*_{\pi,\ell} = \frac{\langle \pi, \ell \rangle^*}{\|\pi\|^* \|\ell\|^*} = \frac{\langle \pi, \ell \rangle}{\|\pi\| \|\ell\|^*}, \tag{19}$$

where $\langle \pi, \ell \rangle^* = \int_\Pi \pi(\theta)\ell(\theta)\,\mathrm{d}\theta$, $\|\ell\|^* = \{\int_\Pi \ell^2(\theta)\,\mathrm{d}\theta\}^{1/2}$, and $\|\pi\|^* = \{\int_\Pi \pi^2(\theta)\,\mathrm{d}\theta\}^{1/2}$. Note that

$$\langle \pi, \ell \rangle^* = \int_\Pi \pi(\theta)\ell(\theta)\,\mathrm{d}\theta = \int_\Theta \pi(\theta)\ell(\theta)\,\mathrm{d}\theta = \langle \pi, \ell \rangle,$$

and thus if $\Pi = \Theta$, then $\kappa^*_{\pi,\ell} = \kappa_{\pi,\ell}$. In practice, we recommend using standard likelihood–prior compatibility (4) instead of its local version (19), with the exception of situations for which the likelihood is square integrable over Π but not over Θ. To illustrate that (19) could be well defined even if (4) is not, suppose $Y \mid \mu, \sigma^2 \sim \mathrm{N}(\mu, \sigma^2)$ with $\mu \sim \mathrm{N}(m, s^2)$ and $\sigma \sim \mathrm{Unif}(a, b)$, for $0 < a < b$. In this pathological single-observation case (4) would not be defined, while it follows that,

$$\kappa^*_{\pi,\ell} = \frac{\int_a^b \int_{-\infty}^{\infty} \phi(\mu \mid m, s^2)/(b-a)\ell(\mu, \sigma)\,\mathrm{d}\mu\,\mathrm{d}\sigma}{[\log(b/a)/\{4\pi s(b-a)\}]^{1/2}}.$$

Since (4) only assesses the level of agreement locally—that is, over $\Pi \subseteq \Theta$—the values of (4) and (19) are not directly comparable. A local $\kappa^*_{\ell,p}$ can be analogously defined to (19).

3.2 Affine-compatibility

We now comment on a version of our geometric setup where one no longer focuses directly on angles between priors, likelihoods, and posteriors, but on functions of these. Specifically, we consider the following measures of agreement,

$$\begin{cases} \kappa_{\sqrt{\pi},\sqrt{\ell}} = \dfrac{\langle \sqrt{\pi}, \sqrt{\ell} \rangle}{\|\sqrt{\ell}\|}, \quad \kappa_{\sqrt{\pi},\sqrt{p}} = \langle \sqrt{\pi}, \sqrt{p} \rangle, \\ \kappa_{\sqrt{\pi_1},\sqrt{\pi_2}} = \langle \sqrt{\pi_1}, \sqrt{\pi_2} \rangle, \quad \kappa_{\sqrt{p_1},\sqrt{p_2}} = \langle \sqrt{p_1}, \sqrt{p_2} \rangle. \end{cases} \tag{20}$$

Some affine-compatibilities in (20) are Hellinger affinities (van der Vaart, 1998, p. 211), and thus have links with Kurtek and Bharath (2015) and Roos et al. (2015). Action does not always takes place at the Hilbert sphere, given the need of considering $\kappa_{\sqrt{\pi},\sqrt{\ell}}$. Local versions of prior–likelihood and likelihood–posterior affine-compatibility, $\kappa_{\sqrt{\pi},\sqrt{\ell}}$ and $\kappa_{\sqrt{\ell},\sqrt{p}}$, can be readily defined using the same principles as in Section 3.1.

