BAYESIAN SENSITIVITY ANALYSIS WITH FISHER-RAO METRIC

SEBASTIAN KURTEK AND KARTHIK BHARATH

Abstract. We propose a geometric framework to assess sensitivity of Bayesian procedures to modelling assumptions based on the nonparametric Fisher–Rao metric. While the framework is general, the focus of this article is on assessing local and global robustness in Bayesian procedures to perturbations of the likelihood and prior, and identification of influential observations. The approach is based on a square-root representation of densities, which enables analytic computation of geodesic paths and distances, facilitating the definition of naturally calibrated local and global discrepancy measures. An important feature of our approach is the definition of a geometric ϵ-contamination class of sampling distributions and priors via intrinsic analysis on the space of probability density functions. We show the applicability of our framework to generalized mixed effects models, and directional and shape data.

Keywords:Fisher–Rao metric; Geodesic; Geometric ϵ-contamination; Influence analysis; Riemannian manifold.

1. Introduction

The main ingredients in a Bayesian model are a likelihood $f(x \mid \theta)$ and a prior distribution $\pi(\theta)$, where x denotes the data and θ denotes a set of unknown parameters. Interest is in performing inference on θ using the posterior distribution $p(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$ or some functional thereof. It is therefore important to develop diagnostic procedures to assess the influence of the data, prior and likelihood on posterior inference, which typically include detection of outlying or influential observations, global sensitivity to the perturbation of the likelihood or the prior over a suitable class, and local sensitivity to perturbations of the likelihood or the prior. Such assessments in the Bayesian setting have received considerable attention over the years; see [23] for a detailed account.

Global Bayesian sensitivity analysis is characterized by derivation of measures from variational properties of posterior functionals, such as their ranges, over a class of prior or likelihood perturbations [5, 4, 6, 40]. Local Bayes robustness methods are based on the derivatives of posterior functionals with respect to a small perturbation of the likelihood or prior; see [34], [21] and [20]. Outlier detection in the Bayesian setting using divergence or other discrepancy measures to ascertain distances between posteriors has been employed by several authors [37, 9, 36, 14, 19]. Local perturbations of the likelihood and prior, and development of case-deletion measures using divergences are intricately related to the geometry of the nonlinear manifold of densities. Geometrical considerations in influence analysis from a frequentist perspective have a rich history starting with the seminal work of [12]; also see, for example, [44], [45] and [47]. In a Bayesian setting, [46] elegantly constructed a Riemannian-geometric Bayesian perturbation model, which provided a background into which different perturbations to a Bayesian model could be embedded.

In this paper we propose a framework for Bayes sensitivity analysis based on the manifold of probability densities using the square-root representation; the framework is comprehensive in the sense that perturbations to the Bayesian model are developed under the same geometric setup. Without striving for utmost generality we ensure that the entirety of our sensitivity analysis, global, local and data perturbation, and subsequent inference, is performed intrinsically on the space of densities under a unified Riemannian metric. The key difference to the approach in [46] is that we use the nonparametric version of the Fisher–Rao Riemannian metric, in contrast to the parametric version employed in their work. The advantage of working with the nonparametric

metric is that under square-root transformation of the densities, the geometry of the space of probability density functions becomes the positive orthant of the Hilbert unit sphere, and the Riemannian metric reduces to the standard $\mathbb{L}^2$ metric. This allows one to develop analytic tools for perturbing density functions and computing geodesic distances, which are bounded above by $\pi/2$. Additionally, unlike the square-root representation used in this paper, the log transformation of densities used by [46] requires them to be strictly positive. Our approach avoids some problems associated with divergences like ϕ-divergences [13], the Kullback–Leibler divergence [28] or the functional Bregman divergence [18], and divergence-based measures: lack of symmetry, violation of the triangle inequality, and unboundedness and absence of natural scale. Several articles have used the Hellinger distance to quantify differences between distributions; see, for example, [22], [32] and [3]. This distance is symmetric and satisfies the triangle inequality, has an upper bound of $2^{1/2}$, and can be viewed as the extrinsic version of the distance we use in this paper. However, in order to use the intrinsic structure of the manifold of densities to define perturbation classes and local sensitivity measures, we choose to work with the intrinsic metric.

As an alternative to existing global sensitivity measures for prior and likelihood perturbations, we propose a novel geometric ϵ-contamination class and develop measures based on geodesic distances. In the local setup, we propose sensitivity measures for the Bayes factor, the geodesic distance, and the posterior mean, which can be easily extended to other posterior functionals. These sensitivity measures are derived using directional derivatives on the space of posterior distributions, giving them a natural geometric calibration. For identifying influential observations, we propose to use the geodesic distance under the Fisher–Rao metric to measure differences between posterior distributions.

An important advantage of the proposed framework is that the geodesic distances are available in closed form and can hence be computed quickly and exactly. This stands in contrast to the geodesic distance used in [46], which requires approximation via Dijkstra's algorithm [15]. This affects situations wherein many such distances need to be computed, and the accuracy of the approximation can affect the inference. In higher dimensions, this issue is exacerbated because the estimated distance depends heavily on the discretization of the space. Also, the approximations suffer from metrication error: roughly, the distance computed using Dijkstra's algorithm does not generally converge to the true geodesic distance with increasing resolution of the grid [11]. A simple example highlighting the practical issues in using the parametric Fisher–Rao metric, proofs of all propositions, and several additional examples, can be found in the Supplementary Material.

2. Fisher–Rao Metric and Representation Space

We restrict our attention to univariate densities on $\mathbb{R}$, though the framework is equally valid for all finite-dimensional distributions. Let $\mathcal{P}$ denote the Banach manifold of probability density functions on $\mathbb{R}$, defined as $\mathcal{P} = \{p : \mathbb{R} \to \mathbb{R}^+ \cup \{0\} : \int_{\mathbb{R}} p(x)dx = 1\}$. The space $\mathcal{P}$ is not a vector space but a manifold with a boundary, because any density function whose value is zero for any $x \in \mathbb{R}$ is a boundary element. For a point p in the interior of $\mathcal{P}$, define the tangent space as $T_p(\mathcal{P}) = \{\delta p : \mathbb{R} \to \mathbb{R} : \int_{\mathbb{R}} \delta p(x)p(x)dx = 0\}$, a vector space. The tangent space at any point p on the manifold $\mathcal{P}$ can be viewed as containing all possible perturbations of p. For any two tangent vectors $\delta p_1, \delta p_2 \in T_p(\mathcal{P})$, the nonparametric version of the Fisher–Rao Riemannian metric [38, 2, 43, 41], referred to as the Fisher–Rao metric hereafter, is defined as

$$\langle\langle \delta p_1, \delta p_2 \rangle\rangle_p = \int_{\mathbb{R}} \delta p_1(x)\delta p_2(x)\frac{1}{p(x)}dx. \tag{2.1}$$

This metric is invariant to re-parameterization [10], and has already proven to be very useful in computer vision, shape analysis and functional data analysis [41, 42, 29]. One drawback in using the Fisher–Rao metric is the difficulty associated with computing geodesic paths and distances, which stems from the fact that the Riemannian metric changes from point to point on the manifold. It

is hence important to choose a representation of the space $\mathcal{P}$ to simplify these computations. Depending on the choice of the representation, the resulting Riemannian structure can have varying degrees of complexity, requiring numerical techniques to approximate geodesics. Choices of representation include the cumulative distribution function and the log density. Unfortunately, none of these representations alleviates the problem of computing geodesics [41].

The square-root representation proposed by [7] provides an elegant solution to this problem. In particular, under this representation, the Fisher–Rao metric becomes the standard $\mathbb{L}^2$ metric and the space of probability density functions becomes the positive orthant of the unit hypersphere in $\mathbb{L}^2$. This leads to the following definition.

Definition 1. Define a continuous mapping $\phi : \mathcal{P} \mapsto \Psi$ where the space Ψ contains the positive square-root of all possible density functions. Using this mapping, define the square-root transform of probability density functions as $\phi(p) = \psi = +p^{1/2}$. The inverse mapping is simply $\phi^{-1}(\psi) = p = \psi^2$.

We omit the + sign from the representation for notational convenience. The space of all square-root transform representations of probability density functions is $\Psi = \{\psi : \mathbb{R} \to \mathbb{R}^+ \cup \{0\} : \int_{\mathbb{R}} |\psi(x)|^2 dx = 1\}$ and represents the positive orthant of the Hilbert sphere [31]. Since the differential geometry of the sphere is well known, one can compute geodesic paths and distances between probability density functions analytically. Our general approach in the remainder of the paper will be to represent probability density functions using their square-root transform representation, compute quantities of interest on Ψ, and then map them back to $\mathcal{P}$ using the inverse mapping provided in Definition 1.

We briefly describe the tools relevant to our analysis based on the geometry of Ψ. The $\mathbb{L}^2$ Riemmanian metric on Ψ is defined as $\langle \delta\psi_1, \delta\psi_2 \rangle = \int_{\mathbb{R}} \delta\psi_1(x)\delta\psi_2(x)dx$, where $\delta\psi_1, \delta\psi_2 \in T_\psi(\Psi)$ and $T_\psi(\Psi) = \left\{\delta\psi : \langle \delta\psi, \psi \rangle = 0\right\}$. Next, we are interested in the geodesic path and distance between two points in Ψ. Observe that since we are on the unit infinite-dimensional sphere, the geodesic distance between any two points equals the angle between them; in other words, the geodesic distance between $\psi_1,\ \psi_2 \in \Psi$ is $d(\psi_1, \psi_2) = \theta = \cos^{-1}(\langle \psi_1, \psi_2 \rangle)$. The geodesic path between ψ_1 and ψ_2, indexed by $\tau \in [0, 1]$, is $\eta(\tau) = \{\sin(\theta)\}^{-1}\{\psi_1 \sin(\theta - \tau\theta) + \psi_2 \sin(\tau\theta)\}$. The restriction to the positive orthant of the unit sphere does not pose any additional difficulties: for two points $\psi_1,\ \psi_2 \in \Psi$ the shortest geodesic between them is entirely contained in Ψ. It is easy to see that θ is bounded above by $\pi/2$, which imposes an upper bound on the geodesic distance between probability densities.

In the proposed framework, the exponential and inverse exponential maps are frequently used. The exponential map at a point $\psi_1 \in \Psi$, denoted by $\exp : T_{\psi_1}(\Psi) \mapsto \Psi$, is defined as $\exp_{\psi_1}(\delta\psi) = \cos(\|\delta\psi\|)\psi_1 + \sin(\|\delta\psi\|)\delta\psi(\|\delta\psi\|)^{-1}$. The purpose of this map is to map points from the tangent space to the representation space. The inverse exponential map, denoted by $\exp^{-1}_{\psi_1} : \Psi \mapsto T_{\psi_1}(\Psi)$, is given by $\exp^{-1}_{\psi_1}(\psi_2) = \theta\{\sin(\theta)\}^{-1}\{\psi_2 - \psi_1 \cos(\theta)\}$, and can be used to map points from the representation space to the tangent space. Figure 1 illustrates the relationship between $\mathcal{P}$ and Ψ.

