Bayesian Model Robustness via Disparities

Giles Hooker
Department of Statistical Science
Cornell University
Ithaca, NY 14850

Anand N. Vidyashankar
Department of Statistics
George Mason University
Fairfax, VA, 22030

Author's Footnote:

Giles Hooker is Assistant Professor, Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14850 (email: giles.hooker@cornell.edu). Anand N. Vidyashankar is Professor, Department of Statistics, George Mason University, VA, 22030 (email: avidyash@gmu.edu).
Giles Hooker's research was supported by NSF grant DEB-0813734 and the Cornell University Agricultural Experiment Station federal formula funds Project No. 150446. Anand Vidyashankar's research was supported in part by a grant from NSF DMS 000-03-07057 and also by grants from the NDCHealth Corporation. All computations were performed in R.

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, 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. The methods are demonstrated on real world data.

KEYWORDS: Deviance test, Kernel density, Hellinger distance, Negative exponential disparity, MCMC, Bayesian Inference, Posterior, Outliers, and Inliers.

1. INTRODUCTION

We focus on 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 either 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 et al., 1994; Albert, 2009; Andrade and O'Hagan, 2006). 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 or to a distortion of prior knowledge. This paper develops a systematic alternative approach based on disparity estimates 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 by using

the minimum Hellinger distance estimator (MHDE). In an i.i.d. context, MHDE estimators are based on minimizing the Hellinger distance between a proposed parametric density $f_\theta(\cdot)$ and a non-parametric estimate $g_n(\cdot)$:

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

over the p-dimensional parameter space Θ. 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). Recently, Hooker and Vidyashankar (2010a) have extended these methods to a non-linear regression framework.

A remarkable property of disparity-based estimates is that while they confer robustness, the 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 the 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. We further establish that the expected D-posterior estimators are asymptotically efficient and the resulting credible 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 and Vidyashankar (2010b). We pursue this extension to hierarchical models and replace various terms in the hierarchy with disparities. This creates a novel "plug-in principle" – 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 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. The divergence measures are computed using MCMC techniques. By contrast, our paper is based on explicitly replacing the likelihood with a disparity in order to provide inherently robust estimation.

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

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:

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

where K is a positive, symmetric, integrable function 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).

We begin by reviewing the class of disparities described in Lindsay (1994). The definition of disparities involves the residual function,

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

defined on the support of $f_\theta(x)$ and a function $G : [-1,\ \infty) \to \mathcal{R}$. $G(\cdot)$ is assumed to be strictly

convex and thrice differentiable with $G(0) = 1$, $G'(0) = 0$ and $G''(0) = 1$. The disparity between f_θ and g_n is defined to be

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

An estimate of θ obtained by minimizing (4) is called a *minimum disparity estimator*; it is equivalent to solving the equation

$$\int A(\delta_\theta(x)) \nabla_\theta f_\theta(x) dx = 0,$$

where $A(\delta) = G'(\delta) - (1+\delta)G(\delta)$ and ∇_θ indicates the derivative with respect to θ.

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

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

for the choice $G(\delta) = -(\delta+1)\log(\delta+1) + a$ for any constant a. We note that the choice of a is arbitrary. In particular, we will assume $a = 1$ so that $A(0) = 1$. The squared Hellinger disparity (HD) corresponds to the choice $G(x) = [(x+1)^{1/2} - 1]^2 + 1$. 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: regions where g_n is small relative to f_θ. An alternative, the negative exponential disparity, based on the choice $G(x) = e^{-x}$ 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). More specifically, if $z_1, \ldots, z_N$ are i.i.d. random samples generated from $g_n(\cdot)$, one can approximate $D(g_n, f_\theta)$ by

$$\tilde{D}(g_n, f_\theta) = \frac{1}{N} \sum_{i=1}^{N} G(\delta_{\theta,g_n}(z_i)) \frac{f_\theta(z_i)}{g_n(z_i)}.$$

In the specific case of Hellinger distance approximation, the above reduces to

$$\tilde{HD}^2(g_n, f_\theta) = 2 - \frac{2}{N} \sum_{i=1}^{N} \frac{f_\theta^{1/2}(z_i)}{g_n^{1/2}(z_i)}. \tag{5}$$

The use of a fixed set of Monte Carlo samples from $g_n(\cdot)$ when optimizing for θ provides a stochastic approximation to an objective function that remains a smooth function of θ and hence avoids the need for complex stochastic optimization. Similarly, in the present paper, we hold the z_i constant

when applying MCMC methods to generate samples from the posterior distribution in order to improve their mixing properties. If f_θ is Gaussian with $\theta = (\mu, \sigma)$, Gauss-Hermite quadrature rules can be used to avoid Monte Carlo integration, leading to improved computational efficiency. In this case we have

$$\tilde{D}(g_n, f_\theta) = \sum_{i=1}^{N} w_i(\theta) G(\delta_{\theta,n}(z_i(\theta))),$$

where the $z_i(\theta)$ and $w_i(\theta)$ are the points and weights for a Gauss-Hermite quadrature scheme for parameters $\theta = (\mu, \sigma)$. However, if $g_n(\cdot)$ contains many local modes surrounded by regions with near-zero density, a small set of quadrature points can result in choosing parameters for which some quadrature point coincides with a local mode. In these cases, it is better to use Monte Carlo samples from $g_n(\cdot)$ even though this increases the computational cost. In this paper, unless stated otherwise, we use the 21-point Gauss-Hermite quadrature rule in all approximations when the proposed distribution is Gaussian.