It is a routine exercise to prove that max-compatible hyperparameters also maximise $\kappa_{\sqrt{\pi},\sqrt{\ell}}$, and thus all comments on Section 2.4 also apply to prior–likelihood affine-compatibility. In terms of affine-compatibility in the exponential family, following the same notation as in Section 2.5, it can be shown that

$$\begin{cases} \kappa_{\sqrt{\pi},\sqrt{\ell}}(\tau, n_0) = \dfrac{\{K(\tau, n_0)K(\sum_{i=1}^n T(Y_i), n)\}^{1/2}}{K(1/2\{\tau + \sum_{i=1}^n T(Y_i)\}, \{n_0 + n\}/2)}, \\ \kappa_{\sqrt{\pi},\sqrt{p}}(\tau, n_0) = \dfrac{\{K(\tau, n_0)K(\tau + \sum_{i=1}^n T(Y_i), n_0 + n)\}^{1/2}}{K(\tau + 1/2\sum_{i=1}^n T(Y_i), n_0 + n/2)}, \\ \kappa_{\sqrt{p},\sqrt{\ell}}(\tau, n_0) = \dfrac{\{K(\tau + \sum_{i=1}^n T(Y_i), n_0 + n)K(\sum_{i=1}^n T(Y_i), n)\}^{1/2}}{K(1/2\tau + \sum_{i=1}^n T(Y_i), n_0/2 + n)}, \end{cases} \tag{21}$$

with $K(\tau, n_0)$ as defined in (16).

Affine-compatibility between priors and posteriors is invariant to monotone increasing parameter transformations, as a consequence of properties of the Hellinger distance (Roos and Held, 2011, p. 267). Affine-compatibility counterparts of all data examples are available from the supplementary materials; the conclusions are tantamount to the ones using compatibility.

4 Posterior and prior mean-based estimators of compatibility

In many situations closed form estimators of κ and $\|\cdot\|$ are not available. This leads to considering algorithmic techniques to obtain estimates. As most Bayes methods resort to MCMC methods it would be appealing to express $\kappa_{\cdot,\cdot}$ and $\|\cdot\|$ as functions of posterior expectations and employ MCMC iterates to estimate them. For example, $\kappa_{\pi,p}$ can be expressed as

$$\kappa_{\pi,p} = E_p\, \pi(\theta) \left[E_p \left\{ \frac{\pi(\theta)}{\ell(\theta)} \right\} E_p\{\ell(\theta)\pi(\theta)\} \right]^{-1/2}, \tag{22}$$

where $E_p(\,\cdot\,) = \int_\Pi \cdot\, p(\theta \mid y)\, \mathrm{d}\theta$ is the expected value with respect to the posterior density. A natural Monte Carlo estimator would then be

$$\hat{\kappa}_{\pi,p} = \frac{1}{B}\sum_{b=1}^B \pi(\theta^b) \left[\left\{ \frac{1}{B}\sum_{b=1}^B \frac{\pi(\theta^b)}{\ell(\theta^b)} \right\} \left\{ \frac{1}{B}\sum_{b=1}^B \ell(\theta^b)\pi(\theta^b) \right\} \right]^{-1/2}, \tag{23}$$

where θ^b denotes the bth MCMC iterate of $p(\theta \mid y)$. Consistency of such an estimator follows trivially by the ergodic theorem and the continuous mapping theorem, but there is an important issue regarding its stability. Unfortunately, (22) includes an expectation that contains $\ell(\theta)$ in the denominator and therefore (23) inherits the undesirable properties of the so-called harmonic mean estimator (Newton and Raftery, 1994). It has been shown that even for simple models this estimator may have infinite variance (Raftery et al. 2007), and has been harshly criticized for, among other things, converging extremely slowly. Indeed, as argued by Wolpert and Schmidler (2012, p. 655):

"the reduction of Monte Carlo sampling error by a factor of two requires increasing the Monte Carlo sample size by a factor of $2^{1/\varepsilon}$, or in excess of $2.5 \cdot 10^{30}$ when $\varepsilon = 0.01$, rendering [the harmonic mean estimator] entirely untenable."