3. Geometric Bayesian Sensitivity Analysis

3.1. **Geometric Perturbation Class.** We develop a geometric framework for assessing prior and likelihood robustness based on the notion of ϵ-contamination. We use the following notation in the rest of this paper: X denotes an observable random variable, which will be assumed to have a density $f(x \mid \theta)$ with respect to Lebesgue measure, where θ is a vector of unknown parameters lying in a compact parameter space Θ. A prior density on Θ is denoted by π and the resulting posterior distribution of θ obtained by the Bayes rule, assuming it exists, is denoted by $p_\pi(\cdot \mid x)$ and is defined by $p_\pi(\theta \mid x) = f(x \mid \theta)\pi(\theta)/m(x \mid \pi)$; here, $m(x \mid \pi)$ is the marginal density of X obtained by averaging over the prior: $m(x \mid \pi) = \int_\Theta f(x \mid \theta)\pi(\theta)d\theta$. In this section, we define the

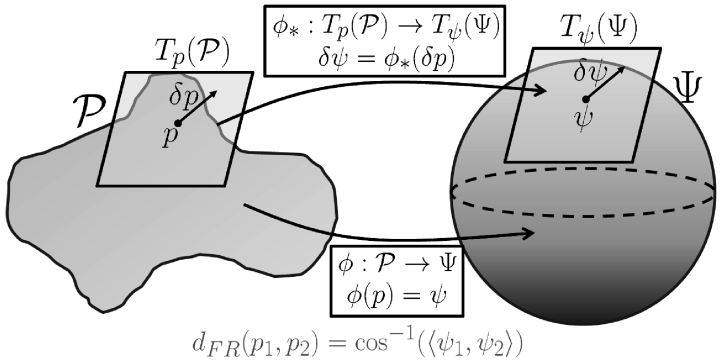

FIGURE 1. Description of the square-root transformation from $\mathcal{P}$ to the positive orthant of the unit Hilbert sphere Ψ. On $\mathcal{P}$, at a point p, its tangent space $T_p(\mathcal{P})$ is shown with the corresponding tangent vector δp. These quantities are mapped to the tangent space of ψ on Ψ and their counterparts are displayed in a similar manner. Note the isometric property: $d_{\mathrm{FR}}(p_1, p_2) = \cos^{-1}(\langle \psi_1, \psi_2 \rangle)$.

geometric perturbation class for the baseline prior and note that the likelihood perturbation class can be formed in a similar manner.

Let π_0 represent a baseline prior probability density on the parameter θ. Also, let $\mathcal{G} = \{g_1, \ldots, g_m\}$ denote a finite class of contaminants. We construct a set of tangent vectors $v_{g_1}, \ldots, v_{g_m} \in T_{\pi_0^{1/2}}(\Psi)$ using the inverse exponential map as $v_{g_i} = \exp^{-1}_{\pi_0^{1/2}}(g_i^{1/2})$ $(i = 1, \ldots, m)$. This provides a finite class of perturbations of the baseline prior, leading to the following definition.

Definition 2. For a class of densities $\mathcal{G} = \{g_1, \ldots, g_m\}$, the geometric ϵ-contamination class corresponding to the baseline prior π_0 is defined as

$$\Gamma = \left\{ \{\exp_{\pi_0^{1/2}}(\epsilon v_{g_i})\}^2 : 0 \leq \epsilon \leq 1,\ g_i \in \mathcal{G}, i = 1, \ldots, m \right\}. \tag{3.1}$$

The interpretation of this set is as follows: for an element $g_i \in \mathcal{G}$, by varying ϵ from 0 to 1, one traces the geodesic path from π_0 to g_i. Thus, if we fix a value for ϵ, we will obtain a finite set of priors that were contaminated in the directions of $g_1, \ldots, g_m$. This is further described in Figure 2. The class $\mathcal{G}$ is appropriately constructed based on the problem of interest and the baseline prior. The choice and size of the perturbations have an effect on the assessment of robustness, and useful guidelines have been suggested by [4]; our approach has been motivated by two of those guidelines: calculation of the minimum and maximum of a posterior functional over the class should be as easy as possible, and the class should correspond to easily elicitable prior information. The examples in Section 4 comment on the choice of the classes employed. We consider finite perturbation classes, although in principle these methods can be extended to infinite classes. An advantage of using such a perturbation class is the natural incorporation of the geometry of the space of densities. As will be seen in later sections, this results in geometrically calibrated local sensitivity measures computed using directional derivatives on Ψ.

Owing to its linear structure, it is easy to see that the class $\Gamma_l = \{(1-\epsilon)\pi + \epsilon g : 0 \leq \epsilon \leq 1,\ g \in \mathcal{G}\}$ induces the same kind of contamination on the marginal and the posterior. One interpretation of ϵ is as a measure of uncertainty regarding the choice of the original prior π [35]. If one were to adopt this interpretation, then under the linear ϵ-contamination class, the amount of uncertainty regarding π carries over exactly to the amount of uncertainty regarding the marginal which is averaged over the prior. If there is uncertainty regarding the choice of the likelihood as well, then such a phenomenon is quite undesirable. On the other hand, under geometric perturbation of the

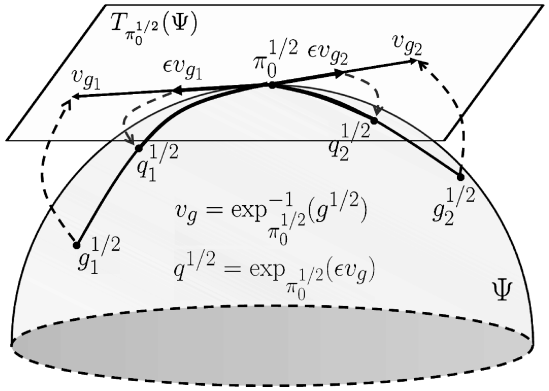

FIGURE 2. Description of the geometric ϵ-contamination class under the square-root transform representation. Using the baseline prior $\pi_0^{1/2}$ and the contamination density $g^{1/2}$, one can form a perturbation vector, v_g, by mapping $g^{1/2}$ to the tangent space of $\pi_0^{1/2}$ with the inverse exponential map. The baseline prior $\pi_0^{1/2}$ is then contaminated by mapping the point ϵv_g onto Ψ using the exponential map.

prior, the interpretation of ϵ does not carry over to the marginal and the posterior densities in the same way due to the nonlinearity in the perturbation.

3.2. **Global Sensitivity Analysis.** In the interest of brevity, we describe the framework for prior perturbations only and note that it is easily extended to handle likelihood perturbations, by defining perturbations on the space of sampling distributions with minimal change in computations. Given a likelihood function $f(x \mid \theta)$ and a baseline prior π_0, one can define the baseline posterior density, when it exists, as $p_0(\theta \mid x) = f(x \mid \theta)\pi_0(\theta)/m(x \mid \pi_0)$. In order to compute distances between posterior probability density functions we will again use the space Ψ. We are now given $p_0(\cdot \mid x)$, the baseline posterior, and $p_{g_1}(\cdot \mid x), \ldots, p_{g_m}(\cdot \mid x)$, the set of posteriors generated from the ϵ-contaminated priors.

Definition 3. For a class of contamination densities $\mathcal{G}$, consider the geometric ϵ-contamination class given by Definition 2. Then, a measure of sensitivity with respect to the geometric perturbation of π_0 is

$$S(\epsilon, \pi_0, \mathcal{G}) = \max\left\{d_{\mathrm{FR}}(p_0, p_{g_i}) : g_i \in \mathcal{G}, i = 1, \ldots, m\right\}. \tag{3.2}$$

Guided by the measure $S(\epsilon, \pi_0, \mathcal{G})$, we can additionally compute posterior functionals with respect to the nearest and farthest posteriors. Using the geodesic distance as a measure of robustness in our framework is meaningful because all of the distances are bounded above by $\pi/2$. Furthermore, this is the intrinsic metric on the space of densities and thus takes into account the geometry of that space. An important issue is that of calibration of the measure S with respect to the choice of $\mathcal{G}$ and ϵ: specification of values of S which imply sensitivity or lack thereof. We are unaware of any such measure currently available which addresses this issue satisfactorily. What is clear, however, is the subtle relationship between the mechanism of perturbation and the calibration of the influence measure. It may be argued that the issue of calibration can be mitigated through a judicious choice of the class $\mathcal{G}$. In our setup, the construction of the geometric ϵ-contamination class using the intrinsic distance on the manifold naturally provides a geometric calibration with an upper bound on the geodesic distance.

3.3. **Local Sensitivity Analysis.** In this section we define first-order local sensitivity measures based on the commonly used Bayes factor and a general posterior functional represented via an integral. We then propose a second-order local sensitivity measure based on the Fisher–Rao geodesic

distance between posterior densities. All of the local sensitivity measures are derived under the geometric ϵ-contamination class. First, we introduce some notation. Let p_0 be the baseline posterior, π_0 be the baseline prior, π_1 be another candidate prior in the model selection setup, f be the likelihood, and $v_g = \exp^{-1}_{\pi_0^{1/2}}(g^{1/2}) \in T_{\pi_0^{1/2}}(\Psi)$ be a perturbation of the baseline prior in the direction of a prior contaminant g. Let $m(x \mid \pi_0)$ denote the marginal density with respect to the baseline prior, and define $m(x \mid \epsilon g) = \int_\Theta f(x \mid \theta)\{\exp_{\pi_0^{1/2}}(\epsilon v_g)(\theta)\}^2 d\theta$ and $\tilde{m}(x \mid v_g) = \int_\Theta f(x \mid \theta)\pi_0^{1/2}(\theta)v_g(\theta)d\theta$. We use F to denote a general functional of interest and $p_{\epsilon g}$ to denote the posterior obtained from a member of the geometric ϵ-contamination class.

