Bayesian model robustness via disparities

Giles Hooker · Anand N. Vidyashankar

Received: 20 March 2013 / Accepted: 10 February 2014 / Published online: 1 March 2014
© Sociedad de Estadística e Investigación Operativa 2014

Abstract This paper develops a methodology for robust Bayesian inference through the use of disparities. Metrics such as Hellinger distance and negative exponential disparity have a long history in robust estimation in frequentist inference. We demonstrate that an equivalent robustification may be made in Bayesian inference by substituting an appropriately scaled disparity for the log likelihood to which standard Monte Carlo Markov Chain methods may be applied. A particularly appealing property of minimum-disparity methods is that while they yield robustness with a breakdown point of 1/2, the resulting parameter estimates are also efficient when the posited probabilistic model is correct. We demonstrate that a similar property holds for disparity-based Bayesian inference. We further show that in the Bayesian setting, it is also possible to extend these methods to robustify regression models, random effects distributions and other hierarchical models. These models require integrating out a random effect; this is achieved via MCMC but would otherwise be numerically challenging. The methods are demonstrated on real-world data.

Keywords Deviance test · Kernel density · Hellinger distance · Negative exponential disparity · MCMC · Bayesian inference · Posterior · Outliers

Mathematics Subject Classification 62F35

Electronic supplementary material The online version of this article (doi:10.1007/s11749-014-0360-z) contains supplementary material, which is available to authorized users.

G. Hooker (✉)
Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, USA
e-mail: gjh27@cornell.edu

A. N. Vidyashankar
Department of Statistics, George Mason University, Fairfax, USA
e-mail: avidyash@gmu.edu

Springer
1 Introduction

In this paper, we develop a new methodology for providing robust inference in a Bayesian context. When the data at hand are suspected of being contaminated with large outliers, it is standard practice to account for these (1) by postulating a heavy-tailed distribution, (2) by viewing the data as a mixture, with the contamination explicitly occurring as a mixture component or (3) by employing priors that penalize large values of a parameter (see Berger 1994; Albert 2009; Andrade and O’Hagan 2006). In the context of frequentist inference, these issues are investigated using methods such as M-estimation, R-estimation, etc. and are part of standard robustness literature (see Hampel et al. 1986; Maronna et al. 2006; Jureckova and Sen 1996). As is the case for Huberized loss functions in frequentist inference, even though these approaches provide robustness they lead to a loss of precision when contamination is not present in (1) and (2) above or to a distortion of prior knowledge in (3). Explicit modeling of outliers as in (2) also requires knowledge of outlier configurations—how many mixture components to use and what distributions to use them in, for example—and may not be robust if these are incorrect. This paper develops an alternative systematic Bayesian approach, based on disparity theory, that is shown to provide robust inference without loss of efficiency for large samples.

In parametric frequentist inference using independent and identically distributed (i.i.d.) data, several authors (Beran 1977; Tamura and Boos 1986; Simpson 1987, 1989; Cheng and Vidyashankar 2006) have demonstrated that the dual goal of efficiency and robustness is achievable using the minimum Hellinger distance estimator (MHDE). In the i.i.d. context, MHDE estimators are defined by minimizing the Hellinger distance between a postulated parametric density \( f_\theta(\cdot) \) and a non-parametric estimate \( g_n(\cdot) \) over the \( p \)-dimensional parameter space \( \Theta \); that is,

\[
\hat{\theta}_{HD} = \arg \inf_{\theta \in \Theta} \int \left( g_n^{1/2}(x) - f_\theta^{1/2}(x) \right)^2 \, dx.
\]

Typically, for continuous data, \( g_n(\cdot) \) is taken to be a kernel density estimate; if the probability model is supported on discrete values, the empirical distribution is used. More generally, Lindsay (1994) introduced the concept of a minimum disparity procedure, developing a class of divergence measures that have similar properties to minimum Hellinger distance estimates. These have been further developed in Basu et al. (1997) and Park and Basu (2004). Hooker (2013) has extended these methods to a regression framework.

A remarkable property of disparity-based estimates is that while they confer robustness, they are also first-order efficient. That is, they obtain the information bound when the postulated density \( f_\theta(\cdot) \) is correct. In this paper, we develop robust Bayesian inference using disparities. We show that appropriately scaled disparities approximate \( n \) times the negative log-likelihood near the true parameter values. We use this as a motivation to replace the log likelihood in Bayes rule with a disparity to create what we refer to as the “D-posterior”. We demonstrate that this technique is readily amenable to Markov Chain Monte Carlo (MCMC) estimation methods. Finally, we establish that the expectation of the D-posterior is asymptotically efficient and the resulting credi-
ble intervals provide asymptotically accurate coverage when the proposed parametric model is correct.

Disparity-based robustification in Bayesian inference can be naturally extended to a regression framework through the use of conditional density estimation as discussed in Hooker (2013). We pursue this extension to hierarchical models and replace various terms in the hierarchy with disparities. This creates a novel “plug-in procedure”—allowing the robustification of inference with respect to particular distributional assumptions in complex models. We develop this principle and demonstrate its utility on a number of examples. The use of a disparity within a Bayesian context imposes an additional computational burden through the estimation of a kernel density estimate and the need to run MCMC methods. Our analysis and simulations demonstrate that while the use of MCMC significantly increases computational costs, the additional cost of the use of disparities is on the order of a factor between 2 and 10, remaining implementable for many applications. These methods require marginalization of an exponentiated disparity with respect to the random effects distribution: a task that can be achieved through MCMC methods, but would otherwise be numerically challenging.

The use of divergence measures for outlier analysis in a Bayesian context has been considered in Dey and Birmiwal (1994) and Peng and Dey (1995). Most of this work is concerned with the use of divergence measures to study Bayesian robustness when the priors are contaminated and to diagnose the effect of outliers. These divergence measures are computed using MCMC techniques. More recently, Zhan and Hettmansperger (2007) and Szpiro et al. (2010) have developed analogues of R-estimates and Bayesian Sandwich estimators. These methods can be viewed to be extensions of robust frequentist methods to the Bayesian context. By contrast, our paper is based on explicitly replacing the likelihood with a disparity to provide a systematic approach to obtain inherently robust and efficient inference.

Within the context of Bayesian analysis, robustness has been studied with respect to the specification of both prior and data distributions. Robustness to outliers as studied in the frequentist literature is referred to as “outlier-rejection” in Bayesian analysis and is studied for example in Dawid (1973), O'Hagan (1979, 1990), Choy and Smith (1997) and Desgagnè and Angers (2007). Here, outlier rejection indicates that as some group of data are moved to infinity, the posterior reverts to the posterior without those observations. This corresponds to a breakdown point of 1: a rather extreme value for frequentist robustness. We also obtain this breakdown point, but additionally develop a notion of an asymptotic breakdown point in which we examine the worst-case displacement as sample-size increases. We are able to show that this notion effectively describes robustness and distinguishes Bayesian methods along with regularized versions of robust estimators from estimators that are trivially made robust by, for example, thresholding their estimates.

The remainder of the paper is structured as follows: we provide a formal definition of the disparities in Sect. 2. Disparity-based Bayesian inference is developed in Sect. 3. Robustness and efficiency of these estimates are demonstrated theoretically and through a simulation for i.i.d. data in Sects. 4 and 5. The methodology is extended to regression models in Sect. 6. The plug-in procedure is presented in Sect. 7 through an application to a one-way random-effects model. Section 8 is devoted to two real-world data sets where we apply these methods to generalized linear mixed models and
random-intercept models for longitudinal data. Proofs of technical results and details of simulation studies are relegated to Online Resource 1.

2 Disparities and their numerical approximations

In this section, we describe a class of disparities and numerical procedures for evaluating them. These disparities compare a proposed parametric family of densities to a non-parametric density estimate. We assume that we have i.i.d. observations $X_i$ for $i = 1, \ldots, n$ from some density $h(\cdot)$. We let $g_n$ be the kernel density estimate:

$$g_n(x) = \frac{1}{nc_n} \sum_{i=1}^{n} K \left( \frac{x - X_i}{c_n} \right)$$  \hspace{1cm} (2)

where the kernel $K$ is the density and $c_n$ is a bandwidth for the kernel. If $c_n \to 0$ and $nc_n \to \infty$, it is known that $g_n(\cdot)$ is an $L_1$-consistent estimator of $h(\cdot)$ Devroye and Györfi (1985). In practice, a number of plug-in bandwidth choices are available for $c_n$ (e.g., Silverman 1982; Sheather and Jones 1991; Engel et al. 1994). For non-i.i.d. data examined in Sects. 6 and 7, plug-in bandwidths can be calculated from the method of moment estimates. We have found our results to be insensitive to the choice of plug-in bandwidth selector.

We review the class of disparities described in Lindsay (1994), see also Basu et al. (2011). The definition of disparities involves the residual function,

$$\delta_{\theta, g}(x) = \frac{g(x) - f_\theta(x)}{f_\theta(x)},$$  \hspace{1cm} (3)

defined on the support of $f_\theta(x)$ and a function $G : [-1, \infty) \to \mathbb{R}$. $G(\cdot)$ is assumed to be strictly convex and thrice differentiable with $G(0) = 0$, $G'(0) = 0$ and $G''(0) = 1$. The disparity between $f_\theta$ and $g_n$ is defined to be

$$D(g_n, f_\theta) = \int_{\mathcal{R}} G(\delta_{\theta, g_n}(x)) f_\theta(x) dx.$$  \hspace{1cm} (4)

An estimate of $\theta$ obtained by minimizing (4) is called a minimum disparity estimator (MDE). Under differentiability assumptions, this is equivalent to solving the equation

$$\nabla_{\theta} D(g_n, f_\theta) = \int_{\mathcal{R}} A(\delta_\theta(x)) \nabla_{\theta} f_\theta(x) dx = 0,$$

where $A(\delta) = G(\delta) - (1 + \delta) G'(\delta)$ and $\nabla_{\theta}$ indicates the derivative with respect to $\theta$.