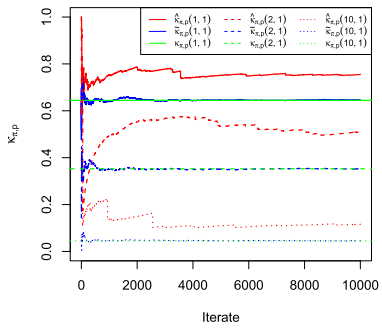


Figure 5: Running point estimates of prior–posterior compatibility, $\kappa_{\pi,p}$, for the on-the-job drug usage toy example. Green lines correspond to the true $\kappa_{\pi,p}$ values computed as in Example 4, blue represents $\tilde{\kappa}_{\pi,p}$ and red denotes $\hat{\kappa}_{\pi,p}$. Notice that $\tilde{\kappa}_{\pi,p}$ converges to the true $\kappa_{\pi,p}$ values quickly while $\hat{\kappa}_{\pi,p}$ will need much more than 10 000 Monte Carlo draws to converge.

An alternate strategy is to avoid writing $\kappa_{\pi,p}$ as a function of harmonic mean estimators and instead express it as a function of posterior and prior expectations. For example, consider

$$\kappa_{\pi,p} = E_p\, \pi(\theta) \left[\frac{E_\pi\{\pi(\theta)\}}{E_\pi\{\ell(\theta)\}}\, E_p\{\ell(\theta)\pi(\theta)\} \right]^{-1/2}, \tag{24}$$

where $E_\pi(\,\cdot\,) = \int_\Pi \cdot\, \pi(\theta)\,\mathrm{d}\theta$. Now the Monte Carlo estimator is

$$\tilde{\kappa}_{\pi,p} = \frac{1}{B}\sum_{b=1}^{B} \pi(\theta^b) \left[\left\{ \frac{\sum_{b=1}^{B} \pi(\theta_b)}{\sum_{b=1}^{B} \ell(\theta_b)} \right\} \left\{ \frac{1}{B}\sum_{b=1}^{B} \ell(\theta^b)\pi(\theta^b) \right\} \right]^{-1/2}, \tag{25}$$

where θ_b denotes the bth draw of θ from $\pi(\theta)$, which can also be sampled within the MCMC algorithm. Although representations (24) and (25) could in principle suffer from numerical instability for diffuse priors, they behave much better in practice than (22) and (23). To see this, Figure 5 contains running estimates of $\kappa_{\pi,p}$ using (23) and (25) for Example 3 with three prior parameter specifications, namely: $(a = 1, b = 1)$, $(a = 2, b = 1)$, and $(a = 10, b = 1)$; the true $\kappa_{\pi,p}$ for each prior specification is also provided. It is fairly clear that $\hat{\kappa}_{\pi,p}$ displays slow convergence and large variance, while $\tilde{\kappa}_{\pi,p}$ converges quickly.

The next proposition contains prior and posterior mean-based representations of geometric quantities that can be readily used for constructing Monte Carlo estimators.

Proposition 4. *Let π be a prior supported on $\Pi = \{\theta : \pi(\theta) > 0\} \subseteq \Theta$, with $\|\ell\|^*$ and $\kappa^*_{\pi,\ell}$ be defined as in (19), and let $E_p(\,\cdot\,) = \int_\Pi \cdot\, p(\theta \mid y)\,\mathrm{d}\theta$ and $E_\pi(\,\cdot\,) = \int_\Pi \cdot\, \pi(\theta)\,\mathrm{d}\theta$.*