Proposition 1. *Under the notation described above, the local sensitivity measures based on the Bayes factor, posterior functional and geodesic distance, are the following.*

(1) *If $F_{\pi_0,\pi_1}(v_g) = m(x \mid \epsilon g)/m(x \mid \pi_1)$ denotes the Bayes factor for comparing the marginals of the contaminated baseline prior and another candidate prior, then the corresponding local sensitivity measure is $dF_{\pi_0,\pi_1}(v_g) \mid_{\epsilon=0} = 2\tilde{m}(x \mid v_g)/m(x \mid \pi_1)$.*

(2) *Suppose $F_{\pi_0,h}(v_g)$ is the expectation of $h(\theta)$ with respect to $p_{\epsilon g}$. Then, $dF_{\pi_0,h}(v_g) \mid_{\epsilon=0} = 2m(x \mid \pi_0)^{-1} \int_\Theta h(\theta) f(x \mid \theta)\pi_0^{1/2}(\theta)v_g(\theta)d\theta - 2\tilde{m}(x \mid v_g)m(x \mid \pi_0)^{-1} \int_\Theta h(\theta)p_0(\theta \mid x)d\theta$.*

(3) *Let $F_{\pi_0}(v_g)$ represent the squared geodesic distance between the posteriors p_0 and $p_{\epsilon g}$. Then, $d^2F_{\pi_0}(v_g) \mid_{\epsilon=0} = 4\tilde{m}(x \mid v_g)m(x \mid \pi_0)^{-1} \int_\Theta v_g(\theta)\pi_0^{-1/2}(\theta)p_0(\theta \mid x)d\theta - 2\int_\Theta v_g(\theta)^2\pi_0(\theta)^{-1}p_0(\theta \mid x)d\theta - 2\tilde{m}(x \mid v_g)^2 m(x \mid \pi_0)^{-2}$.*

These local sensitivity measures are defined using directional derivatives, where the directions are the perturbations defined using the proposed geometric ϵ-contamination method. We are hence able to incorporate the geometry of the space of densities in the definition of the measure. In other words, this approach unifies the local diagnostic measures with the geometry of the space under consideration.

3.4. **Influential Observations.** The general methodology in identifying influential observations is similar to that introduced in Section 3.2 in the sense that we define the influence measure based on distances between posteriors. We again denote the baseline posterior as p_0. One can evaluate the influence of the kth observation on the posterior distribution by removing it from the observation set and estimating the posterior distribution using the remaining observations. This results in a new posterior distribution p_k, leading to the following definition.

Definition 4. Given the baseline posterior p_0 and the posterior under case deletion p_k, the influence of observation k is defined as $I(k) = d_{\mathrm{FR}}(p_0, p_k)$.

This distance is symmetric and has an upper bound of $\pi/2$, which avoids the ambiguity present in many divergence measures, and provides a natural scale for evaluating influence. When the posterior density is unavailable in closed-form, and computing the marginal likelihood numerically is infeasible, it becomes necessary to estimate the quantity in Definition 4 using Monte Carlo methods. We propose an estimator based on samples from the baseline posterior, generated using either direct sampling or Markov chain Monte Carlo, to evaluate the Fisher–Rao distance between the baseline posterior and the posterior under case-deletion.

Proposition 2. *Suppose p_k is the posterior density under case-deletion and p_0 is the baseline posterior density. Correspondingly, let f_k and f_0 be the case-deletion and baseline likelihoods with π representing the prior on the parameters. Let x_k denote the kth observation and $x_{(k)}$ denote the set of observations not containing the kth one. Then $I(k) = \cos^{-1}\{L(k)\}$, where*

$$L(k) = \left\{ \int_\Theta f(x_k \mid x_{(k)}, \theta)^{-1} p_0(\theta \mid x) d\theta \right\}^{-1/2} \int_\Theta \left\{ f_k(x \mid \theta) f(x \mid \theta)^{-1} \right\}^{1/2} p_0(\theta \mid x) d\theta.$$

Given a sample from the baseline posterior density, θ_i $(i = 1, \ldots, N)$, the Monte Carlo estimate of $I(k)$ is given by $\hat{I}(k) = \cos^{-1}\left\{(b/N)\sum_{i=1}^{N} a_i\right\}$, where

$$a_i = f_k(x \mid \theta_i)^{1/2} f(x \mid \theta_i)^{-1/2}, \quad b = \left\{ N^{-1} \sum_{i=1}^{N} f(x_k \mid x_{(k)}, \theta_i)^{-1} \right\}^{-1/2}. \tag{3.3}$$

It is routine to show that the estimate $\hat{I}$ is a consistent estimator of I using the ergodic theorem and the continuous mapping property. Only one posterior sample needs to be generated to evaluate the influence measure for all observations, making this approach computationally tractable.

3.5. **Properties of Fisher–Rao Metric for Bayes Robustness.** The Fisher–Rao metric satisfies two fundamental properties in Bayes robustness analysis: first, any perturbation of the baseline prior should not affect the sampling distribution; second, when considering simultaneous perturbations of the prior and likelihood, one should be able to separate their effects on the joint distribution. We now establish these properties. Let f be the likelihood function, π_0 be the baseline prior, and let $g \in \mathcal{G}$ represent a contamination density. Under the geometric perturbation class Γ, we write a perturbation of the baseline prior and the corresponding contaminated prior, using the square-root transform representation, as

$$\delta_g \pi_0^{1/2} = \exp^{-1}_{\pi_0^{1/2}}(g^{1/2}), \quad \pi_g^{1/2} = \exp_{\pi_0^{1/2}}(\epsilon \delta_g \pi_0^{1/2}).$$

The square-root transform representation of the contaminated joint density is

$$p_g^{1/2}(x, \theta) = \{f(x \mid \theta)\pi_g(\theta)\}^{1/2}.$$

Thus, the perturbation vector on the space of square-root transform representations of joint densities is

$$\begin{aligned}
v_g(x,\theta) &= \frac{d}{d\epsilon}\left\{ f^{1/2}(x \mid \theta)\pi_g^{1/2}(\theta)\right\}\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\left\{ f^{1/2}(x \mid \theta)\exp_{\pi_0^{1/2}}(\epsilon\delta_g\pi_0^{1/2})(\theta)\right\}\Big|_{\epsilon=0} \\
&= \frac{d}{d\epsilon}\left[f^{1/2}(x \mid \theta)\left\{ \cos(\epsilon\|\delta_g\pi_0^{1/2}(\theta)\|)\pi_0^{1/2}(\theta) + \sin(\epsilon\|\delta_g\pi_0^{1/2}(\theta)\|)\frac{\delta_g\pi_0^{1/2}(\theta)}{\|\delta_g\pi_0^{1/2}(\theta)\|}\right\}\right]\Bigg|_{\epsilon=0} \\
&= f^{1/2}(x \mid \theta)\Big\{ -\sin(\epsilon\|\delta_g\pi_0^{1/2}(\theta)\|)\pi_0^{1/2}(\theta)\|\delta_g\pi_0^{1/2}(\theta)\| + \cos(\epsilon\|\delta_g\pi_0^{1/2}(\theta)\|)\delta_g\pi_0^{1/2}(\theta)\Big\}\Big|_{\epsilon=0} \\
&= f^{1/2}(x \mid \theta)\delta_g\pi_0^{1/2}(\theta).
\end{aligned}$$

Given two geometric perturbations of the baseline prior $\delta_{g_1}\pi_0^{1/2},\ \delta_{g_2}\pi_0^{1/2}$ $(g_1,\ g_2 \in \mathcal{G})$, we compute the corresponding perturbations of the joint density under the square-root transform representation and derive the corresponding Riemannian metric on that space,

$$\langle v_{g_1}, v_{g_2}\rangle = \int_\Theta \int_{\mathbb{R}} \delta_{g_1}\pi_0^{1/2}(\theta)\delta_{g_2}\pi_0^{1/2}(\theta) f(x \mid \theta) dx d\theta = \int_\Theta \delta_{g_1}\pi_0^{1/2}(\theta)\delta_{g_2}\pi_0^{1/2}(\theta) d\theta, \tag{3.4}$$

which is independent of the sampling distribution, verifying our claim. Furthermore, consistent with intuition, if the sampling distribution is fixed and the geometric perturbation model is used, the Riemannian metric on the space of joint densities is the same as that on the space of priors.

Consider simultaneous perturbations of the prior and likelihood. Let f_0 and π_0 be the baseline likelihood and prior, and let q and g represent a likelihood contaminant density and a prior contaminant density. Then, as before, the square-root transform representations of the contaminated likelihood and contaminated prior are

$$f_q^{1/2} = \exp_{f^{1/2}}(\epsilon\delta_q f^{1/2}), \quad \pi_g^{1/2} = \exp_{\pi_0^{1/2}}(\epsilon\delta_g\pi_0^{1/2}),$$

where

$$\delta_q f^{1/2} = \exp^{-1}_{f^{1/2}}(q^{1/2}), \quad \delta_g \pi_0^{1/2} = \exp^{-1}_{\pi_0^{1/2}}(g^{1/2}).$$

The resulting perturbations on the space of square-root transform representations of joint densities are

$$v_q = \frac{d}{d\epsilon}\left(f_q \pi_0\right)^{1/2}\Big|_{\epsilon=0} = \delta_q f_0^{1/2} \pi_0^{1/2}, \quad v_g = \frac{d}{d\epsilon}(f_0 \pi_g)^{1/2}\Big|_{\epsilon=0} = f_0^{1/2} \delta_g \pi_0^{1/2}.$$

Then,

$$\begin{aligned} \langle v_g, v_q \rangle &= \int_\Theta \int_{\mathbb{R}} f_0^{1/2}(x \mid \theta) \delta_g \pi_0^{1/2}(\theta) \delta_q f_0^{1/2}(x \mid \theta) \pi_0^{1/2}(\theta) dx d\theta \\ &= \int_\Theta \delta_g \pi_0^{1/2}(\theta) \pi_0^{1/2}(\theta) \int_{\mathbb{R}} f_0^{1/2}(x \mid \theta) \delta_q f_0^{1/2}(x \mid \theta) dx d\theta = 0, \end{aligned} \tag{3.5}$$

because $\langle f_0^{1/2}, \delta_q f_0^{1/2} \rangle = 0$, since perturbations are orthogonal to the representation space. This leads to a natural decomposition of the metric on the space of joint densities. To show this, consider simultaneous perturbations of the likelihood and prior. The resulting perturbation on the space of square-root transform representations of joint densities is $v = \delta_q f_0^{1/2} \pi_0^{1/2} + f_0^{1/2} \delta_g \pi_0^{1/2}$. If we are given two such simultaneous perturbations, the Riemannian metric is:

$$\begin{aligned} \langle v_1, v_2 \rangle &= \langle f_0^{1/2} \delta_{g_1} \pi_0^{1/2}, f_0^{1/2} \delta_{g_2} \pi_0^{1/2} \rangle + \langle \delta_{q_1} f_0^{1/2} \pi_0^{1/2}, f_0^{1/2} \delta_{g_2} \pi_0^{1/2} \rangle \\ &\quad + \langle f_0^{1/2} \delta_{g_1} \pi_0^{1/2}, \delta_{q_2} f_0^{1/2} \pi_0^{1/2} \rangle + \langle \delta_{q_1} f_0^{1/2} \pi_0^{1/2}, \delta_{q_2} f_0^{1/2} \pi_0^{1/2} \rangle \\ &= \langle f_0^{1/2} \delta_{g_1} \pi_0^{1/2}, f_0^{1/2} \delta_{g_2} \pi_0^{1/2} \rangle + \langle \delta_{q_1} f_0^{1/2} \pi_0^{1/2}, \delta_{q_2} f_0^{1/2} \pi_0^{1/2} \rangle, \end{aligned} \tag{3.6}$$

where the last equality holds due to (3.5). Thus, the Riemannian metric on the space of joint densities can be written as a sum of likelihood and prior perturbation terms.