This framework contains Kullback–Leibler divergence as approximation to the likelihood:

$$KL(g_n, f_\theta) = -\int (\log f_\theta(x)) g_n(x) dx \approx -\frac{1}{n} \sum_{i=1}^{n} \log f_\theta(x_i)$$

\(\square\) Springer
for the choice $G(\delta) = (\delta + 1) \log(\delta + 1)$ up to a constant. The squared Hellinger disparity (HD) corresponds to the choice $G(x) = [(x + 1)^{1/2} - 1]^2 - 1$. While robust statistics is typically concerned with the impact of outliers, the alternate problem of inliers—defined as nominally-dense regions that lack empirical data and consequently small values of $\delta_{\theta,g_n}(x)$—can also cause instability. It has been illustrated in the literature that HD down weighs the effect of large values of $\delta_{\theta,g_n}(x)$ (outliers) relative to the likelihood but magnifies the effect of inliers. An alternative, the negative exponential disparity (NED), based on the choice $G(\delta) = e^{-\delta} - 1$ down weighs the effect of both outliers and inliers.

The integrals involved in (4) are not analytically tractable and the use of Monte Carlo integration to approximate the objective function has been suggested in Cheng and Vidyashankar (2006), in specific cases numerical quadrature methods can also be employed. The trade-offs between these two and their computational costs are discussed in detail in Online Appendix D.

3 The D-posterior and MCMC methods

We begin this section by a heuristic description of the second-order approximation of $KL(f_\theta, g_n)$ by $D(f_\theta, g_n)$. A Taylor expansion of $KL(f_\theta, g_n)$ about $\theta$ has the following first two terms:

$$
\nabla_\theta^2 KL(g_n, f_\theta) = \int \left[ \frac{1}{f_\theta(x)} \left( \nabla_\theta f_\theta(x) \right) \left( \nabla_\theta f_\theta(x) \right)^T - \nabla_\theta^2 f_\theta(x) \right] (\delta_{\theta,g_n}(x) + 1) dx.
$$

where the second term approximates the observed Fisher Information when the bandwidth is small. The equivalent terms for $D(g_n, f_\theta)$ are

$$
\nabla_\theta^2 D(g_n, f_\theta) = \int \left[ \left( \nabla_\theta f_\theta(x) \right) \left( \frac{\nabla_\theta f_\theta(x)}{f_\theta(x)} \right)^T - \nabla_\theta^2 f_\theta(x) \right] g_n(x) dx
$$

where the second term approximates the observed Fisher Information when the bandwidth is small. The equivalent terms for $D(g_n, f_\theta)$ are

$$
\nabla_\theta^2 D(g_n, f_\theta) = \int \nabla_\theta^2 f_\theta(x) A(\delta_{\theta,g_n}(x)) dx
$$

Now, if $g_n$ is consistent, $\delta_{\theta,g_n}(x) \to 0$ almost surely (a.s.). Observing that $A(0) = 0$, $A'(0) = -1$ from the conditions on $G$ and observing $\int \nabla_\theta^2 f_\theta(x) dx = 0$, we obtain the asymptotic equality of (5) and (6). The fact that these heuristics yield efficiency was first noticed by Beran (1977) (Eq. 1.1).

In the context of Bayesian methods, inference is based on the posterior

$$
P(\theta|x) = \frac{P(x|\theta)\pi(\theta)}{\int P(x|\theta)\pi(\theta)d\theta}.
$$
Bayesian model robustness via disparities

Fig. 1  Left a comparison of log posteriors for $\mu$ with data generated from $N(\mu, 1)$ with $\mu = 1$ using an $N(0, 1)$ prior for $\mu$. Middle influence of an outlier on expected D-a posteriori (EDAP) estimates of $\mu$ as the value of the outlier is changed from 0 to 20. Right influence of the prior as the prior mean is changed from 0 to −10

where $P(x|\theta) = \exp(\sum_{i=1}^{n} \log f_\theta(x_i))$ and $\pi$ a prior density which we assume has a first moment and which we hold fixed throughout our analysis. Following the heuristics above, in this paper, we propose the simple expedient of replacing the log likelihood, log $P(x|\theta)$, in (7) with a disparity:

$$P_D(\theta|g_n) = \frac{e^{-nD(g_n, f_\theta)}\pi(\theta)}{\int e^{-nD(g_n, f_\theta)}\pi(\theta)d\theta}.$$  \hspace{1cm} (8)

In the case of Hellinger distance, the appropriate disparity is $2HD^2(g_n, f_\theta)$ and we refer to the resulting quantity as the H-posterior. When $D(g_n, f_\theta)$ is based on NED, we refer to it as N-posterior, and D-posterior more generally. These choices are illustrated in Fig. 1 where we show the approximation of the log likelihood by Hellinger and negative exponential disparities and the effect of adding an outlier to these in a simple normal-mean example.

Throughout the examples below, we employ a Metropolis algorithm based on a symmetric random walk to draw samples from $P_D(\theta|g_n)$. While the cost of evaluating $D(g_n, f_\theta)$ is greater than the cost of evaluating the likelihood at each Metropolis step, we have found these algorithms to be computationally feasible and numerically stable. Furthermore, the burn-in periods for sampling from $P_D(\theta|g_n)$ and the posterior are approximately the same, although the acceptance rate of the former is around ten percent higher.

After substituting $-nD(g_n, f_\theta)$ for the log likelihood, it will be useful to define summary statistics of the D-posterior to demonstrate their asymptotic properties. Since the D-posterior (8) is a proper probability distribution, the expected D-a posteriori (EDAP) estimates exist and are given by

$$\theta^*_n = \int_{\Theta} \theta P_D(\theta|g_n)d\theta.$$

and credible intervals for $\theta$ can be based on the quantiles of $P_D(\theta|g_n)$. These quantities are calculated via Monte Carlo integration using the output from the Metropolis algorithm. We similarly define the maximum D-a posteriori (MDAP) estimates by

$$\theta^+_n = \arg\max_{\theta \in \Theta} P_D(\theta|g_n).$$
In Sects. 4 and 5, we describe the asymptotic properties of EDAP and MDAP estimators. In particular, we establish the posterior consistency, posterior asymptotic normality and efficiency of these estimators and their robustness properties. Differences between $P_D(\theta | g_n)$ and the posterior do exist and are described below:

1. The disparities $D(g_n, f_\theta)$ have strict upper bounds; in the case of Hellinger distance $0 \leq HD^2(g_n, f_\theta) \leq 2$, the upper bound for NED is $e$. This implies that the likelihood part of the D-posterior, $\exp(-nD(g_n, f_\theta))$, is bounded away from zero. Consequently, a proper prior $\pi(\theta)$ is required to normalize $P_D(\theta | g_n)$. A random $\theta$ from $\pi(\theta)$ must also have finite expectation in order for the EDAP to be defined. In particular, uniform priors on unbounded ranges, along with most reference priors, cannot be employed here. Further, the tails of $P_D(\theta | g_n)$ are proportional to that of $\pi(\theta)$. As a consequence, the breakdown point for the EDAP, as traditionally defined, is 1. Although note that in Sect. 4, we propose an modified definition of breakdown which is appropriate for regularized and Bayesian estimators under which EDAP has a breakdown of 1/2.

The tail weakness of disparities can be observed in Fig. 1 where in the left panel we see that the tails diverge towards those of the log prior for disparities and in the right panel, the prior begins to dominate when the data are more than 7 prior standard deviations from the prior mean. However, these effects all occur at the extreme tails of the prior and a simple diagnostic that “you get the prior distribution back” is a good indicator that this is indeed the case.

These results do not affect the asymptotic behavior of $P_D(\theta | g_n)$ since the lower bounds decrease with $n$. Modified D-posteriors based on a transformation $m(D(g_n, f_\theta))$ that removes the upper bound can be defined without affecting either the efficiency or robustness of the resulting EDAP estimates. However, appropriate transformations $m$ will depend on the parametric family $f_\theta$ and are beyond the scope of this paper. Sequences of priors such as $g$-priors for regression models can also be employed here. We speculate that so long as the prior variance does not grow faster than $n$, the method will retain good asymptotic properties, but this analysis is beyond the scope of this paper.

2. In Bayesian inference for i.i.d. random variables, the log likelihood is a sum of $n$ terms. This implies that if new data $X_{n+1}, \ldots, X_{n^*}$ are obtained, the posterior for the combined data $X_1, \ldots, X_{n^*}$ can be obtained using posterior after $n$ observations, $P(\theta | X_1, \ldots, X_n)$ as a prior $\theta$:

$$P(\theta | X_1, \ldots, X_{n^*}) \propto P(X_{n+1}, \ldots, X_{n^*} | \theta) P(\theta | X_1, \ldots, X_n).$$

By contrast, $D(g_n, f_\theta)$ is generally not additive in $g_n$; hence $P_D(\theta | g_n)$ cannot be factored as above. Extending arguments in Park and Basu (2004), we conjecture that no disparity that is additive in $g_n$ will yield both robust and efficient posteriors.

3. While we have found that the same Metropolis algorithms can be effectively used for the D-posterior as would be used for the posterior, it is not possible to use conjugate priors with disparities. This removes the possibility of using conjugacy to provide efficient sampling methods within a Gibbs sampler, although these could be approximated by combining sampling from a conditional distribution with a rejection step.
4. It is evident that the choice of $g_n$ can be important. In particular $c_n$ needs to be chosen so that $\delta_{\theta, 0} g_n \to 0$ to obtain efficiency. If $c_n$ is too large $\delta_{\theta, 0} g_n$ will not be stable at finite sample sizes hence compromising efficiency. On the other hand, when $c_n$ is large, $g_n$ tends to smooth over the distinction between the main part of the data and outliers, reducing the robustness of parameter estimates. In practice, we have found that all plug-in estimators we have tried have achieved this trade-off without noticeable differences between them.

The idea of replacing log likelihood in the posterior with an alternative criterion occurs in other settings and these have been grouped under the name of “substitution likelihood methods”. See Sollich (2002), for example, in developing Bayesian methods for support vector machines, Jiang and Tanner (2008) in a more general classification setting, Dunson and Taylor (2005) employ similar ideas in quantile regression and Hoff (2007) considers the use of ranks in Bayesian copula estimation. However, we replace the log likelihood with an approximation that is explicitly designed to be both robust and efficient, rather than as a convenient sampling tool for a non-probabilistic model.