Then,

$$\|p\| = \left\{ \frac{E_p\{\ell(\theta)\pi(\theta)\}}{E_\pi\, \ell(\theta)} \right\}^{1/2}, \quad \|\pi\| = \{E_\pi\, \pi(\theta)\}^{1/2},$$

$$\|\ell\|^* = \left\{ E_\pi\, \ell(\theta)\, E_p \left\{ \frac{\ell(\theta)}{\pi(\theta)} \right\} \right\}^{1/2},$$

$$\kappa^*_{\pi,\ell} = E_\pi\, \ell(\theta) \left[E_\pi\, \pi(\theta)\, E_\pi\, \ell(\theta)\, E_p \left\{ \frac{\ell(\theta)}{\pi(\theta)} \right\} \right]^{-1/2},$$

$$\kappa_{\pi,p} = E_p\, \pi(\theta) \left[\frac{E_\pi\, \pi(\theta)}{E_\pi\, \ell(\theta)} E_p\, \{\ell(\theta)\pi(\theta)\} \right]^{-1/2},$$

$$\kappa_{\pi_1,\pi_2} = E_{\pi_1}\, \pi_2(\theta) \left[E_{\pi_1}\, \pi_1(\theta)\, E_{\pi_2}\, \pi_2(\theta) \right]^{-1/2},$$

$$\kappa^*_{\ell,p} = E_p\, \ell(\theta) \left[E_p \left\{ \frac{\ell(\theta)}{\pi(\theta)} \right\} E_p\, \{\ell(\theta)\pi(\theta)\} \right]^{-1/2}.$$

Similar derivations can be used to obtain posterior and prior mean-based estimators for affine-compatibility; see supplementary materials. In the next section we provide an example that requires the use of Proposition 4 to estimate κ and $\|\cdot\|$.

5 Example: Regression shrinkage priors

5.1 Compatibility of Gaussian and Laplace priors

The linear regression model is ubiquitous in applied statistics. In vector form, the model is commonly written as

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 I), \tag{26}$$

where $y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, X is a $n \times p$ design matrix, β is a p-vector of regression coefficients, and σ^2 is an unknown idiosyncratic variance parameter; the experiments below employ $\sigma \sim \mathrm{Unif}(0, 2)$. We consider Gaussian and Laplace prior distributions for β. As documented in Park and Casella (2008) and Kyung et al. (2010) ridge regression and $\beta_j \sim \mathrm{N}(0, \lambda^2)$ (iid) produce the same regularization on β while the lasso produces the same regularization on β as assuming $\beta_j \sim \mathrm{Laplace}(0, b)$ (iid, with $\mathrm{var}(\beta_j) = 2b^2$). Below, we use π_1 to denote a Gaussian prior and π_2 a Laplace. Furthermore, we set $b = \sqrt{0.5\lambda^2}$ which ensures that $\mathrm{var}_{\pi_1}(\beta_j) = \mathrm{var}_{\pi_2}(\beta_j) = \lambda^2$ for all j.

5.2 Prostate cancer data example

We now consider the prostate cancer data example found in Hastie, Tibshirani and Friedman (2008, Section 3.4) to explore the 'informativeness' of and various compatibility measures for π_1 and π_2. In this example the response variable is the level of prostate-specific antigens measured on 97 males. Eight other clinical measurements (such as age and log prostate weight) were also measured and are used as covariates.