4. Examples

4.1. **Local and global robustness in a directional data model.** In this example, we use a Bayesian model to analyze directional data on the unit circle $\mathbb{S}^1$. This dataset consists of 76 directions of turtle movement after a certain treatment is applied; the raw data is displayed in Figure 3(a). Assuming independent and identically distributed observations, [19] considered the following baseline model:

$$x_i \mid \theta \sim f = \mathrm{vM}(\theta, \hat{\kappa}), \quad \theta \sim \pi_0 = \mathrm{vM}(0, 0.01) \quad (i = 1, \ldots, 76),$$

where $\mathrm{vM}(\mu, \kappa)$ is the von Mises distribution with mean μ and concentration κ, and $\hat{\kappa} = 1.1423$ is the maximum likelihood estimator of the concentration parameter based on the given data. [19] considered identification of influential observations based on the functional Bregman divergence. For simplicity, they set the unknown likelihood concentration parameter to equal the maximum likelihood estimator. We assess the global sensitivity of the posterior distribution of μ to this choice via perturbations of the concentration parameter. Two distinct groups are apparent from Figure 3(a). This suggests that if the concentration κ were to be small compared to $\hat{\kappa}$, the posterior of μ at the perturbed likelihood would significantly move away from the base posterior with κ set to its maximum likelihood estimate; although not as drastic, similar behaviour can be expected for large values of κ. For this purpose, we assess sensitivity of the posterior of μ at 100 different values of κ in the likelihood function ranging from 0.01 to 10. The global sensitivity measure is the Fisher–Rao distance between the baseline posterior and the posterior under the perturbed likelihood. Since this is a conjugate model all of the posteriors are von Mises densities and we use numerical integration to compute the distance. The results are shown in Figure 3(b), and are as expected with regards to the values of κ. The range of the global measure based on the geodesic distance between the base and perturbed posteriors is approximately 1.2. Since $\pi/2$ is the upper

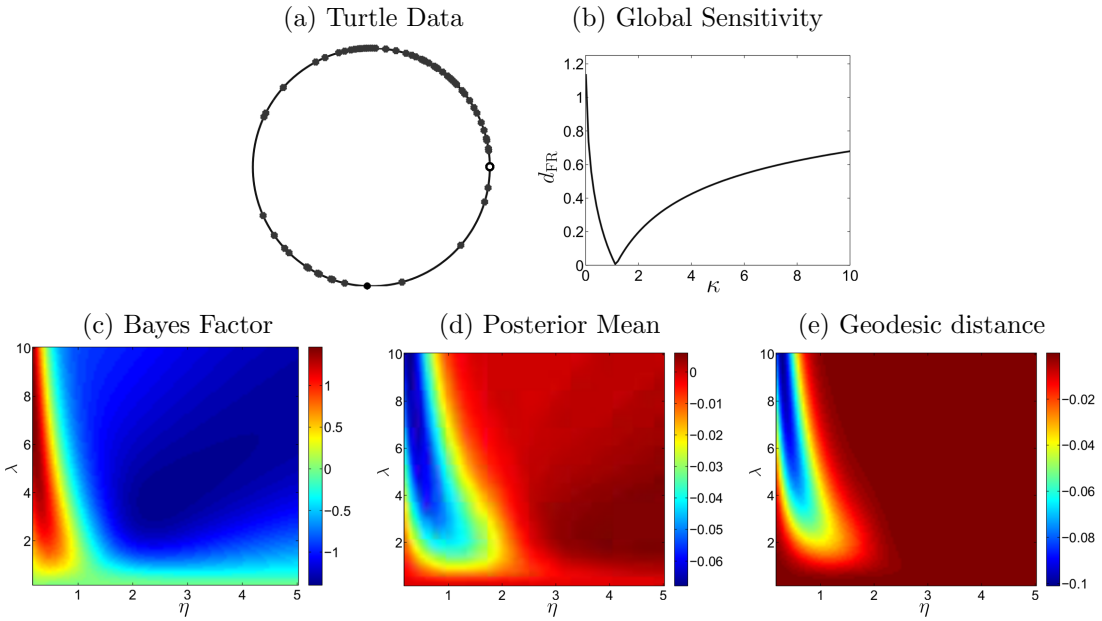

FIGURE 3. Results of sensitivity analysis for a directional data model. (a) Turtle directional data with the origin marked by an empty bullet symbol. (b) Global sensitivity to perturbations of the likelihood concentration parameter (x-axis) away from $\hat{\kappa}$ based on the Fisher–Rao distance. (c) Bayes factor, (d) posterior mean, and (e) geodesic distance local influence analysis to perturbations of the baseline von Mises prior density by a class of wrapped Laplace priors with different concentration (y-axis) and skewness (x-axis) parameters.

bound for our distance, we can conclude that the posterior of μ is quite sensitive to the choice of κ, advocating the use of a prior distribution for κ instead of choosing it based on the data.

Next, we assess local sensitivity to ϵ-contamination of the prior. The clustering of the data suggests that we check for the assumption of a symmetric distribution for the prior. The contamination class we consider is a family of wrapped Laplace distributions with zero mode [24], which can be parameterized by a concentration parameter λ and a skewness parameter η. For $\eta < 1$, the wrapped Laplace distribution is skewed in the counter-clockwise direction, and for $\eta > 1$ it is skewed in the clockwise direction. When $\eta = 1$ we obtain the symmetric wrapped Laplace distribution. Thus, our contamination class is formed by jointly varying the parameters λ from 0.2 to 10 and η from 0.2 to 5. For the local measure for Bayes factor, we set $\pi_1 = vM(\pi/2, 0.01)$; Figure 3(c)–(e) displays the results. The local Bayes factor measure is insensitive to perturbations of the prior only when the contaminant is approximately symmetric, i.e. $\eta = 1$. When the contaminant prior is highly skewed in the counter-clockwise direction, the Bayes factor sensitivity measure is positive and vice versa when $\eta > 1$. For the concentration parameter, the local measure for Bayes factor tends to zero as the concentration goes to zero. This is reasonable since both the von Mises and the wrapped Laplace distributions converge to the uniform distribution. As a result, the perturbations have little effect on the baseline model. The posterior mean, however, is fairly sensitive for moderately concentrated and highly counter-clockwise skewed wrapped Laplace distributions. In these cases, the posterior mean decreases, with respect to the baseline posterior mean of 1.1198 radians, as indicated by the negative sign of the local sensitivity measure. On the other hand, when we perturb the prior using highly clockwise skewed wrapped Laplace distributions, the posterior mean is fairly insensitive with the local sensitivity measure close to zero. Finally, the second-order

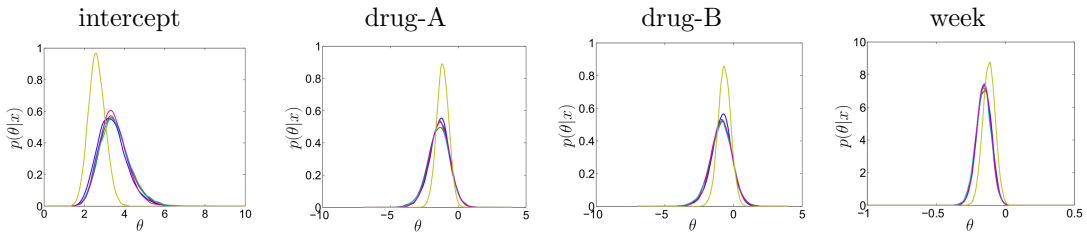

FIGURE 4. Kernel density estimates of marginal posteriors under baseline and perturbed models (i)–(v). Baseline is marked in blue, (i) in green, (ii) in red, (iii) in cyan, (iv) in purple and (v) in yellow.

measure for the geodesic distance shows a similar trend to the first-order measure for the posterior mean. The relative ease of implementation of the sensitivity measure is due to the fact that all geometric quantities are available analytically.

4.2. **Global robustness in a generalized mixed effects model.** This example considers the presence or absence of bacteria in persons monitored through a fixed time window using a Bayesian generalized linear mixed effects model; the data, available in the `MASS` package in `R`, was previously used in [8]. The predictors are the week of test and the treatment groups, placebo, drug-A, drug-B. We use the baseline logistic mixed effects model

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}),$$

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \sum_{k=1}^{3} x_{ij}^k \beta^k + V_i,$$

$$\mu \sim N(0,100), \qquad \beta^k \sim N(0,100), \qquad V_i \sim N(0,\sigma^2), \qquad \tau = \frac{1}{\sigma^2} \sim \Gamma(0.01, 0.01),$$

where the response Y_{ij} indicates the presence or absence of bacteria in person i at week j, x_{ij} are the week of test and indicator variables for the treatment, p_{ij} is the probability of bacteria presence, and V_i are independent and identically distributed subject random effects; $\Gamma(a,b)$ denotes the Gamma distribution with a and $1/b$ as shape and scale parameters respectively. For the precision parameter of the random effects, [39] and [33] argue against the use of a Gamma prior with small shape and large scale parameters; instead, they advocate the use of a half-normal or a half-Cauchy prior on the standard deviation of the random effects. They note that the effect of the choice of prior for the precision parameter of the random effects on the posterior distribution of the fixed effects is of particular interest. Consequently, in this example, our interest is in assessing global robustness to such choices of prior. Aside from the baseline prior, we consider five other choices for the precision parameter of the mixed effects: (i) half normal with variance 100, (ii) half Cauchy with scale parameter 100, (iii) uniform on $(0, 100)$, (iv) $\Gamma(1, 2)$, and (v) $\Gamma(9, 0.5)$. The prior (v) is included here for comparison only. For all models, we use Markov chain Monte Carlo simulations to generate 9500 samples from the posterior, after a burn-in of 1000, and use these samples to generate individual kernel density estimates for the marginal posteriors for all of the coefficients of the fixed effects, which are displayed in Figure 4. We then compute the Fisher–Rao distance between each baseline marginal posterior and the corresponding posterior resulting from the perturbation of the prior. These results are reported in Table 4.2. In this case, the baseline model and models with priors (i)–(iv) all yield very similar marginal posteriors, as confirmed by the very small Fisher–Rao distances. For comparison, the unreasonable prior choice in (v) results in much larger distances between the marginal posteriors. Thus, we conclude that, for the available dataset, the Bayesian model is insensitive to any reasonable choice of the prior for the precision of the random effects.