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 θ has as its first two terms:

$$\begin{aligned} \nabla_\theta KL(g_n, f_\theta) &= \int \left[\nabla_\theta f_\theta(x)\right] (\delta_{\theta,g_n}(x) + 1) dx \\ \nabla^2_\theta KL(g_n, f_\theta) &= \int \left[\nabla^2_\theta f_\theta(x) - \frac{1}{f_\theta(x)} \left(\nabla_\theta f_\theta(x)\right) \left(\nabla_\theta f_\theta(x)\right)^T\right] (\delta_{\theta,g_n}(x) + 1) dx. \end{aligned} \tag{6}$$

The equivalent terms for $D(g_n, f_\theta)$ are:

$$\begin{aligned} \nabla_\theta D(g_n, f_\theta) &= \int \left[\nabla_\theta f_\theta(x)\right] A(\delta_{\theta,g_n}(x)) dx \\ \nabla^2_\theta D(g_n, f_\theta) &= \int \nabla^2_\theta f_\theta(x) A(\delta_{\theta,g_n}(x)) dx \\ &\quad - \int \frac{1}{f_\theta(x)} \left(\nabla_\theta f_\theta(x)\right) \left(\nabla_\theta f_\theta(x)\right)^T (\delta_{\theta,g_n}(x) + 1) A'(\delta_{\theta,g_n}(x)) dx. \end{aligned} \tag{7}$$

Now, letting $\delta_{\theta,g_n}(x) \to 0$ and observing that $A(0) = 1$, $A'(0) = 1$ from the conditions on G, we obtain the equality of (6) and (7). The fact that these heuristics yield efficiency was first noticed by Beran (1977) (eq. 1.1).

In the context of the Bayesian methods examined in this paper, inference is based on the posterior

$$P(\theta|x) = \frac{P(x|\theta)\pi(\theta)}{\int P(x|\theta)\pi(\theta)d\theta}, \tag{8}$$

where $P(x|\theta) = \exp(\sum_{i=1}^{n} \log f_\theta(x_i))$. Following the heuristics above, we propose the simple expedient of replacing the log likelihood, $\log P(x|\theta)$, in (8) with a disparity:

$$P_D(\theta|g_n) = \frac{e^{-nD(g_n,f_\theta)}\pi(\theta)}{\int e^{-nD(g_n,f_\theta)}\pi(\theta)d\theta}. \tag{9}$$

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*. For the Negative Exponential disparity, we refer to the *N-posterior*. These choices are illustrated in Figure 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 period for sampling from $P_D(\theta|g_n)$ and the posterior are approximately the same, although the acceptance rate of the former is approximately around ten percent higher.

Since the D-posterior 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 θ 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 the next section 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 $P_D(f_\theta, g_n)$ 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 negative exponential disparity 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 in order to normalize $P_D(\theta|g_n)$. Further, the tails of $P_D(\theta|g_n)$ are proportional to that of $\pi(\theta)$. This leads to a breakdown point of 1 (see below). However, these results do not affect the asymptotic behavior of $P_D(\theta|g_n)$ since the lower bounds decrease with n, but they do suggest a potential for alternative disparities that allow $D(g_n, f_\theta)$ to diverge at a sufficiently slow rate to retain robustness.

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 by using posterior after n observations, $P(\theta|X_1, \ldots, X_n)$ as a prior θ:

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

   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. In that respect, disparity-based methods can incur additional computational cost.

The idea of replacing log likelihood in the posterior with an alternative criterion occurs in other settings. See Sollich (2002), for example, in developing Bayesian methods for support vector machines. 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.