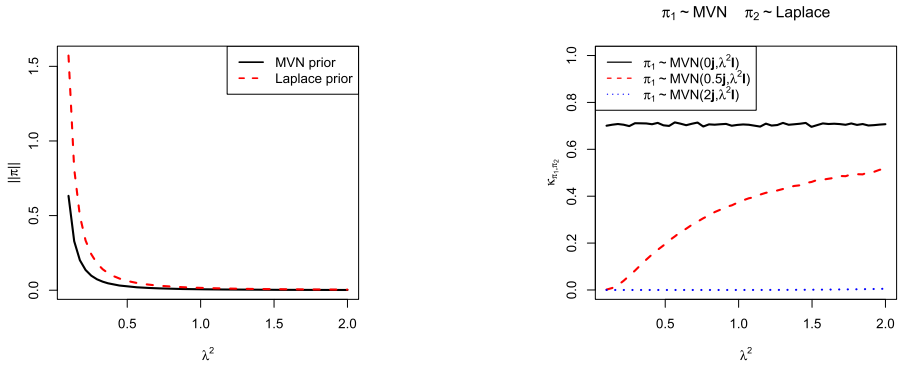


Figure 6: A comparison of priors associated with Ridge (MVN, π_1) and Lasso (Laplace, π_2) regularization in regression models in terms of $\|\pi\|$ and κ_{π_1,π_2}. The left plot depicts $\|\cdot\|$ as a function of λ^2 for both π_1 and π_2. The right compares κ_{π_1,π_2} values as a function of λ^2 when π_1 and π_2 are centered at zero to that when the center of π_1 moves away from zero.

We first evaluate the 'informativeness' of the two priors by computing $\|\pi_1\|$ and $\|\pi_2\|$ and then their compatibility using κ_{π_1,π_2}. All calculations employed Proposition 4 and results for a sequence of λ^2 values are provided in Figure 6. Focusing on the left plot of Figure 6 it appears that for small values of the λ^2, $\|\pi_1\| < \|\pi_2\|$, indicating that the Laplace prior is more peaked than the Gaussian. Thus, even though the Laplace has thicker tails, it is more 'informative' relative to the Gaussian. This corroborates the lasso penalization's ability to shrink coefficients to zero (something ridge regulation lacks). As λ^2 increases the two norms converge as both spread their mass more uniformly. The right plot of Figure 6 depicts κ_{π_1,π_2} as a function of λ^2. When π_1 is centered at zero, then κ_{π_1,π_2} is constant over values of λ^2 which means that mass intersection when both priors are centered at zero is not influenced by tail thickness. Compare this to κ values when π_1 is not centered at zero [i.e., $\pi_1 \sim \text{MVN}(0.5j, \lambda^2 I)$ or $\pi_1 \sim \text{MVN}(2j, \lambda^2 I)$]. For the former, κ increases as intersection of prior and posterior mass increases. For the latter, λ^2 must be greater than two for there to be any substantial mass intersection as κ_{π_1,π_2} remains essentially at zero.

We now fit model (26) to the cancer data and use Proposition 4 to calculate various measures of compatibility. Without loss of generality we centered the y so that β does not include an intercept and standardized each of the eight covariates to have mean zero and standard deviation one. The results are available from Figure 7.

Focusing on the left plot of Figure 7 the small values of $\kappa_{\pi_1,\ell}$ and $\kappa_{\pi_2,\ell}$ indicate the existence of prior–data incompatibility. For small values of λ^2, $\kappa_{\pi_1,\ell} > \kappa_{\pi_2,\ell}$ indicating more compatibility between prior and data for the Gaussian prior. Prior–posterior compatibility ($\kappa_{\pi,p}$) is very similar for both priors with that for π_2 being slightly bigger when λ^2 is close to 10^{-4}. The slightly higher $\kappa_{\pi,p}$ value for the Laplace prior implies that it has slightly more influence on the posterior than the Gaussian. Similarly, the Laplace

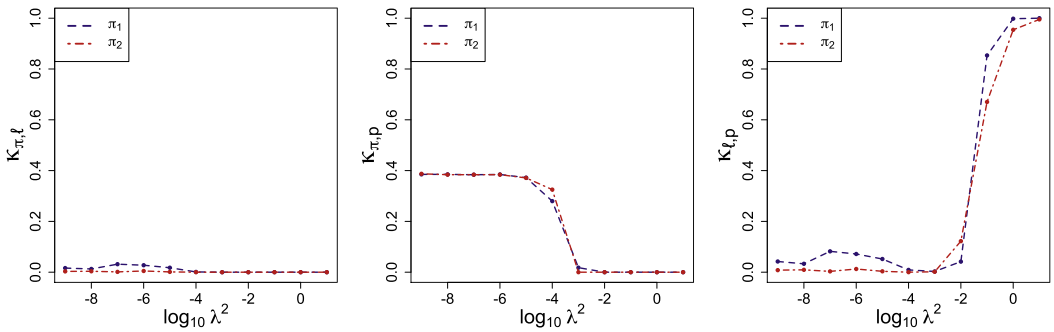


Figure 7: Compatibility (κ) for linear regression model in (26), with shrinkage priors [Ridge (MVN, π_1) and Lasso (Laplace, π_2)], applied to the prostrate cancer data from Hastie, Tibshirani and Friedman (2008, Section 3.4). The κ estimates were computed using Proposition 4.