Fisher–Rao distances between the baseline and perturbed marginal posteriors

Fixed Effect	Model				
	(i)	(ii)	(iii)	(iv)	(v)
intercept	0.1054	0.0864	0.0982	0.0740	0.6716
drug-A	0.0716	0.0499	0.0590	0.0435	0.3835
drug-B	0.0666	0.0580	0.0683	0.0445	0.3432
week	0.0524	0.0572	0.0630	0.0311	0.3670

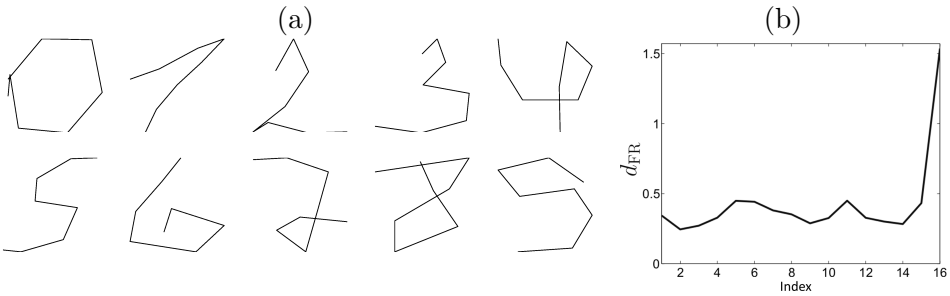

FIGURE 5. Handwritten digit dataset. (a) One example of each digit. (b) Influence measures for 15 shapes of digit 0 (first 15 cases), and one shape of digit 1 (16th case).

4.3. **Identification of influential observations in a shape data model.** In the final example, we consider the effects of deleting an observation from a sample of shapes on the estimated posterior distribution of the modal shape using a handwritten digit dataset from [1]. The full database was created by collecting 250 writing samples of digits from 44 writers. The raw data is provided as (x, y) coordinates of 8 landmark points on each digit. An example of each digit is shown in Figure 5(a), with the landmarks in the plot connected by straight lines for improved visualization. [26] defined shape as a mathematical property that remains unchanged under rotation, translation, and global scaling. In this example, we use his definition and use a Bayesian approach to estimate the modal shape of a digit class. Throughout this description, we use material from [16]. Let a configuration of landmarks denoting a digit be represented using a complex vector $x \in \mathbb{C}^8$. In order to remove the translation variability from the representation space, we pre-multiply each of the landmark configurations with a Helmert submatrix, resulting in $x_H = Hx \in \mathbb{C}^7$. Then, the pre-shape of a landmark configuration is defined as $z = x_H/\|x_H\| \in \mathbb{C}S^6$, where $\mathbb{C}S^6$ is the complex sphere in seven dimensions. The pre-shape is invariant to translations and scalings of the original landmark configurations. In order to remove rotational variability from the data, we align all of the pre-shapes to a randomly chosen observation. Then, given a sample of digit shapes $z_1, \ldots, z_n \in \mathbb{C}S^6$, we define the likelihood as the complex Watson distribution with mode μ and a known concentration parameter κ. In this example, we estimate κ from the data using Equation (6.14) in [16]. As the prior distribution for μ we choose the complex Bingham distribution with parameter matrix $A = I_7$. The advantage of using this distribution as a prior is that it is invariant to rotation and it is a conjugate prior for the complex Watson distribution. The resulting posterior distribution for the mode μ is a complex Bingham distribution with parameter matrix $\kappa \sum_{i=1}^{n} z_i z_i^* + I_7$, where z^* is the conjugate transpose of z.

As mentioned earlier, we are interested in identifying observations that have a high influence on the posterior distribution of μ. We again use the influence measure from Definition 4 for this purpose, and compute the estimator of the Fisher–Rao distance given in Proposition 2. Because the baseline posterior distribution for μ is a complex Bingham distribution, we can sample from it directly using the methods given in [27]. In other cases, we would have to resort to Markov chain Monte Carlo methods. The right panel of Figure 6 provides plots of sorted influence measures for

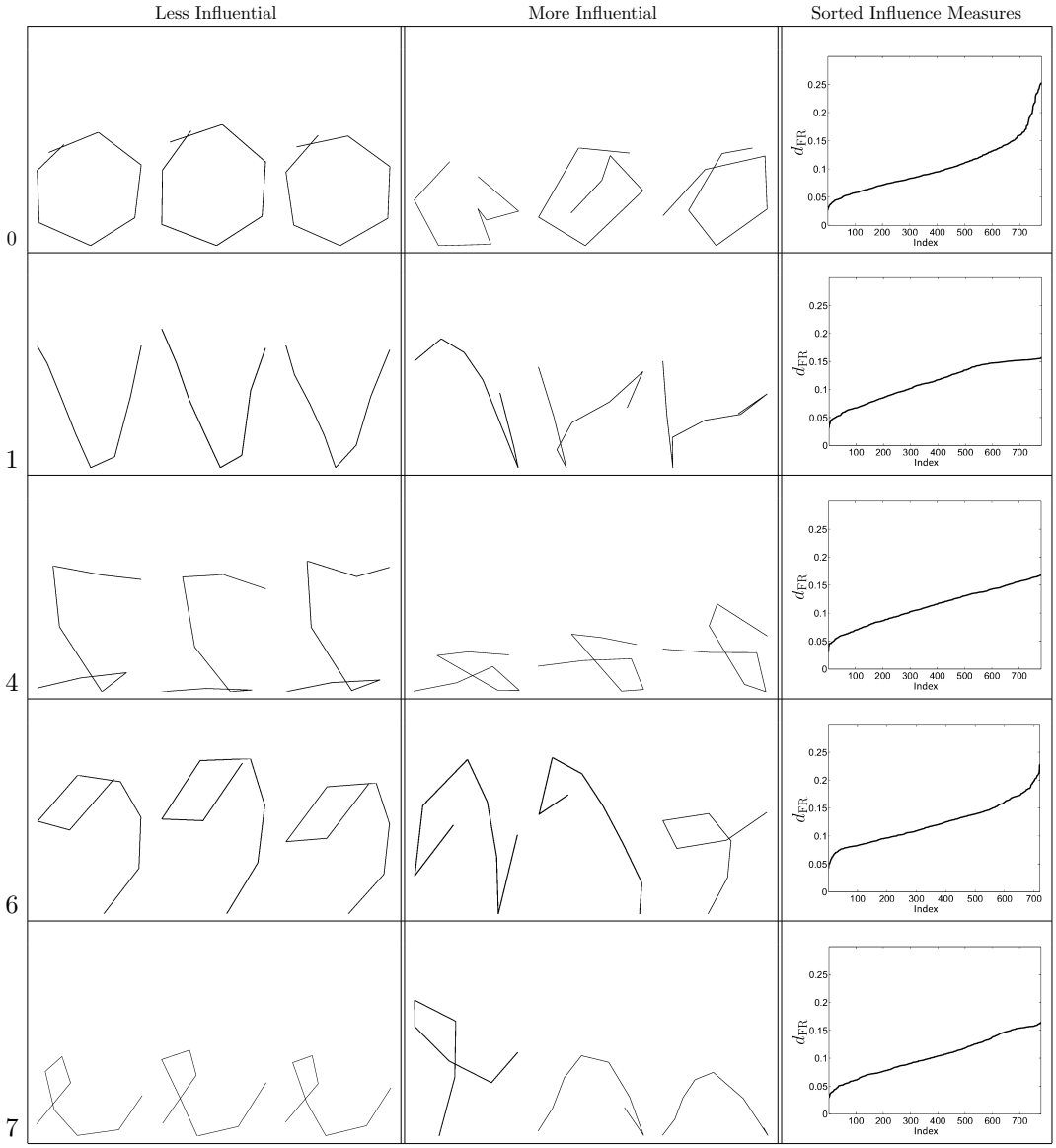

FIGURE 6. Identification of influential digit shapes. Left: Least and most influential digit shapes for digits $0, 1, 4, 6, 7$. Note that the shapes are invariant to rotations. Right: Sorted influence measures for all instances in each digit class.

all observations in the case of digits $0, 1, 4, 6, 7$; each digit class was considered separately. First, we observe that there is a lot of variability in each of the digit classes, due to significantly different handwriting styles of the subjects. As a result, none of the observations is highly influential on its own; all influence measures were less than 0.3. We hypothesize that by removing blocks of observations, this result would change. Nonetheless, we plot the three least and three most influential observations in each of the considered digit classes in Figure 6. Consistent with intuition, the shapes of the three least influential digits look very similar, while the three most influential digits look outlying. To assess the quality of the proposed estimator in this setting, we estimate the influence measure for 50 randomly chosen shapes of digits $0, 1, 4, 6, 7$ based on 50 samples of size 100000 from the baseline posterior. The mean variances of our estimator across 50 randomly

chosen shapes, corresponding to digits $0, 1, 4, 6, 7$ respectively, were $6.9 \times 10^{-8}, 2.3 \times 10^{-4}, 1.5 \times 10^{-4}, 1.5 \times 10^{-5}$ and 4.9×10^{-4}.

In order to further assess the effectiveness of the proposed influence measure, we generated a dataset consisting of the fifteen least influential shapes for digit 0 and a random shape of digit 1, and computed the influence measures for this data. We expect the shape of digit 1, the 16th shape, to be highly influential. Figure 5(b) shows the influence measures of all 16 shapes used in this simulation. While some of the 0 digit shapes have influence scores close to 0.5, there is one clear highly influential shape with a score close to $\pi/2$, which is the maximum of the scale.

5. Discussion

The square-root transform representation should make a seamless transition to the Bayesian nonparametric setting since the manifold of parametric densities is a submanifold of $\mathcal{P}$ considered here; expressions for geodesic paths and distances remain unaltered. Additionally, it would be interesting to examine issues of posterior consistency in a geometric neighbourhood such as the one considered in this paper; much work remains to be done in this direction.

When posterior densities are unavailable in closed-form, good estimators of the geodesic distance are imperative. Excepting the setting of influence analysis under case-deletion, they have not been explored in this article. Methods of incorporating the calculation of geodesics into existing Markov chain Monte Carlo procedures would be greatly beneficial. However, under the parametric setting when the unknown parameter vector is of small dimension, similar to the settings considered in this article, the geodesic distances can be calculated with fair accuracy.

Acknowledgements

We are indebted to Steve MacEachern for encouragement and discussions. We thank Dipak Dey, Anuj Srivastava and Huiling Le for suggestions. We also thank the Editor, Associate Editor and two referees for valuable comments and suggestions which improved the overall exposition.

Supplementary Material

Supplementary Material available at *Biometrika* online includes an example demonstrating practical issues with the parametric Fisher–Rao metric, examples on calculation of geodesic distances with comparisons to the Kullback–Leibler divergence, additional examples for global/local sensitivity analysis and identification of influential observations, and proofs of propositions.

References

[1] F. Alimoglu and E. Alpaydin. Methods of combining multiple classifiers based on different representations for pen-based handwriting recognition. In *Proceedings of the Fifth Turkish Artificial Intelligence and Artificial Neural Networks Symposium (TAINN 96)*, Istanbul, 1996.

[2] S. Amari. *Differential-Geometric Methods in Statistics*. Springer, New York, 1985.

[3] R. Beran. Minimum Hellinger distance estimators for parametric models. *Annals of Statistics*, 5:445–463, 1977.

[4] J. O. Berger. Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25:303–328, 1990.

[5] J. O. Berger. An overview of robust Bayesian analysis. *TEST*, 3:5–58, 1994.