4 Robustness

The appeal of disparity-based methods is that in addition to the statistical efficiency of the estimators defined above when the parametric model is correctly specified, these estimators are also robust to contamination by data taking large values. EDAP estimators behave similarly to their minimum-disparity counterparts at finite levels of contamination at large but finite values. However, classical measures of robustness— influence functions and breakdown points—are based at limiting values, either of infinitesimal contamination levels or contaminating values at infinity where the convergence between EDAP and minimum-disparity estimators fails. This is due to a lack of uniformity and we argue that the direct application of these robustness measures does not provide an accurate description of the behavior of EDAP estimates. Instead, robustness should be measured by the properties of the pointwise limit of $\alpha$-level influence functions.

As noted in the introduction, robustness to outliers is treated under title of “outlier rejection” in Bayesian analysis and generally corresponds to a breakdown point of 1. As we show below, the analysis of robustness we propose also reconciles Bayesian and other regularized estimates with the traditional description of a robust estimator as having breakdown point of 1/2. A Bayesian analysis of outlier rejection for our methods can be undertaken using the analysis techniques developed here; it is omitted for the sake of brevity.

To describe robustness, we view our estimates as functionals $T_n(h)$ mapping the space of densities to $\mathbb{R}^p$. In particular, we examine the EDAP estimate

$$T_n(h) = \frac{\int \theta e^{-n D(h, f_\theta)} \pi(\theta) d\theta}{\int e^{-n D(h, f_\theta)} \pi(\theta) d\theta}$$

and note that in contrast to classical approaches to analyzing robustness the interaction between the disparity and the prior requires us to make the dependence of $T_n$ on
n explicit. This dependence is shared by any estimator that incorporates priors—
including all classical Bayesian methods—and affects the traditional measures of
robustness as examined below. Note that here, \( T_n \) is taken to be a deterministic sequence
of maps from the space of densities to \( \Theta \).

We analyze the behavior of \( T_n(h) \) under the sequence of perturbations \( h_{z,\alpha}(x) = (1-\alpha)g(x) + \alpha t_z(x) \) for a sequence of densities \( t_z(\cdot) \) and \( 0 \leq \alpha \leq 1 \). Here, we assume that \( t_z(\cdot) \) is a contaminating sequence defined so that it becomes orthogonal to both \( g \) and the parametric family for large \( z \). Note that unlike our examination of efficiency
below, in these analyzes we do not require that \( g \) belongs to the parametric family;
thus \( h_{\alpha,z} \) describes the effect of adding outliers to a fixed kernel density estimate. We
also assume that \( f_{\theta} \) and \( g \) become orthogonal at large values of \( \theta \).

\[
\lim_{z \to \infty} \int t_z(x) g(x) dx = 0 \tag{10}
\]

\[
\lim_{z \to \infty} \int t_z(x) f_{\theta}(x) dx = 0, \quad \forall \theta \in \Theta \tag{11}
\]

\[
\lim_{\theta^* \to \infty} \sup_{\|\theta\| > \theta^*} \int g(x) f_{\theta}(x) dx = 0. \tag{12}
\]

Typically, \( t_z(\cdot) \) is taken to be a uniform distribution on a small neighborhood centered
at \( z \); but these conditions are clearly more general. They extend those given in Park
and Basu (2004) in not requiring \( g \) to be a member of the parametric family \( f_{\theta} \). The
\( \alpha \)-level influence function is then defined analogously to Beran (1977) by

\[
\text{IF}_{\alpha,n}(z) = \alpha^{-1} \left[ T_n(h_{z,\alpha}) - T_n(h) \right] \tag{13}
\]

where we again note that the dependence of \( \text{IF}_{\alpha,n}(z) \) on \( n \) is induced by the prior.

Equation (13) represents a complete description of the behavior of our estimator in
the presence of contamination, up to the shape of the contaminating density. While for
EDAP estimators, it contains an explicit dependence on \( n \), we begin by observing its
limit for large \( n \). Firstly, as with more classical Bayesian estimates, EDAP estimators
approach their frequentist counterparts at a \( n^{-1} \) rate.

**Theorem 1** Assume that \( G \) has four continuous derivatives, that \( f_{\theta} \) is four times
continuously differentiable in \( \theta \) and that the third derivatives of \( \pi \) are bounded. Define
\( T_n(h) \) as in (9) and the MDE as

\[
\hat{\theta}(h) = \arg\min_{\theta \in \Theta} D(h, f_{\theta}) \tag{14}
\]

then \( T_n(h) - \hat{\theta}(h) = o_p(n^{-1}) \).

As an immediate corollary, the \( \alpha \)-level influence functions converge at the same rate

\( \square \) Springer
**Corollary 1** Under the conditions of Theorem 1, define

\[
IF_{\alpha, \infty}(z) = \alpha^{-1} \left[ \hat{\theta}(h_{z, \alpha}) - \hat{\theta}(h) \right]
\]

then for every \( \alpha \) and \( z \),

\[
IF_{\alpha, n}(z) - IF_{\alpha, \infty}(z) = o_p(n^{-1}).
\] (15)

That is, the influence function for \( \hat{\theta}(\cdot) \) represents a reasonable description of the behavior of \( T_n(\cdot) \). In particular, in the case of Hellinger distance methods, Theorems 5 and 6 in Beran (1977) have direct analogues for EDAP and MDAP estimators, respectively.

While Corollary 1 motivates using the properties of \( IF_{\alpha, \infty}(z) \) to describe the robustness of \( T_n(\cdot) \), we note that the convergence in (15) need not be uniform in \( z \). The classical summaries of robustness properties investigated below are focussed on extremal values of \( IF_{\alpha, n}(z) \); the breakdown point at large \( z \) and the classical influence function at small \( \alpha \). The lack of uniformity in (15) means that these summaries when applied at finite values of \( n \) need not reflect the asymptotic properties as described in \( IF_{\alpha, \infty}(\theta) \). We explore this discrepancy below and argue that the asymptotic influence measure is more appropriate in the sense of representing a minimax approach to robustness. The arguments employed here are broadly applicable to robust estimators that depend on \( n \); in particular, regularized versions of robust estimators are susceptible to the same discrepancies and our analysis provides a framework for describing robustness in this context as well.

4.1 Breakdown point

We begin by motivating a version of the breakdown point for \( n \)-dependent estimators. Classically, the breakdown point is defined to be

\[
B(T_n) = \sup \left\{ \alpha : \sup_z |IF_{\alpha, n}(z)| < \infty \right\},
\] (16)

(see Huber 1981). Beran (1977) and Park and Basu (2004) demonstrated that the MDE has a breakdown point of \( 1/2 \) in the case of Hellinger distance and when \( G(\cdot) \) and \( G'(\cdot) \) are bounded, respectively. In contrast, we show in Theorem 3 and Corollary 2 below that for each fixed \( n \), \( B(T_n) = 1 \). The distinction between these cases motivates an alternative measure that captures the intuition of the classical breakdown point when applied to a sequence of estimators that changes over \( n \). We define the asymptotic breakdown point of the sequence \( \{T_n\}_{n=1}^\infty \), as follows

\[
B^* (\{T_k\}) = \sup \left\{ \alpha : \lim_{n} \sup_z |IF_{\alpha, n}(z)| < \infty \right\}.
\] (17)

That is, for each \( n \) we consider the maximal displacement under \( \alpha \)-level contamination and declare a breakdown if the limit of these displacements is unbounded.
It is easy to see that MDAP estimators have asymptotic breakdown 1/2 if their MDE counterparts do and \( \log \pi(\theta) \) is convex. Writing the MDAP estimator as

\[
\tilde{T}_n(h) = \arg\max_{\theta \in \Theta} (nD(h, f_\theta) - \log \pi(\theta))
\]  

(18)

it is readily seen that if \( \theta^* \) maximizes \( \pi(\theta) \) then \( |\tilde{T}_n(h) - \theta^*| < |\hat{\theta}(h) - \theta^*| \) for \( \hat{\theta}(h) \) sufficiently large and thus if \( IF_{\alpha,\infty}(z) \) is uniformly bounded, so is the influence function of \( \tilde{T}_n \). Nonetheless, the convergence of \( \tilde{T}_n(h) \) to \( \hat{\theta}(h) \) means that if \( \hat{\theta}(h_{\alpha,z_k}) \to \infty \) there is a sequence \( n_k \) so that \( T_{n_k}(h_{\alpha,z_k}) \to \infty \) as \( k \to \infty \).

For EDAP estimators, a uniform identifiability condition is required. We also impose boundedness on \( G \) and \( G' \), but note these conditions do not hold for Hellinger distance; however, a direct argument can be given and hence in Theorem 2 below and in other results we will sometimes state that “The results also hold for Hellinger Distance”.

**Theorem 2** Under the conditions of Theorem 1, additionally assume \( G(\cdot) \) and \( G'(\cdot) \) are bounded and let \( \int \|\theta\|_2 \pi(\theta) d\theta < \infty \), then under conditions (10–12) if,

\[
\inf_{z} \inf_{\theta \in \Theta} D(\alpha t_z, f_\theta) > \inf_{\theta \in \Theta} D((1-\alpha)g, f_\theta) + \delta,
\]  

(19)

for \( \alpha \leq 1/2 \), the asymptotic breakdown point of \( T_1, T_2, \ldots \) is equal to the breakdown point of \( \hat{\theta}(h) \). The result continues to hold when \( D \) is given by Hellinger distance.

The identifiability condition imposed above ensures that \( t_z \) does not more closely resemble the family \( f_\theta \) than \( g \). In the analysis of Park and Basu (2004), \( g = f_\theta \) is assumed, under these conditions \( B^*([T_n]) = 1/2 \). If we replace \( g \) by the estimate \( g_n \), it could happen that the inequality in (19) is reversed (ie for every \( z \) there is a \( f_{\hat{\theta}_z} \) closer to \( t_z \) than any member of \( f_\theta \) is to \( g_n \)). In this case, the breakdown point of \( \hat{\theta}(h) \) could be strictly less than 1/2.