prior seems to produce smaller $\kappa_{\ell,p}$ values than that of the Gaussian prior and κ_{ℓ,p_1} approaches one quicker than κ_{ℓ,p_2} indicating a larger amount of posterior-data compatibility. Overall, it appears that the Gaussian prior has less influence on the resulting posterior distribution relative to the Laplace when updating knowledge via Bayes theorem. Similar conclusions as above would be reached by considering affine-compatibility; see supplementary materials.

6 Discussion

Bayesian inference is regarded from the viewpoint of the geometry of Hilbert spaces. The framework offers a direct connection to Bayes theorem, and a unified treatment that can be used to quantify the level of agreement between priors, likelihoods, and posteriors—or functions of these. The possibility of developing new probabilistic models, obeying the geometrical principles discussed here, offering alternative ways to recast the prior vector using the likelihood vector remains to be explored. In terms of high-dimensional extensions, one could anticipate that as the dimensionality increases, there is increased potential for disagreement between two distributions. Consequently, κ would generally diminish as additional parameters are added, *ceteris paribus*, but a suitable offsetting transformation of κ could result in a measure of 'per parameter' agreement.

Some final comments on related constructions are in order. Compatibility as set in Definition 2 includes as a particular case the measures of niche overlap in Slobodchikoff and Schulz (1980). Peakedness as discussed in here should not be confused with the concept of Birnbaum (1948). The geometry in Definition 1 has links with the so-called affine space and thus the geometrical framework discussed above is different but has many similarities with that of Marriott (2002) and also with the mixture geometry of Amari (2016). A key difference is that the latter approaches define an inner product with respect to a density which is the basis of the construction of the Fisher information while here we define it simply as the product of two functions in $L_2(\Theta)$, and connect the

construction with Bayes theorem and with Pearson's correlation coefficient. While here we deliberately focus on positive $g, h \in L_2(\Theta)$, the case of a positive $m \equiv g(\theta) + kh(\theta) \in L_2(\Theta)$—but with g always positive and with h negative on a part of Θ—is of interest in itself, as well as the set values of k ensuring positivity of m for all θ. Some further interesting setups would be naturally allowed by slightly extending our geometry, say to include 'mixtures' with negative weights. Indeed, the parameter λ in (6) might in some cases be allowed to take some negative values while the resultant function is still positive; see Anaya-Izquierdo and Marriott (2007).

While not explored here, the use of compatibility as a means of assessing the suitability of a given sampling model, is a natural inquiry for future research.

Supplementary Material

Supplementary Material to "On the Geometry of Bayesian Inference" (DOI: 10.1214/18-BA1112SUPP; .pdf). The online supplementary materials include the counterparts of the data examples in the paper for the case of affine-compatibility as introduced in Section 3.2, technical derivations, and proofs of propositions.

References

Agarawal, A. and Daumé, III, H. (2010). "A geometric view of conjugate priors." *Machine Learning* **81**, 99–113. MR3108176. doi: https://doi.org/10.1007/s10994-010-5203-x. 1014

Aitchison, J. (1971). "A geometrical version of Bayes' theorem." *The American Statistician* **25**, 45–46. 1014

Al Labadi, L. and Evans, M. (2016). "Optimal robustness results for relative belief inferences and the relationship to prior–data conflict." *Bayesian Analysis* **12**, 705–728. MR3655873. doi: https://doi.org/10.1214/16-BA1024. 1013