[6] J. O. Berger and L. M. Berliner. Robust Bayes and empirical Bayes analysis with ϵ-contaminated prior. *Annals of Statistics*, 14:461–486, 1986.

[7] A. Bhattacharya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[8] P. Brown and L. Zhou. MCMC for generalized linear mixed models with glmmBUGS. *The R Journal*, 2:13–17, 2010.

[9] B. P. Carlin and N. G. Polson. An expected utility approach to influence diagnostics. *Journal of the American Statistical Association*, 86:1013–1021, 1991.
[10] N. N. Čencov. *Statistical decision rules and optimal inference*. American Mathematical Society, Providence, 1982.
[11] L. D. Cohen and R. Kimmel. Global minimum for active contour models: A minimal path approach. *International Journal of Computer Vision*, 24:57–78, 1997.
[12] R. D. Cook. Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, Series B*, 48:133–169, 1986.
[13] I. Cziszar. Information measures: A critical survey. In *Transactions of 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pages 73–86. Academia, Prague, 1974.
[14] D. K. Dey and L. R. Birmiwal. Robust Bayesian analysis using divergence measures. *Statistics and Probability Letters*, 20:287–294, 1994.
[15] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.
[16] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. Wiley, Chichester, 1998.
[17] D. J. Finney. The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, 34:320–334, 1947.
[18] A. B. Frigyik, S. Srivastava, and M. R. Gupta. Functional Bregman divergences and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54:5130–5139, 2008.
[19] G. Goh and D. K. Dey. Bayesian model diagnostics using functional Bregman divergence. *Journal of Multivariate Analysis*, 124:371–383, 2014.
[20] P. Gustafson. Local sensitivity of inferences to prior marginals. *Journal of American Statistical Association*, 91:774–781, 1996.
[21] P. Gustafson and L. Wasserman. Local sensitivity diagnostics for Bayesian inference. *Annals of Statistics*, 23:2153–2167, 1995.
[22] G. Hooker and A. N. Vidyashankar. Bayesian model robustness via disparities. *TEST*, 23:556–584, 2013.
[23] D. R. Insua and F. Ruggeri, editors. *Robust Bayesian Analysis*, volume 152 of *Lecture Notes in Statistics*. Springer, New York, 2000.
[24] S. R. Jammalamadaka and T. J. Kozubowski. New families of wrapped distributions for modeling skew circular data. *Communications in Statistics-Theory and Methods*, 33:2059–2074, 2004.
[25] R. E. Kass. The geometry of asymptotic inference. *Statistical Science*, 3:188–219, 1989.
[26] D. G. Kendall. Shape manifolds, Procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society*, 16:81–121, 1984.
[27] J. T. Kent, P. D. L. Constable, and F. Er. Simulation for the complex Bingham distribution. *Statistics and Computing*, 14:53–57, 2004.
[28] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.
[29] S. Kurtek, A. Srivastava, E. Klassen, and Z. Ding. Statistical modeling of curves using shapes and related features. *Journal of the American Statistical Association*, 107:1152–1165, 2012.
[30] M. H. Kutner, C. J. Nachtsheim, and J. Neter. *Applied Linear Regression Models*. McGraw-Hill/Irwin, Burr Ridge, 2004.
[31] S. Lang. *Fundamentals of Differential Geometry*. Springer, New York, 2nd edition, 1999.
[32] B. Lindsay. Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, 2:1081–1114, 1994.
[33] D. Lunn, D. J. Spiegelhalter, A. Thomas, and N. Best. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28:3049–3067, 2009.

[34] R. E. McCulloch. Local model influence. *Journal of American Statistical Association*, 84:473–478, 1984.
[35] E. Moreno. Global Bayesian robustness for some classes of prior distributions. In D. R. Insua and F. Ruggeri, editors, *Robust Bayesian Analysis*, volume 152 of *Lecture Notes in Statistics*, pages 45–70. Springer, New York, 2000.
[36] D. Pena and I. Guttman. Comparing probabilistic methods for outlier detection in linear models. *Biometrika*, 80:603–610, 1993.
[37] F. Peng and D. K. Dey. Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics*, 23:199–213, 1995.
[38] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37:81–91, 1945.
[39] M. Roos and L. Held. Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6:259–278, 2011.
[40] F. Ruggeri and S. Sivaganesan. On a global sensitivity measure for Bayesian inference. *Sankhya*, 62:110–127, 2000.
[41] A. Srivastava, I. H. Jermyn, and S. H. Joshi. Riemannian analysis of probability density functions with applications in vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, Minneapolis, 2007.
[42] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1415–1428, 2011.
[43] P. W. Vos and R. E. Kass. *Geometrical Foundations of Asymptotic Inference*. Wiley, New York, 1997.
[44] H. Zhu, J. G. Ibrahim, S. Lee, and H. Zhang. Perturbation selection and influence measures in local influence analysis. *Annals of Statistics*, 35:2565–2588, 2007.
[45] H. Zhu, J. G. Ibrahim, and N. Tang. Bayesian sensitivity analysis of statistical models with missing data. *Statistica Sinica*, 24:871–896, 2014.
[46] H. Zhu, J. G. Ibrahim, and N. Teng. Bayesian influence analysis: A geometric approach. *Biometrika*, 98(2):307–323, 2011.
[47] H. Zhu and S. Lee. Local influence for incomplete-data models. *Journal of the Royal Statistical Society, Series B*, 63:111–126, 2001.

SUPPLEMENTARY MATERIAL

6. Additional Example for Global Sensitivity Analysis

Consider the following baseline model for independent and identically distributed observations:

$$x_i \mid \theta \sim f = N(\theta, 1), \quad \theta \sim \pi_0 = N(0, 1) \quad (i = 1, \ldots, 50).$$

We consider a skew normal contamination class, parameterized by a shape parameter $\alpha \in [-5, 5]$. Figure 7 displays the considered ϵ-contaminated prior set under the linear and geometric frameworks by fixing $\epsilon = 0.5$ and α in the set $\{\pm 1, \pm 2, \pm 3, \pm 4, \pm 5\}$. We begin by simulating data $x_1, \ldots, x_{50}$ from the baseline model and generating a set of contaminated priors for 31 equally spaced values $\epsilon \in [0, 1]$ and 101 equally spaced values $\alpha \in [-5, 5]$ using the two different types of contamination methods. First, the baseline posterior p_0 is computed, where the normalizing constant is calculated numerically. In similar fashion, we compute the posterior densities resulting from the contaminated priors. Denote any of the contaminated posteriors by $p(\theta \mid x)$. In this example, we use two approaches for global analysis: geometric contamination with d_{FR} between posteriors as a sensitivity measure, and linear contamination with the Kullback–Leibler divergence as a sensitivity measure; the results are presented in Figure 8(a)–(c). Since the Kullback–Leibler divergence is not symmetric, we compute it in both directions. We also compute the posterior mean for $\epsilon = 0.5$ based on the

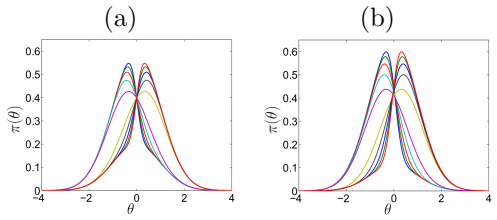

FIGURE 7. Normal prior contaminated using the skew normal contamination class under the (a) linear and (b) geometric frameworks. The tails are more separated under the latter.

same set of contaminated models; when $\alpha = 0$, the contaminated posterior is the same as the baseline posterior. This procedure is repeated on three simulated datasets corresponding to the three rows in Figure 8.

Based on Figure 8, the Kullback–Leibler divergence is asymmetric discouraging its use as a global robustness measure. In all cases, the Fisher–Rao distance suggests that the posterior is fairly robust to geometric contamination of the Gaussian prior using skew normal distributions, especially if one takes ϵ to be small, i.e. $\epsilon < 0.5$. The distance becomes relatively high only when ϵ approaches one and for large α. When ϵ equals one, the baseline Gaussian prior is replaced by the skew normal distribution. By considering the rate of change of the distance measures, it can also be seen that the Fisher–Rao distance is more sensitive to departures from $N(0, 1)$ than the Kullback–Leibler divergence; the Kullback–Leibler divergence appears to pick up departures only for ϵ exceeding 0.75.

We also notice an interesting result from panel (d). When the baseline posterior mean is close to zero, the geometric and linear contamination methods result in similar values of the contaminated posterior mean; the difference is less than 0.02. When the baseline posterior mean is greater than zero, the linear and geometric contamination classes yield very similar contaminated posterior means in the positive α direction. In the negative α direction, geometric contamination yields posterior means that depart more severely from the baseline than under linear contamination; the opposite result is observed in the third row of panel (d). We posit that this phenomenon is due to the nonlinear structure of the geometric contamination class and is consistent with intuition.

7. Additional Example for Local Sensitivity Analysis

Consider the following data generating model for independent and identically distributed observations:

$$x_i \mid \theta \sim f = N(\theta, 1), \quad \theta \sim \pi_0 = N(0, 1) \quad (i = 1, \ldots, 50).$$

We simulate 50 observations from this model to use as the given data. We consider a family of t prior contaminations, parameterized by the degrees of freedom, $df = 3, \ldots, 100$. We compute the three different local sensitivity measures defined in Section 3.3 in the paper; the plots of these sensitivity measures are provided in Figure 9. For the Bayes factor, we use $\pi_1 = N(0, 5)$. This simulation example allows for an easy interpretation of the effectiveness of the proposed method. Suppose the true model is the baseline model. As we perturb away from the baseline model, the Bayes factor should decrease, indicated by a negative sign in its local sensitivity measure for all degrees of freedom of the contaminating t distribution. Furthermore, as the degrees of freedom increase, this local measure should tend to zero, because the contaminating densities look more and more like the baseline prior. The same trend should hold for the second order local sensitivity measure based on the geodesic distance. This is easy to see because as one increases the degrees of freedom of the t, the perturbed posterior approaches the baseline posterior, collapsing the geodesic