These results are in contrast to the treatment of \( T_n \) for fixed \( n \). Here, we follow Beran (1977) in evaluating \( \lim_{z \to \infty} T_n(h_{\alpha,z}) \) for each \( z \) and we can extend the result to the entire posterior distribution, mimicking the outlier rejection phenomenon found in Bayesian inference.

**Theorem 3** Under the conditions of Theorem 2, for each \( \theta \)

\[
\lim_{z \to \infty} P_D(\theta|h_{\alpha,z}) = P_D(\theta|(1-\alpha)g)
\]

and

\[
\lim_{z \to \infty} \int |P_D(\theta|h_{\alpha,z}) - P_D(\theta|(1-\alpha)g)| d\theta \to 0
\]

This result also holds when \( D \) is given by Hellinger distance.
As a direct consequence, the EDAP estimators also converge: \( \lim_{z \to \infty} T_n(h_{\alpha z}) = T_n((1 - \alpha)g) \). For MDE’s, taking \( g = f_{\theta_0} \) yields \( T_n((1 - \alpha)f_{\theta_0}) = \theta_0 \). For EDAP estimators, the \((1 - \alpha)\) factor generally results in a reduction in strength in the disparity relative to the prior. For Hellinger distance

\[
H_n(\alpha, f) = 4 - 4\sqrt{1 - \alpha} \int \sqrt{g(x)f_{\theta}(x)}dx
\]

since the second term is canceled in normalizing the D-posterior, this is equivalent to reducing \( n \) by a factor \( \sqrt{1 - \alpha} \).

We note here that while the above discussion examines the behavior of \( T_n(h_{\alpha z}) \) for small \( \alpha \), it can readily be extended to the following corollary

**Corollary 2** Let \( D(g, f_{\theta}) \) be bounded for all \( \theta \) and all densities \( g \) and let \( \int \|\theta\|_2\pi(\theta)d\theta < \infty \), then the breakdown point of the EDAP is 1.

A simple direct proof is given for this in Online Appendix A. We observe that \( D(0, f_{\theta}) = G(-1) \) is independent of \( \theta \), yielding \( T_n(0) = \int \theta \pi(\theta)d\theta \): the prior mean. The results at fixed \( n \) indicate an extreme form of robustness that results from the fact that the disparity approximation to the likelihood is weak in its tails. This produces a lack of equivariance in the resulting estimator that appears in the third term of the asymptotic expansion as shown in Eq. A.1.1 of the online appendix. The fixed \( n \) result does not distinguish our estimator from alternative estimators that are clearly problematic. In particular, the threshold estimator of the mean defined by

\[
m_n(f) = \begin{cases} \int xdF(x) \\ 1,000n \times \text{sign} \left( \int xdF(x) \right) \end{cases} \quad \begin{cases} < 1,000n \\ \text{otherwise} \end{cases}
\]

is also efficient and has breakdown point 1. However, by considering contamination with \( r_{zn}(x) \) taken to be uniform on \([1,000n - 1, 1,000n]\), it is readily seen that the asymptotic breakdown point is \( B^*([m_n]) = 0 \). We might also contemplate a mean estimate based on a penalized Huber loss:

\[
h_n(f) = \arg\min_{\mu} n \int H(x - \mu)dF(x) + \lambda_n \mu^2,
\]

where \( H \) is the Huber loss function (see Huber 1981) and \( \lambda_n \to 0 \). Since \( H(z) \) increases linearly for \( z \) sufficiently large, the breakdown point of \( h(f) \) is also 1, but \( B^*([h_n]) = 1/2 \). Of course in this case, \( h_n(f) \) will not be efficient.

While we have suggested that the distinction between the finite \( n \) and asymptotic breakdown point is a reflection more on the definition (16) than the properties of EDAP, it does leave considerable room for the design of unbounded disparities that are nonetheless robust and which would, therefore, also allow the use of improper priors while still obtaining proper D-a posteriori distributions.
4.2 Influence function

An alternative measure of robustness is given by the influence function Hampel (1974):

$$\text{IF}_{0,n}(z) = \lim_{\alpha \to 0} \text{IF}_{\alpha,n}(z)$$  (20)

That this function need not always provide a useful guide to the behavior of $T_n(h_{\alpha,z})$ was observed in Beran (1977) and further expanded in Lindsay (1994) who demonstrated that all MDEs that yield efficiency share the same influence function as the maximum likelihood estimator (MLE) whatever their behavior at gross levels of contamination. The analysis in Lindsay (1994) implicitly assumes an equivariant estimator so that $T_n(f_\theta) = \theta$ for any $\theta$. When the effect of a prior is included in the analysis, a different result is obtained at finite samples, but an equivalent limiting result can be derived.

To examine the influence function for EDAP estimators, we assume that the limit may be taken inside all integrals in (20) and obtain

$$\text{IF}_{0,n}(z) = nE_{PD(\theta|g)}[\theta C_z(\theta, g)] - n\left[ E_{PD(\theta|g)}\theta \right] \left[ E_{PD(\theta|g)}C_z(\theta, g) \right]$$

where $E_{PD(\theta|g)}$ indicates expectation with respect to the D-posterior with fixed density $g$ and

$$C_z(\theta, g) = \frac{d}{d\epsilon} \int G\left( \frac{h_{z,\epsilon}(x)}{f_\theta(x)} - 1 \right) f_\theta(x)dx \bigg|_{\epsilon=0}$$

$$= \int G'(\frac{g(x)}{f_\theta(x)} - 1) (t_z(x) - g(x))dx.$$

Here we observe that $\text{IF}_{0,n}(z)$ depends on the prior $\pi$. This is the case for any a posteriori estimate. $\text{IF}_{0,n}(z)$ also depends on the disparity employed as demonstrated in

**Theorem 4** Let $D(g, f_\theta)$ be bounded and assume that

$$e_0 = \sup_x \int \left| G'\left( \frac{g(x)}{f_\theta(x)} - 1 \right) \pi(\theta) \right| d\theta < \infty$$  (21)

and

$$e_1 = \sup_x \int \left| \theta G'\left( \frac{g(x)}{f_\theta(x)} - 1 \right) \pi(\theta) \right| d\theta < \infty$$

then $|IF(\theta; g, t_z)| < \infty$.

In the case of Hellinger distance, the conditions of Theorem 4 require the boundedness of $r(x) = \int (\sqrt{f_\theta(x)}/\sqrt{g(x)})\pi(\theta)d\theta$ which may not always hold (e.g., Beran 1977).
Despite this strong result, in an asymptotic sense the choice of disparity and prior is indistinguishable when \( g \) is assumed to lie within the model class. Expanding \( C_z(\theta, g) \) about \( \bar{\theta} = EPD(\theta|g)\theta \) provides

\[
IF_{0,n}(z) = nEPD(\theta|g)(\theta - \bar{\theta}) C_z'(\bar{\theta}) + \frac{n}{2}EPD(\theta|g)(\theta - \bar{\theta}) C_z''(\theta^*)
\]

\[
= ID(\theta_g) C_z'(\theta_g) + (\bar{\theta} - \theta_g) C_z'(\theta^*) + o(n^{-1/2})
\]

\[
= ID(\theta_g) C_z'(\theta_g) + o(n^{-1/2}),
\]

where \( \theta^* \) lies between \( \theta \) and \( \bar{\theta} \) and \( \theta^+ \) between \( \bar{\theta} \) and \( \theta_g \), since \( (\theta - \theta_g) \) and \( (\bar{\theta} - \theta_g) \) are \( O_p(n^{-1/2}) \) and \( o(n^{-1/2}) \), respectively. We now observe that when \( g = f_{\theta_0} \) is in the model class,

\[
C_z'(\theta_0) = \int \frac{\nabla_\theta f_{\theta_0}(x)}{f_{\theta_0}(x)} (t(x) - g(x)) dx
\]

is independent of both the disparity and the prior, as is \( ID(\theta_0) \) and the limiting value of \( IF_{0,n}(z) \) coincides with the influence function of the MLE and the results in Lindsay (1994).

Expanding on this, in an asymptotic sense, our results continue to mimic the second-order diagnostics in Lindsay (1994); the next leading term in the expansion above is \( C_z''(\theta) \), which coincides with the second-order approximation in (Lindsay 1994, Eq. 7). However, we note that at finite \( n \) this second-order term does have an effect, although whether it results in bounded influence functions depends on the parametric family, disparity, and the prior being employed.

5 Efficiency and numerical results

While there is a large literature on robust estimation methods, disparity-based estimation methods also achieve statistical efficiency when \( g \) is a member of the parametric family \( f_\theta \). In this section, we present theoretical results for i.i.d. data to demonstrate that inference based on the D-posterior is also asymptotically efficient. We also conduct a simulation study to demonstrate the finite-sample performance of these estimators.

5.1 Efficiency

We recall that under suitable regularity conditions, expected a posteriori estimators (EAPs) are strongly consistent, asymptotically normal and are statistically efficient (see Ghosh et al. 2006, Theorems 4.2–4.3). Our results in this section show that this property continues to hold for EDAP estimators under regularity conditions on \( G(\cdot) \) when the model \( \{f_\theta : \theta \in \Theta\} \) contains the true distribution. We define

\[
I^D(\theta) = \nabla^2_\theta D(g, f_\theta), \quad \text{and} \quad I^D_n(\theta) = \nabla^2_\theta D(g_n, f_\theta)
\]
as the disparity information and $\theta_g$ the parameter that minimizes $D(g, f_{\theta})$ (note that $\theta_g$ here depends on $g$). We note that if $g = f_{\theta_g}$, $I^D(\theta_g)$ is exactly equal to the Fisher information for $\theta_g$.