Amari, S.-i. (2016). *Information Geometry and its Applications*. New York: Springer. MR3495836. doi: https://doi.org/10.1007/978-4-431-55978-8. 1032

Anaya-Izquierdo, K. and Marriott, P. (2007). "Local mixtures of the exponential distribution." *Annals of the Institute of Statistical Mathematics* **59** 111–134. MR2405289. doi: https://doi.org/10.1007/s10463-006-0095-z. 1033

Berger, J. (1991). "Robust Bayesian analysis: Sensitivity to the prior." *Journal of Statistical Planning and Inference* **25**, 303–328. MR1064429. doi: https://doi.org/10.1016/0378-3758(90)90079-A. 1013

Berger, J. and Berliner, L. M. (1986). "Robust Bayes and empirical Bayes analysis with ε-contaminated priors." *Annals of Statistics* **14**, 461–486. MR0840509. doi: https://doi.org/10.1214/aos/1176349933. 1013

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*. In *IMS Lecture Notes*, Ed. Gupta, S. S., Institute of Mathematical Statistics, vol. 6. MR0773665. 1018

Birnbaum, Z. W. (1948). "On random variables with comparable peakedness." *Annals of Mathematical Statistics* **19** 76–81. MR0024099. 1032

Christensen, R., Johnson, W. O., Branscum, A. J. and Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis*. Boca Raton: CRC Press. MR2682928. 1014, 1020, 1021, 1022

Cheney, W. (2001). *Analysis for Applied Mathematics*. New York: Springer. MR1838468. doi: https://doi.org/10.1007/978-1-4757-3559-8. 1015

de Carvalho, M., Page, G. L., and Barney, B. J. (2018). "Supplementary Material to "On the Geometry of Bayesian Inference"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1112SUPP. 1015

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families," *Annals of Statistics* **7** 269–281. MR0520238. 1025, 1026

Evans, M. and Jang, G. H. (2011). "Weak informativity and the information in one prior relative to another." *Statistical Science* **26**, 423–439. MR2917964. doi: https://doi.org/10.1214/11-STS357. 1013

Evans, M. and Moshonov, H. (2006). "Checking for prior–data conflict." *Bayesian Analysis* **1**, 893–914. MR2282210. doi: https://doi.org/10.1016/j.spl.2011.02.025. 1013

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y. S. (2008). "A weakly informative default prior distribution for logistic and other regression models." *Annals of Applied Statistics* **2**, 1360–1383. MR2655663. doi: https://doi.org/10.1214/08-AOAS191. 1013

Gutiérrez-Peña, E. and Smith, A. F. M. (1995). "Conjugate parametrizations for natural exponential families." *Journal of the American Statistical Association* **90**, 1347–1356. MR1379477. 1026

Giné, E. and Nickl, R. (2008). "A simple adaptive estimator of the integrated square of a density." *Bernoulli* **14**, 47–61. MR2401653. doi: https://doi.org/10.3150/07-BEJ110. 1020

Hartigan, J. A. (1998). "The maximum likelihood prior." *Annals of Statistics* **26** 2083–2103. MR1700222. doi: https://doi.org/10.1214/aos/1024691462. 1025, 1026

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *Elements of Statistical Learning*. New York: Springer. MR2722294. doi: https://doi.org/10.1007/978-0-387-84858-7. 1030, 1032

Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*. New York: Springer. MR2648134. doi: https://doi.org/10.1007/978-0-387-92407-6. 1023

Hunter, J. and Nachtergaele, B. (2005). *Applied Analysis*. London: World Scientific Publishing. MR1829589. doi: https://doi.org/10.1142/4319. 1020

Kurtek, S. and Bharath, K. (2015). "Bayesian sensitivity analysis with the Fisher–Rao metric." *Biometrika* **102**, 601–616. MR3394278. doi: https://doi.org/10.1093/biomet/asv026. 1014, 1027