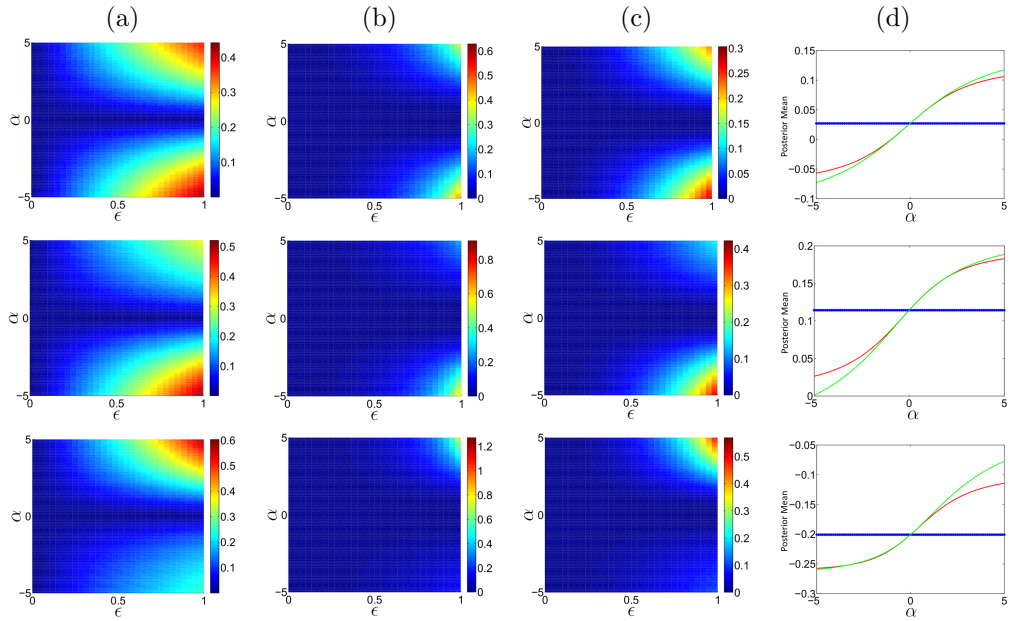

FIGURE 8. Assessment of Bayesian prior robustness to ϵ-contamination of a Gaussian prior with a skew normal distribution. (a) Image of Fisher–Rao distances between baseline and geometrically contaminated posteriors for different values of ϵ (x-axis) and α (y-axis) for 3 simulated datasets. (b) Image of Kullback–Leibler divergences (expectation computed with respect to p_0) between baseline and linearly contaminated posteriors. (c) Same as (b) but the expectation was computed with respect to p. (d) Posterior means for varying values of α and $\epsilon = 0.5$ where baseline=blue, geometric contamination=green, linear contamination=red.

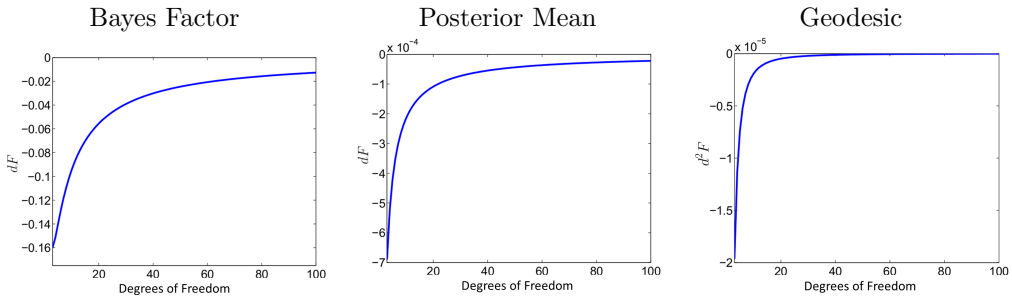

FIGURE 9. Local influence analysis based on the Bayes factor, posterior mean and geodesic distance. Here, we consider perturbing the baseline standard normal prior with a t with increasing degrees of freedom (x-axis, $df = 3, \ldots, 100$).

to a single point. The local sensitivity of the posterior mean is harder to interpret in this case because the trend is dependent on the simulated data. In general, the sign of this measure should be opposite from that of the sample mean of the simulated data, and the measure should tend to zero for increasing degrees of freedom of the t. Overall, we expect the local sensitivity of the posterior mean to be small since both the standard normal and the t have mean zero.

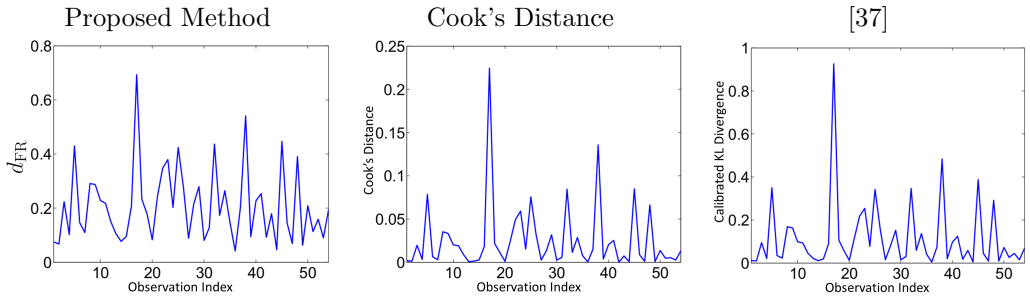

FIGURE 10. Influence analysis in Bayesian multiple linear regression. The 54 observations are listed on the x-axis with the corresponding measure of influence on the y-axis.

8. Additional Examples for Identifying Influential Observations

8.1. **Bayesian Linear Regression.** The data analyzed here comes from [30] containing 54 test cases. The response y is the natural logarithm of survival time. There are eight predictors: blood-clotting score, prognostic index, enzyme test, liver test, age, gender, moderate alcohol use, and heavy alcohol use. Due to large differences in predictor scales, we standardize the response and predictor variables. We use X to denote the standard design matrix and θ to denote the nine-dimensional vector of unknown regression coefficients. We use the following baseline Bayesian model:

$$y \mid \theta, X \sim f = N(X^T\theta, \sigma^2 I_{54}), \quad \theta \sim \pi = N(0, 1000 I_9).$$

For simplicity, instead of placing a prior on σ, we estimate it from the given data. Because we have chosen a conjugate prior for θ, the posterior density is also a Gaussian distribution; if one deletes a case from the data, the resulting posterior distribution is again Gaussian. Under this setup, we are faced with computing the Fisher–Rao distance between two multivariate Gaussian posteriors. This requires the computation of a high-dimensional integral and we will approximate it using importance sampling. Since it is easy to sample from the baseline posterior we can use it as a natural importance sampling density. We rewrite the Fisher–Rao inner product between the baseline posterior and the posterior under case-deletion as

$$\langle p_0^{1/2}, p_k^{1/2} \rangle = \int_\Theta p_0^{1/2}(\theta \mid y, X) p_k^{1/2}(\theta \mid y, X) d\theta = \int_\Theta \left\{ \frac{p_k(\theta \mid y, X)}{p_0(\theta \mid y, X)} \right\}^{1/2} p_0(\theta \mid y, X) d\theta,$$

where p_k is the posterior after deletion of observation k. Our approach is to generate a large sample, $\{\theta_1, \ldots, \theta_N\}$, from the baseline posterior and then estimate the Fisher–Rao distance using Monte Carlo integration:

$$\hat{I}(k) = \hat{d}_{\mathrm{FR}}(p_0, p_k) = \cos^{-1}\left[\frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{p_k(\theta_i | y, X)}{p_0(\theta_i | y, X)} \right\}^{1/2} \right].$$

In this example, we set $N = 100000$. The given approximation is possible because we can easily evaluate the posterior density of each θ under p_k and p_0; in other cases, we would be forced to resort to the estimate given in Proposition 2 in the main article.

In the left panel of Figure 10, we display Fisher–Rao distances between the baseline posterior and the posterior under deletion of each case; the middle panel displays the standard Cook's distance in a frequentist setting. Finally, in the right panel we have computed the influence measure proposed by [37] based on the Kullback–Leibler divergence. Based on the F statistic, Cook's distance does not flag any of the observations as influential, even though visually, observation 17

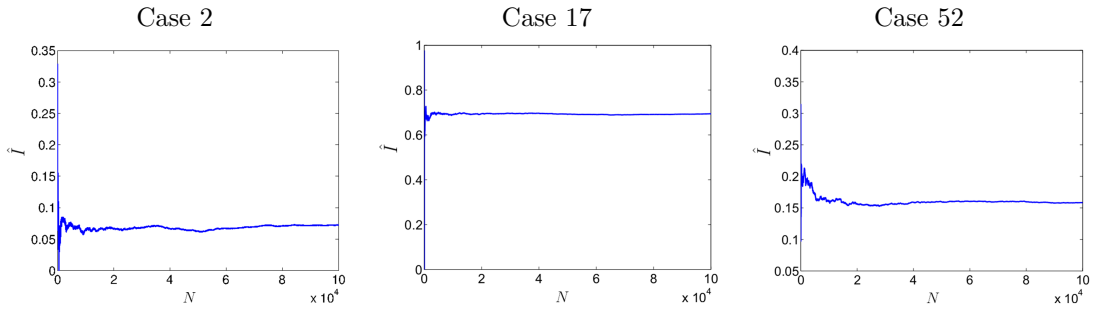

FIGURE 11. Convergence plots for the Monte Carlo estimate of the Fisher–Rao distance for cases 2, 17 and 52.

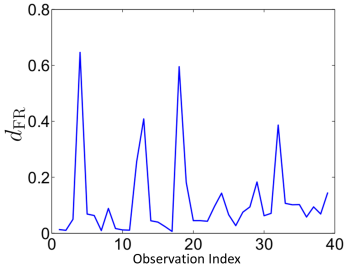

FIGURE 12. Influence analysis in Bayesian logistic regression. Each of the 39 observations is listed on the x-axis with the corresponding measure of influence on the y-axis.

appears highly influential. [37] suggest flagging all observations, which yield a distance greater than 0.25 as influential; under their framework, one would consider seven observations as influential, with 17 being highly influential. A similar result is observed when using the proposed influence measure. Observation 17 is highly influential with a distance greater than 0.7, and there are eight other possibly influential observations with distances greater than 0.3.

To numerically assess the convergence of the Monte Carlo estimator of the Fisher–Rao distance we plot the estimate as a function of the number of samples from the baseline posterior for observations 2, 17 and 52. It is evident in Figure 11 that the estimator used in this example has good convergence properties and can be reliably used for detecting influential observations. For all of the presented cases, the estimate of the Fisher–Rao distance has converged with approximately 30000 samples from the baseline posterior. We further assess the quality of our estimator by generating 50 different samples of size 100000 from the baseline posterior and reporting the variance of the estimated Fisher–Rao distances; for all observations, the variance of the estimates was smaller than 1×10^{-5}.

8.2. **Bayesian Logistic Regression.** In this illustration, we identify influential observations under the Bayesian logistic regression setup. The dataset used here was previously analyzed by [17] and was studied by [37] in the influence analysis setting. There are 39 cases in this data, where the response y is a vector of binary outcomes indicating whether or not vasoconstriction occurred. The two predictor variables are the volume of air inspired and the rate of air inspiration. We use X to denote the standard design matrix and consider the logistic model for this data: $P(Y_i = 1) = \exp(X_i^T\theta)\{1 + \exp(X_i^T\theta)\}^{-1}$, where θ is the unknown vector of regression coefficients. We assume a multivariate normal prior on θ, $\pi = N(1, 1000I_3)$. Then, the baseline posterior distribution is

$$p_0(\theta|y,X) \propto \exp\left[-\frac{1}{2000}(\theta-1)^T(\theta-1) + \sum_{i=1}^{39}\{y_i X_i^T\theta - \log(1+e^{X_i^T\theta})\}\right].$$

Since θ is only three-dimensional we use numerical integration to compute the Fisher–Rao distance. Figure 12 presents the results of our analysis. The proposed influence measure indicates four influential observations: 4, 18, 13 and 32 in order of decreasing influence. Observations 4 and 18 appear to have the most severe effect on the posterior distribution with resulting distances close to 0.6, which is nearly half of the maximum Fisher–Rao distance. The remaining 35 observations yield influence measures lower than 0.2, which we consider as having low influence. We compare our result to that provided in [37]; we refer the reader to their paper for a figure similar to Figure 12. Their influence measure, based on a divergence, is asymmetric and possesses no natural scale. The authors suggest a calibration strategy but a choice of this calibration is rather arbitrary in general. Their method flags observations 4 and 18 as influential, in decreasing order of influence, and many other observations as weakly influential. Our approach provides a clearer separation of the influential versus the non-influential observations in this example.