The proofs of our asymptotic results rely on the assumptions listed below. Among these are that MDEs are strongly consistent and efficient; this in turn relies on further assumptions, some of which make those listed below redundant. They are given here to maximize the mathematical clarity of our arguments. We assume that $X_1, \ldots, X_n$ are i.i.d. generated from some distribution $g(x)$ and that a parametric family, $f_{\theta}(x)$ has been proposed for $g(x)$ where $\theta$ has distribution $\pi$. To demonstrate efficiency, we assume

(A1) $g(x) = f_{\theta_g}(x)$; i.e., $g$ is a member of the parametric family.
(A2) $G$ has three continuous derivatives with $G'(0) = 0$, $G''(0) = 1$ and $|G'''(0)| \leq \infty$.
(A3) There exists $C > 0$ such that for all $g$ and $h$

$$\sup_{\theta \in \Theta} |D(g, f_{\theta}) - D(h, f_{\theta})| \leq C \int |g(x) - h(x)| \, dx.$$

(A4) $\nabla^2 D(g, f_{\theta})$ is positive definite and continuous in $\theta$ at $\theta_g$ and continuous in $g$ with respect to the $L_1$ metric.
(A5) For any $\delta > 0$, there exists $\epsilon > 0$ such that

$$\sup_{|\theta - \theta_g| > \delta} (D(g, f_{\theta}) - D(g, f_{\theta_g})) > \epsilon.$$

(A6) The MDE, $\hat{\theta}_n$, satisfies $\hat{\theta}_n \rightarrow \theta_g$ almost surely and $\sqrt{n}(\hat{\theta}_n - \theta_g) \overset{d}{\rightarrow} N(0, I^D(\theta_g)^{-1})$.

Our first result concerns the limit distribution for the posterior density of $\sqrt{n}(\theta - \hat{\theta}_n)$, which demonstrates that the D-posterior converges in $L_1$ to a Gaussian density centered on the MDE $\hat{\theta}_n$ with variance $\left[ nI^D(\hat{\theta}_n) \right]^{-1}$. This establishes that credible intervals based on either $P_D(\theta|x_1, \ldots, x_n)$ or from $N(\hat{\theta}_n, I^D_n(\hat{\theta}_n)^{-1})$ will be asymptotically accurate.

**Theorem 5** Let $\hat{\theta}_n$ be the MDE of $\theta_g$, $\pi(\theta)$ be any prior that is continuous and positive at $\theta_g$ with $\int_{\Theta} \|\theta\|_2 \pi(\theta) \, d\theta < \infty$ where $\| \cdot \|_2$ is the usual 2-norm, and $\pi^D_n(t)$ be the D-posterior density of $t = (t_1, \ldots, t_p) = \sqrt{n}(\theta - \hat{\theta}_n)$. Then, under conditions (A2)–(A6),

$$\lim_{n \rightarrow \infty} \int \pi^D_n(t) - \left( \frac{I^D(\theta_g)}{2\pi} \right)^{p/2} e^{-\frac{1}{2} t' I^D(\theta_g)^{-1} t} \, dt \overset{a.s.}{\rightarrow} 0. \quad (22)$$

Furthermore, (22) also holds with $I^D(\theta_g)$ replaced with $I^D_n(\hat{\theta}_n)$.

Our next theorem is concerned with the efficiency and asymptotic normality of EDAP estimates.
Theorem 6 Assume conditions (A2)–(A6) and \( \int \| \theta \|_2 \pi(\theta) d\theta < \infty \), then \( \sqrt{n} (\theta_n^* - \hat{\theta}_n) \xrightarrow{a.s.} 0 \) where \( \theta_n^* \) is the EDAP estimate. Further, \( \sqrt{n} (\theta_n^* - \theta_g) \xrightarrow{d} N(0, I^D(\theta_g)) \).

The proofs of these theorems are deferred to the online appendix B, but the following remarks concerning the assumptions (A1)–(A6) are in order:

1. Assumption A1 states that \( g \) is a member of the parametric family. When this does not hold, a central limit theorem can be derived for \( \hat{\theta}_n \) but the variance takes a sandwich-type form; see Beran (1977) in the case of Hellinger distance. For brevity, we have followed Basu et al. (1997) and Park and Basu (2004) in restricting to the parametric case.

2. Assumptions A2–A5 are required for the regularity and identifiability of the parametric family \( f_\theta \) in the disparity \( D \). Note that A3 holds for Hellinger distance and if \( G'(\cdot) \) is bounded from arguments in Park and Basu (2004); other disparities may require specialized demonstrations. Specific conditions for A6 to hold are given in various forms in Beran (1977), Basu et al. (1997), Park and Basu (2004) and Cheng and Vidyashankar (2006), see conditions in Online Appendix B.

3. The proofs of these results employ the same strategies as those for posterior asymptotic efficiency (see Ghosh et al. 2006, for example). However, here we rely on the second-order convergence of the disparity to the likelihood at appropriate rates and the consequent asymptotic efficiency of minimum-disparity estimators, which in turn is based on a careful analysis of non-parametric density estimates.

4. Since the structure of the proof only requires second-order properties and appropriate rates of convergence, we can replace \( D(g_n, f_\theta) \) for i.i.d. data with an appropriate disparity-based term for more complex models as long as A6 can be shown hold. In particular, the results in Hooker (2013) suggest that the disparity methods for regression problems detailed in Sect. 6 will also yield efficient estimates.

5.2 Simulation studies

To illustrate the small sample performance of D-posteriors, we undertook a simulation study for i.i.d. data from Gaussian distribution. 1,000 sample data sets of size 20 from a \( N(5, 1) \) population were generated. For each sample data set, a random walk Metropolis algorithm was run for 20,000 steps using a \( N(0, 0.5) \) proposal distribution and a \( N(0, 25) \) prior, placing the true mean one prior standard deviation above the prior mean. The kernel bandwidth was selected by the bandwidth selection in Sheather and Jones (1991). H- and N-posteriors were easily calculated by combining the KernSmooth (Wand and Ripley 2009) and LearnBayes (Albert 2008) packages in R. We also report an experiment in which the normal log likelihood is replaced in the posterior with Tukey’s biweight objective function using a cut-point of 4.685 as a comparison to alternative robust estimators. To compare computational cost, we have run an MCMC chain for the normal log likelihood and report these below, even though analytic posteriors are available.

Expected posteriori estimates for the sample mean were obtained along with 95% credible intervals from every second sample in the second half of the MCMC chain.
Table 1 A simulation study for a normal mean using the usual posterior, the Hellinger posterior and the negative exponential posterior

|                | Bias  | SD    | Coverage | Length | CPU time |
|----------------|-------|-------|----------|--------|----------|
| Posterior      | -0.015| 0.222 | 0.956    | 0.873  | 3.393    |
| Hellinger      | -0.015| 0.225 | 0.954    | 0.920  | 7.669    |
| Negative exp.  | -0.018| 0.229 | 0.973    | 1.022  | 7.731    |
| Biweight       | -0.017| 0.228 | 0.977    | 1.007  | 3.523    |

Columns give the bias and variance of the posterior mean, coverage and average CPU time of the central 95% credible interval based on 1,000 simulations. These are recorded for the posterior, Hellinger distance (HD), negative exponential disparity (NED) and Tukey’s biweight objective used in place of the log likelihood (Biweight).

Table 2 Results for contaminating the data sets used in Table 1 with outliers

| Loc | 1 outlier | 2 outliers | 5 outliers |
|-----|-----------|------------|------------|
|     | Bias      | SD         | Coverage   | Bias      | SD         | Coverage   |
|     |           |            |            | Bias      | SD         | Coverage   |
|     |           |            |            | Bias      | SD         | Coverage   |
| Posterior | -0.164 | 0.219 | 0.883 | -0.300 | 0.206 | 0.722 | -0.637 | 0.182 | 0.100 |
|         | -0.264 | 0.219 | 0.778 | -0.490 | 0.206 | 0.375 | -1.053 | 0.182 | 0.001 |
|         | -0.513 | 0.219 | 0.360 | -0.965 | 0.207 | 0.004 | -2.093 | 0.182 | 0.000 |
| HD    | -0.109 | 0.246 | 0.920 | -0.194 | 0.275 | 0.859 | -0.237 | 0.299 | 0.770 |
|         | -0.027 | 0.238 | 0.942 | -0.040 | 0.257 | 0.928 | -0.024 | 0.305 | 0.865 |
|         | -0.014 | 0.234 | 0.948 | -0.019 | 0.249 | 0.935 | 0.018  | 0.286 | 0.883 |
| NED   | -0.080 | 0.256 | 0.959 | -0.133 | 0.279 | 0.933 | -0.166 | 0.308 | 0.893 |
|         | -0.020 | 0.238 | 0.977 | -0.025 | 0.243 | 0.968 | -0.015 | 0.264 | 0.948 |
|         | -0.017 | 0.237 | 0.973 | -0.020 | 0.241 | 0.970 | -0.007 | 0.260 | 0.952 |
| Biweight | -0.091 | 0.246 | 0.954 | -0.175 | 0.252 | 0.915 | -0.443 | 0.275 | 0.645 |
|         | -0.018 | 0.237 | 0.974 | -0.019 | 0.236 | 0.972 | -0.022 | 0.243 | 0.967 |
|         | -0.017 | 0.236 | 0.977 | -0.018 | 0.234 | 0.971 | -0.018 | 0.236 | 0.969 |

1, 2, and 5 outliers (large columns) are added at locations −3, −5 and −10 (column Loc) for the posterior, Hellinger distance (HD), negative exponential disparity (NED) and Tukey’s biweight objective used in place of the log likelihood (Biweight).

Outlier contamination was investigated by reducing the last one, two or five elements in the data set by 3, 5 or 10. This choice was made so that both outliers and prior influence the EDAP in the same direction. The analytic posterior without the outliers is normal with mean 4.99 (equivalently, bias of −0.01) and standard deviation 0.223.

The results of this simulation are summarized in Tables 1 (uncontaminated data) and 2 (contaminated data). As can be expected, the standard Bayesian posterior suffers from sensitivity to large negative values, whereas the disparity-based methods remain nearly unchanged. Tukey’s biweight also ignored large outliers, but was more sensitive.
than the disparity methods to larger amounts of contamination. Near-outliers at the smaller value of $-3$ resulted in similar biases across all methods. We observe that all robust estimates have slightly larger standard deviations than the EAP corresponding to a loss of efficiency of 2% for the H-posterior and 5% for the N-posterior and Tukey estimates. We speculate that the increased variance from the N-posterior is due to its relatively heavy tails (the maximal value of NED is $e^{-1}$ compared to 4 for 2HD).