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010). "Penalized regression, standard errors and Bayesian lassos." *Bayesian Analysis* **5**, 369–412. MR2719657. doi: https://doi.org/10.1214/10-BA607. 1030

Lavine, M. (1991). "Sensitivity in Bayesian statistics: The prior and the likelihood." *Journal of the American Statistical Association* **86** 396–399. MR1137121. 1013

Lopes, H. F. and Tobias, J. L. (2011). "Confronting prior convictions: On issues of prior sensitivity and likelihood robustness in Bayesian analysis." *Annual Review of Economics* **3**, 107–131. 1013

Marriott, P. (2002). "On the local geometry of mixture models." *Biometrika* **89** 77–93. MR1888347. doi: https://doi.org/10.1093/biomet/89.1.77. 1016, 1032

Millman, R. S. and Parker, G. D. (1991). *Geometry: A Metric Approach with Models*. New York: Springer. MR1083550. doi: https://doi.org/10.1007/978-1-4612-4436-3. 1016

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference with the weighted likelihood Bootstrap (With Discussion)." *Journal of the Royal Statistical Society, Series* B, **56**, 3–26. MR1257793. 1028

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association* **103**, 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 1030

Raftery, A. E., Newton, M. A., Satagopan, J. M. and Krivitsky, P. N. (2007). "Estimating the integrated likelihood via posterior simulation using the harmonic mean identity." In *Bayesian Statistics*, Eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M. and West, M., Oxford University Press, vol. 8. MR2433201. 1028

Roos, M. and Held, L. (2011). "Sensitivity analysis for Bayesian hierarchical models." *Bayesian Analysis* **6**, 259–278. MR2806244. doi: https://doi.org/10.1214/11-BA609. 1028

Roos, M., Martins T. G., Held, L. and Rue, H. (2015). "Sensitivity analysis for Bayesian hierarchical models." *Bayesian Analysis* **10**, 321–349. MR3420885. doi: https://doi.org/10.1214/14-BA909. 1027

Slobodchikoff, C. N. and Schulz, W. C. (1980). "Measures of niche overlap." *Ecology* **61** 1051–1055. 1032

Scheel, I., Green, P. J. and Rougier, J. C. (2011). "A graphical diagnostic for identifying influential model choices in Bayesian hierarchical models." *Scandinavian Journal of Statistics* **38**, 529–550. MR2833845. doi: https://doi.org/10.1111/j.1467-9469.2010.00717.x. 1013

Shortle, J. F. and Mendel, M. B. (1996). "The geometry of Bayesian inference." In *Bayesian Statistics*. eds. Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M., Oxford University Press, vol. 5, pp. 739–746. MR1425443. 1014

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press. MR1652247. doi: https://doi.org/10.1017/CBO9780511802256. 1027

Walter, G. and Augustin, T. (2009). "Imprecision and prior-data conflict in generalized Bayesian inference." *Journal of Statistical Theory and Practice* **3**, 255–271. MR2667666. doi: https://doi.org/10.1080/15598608.2009.10411924. 1013

Wolpert, R. and Schmidler, S. (2012). "α-stable limit laws for harmonic mean estimators of marginal likelihoods." *Statistica Sinica* **22**, 655–679. MR2987490. doi: https://doi.org/10.5705/ss.2010.221. 1028

Zhu, H., Ibrahim, J. G. and Tang, N. (2011). "Bayesian influence analysis: A geometric approach." *Biometrika* **98**, 307–323. MR2806430. doi: https://doi.org/10.1093/biomet/asr009. 1014

Acknowledgments

We thank the Editor, the Associate Editor, and a Reviewer for insightful comments on a previous version of the paper. We extend our thanks to J. Quinlan for research assistantship and discussions, and to V. I. de Carvalho, A. C. Davison, D. Henao, W. O. Johnson, A. Turkman, and F. Turkman for constructive comments. The research was partially supported by Fondecyt 11121186 and 11121131 and by FCT (Fundação para a Ciência e a Tecnologia) through UID/MAT/00006/2013.