9. Proofs

SRT representation: Let r be a small positive scalar and $\delta p \in T_p(\mathcal{P})$. We begin by computing the differential of the mapping ϕ, $\phi_* : T_p(\mathcal{P}) \to T_{\phi(p)}(\Psi)$:

$$\phi_*(\delta p) = \frac{d}{dr}\phi(p+r\delta p)\Big|_{r=0} = \frac{d}{dr}(p+r\delta p)^{1/2}\Big|_{r=0} = \frac{\delta p}{2(p+r\delta p)^{1/2}}\Big|_{r=0} = \frac{\delta p}{2p^{1/2}}.$$

Plugging this expression into the standard $\mathbb{L}^2$ metric, for two tangent vectors $\delta p_1,\ \delta p_2 \in T_p(\mathcal{P})$, we obtain:

$$\langle \phi_*(\delta p_1), \phi_*(\delta p_2)\rangle = \left\langle \frac{\delta p_1}{2p^{1/2}}, \frac{\delta p_2}{2p^{1/2}}\right\rangle = \frac{1}{4}\int_{\mathbb{R}} \delta p_1(x)\delta p_2(x)\frac{1}{p(x)}dx = \frac{1}{4}\langle\langle \delta p_1, \delta p_2\rangle\rangle_p,$$

where the metric on the right is the Fisher–Rao metric.

Proof of Proposition 1: We use the same notation as in the main text and set $e(\theta) = \exp_{\pi_0^{1/2}}(\epsilon v_g)(\theta)$, $de(\theta) = \frac{d}{d\epsilon}\exp_{\pi_0^{1/2}}(\epsilon v_g)(\theta)$, and $d^2e(\theta) = \frac{d^2}{d\epsilon^2}\exp_{\pi_0^{1/2}}(\epsilon v_g)(\theta)$ for convenience. We use the following results:

$$\begin{aligned}
\exp_{\pi_0^{1/2}}(\epsilon v_g)\Big|_{\epsilon=0} &= \cos(\epsilon\|v_g\|)\pi_0^{1/2} + \sin(\epsilon\|v_g\|)\frac{v_g}{\|v_g\|}\Big|_{\epsilon=0} = \pi_0^{1/2},\\
\frac{d}{d\epsilon}\exp_{\pi_0^{1/2}}(\epsilon v_g)\Big|_{\epsilon=0} &= -\sin(\epsilon\|v_g\|)\pi_0^{1/2}\|v_g\| + \cos(\epsilon\|v_g\|)v_g\Big|_{\epsilon=0} = v_g,\\
\frac{d^2}{d\epsilon^2}\exp_{\pi_0^{1/2}}(\epsilon v_g)\Big|_{\epsilon=0} &= -\cos(\epsilon\|v_g\|)\pi_0^{1/2}\|v_g\|^2 - \sin(\epsilon\|v_g\|)v_g\|v_g\|\Big|_{\epsilon=0} = -\pi_0^{1/2}\|v_g\|^2.
\end{aligned}$$

Proof of 1:

$$\frac{d}{d\epsilon}F_{\pi_0,\pi_1}(v_g)\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\frac{m(x\mid \epsilon g)}{m(x\mid \pi_1)}\Big|_{\epsilon=0} = 2\frac{\int_\Theta f(x\mid\theta)de(\theta)e(\theta)d\theta}{m(x\mid\pi_1)}\Big|_{\epsilon=0} = 2\frac{\tilde{m}(x\mid v_g)}{m(x\mid\pi_1)}.$$

Proof of 2:

$$\begin{aligned}
&\frac{d}{d\epsilon}F_{\pi_0}(v_g)\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\int_\Theta h(\theta)\frac{f(x\mid\theta)e(\theta)^2}{\int_\Theta f(x\mid\theta)e(\theta)^2d\theta}d\theta\Big|_{\epsilon=0}\\
&= 2\int_\Theta h(\theta)\frac{f(x\mid\theta)e(\theta)de(\theta)\int_\Theta f(x\mid\theta)e(\theta)^2d\theta - f(x\mid\theta)e(\theta)^2\int_\Theta f(x\mid\theta)e(\theta)de(\theta)d\theta}{\{\int_\Theta f(x\mid\theta)e(\theta)^2d\theta\}^2}d\theta\Big|_{\epsilon=0}\\
&= \frac{2}{m(x\mid\pi_0)}\int_\Theta h(\theta)f(x\mid\theta)\pi_0^{1/2}(\theta)v_g(\theta)d\theta - \frac{2\tilde{m}(x\mid v_g)}{m(x\mid\pi_0)}\int_\Theta h(\theta)p_0(\theta\mid x)d\theta.
\end{aligned}$$

Proof of 3: Since we are dealing with infinitesimal quantities, we make a simplification using the local Euclidean structure of Ψ, i.e. we approximate the arc-length (Fisher–Rao) distance using a chord-length (Hellinger) distance, which locally are essentially the same (see equation 2.9 in [25]). Therefore,

$$\begin{aligned}
&\frac{d}{d\epsilon}F_{\pi_0}(v_g)\Big|_{\epsilon=0} = \frac{d}{d\epsilon}\int_\Theta \left[p_0^{1/2}(\theta \mid x) - \left\{\frac{f(x \mid \theta)e(\theta)^2}{m(x \mid \epsilon g)}\right\}^{1/2}\right]^2 d\theta\Big|_{\epsilon=0}\\
&= \int_\Theta \left[p_0^{1/2}(\theta \mid x) - \left\{\frac{f(x \mid \theta)e(\theta)^2}{m(x \mid \epsilon g)}\right\}^{1/2}\right]\left\{\frac{m(x \mid \epsilon g)}{f(x \mid \theta)e(\theta)^2}\right\}^{1/2}\\
&\quad\left\{\frac{f(x \mid \theta)e(\theta)de(\theta)\int_\Theta f(x \mid \theta)e(\theta)^2 d\theta - f(x \mid \theta)e(\theta)^2\int_\Theta f(x \mid \theta)e(\theta)de(\theta)d\theta}{m(x \mid \epsilon g)^2}\right\}d\theta\Big|_{\epsilon=0}\\
&= \int_\Theta \left[p_0^{1/2}(\theta \mid x)\left\{\frac{m(x \mid \epsilon g)}{f(x \mid \theta)e(\theta)^2}\right\}^{1/2} - 1\right]T(\theta)d\theta\Big|_{\epsilon=0}\\
&= \int_\Theta \left[\left\{\frac{p_0(\theta \mid x)}{p_0(\theta \mid x)}\right\}^{1/2} - 1\right]\left\{\frac{f(x \mid \theta)\pi_0^{1/2}(\theta)v_g(\theta) - p_0(\theta \mid x)\tilde{m}(x \mid v_g)}{m(x \mid \pi_0)}\right\}d\theta = 0,
\end{aligned}$$

where $T(\theta) = \frac{f(x|\theta)e(\theta)de(\theta)\int_\Theta f(x|\theta)e(\theta)^2d\theta - f(x|\theta)e(\theta)^2\int_\Theta f(x|\theta)e(\theta)de(\theta)d\theta}{m(x|\epsilon g)^2}$. This result is expected since the distance is minimized at 0. Thus, we consider the second derivative with respect to ϵ:

$$\begin{aligned}
&\frac{d^2}{d\epsilon^2}F_{\pi_0}(v_g)\Big|_{\epsilon=0} = \frac{d^2}{d\epsilon^2}\int_\Theta \left[p_0^{1/2}(\theta \mid x) - \left\{\frac{f(x \mid \theta)e(\theta)^2}{m(x \mid \epsilon g)}\right\}^{1/2}\right]^2 d\theta\Big|_{\epsilon=0}\\
&= 4\frac{\tilde{m}(x \mid v_g)}{m(x \mid \pi_0)}\int_\Theta \frac{v_g(\theta)}{\pi_0^{1/2}(\theta)}p_0(\theta \mid x)d\theta - 2\int_\Theta \frac{v_g(\theta)^2}{\pi_0(\theta)}p_0(\theta \mid x)d\theta - \frac{2\tilde{m}(x \mid v_g)^2}{m(x \mid \pi_0)^2}.
\end{aligned}$$

Proof of Proposition 2: Under the case deletion setup, the prior on the parameters does not change. Thus, we have the following:

$$\begin{aligned}
L(k) &= \int_\Theta p_k^{1/2}(\theta \mid x)p_0^{1/2}(\theta \mid x)d\theta = \int_\Theta p_k^{1/2}(\theta \mid x)\frac{1}{p_0^{1/2}(\theta \mid x)}p_0(\theta \mid x)d\theta\\
&= \int_\Theta \left\{\frac{f_k(x \mid \theta)\pi(\theta)}{\int_\Theta f_k(x \mid \theta)\pi(\theta)d\theta}\right\}^{1/2}\left\{\frac{\int_\Theta f_0(x \mid \theta)\pi(\theta)d\theta}{f_0(x \mid \theta)\pi(\theta)}\right\}^{1/2} p_0(\theta \mid x)d\theta\\
&= \int_\Theta \left\{\frac{f_k(x \mid \theta)}{f_0(x \mid \theta)}\right\}^{1/2}\left\{\frac{\int_\Theta f(y \mid \theta)\pi(\theta)d\theta}{\int_\Theta f_k(y \mid \theta)\pi(\theta)d\theta}\right\}^{1/2} p(\theta \mid y)d\theta\\
&= \int_\Theta \left\{\frac{f_k(x \mid \theta)}{f_0(x \mid \theta)}\right\}^{1/2}\left\{\int_\Theta \frac{1}{f(x_k \mid x_{(k)}, \theta)}p_0(\theta \mid x)d\theta\right\}^{-1/2} p_0(\theta \mid x)d\theta.
\end{aligned}$$

Department of Statistics; 404 Cockins Hall; 1958 Neil Avenue; The Ohio State University; Columbus, OH 43210.

E-mail address: kurtek.1@stat.osu.edu

School of Mathematical Sciences; University Park; University of Nottingham; Nottingham NG72RD.

E-mail address: Karthik.Bharath@nottingham.ac.uk