A comparison of CPU time indicates that the use of disparity methods required a little more than twice the computational effort as compared to using the likelihood within an MCMC method. Further details from this simulation including comparisons with Huber estimators are given in Online Appendix E.1.

The influence of the prior is investigated in the right-hand plot of Fig. 1 where we observe that the EAP and EDAP estimates are essentially identical until the prior is about 9 standard deviations from the mean of the data: at this point the prior dominates. However, we note that this picture will depend strongly on the prior chosen; a less informative prior will have a smaller range of dominance.

Because the normal distribution is symmetric, estimating its mean is relatively easy. We, therefore, also conducted a simulation to estimate both shape and scale parameters in an exponential-Gamma distribution (i.e., $\exp(X_i)$ has a Gamma distribution). The details and results of this simulation are reserved in Online Appendix E.1. We observed the expected behavior: EDAP estimates remained insensitive to outliers, whereas they significantly distorted the EAP. However, in this case, the H-posterior demonstrated larger variance than the N-posterior which we explain as being due to the tendency of nonparametric density estimates from exponential-Gamma data to become bimodal: producing inliers where a large value of the parametric density is compared to a relatively small value of the nonparametric estimate.

### 6 Disparities based on conditional density for regression models

The discussion above, along with most of the literature on disparity estimation, has focussed on i.i.d. data in which a kernel density estimate may be calculated. The restriction to i.i.d. contexts severely limits the applicability of disparity-based methods. We extend these methods to non-i.i.d. data settings via the use of conditional density estimates. This extension is studied in the frequentist context in the case of minimum-disparity estimates for parameters in non-linear regression in Hooker (2013).

Consider the classical regression framework in which data $(Y_1, X_1), \ldots, (Y_n, X_n)$ are collections of i.i.d. random variables where inference is made conditionally on $X_i$. For continuous $X_i$, a non-parametric estimate of the conditional density of $y|x$ is given by Hansen (2004) and Li and Racine (2007):

$$
 g_n^{(c)}(y|x) = \frac{1}{nc_1c_2} \sum_{i=1}^n K \left( \frac{y-Y_i}{c_1} \right) K \left( \frac{\|x-X_i\|}{c_2} \right)
$$

Under a parametric model $f_\theta(y|X_i)$ assumed for the conditional distribution of $Y_i$ given $X_i$, we define a disparity between $g_n^{(c)}$ and $f_\theta$ as follows:
\[ D^{(c)}(g^{(c)}, f_\theta) = \sum_{i=1}^{n} D\left( g^{(c)}_n(\cdot|X_i), f_\theta(\cdot|X_i) \right). \] (24)

As before, for Bayesian inference, we replace the log likelihood by negative of the conditional disparity (24); that is,

\[ e^{l(Y_i|X_i, \theta)} \pi(\theta) \approx e^{-D^{(c)}(g^{(c)}_n, f_\theta)} \pi(\theta). \]

In the case of simple linear regression, \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \theta = (\beta_0, \beta_1, \sigma^2) \) and \( f_\theta(\cdot|X_i) = \phi_{\beta_0+\beta_1 X_i, \sigma^2}(\cdot) \) where \( \phi_{\mu, \sigma^2} \) is Gaussian density with mean \( \mu \) and variance \( \sigma^2 \).

The use of a conditional formulation, involving a density estimate over a multidimensional space, produces an asymptotic bias in MDAP and EDAP estimates similar to that found in Tamura and Boos (1986), who also note that this bias is generally small. Online Appendix C proposes two alternative formulations that reduce the dimension of the density estimate and the bias. In a minimum-disparity context, Hooker (2013) demonstrates that a bootstrap method can be employed to reduce bias, although this is less feasible here.

When the \( X_i \) are discrete, (23) reduces to a distinct conditional density for each level of \( X_i \). For example, in a one-way ANOVA model \( Y_{ij} = X_i + \epsilon_{ij}, j = 1, \ldots, n_i, \) \( i = 1, \ldots, N, \) this reduces to

\[ g^{(c)}_n(y|X_i) = \frac{1}{n_i c_n} \sum_{i=1}^{n_i} K\left( \frac{y - Y_{ij}}{c_n} \right). \]

We note that in this case the bias noted above does not appear. However, when the \( n_i \) are small, or for high-dimensional covariate spaces the non-parametric estimate \( g_n(y|X_i) \) can become inaccurate. The marginal methods discussed in Online Appendix C can also be employed in this case.

Online Appendix E.2 gives details of a simulation study of this method as well as those described in Online Appendix C for a regression problem with a three-dimensional covariate. All disparity-based methods perform similarly to using the posterior with the exception of the conditional form in Sect. 6 when Hellinger distance is used which demonstrates a substantial increase in variance. We speculate that this is due to the sparsity of the data in high dimensions creating inliers; NED is less sensitive to this problem Basu et al. (1997).

### 7 Disparity metrics and the plug-in procedure

The disparity-based techniques developed above can be extended to hierarchical models. In particular, consider the following structure for an observed data vector \( Y \) along with an unobserved latent effect vector \( Z \) of length \( n \):

\[ P(Y, Z, \theta) = P_1(Y|Z, \theta)P_2(Z|\theta)P_3(\theta) \] (25)
where $P_1$, $P_2$ and $P_3$ are the conditional distributions of $Y$ given $Z$ and $\theta$ the distribution of $Z$ given $\theta$ and the prior distribution of $\theta$. Any term in this factorization that can be expressed as the product of densities of i.i.d. random variables can now be replaced by a suitably chosen disparity. This creates a plug-in procedure in which particular terms of a complete data log likelihood are replaced by disparities. For example, if the middle term is assumed to be a product:

$$P(Z|\theta) = \prod_{i=1}^{n} p(Z_i|\theta),$$

inference can be robustified for the distribution of the $Z_i$ by replacing (25) with

$$P_{D_1}(Y, Z, \theta) = P(Y|Z, \theta)e^{-2D(g_n(\cdot; Z), P_2(\cdot|\theta))} P_3(\theta)$$

where

$$g_n(z; Z) = \frac{1}{nc_n} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{c_n}\right).$$

In an MCMC scheme, the $Z_i$ will be imputed at each iteration and the estimate $g_n(\cdot; Z)$ will change accordingly. If the integral is evaluated using Monte Carlo samples from $g_n$, these will also need to be updated. The evaluation of $D(g_n(\cdot; Z), P_2(\cdot|\theta))$ creates additional computational overhead, but we have found this to remain feasible for moderate $n$. A similar substitution may also be made for the first term using the conditional approach suggested above.

To illustrate this principle in a concrete example, consider a one-way random-effects model:

$$Y_{ij} = Z_i + \epsilon_{ij}, \ i = 1, \ldots, n, \ j = 1, \ldots, n_i$$

under the assumptions

$$\epsilon_{ij} \sim N(0, \sigma^2), \ Z_i \sim N(\mu, \tau^2)$$

where the interest is in the value of $\mu$. Let $\pi(\mu, \sigma^2, \tau^2)$ be the prior for the parameters in the model; an MCMC scheme may be conducted with respect to the probability distribution

$$P(Y, Z, \mu, \sigma^2, \tau^2) = \prod_{i=1}^{n} \left(\prod_{j=1}^{n_i} \phi_{0, \sigma^2}(Y_{ij} - Z_i)\right) \prod_{i=1}^{n} \phi_{\mu, \tau^2}(Z_i) \pi(\mu, \sigma^2, \tau^2) \quad (26)$$

where $\phi_{\mu, \sigma^2}$ is the $N(\mu, \sigma^2)$ density. There are now two potential sources of distributional errors: either in individual observed $Y_{ij}$, or in the unobserved $Z_i$. Either (or both) possibilities can be dealt with via the plug-in procedure described above.
If there are concerns that the distributional assumptions on the $\epsilon_{ij}$ are not correct, we observe that the statistics $Y_{ij} - Z_i$ are assumed to be i.i.d. $N(0, \sigma^2)$. We may then form the conditional kernel density estimate:

$$g_n(c)(t|Z_i; Z) = \frac{1}{nc_n1} \sum_{j=1}^{n_i} K\left(\frac{t - (Y_{ij} - Z_i)}{c_{n1}}\right)$$

and replace (26) with

$$P_{D_2}(Y, Z, \mu, \sigma^2, \tau^2) = e^{-\sum_{i=1}^{n} n_i D(g_n^{(c)}(t|Z_i; Z), \phi_{0,\sigma^2}(\cdot))} \prod_{i=1}^{n} \phi_{\mu,\tau^2}(Z_i) \pi(\mu, \sigma^2, \tau^2).$$

(27)

On the other hand, if the distribution of the $Z_i$ is miss-specified, we form the estimate

$$g_n(z; Z) = \frac{1}{nc_n2} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{c_{n2}}\right)$$

and use

$$P_{D_1}(X, Y, \mu, \sigma^2, \tau^2) = \prod_{i=1}^{n} \left( \prod_{j=1}^{n_i} \phi_{0,\sigma^2}(Y_i - Z_i) \right) e^{-n D(g_n(\cdot; Z), \phi_{\mu,\tau^2}(\cdot))} \pi(\mu, \sigma^2, \tau^2)$$

(28)

as the D-posterior. For inference using this posterior, both $\mu$ and the $Z_i$ will be included as parameters in every iteration, necessitating the update of $g_n(\cdot; Z)$ or $g_n^{(c)}(\cdot|z; Z)$. Naturally, it is also possible to substitute a disparity in both places:

$$P_{D_2}(Z, Y, \mu, \sigma^2, \tau^2) = e^{-\sum_{i=1}^{n} n_i D(g_n^{(c)}(\cdot|Z_i; Z), \phi_{0,\sigma^2}(\cdot))} e^{-n D(g_n(\cdot; Z), \phi_{\mu,\tau^2}(\cdot))} \pi(\mu, \sigma^2, \tau^2).$$

(29)

A simulation study considering all these approaches with Hellinger distance chosen as the disparity is described in Online Appendix E.3. Our results indicate that all replacements with disparities perform well, although some additional bias is observed in the estimation of variance parameters which we speculate to be due to the interaction of the small sample size with the kernel bandwidth. Methods that replace the random effect likelihood with a disparity remain largely unaffected by the addition of an outlying random effect, while methods that do not estimate both the random effect mean and variance are substantially biased.

While a formal analysis of this method is beyond the scope of this paper we remark that the use of density estimates of latent variables requires significant theoretical development in both Bayesian and frequentist contexts. In particular, in the context...
Bayesian model robustness via disparities

of using $P_{D_1}$ appropriate inference on $\theta$ will require local agreement in the integrated likelihoods

$$\int \cdots \int \prod_{i=1}^{n} \left( \prod_{j=1}^{n_i} \phi_{0, \sigma^2}(Y_i - Z_i) \right) e^{-nD(g_0(\cdot; Z), \phi_{\mu, \tau^2}(\cdot))} dZ_1, \ldots, dZ_n$$

$$\approx \int \cdots \int \prod_{i=1}^{n} \left( \prod_{j=1}^{n_i} \phi_{0, \sigma^2}(Y_{ij} - Z_i) \right) \prod_{i=1}^{n} \phi_{\mu, \tau^2}(Z_i) dZ_1, \ldots, dZ_n.$$ 

This can be demonstrated if the $n_i \rightarrow \infty$ and hence the conditional variance of the $Z_i$ is made to shrink at an appropriate rate.

We note here that the Bayesian methods developed in this paper are particularly relevant in allowing the use of MCMC for these problems. A frequentist analysis could be obtained by marginalizing over the $Z_i$ in (27), (28), or (29). However, this marginalization is numerically challenging while it can be very readily obtained in a Bayesian context via MCMC methods.

8 Real data examples

8.1 Parasite data

We begin with a one-way random effect model for binomial data. These data come from one equine farm participating in a parasite control study in Denmark in 2008. Fecal counts of eggs of the Equine Strongyle parasites were taken pre- and post-treatment with the drug Pyrantol for each of 7 horses; the full study is presented in Nielsen et al. (2013). The data used in this example are reported in Online Appendix F.

For our purposes, we model the post-treatment data from each horse as binomial with probabilities drawn from a logit normal distribution. Specifically, we consider the following model:

$$k_i \sim \text{Bin}(N_i, p_i), \text{logit}(p_i) \sim N(\mu, \sigma^2), i = 1, \ldots, n,$$

where $N_i$ are the pre-treatment egg counts and $k_i$ are the post-treatment egg counts. We observe the data $(k_i, N_i)$ and desire an estimate of $\mu$ and $\sigma$. The likelihood for these data is

$$l(\mu, \sigma | k, N) = -\sum_{i=1}^{n} \left[ k_i \log p_i + (N_i - k_i) \log(1 - p_i) \right] - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\log(p_i) - \mu)^2.$$ 

We cannot use conditional disparity methods to account for outlying $k_i$ since we have only one observation per horse. However, we can consider robustifying the $p_i$
To perform a Bayesian analysis, $\mu$ was given a $N(0, 5)$ prior and $\sigma^2$ an inverse Gamma prior with shape parameter 3 and scale parameter 0.5. These were chosen as conjugates to the assumed Gaussian distribution and are diffuse enough to be relatively uninformative while providing reasonable density at the maximum likelihood estimates. A random walk Metropolis algorithm was run for this scheme with parameterization $(\mu, \log(\sigma), \logit(p_1), \ldots, \logit(p_n))$ for 200,000 steps with posterior samples collected every 100 steps in the second half of the chain. $c_n$ was chosen via the method in Sheather and Jones (1991) treating the empirical probabilities as data.

The resulting posterior distributions, given in Fig. 2, indicate a substantial difference between the two posteriors, with the N-posterior having higher mean and smaller variance. This suggests some outlier contamination and a plot of a sample of densities $g_n$ on the right of Fig. 2 suggests a lower-outlier with logit($p_i$) around $-4$. In fact, this corresponds to observation 5 which had unusually high efficacy in this horse. Removing the outlier results in good agreement between the posterior and the N-posterior. We note that, as also observed in Stigler (1973), trimming observations in this manner, unless done carefully, may not yield accurate credible intervals.

8.2 Class survey data

Our second data set is from an in-class survey in an introductory statistics course held at Cornell University in 2009. Students were asked their expected income at ages 35, 45, 55 and 65. Responses from 10 American-born and 10 foreign-born students in the class are used as data in this example; the data are presented and plotted in Online Appendix F. Our object is to examine the expected rate of increase in income and any
differences in this rate or in the overall salary level between American and foreign students. From Figure 4 in Online Appendix F, some potential outliers in both overall level of expected income and in specific deviations from income trend are evident. This framework leads to a longitudinal data model. We begin with a random intercept model

$$Y_{ijk} = b_{0ij} + b_{1j} t_k + \epsilon_{ijk}$$

(30)

where $Y_{ijk}$ is log income for the $i$th student in group $j$ [American ($a$) or foreign ($f$)] at age $t_k$. We extend to this the distributional assumptions

$$b_{0ij} \sim N(\beta_{0j}, \tau_0^2), \epsilon_{ijk} \sim N(0, \sigma^2)$$

leading to a complete data log likelihood given up to a constant by

$$l(Y, \beta, \sigma^2, \tau_0^2) = -\sum_{i=1}^{n} \sum_{j \in \{a, f\}} \sum_{k=1}^{4} \frac{1}{2\sigma^2} (Y_{ijk} - b_{0ij} - \beta_{1j} t_k)^2$$

$$-\sum_{i=1}^{n} \sum_{j \in \{a, f\}} \frac{1}{2\tau_0^2} (b_{0ij} - \beta_{0j})^2$$

(31)

to which we attach Gaussian priors centered at zero with standard deviations 150 and 0.5 for the $\beta_{0j}$ and $\beta_{1j}$, respectively, and Gamma priors with shape parameter 3 and scale 0.5 and 0.05 for $\tau_0^2$ and $\sigma^2$. These are chosen to correspond to the approximate orders of magnitude observed in the maximum likelihood estimates of the $b_{0ij}$, $\beta_{1j}$ and residuals.

As in Sect. 7, we can robustify this likelihood in two different ways: either against the distributional assumptions on the $\epsilon_{ijk}$ or on the $b_{0ij}$. In the latter case, we form the density estimate

$$g_n(b; \beta) = \frac{1}{-2nc_n} \sum_{i=1}^{n} \sum_{j \in \{a, f\}} K \left( \frac{b - b_{0ij} + \beta_{0j}}{c_n} \right)$$

and replace the second term in (31) with $-2nD(g_n(\cdot; \beta), \phi_{0, \tau_0^2}(\cdot))$. Here, we have used $\beta = (\beta_{0a}, \beta_{f0}, \beta_{0a1}, \beta_{f1}, b_{0a1}, b_{0f1}, \ldots, b_{0an}, b_{0fn})$ as an argument to $g_n$ to indicate its dependence on the estimated parameters. We have chosen to combine the $b_{0ai}$ and the $b_{0fi}$ together to obtain the best estimate of $g_n$, rather than using a sum of disparities, one for American and one for foreign students.

To robustify the residual distribution, we observe that we cannot replace the first term with a single disparity based on the density of the combined $\epsilon_{ijk}$ since the $b_{0ij}$ cannot be identified marginally. Instead, we estimate a density at each $ij$:

$$g_{ij,n}(e; \beta) = \frac{1}{4nc_n} \sum_{k=1}^{4} K \left( \frac{e - (Y_{ijk} - b_{0ij} - \beta_{1j} t_k)}{c_n} \right)$$
Fig. 3 Analysis of the class survey data using a random intercept model with Hellinger distance replacing the observation likelihood, the random effect likelihood or both. Top differences in intercepts between foreign and American students (left) and differences in slopes (right). Bottom random effect variance (left) and observation variance (right). Models robustifying the random effect distribution do not show a significant difference in the slope parameters. Those robustifying the observation distribution estimate a significantly smaller observation variance.

and replace the first term with $-\sum_{i=1}^{n} \sum_{j \in \{a, f\}} 4D(g_{ij,n}^{(e)}(\cdot; \beta), \phi_{0, \sigma^2}(\cdot))$. This is the conditional form of the disparity. Note that this reduces us to four points for each density estimate; the limit of what could reasonably be employed. Naturally, both replacements can be made.

Throughout our analysis, we used Hellinger distance as a disparity; we also centered the $t_k$, resulting in $b_{0ij}$ representing the expected salary of student $ij$ at age 50. Bandwidths were fixed within a Metropolis sampling procedures. These were chosen by estimating the $\hat{b}_{0ij}$ and $\hat{\beta}_{1j}$ via least squares, and using these to estimate residuals and all other parameters:

$$\hat{b}_{0j} = \frac{1}{n} \sum_{i} \hat{b}_{0ij}, \quad e_{ijk} = Y_{ijk} - \hat{b}_{0ij} - \hat{\beta}_{1j} t_k,$$

$$\hat{\sigma}^2 = \frac{1}{8n-1} \sum_{ijk} e_{ijk}^2, \quad \hat{\tau}_0^2 = \frac{1}{2n-1} \sum_{ij} (\hat{b}_{0ij} - \hat{\beta}_{0j})^2.$$

The bandwidth selector in Sheather and Jones (1991) was applied to the $\hat{b}_{0ij} - \hat{b}_{0j}$ to obtain a bandwidth for $g_n(b; \beta)$. The bandwidth for $g_{ij,n}^{(e)}(e; \beta)$ was chosen as the average of the bandwidths selected for the $e_{ijk}$ for each $i$ and $j$. For each analysis,
a Metropolis algorithm was run for 200,000 steps and every 100th sample was taken from the second half of the resulting Markov chain. The results of this analysis can be seen in Fig. 3. Here, we have plotted only the differences $\beta_{f0} - \beta_{a0}$ and $\beta_{f1} - \beta_{a1}$ along with the variance components. We observe that for posteriors that have not robustified the random effect distribution, there appears to be a significant difference in the rate of increase in income ($P(\beta_{f1} < \beta_{a1}) < 0.02$ for both posterior and replacing the observation likelihood with Hellinger distance); however, when the random effect likelihood is replaced with Hellinger distance, the difference is no longer significant ($P(\beta_{f1} < \beta_{a1}) > 0.145$ in both cases). We also observe that the estimated observation variance for the model is significantly reduced for posteriors in which the observation likelihood is replaced by Hellinger distance, but that uncertainty in the difference $\beta_{f0} - \beta_{a0}$ is increased.

Investigating these differences, there were two foreign students whose overall expected rate of increase is negative and separated from the least-squares slopes for all the other students. Removing these students increased the posterior probability of $\beta_{a1} > \beta_{f1}$ to 0.11 and decreased the estimate of $\sigma$ from 0.4 to 0.3. Removing the evident high outlier with a considerable departure from trend at age 45 in Figure 4 in Online Appendix F further reduced the EAP of $\sigma$ to 0.185, in the same range as those obtained from robustifying the observation distribution.

Online Appendix F.1 explores the use of a random slope model with additional modeling techniques, where a distinction in average slope between American and foreign students does not appear significant when the slope distribution is robustified via Hellinger distance.

### 9 Conclusions

This paper combines disparity methods with Bayesian analysis to provide robust and efficient inference across a broad spectrum of models. In particular, these methods allow the robustification of any portion of a model for which the likelihood may be written as a product of distributions for i.i.d. random variables. This can be done without the need to modify either the assumed data-generating distribution or the prior. In our experience, Metropolis algorithms developed for the parametric model can be used directly to evaluate the D-posterior and generally incur a modest increase in the acceptance rate and computational cost. Our use of Metropolis algorithms in this context is deliberately naive to demonstrate the immediate applicability of our methods in combination with existing computational tools. We expect that a careful study of the properties of these methods will yield considerable improvements in both computational and sampling efficiency.

The methods in this paper can be employed as a tool for model diagnostics; differences in results by an application of posterior and D-posterior can indicate problematic components of a hierarchical model. Further, estimated densities can indicate how the current model may be improved by, for example, the addition of mixture components. However, the D-posterior can also be used directly to provide robust inference in an automated form.
Our mathematical results are given solely for i.i.d. data; ideas from Hooker (2013) can be used to extend these to the regression framework. Our proposal of hierarchical models remains under mathematical investigation; we expect that similar results can be established in this case. The framework can also be applied within a frequentist context to define a robust marginal likelihood for random effects models, at the cost of numerical complexity. Within hierarchical models, the choice of bandwidth $c_n$ can become difficult. We have employed initial least-squares estimates above, but robust estimators could also be used. Empirically, we have found our results to be relatively insensitive to the choice of bandwidth.

An opportunity for further development of the proposed methodology lies in removing the boundedness of many disparities in common use. These yield EDAP estimates with finite-sample breakdown points of 1, indicating hyper-insensitivity to outliers. Theoretically, some form of boundedness in $G$ has been used within proofs of the robustness of MDEs. However, transformations of $D(g_n, f_\theta)$ can yield non-bounded replacements for the log likelihood which retain both robustness and efficiency properties and this suggests an investigation of the relationship between appropriate transformations and the structure of the parameter space $f_\theta$.

Disparity methods rely on kernel density estimates which are most easily analyzed on unbounded domains. When the data lie in a bounded subset, standard kernel density estimation is non-optimal. One approach to this is to employ a transformation of the data, as we have done for the Gamma distribution in the Online Appendix. This does require that the transformed data have appropriate tail behavior and examining alternative density estimators is an important topic of future study.

The use of a kernel density estimate may also be regarded as inconsistent with a Bayesian analysis and it may be desirable to employ non-parametric Bayesian density estimates as an alternative. Results for disparity estimation are dependent on the properties of kernel density estimates and this extension will require significant mathematical development; an initial study of the use of Dirichlet-process priors for density estimates in this context can be found in Wu and Hooker (2013).

There is considerable scope to extend these methods to further problems. Robustification of the innovation distribution in time-series models, for example, can be readily carried out through disparities and the hierarchical approach will extend this to either the observation or the innovation process in state-space models. The extension to continuous-time models such as stochastic differential equations, however, remains an open and interesting problem. More challenging questions arise in spatial statistics in which dependence decays over some domain and where a collection of i.i.d. random variables may not be available. There are also open questions in the application of these techniques to non-parametric smoothing, and in functional data analysis.

Acknowledgments Giles Hooker’s research was supported by National Science Foundation grant DEB-0813734 and the Cornell University Agricultural Experiment Station federal formula funds Project No. 150446. Anand N. Vidyashankar’s research was supported in part by a grant from National Science Foundation, DMS-000-03-07057.
Appendices A–F are available as Online Resource 1.

References

Albert J (2008) LearnBayes: functions for learning Bayesian inference. R package version 2
Albert J (2009) Bayesian computation with R. Springer, New York
Andrade JAA, O’Hagan A (2006) Bayesian robustness modeling using regularly varying distributions. Bayesian Anal 1(1):169–188
Basu A, Sarkar S, Vidyashankar AN (1997) Minimum negative exponential disparity estimation in parametric models. J Stat Plan Inference 58:349–370
Basu A, Shioya H, Park C (2011) Statistical inference, monographs on statistics and applied probability, vol 120. CRC Press, Boca Raton (the minimum distance approach)
Beran R (1977) Minimum Hellinger distance estimates for parametric models. Ann Stat 5:445–463
Berger JO (1994) An overview of robust Bayesian analysis. TEST 3:5–124
Chern WA, Vidyashankar AN (2006) Minimum Hellinger distance estimation for randomized play the winner design. J Stat Plan Inference 136:1875–1910
Choy STB, Smith AFM (1997) On robust analysis of a normal location parameter. J Royal Stat Soc B 59:463–474
Dawid AP (1973) Posterior expectations for large observations. Biometrika 60:664–667
Desgagnè A, Angers JF (2007) Conflicting information and location parameter inference. Metron 65:67–97
Devroye L, Györfi G (1985) Nonparametric density estimation: the L1 view. Wiley, New York
Dey DK, Birmiwal LR (1994) Robust Bayesian analysis using divergence measures. Stat Prob Lett 20:287–294
Dunson DB, Taylor JA (2005) Approximate bayesian inference for quantiles. J Nonparametr Stat 17(3):385–400
Engel J, Herrmann E, Gasser T (1994) An iterative bandwidth selector for kernel estimation of densities and their derivatives. J Nonparametr Stat 4:2134
Ghosh JK, Delampady M, Samanta T (2006) An introduction to Bayesian analysis. Springer, New York
Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69:383–393
Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. Wiley Series in probability and mathematical statistics: probability and mathematical statistics. Wiley, New York (the approach based on influence functions)
Hansen BE (2004) Nonparametric conditional density estimation. http://www.ssc.wisc.edu/~bhansen/papers/ncde (Unpublished Manuscript)
Hoff PD (2007) Extending the rank likelihood for semiparametric copula estimation. Ann Appl Stat 1(1):265–283
Hooker G (2013) Consistency, efficiency and robustness of conditional disparity methods. arXiv:1307.3730
Huber P (1981) Robust statistics. Wiley, New York
Jiang W, Tanner MA (2008) Gibbs posterior for variable selection in high-dimensional classification and data mining. Ann Stat 26(5):2207–2231
Jureckova J, Sen PK (1996) Robust statistical procedures. Wiley Series in probability and statistics: applied probability and statistics. Wiley, New York (asymptotics and interrelations, A Wiley-Interscience Publication)
Li Q, Racine JS (2007) Nonparametric econometrics. Princeton University Press, Princeton
Lindsay BG (1994) Efficiency versus robustness: the case for minimum Hellinger distance and related methods. Ann Stat 22:1081–1114
Maronna RA, Martin RD, Yohai VJ (2006) Robust statistics. Wiley Series in probability and statistics. Theory and methods. Wiley, Chichester
Nielsen M, Vidyashankar A, Hanlon B, Diao G, Petersen S, Kaplan R (2013) Hierarchical model for evaluating pyrantel efficacy against strongyle parasites in horses. Vet Parasitol 197(3):614–622
O’Hagan A (1979) On outlier rejection phenomena in bayes inference. J Royal Stat Soc B 41:358–367
O’Hagan A (1990) Outliers and credence for location parameter inference. J Am Stat Assoc 85:172–176
Park C, Basu A (2004) Minimum disparity estimation: asymptotic normality and breakdown point results. Bull Inf Cybernet 36:19–34
Peng F, Dey DK (1995) Bayesian analysis of outlier problems using divergence measures. Can J Stat 23:199–213
Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. J Royal Stat Soc Ser B 53:683–690
Silverman BW (1982) Density estimation. Chapman and Hall, Boca Raton
Simpson DG (1987) Minimum Hellinger distance estimation for the analysis of count data. J Am Stat Assoc 82:802–807
Simpson DG (1989) Hellinger deviance test: efficiency, breakdown points and examples. J Am Stat Assoc 84:107–113
Sollich P (2002) Bayesian methods for support vector machines: evidence and predictive class probabilities. Mach Learn 46:21–52
Stigler SM (1973) The asymptotic distribution of the trimmed mean. Ann Stat 1:427–477
Szpiro AA, Rice KM, Lumley T (2010) Model-robust regression and a Bayesian “sandwich” estimator. Ann Appl Stat 4:2099–2113
Tamura RN, Boos DD (1986) Minimum Hellinger distances estimation for multivariate location and covariance. J Am Stat Assoc 81:223–229
Wand M, Ripley B (2009) KernSmooth: functions for kernel smoothing. R package version 2.23-3
Wu Y, Hooker G (2013) Bayesian model robustness via disparities. arXiv:1112.4213
Zhan X, Hettmansperger TP (2007) Bayesian R-estimates in two-sample location models. Comput Statist Data Anal 51(10):5077–5089