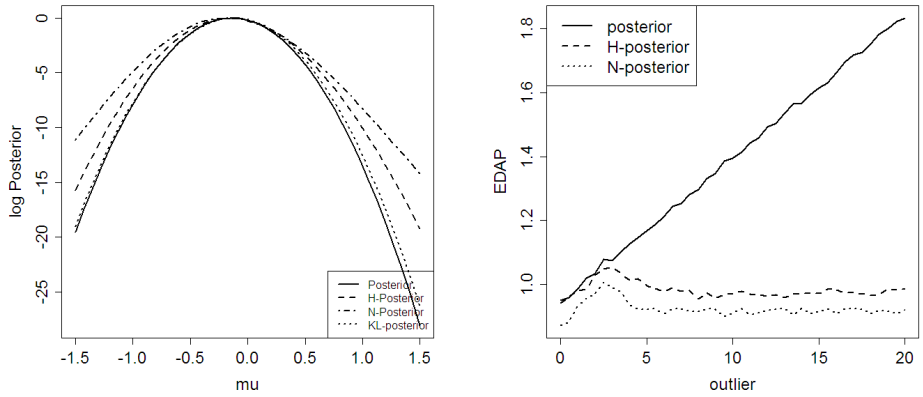

Figure 1: Left: A comparison of log posteriors for μ with data generated from $N(\mu, 1)$ and $N(0, 1)$ prior for μ. The exact log posterior is very well approximated by the Kullback-Leibler divergence between the proposed density and the kernel density estimate. Both Hellinger distance and negative exponential disparity provide good local approximation. Right: influence of an outlier on EDAP estimates of μ as the value of the outlier is changed from 0 to 20.

4. ROBUSTNESS AND EFFICIENCY

In this section, we present theoretical results for i.i.d. data to demonstrate that inference based on the D-posterior is both asympotically efficient and robust. Results for maximum D-*a posteriori* estimates naturally inherit the properties of minimum disparity estimators and hence we focus on EDAP estimators only.

4.1 Efficiency

We recall that under suitable regularity conditions, Expected *a posteriori* estimators 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 $G(\cdot)$ and the model $\{f_\theta : \theta \in \Theta\}$. These are given without requiring that the generating density g is a member of the family f_θ. We therefore define

$$I^D(\theta) = \nabla^2_\theta D(g, f_\theta), \text{ and } \hat{I}^D_n(\theta) = \nabla^2_\theta D(g_n, f_\theta)$$

as the disparity information and θ_g the parameter that minimizes $D(g, f_\theta)$ (note that θ_g here depends on g). We note that if $g = f_{\theta_0}$, $I^D(\theta_g)$ is exactly equal to the Fisher information for θ_0.

The proofs of our asymptotic results rely on the assumptions listed below. Among these are that minimum disparity estimators 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 θ has distribution π. We assume

(A1) G has three continuous derivatives with $G'(0) = 0$, $G''(0) = 1$ and $|G'''(0)| \leq \infty$.

(A2) $\nabla^2_\theta D(g, f_\theta)$ is positive definite and continuous in θ at θ_g and continuous in g with respect to the L_1 metric.

(A3) For any $\delta > 0$, there exists $\epsilon > 0$ such that

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

(A4) The parameter space Θ is compact.

(A5) The minimum disparity estimator, $\hat{\theta}_n$, satisfies $\hat{\theta}_n \to \theta_g$ almost surely and $\sqrt{n}(\hat{\theta}_n - \theta_g) \xrightarrow{d} N(0, I^D(\theta)^{-1})$.

The first three of these are required for the regularity and identifiability of the parametric family f_θ in the disparity D. Specific conditions for (A5) 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 Appendix A.

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 minimum disparity estimator $\hat{\theta}_n$ with variance $\left[nI^D(\hat{\theta}_n)\right]^{-1}$.

Theorem 1. *Let $\hat{\theta}_n$ be the minimum disparity estimator of θ_g, $\pi(\theta)$ be any prior that is continuous and positive at θ_g with $\int_\Theta \|\theta\|_2 \pi(\theta) d\theta < \infty$ and $\pi_n^{*D}(t)$ be the D-posterior density of*

$t = (t_1, \ldots, t_p) = \sqrt{n}(\theta - \hat{\theta}_n)$. *Under conditions (A1)-(A5),*

$$\lim_{n\to\infty} \int \left| \pi_n^{*D}(t) - \left(\frac{|I^D(\theta_g)|}{2\pi} \right)^{p/2} e^{-\frac{1}{2} t' I^D(\theta_g) t} \right| dt \xrightarrow{a.s.} 0. \tag{10}$$

Furthermore, (10) *also holds with* $I^D(\theta_g)$ *replaced with* $\hat{I}_n^D(\hat{\theta}_n)$.

This indicates that credible intervals based on either $P_D(\theta|x_1, \ldots, x_n)$ or from $N(\hat{\theta}_n, I_n^D(\hat{\theta}_n)^{-1})$ will be asymptotically accurate. Our next theorem is concerned with the efficiency and asymptotic normality of EDAP estimates.

Theorem 2. *Assume conditions (A1)-(A5) and* $\int_\Theta \|\theta\|_2 \pi(\theta) d\theta < \infty$, *then* $\sqrt{n}\left(\theta_n^* - \hat{\theta}_n\right) \xrightarrow{a.s.} 0$ *where* θ_n^* *is the EDAP estimate. Further,* $\sqrt{n}\left(\theta_n^* - \theta_g\right) \xrightarrow{d} N\left(0, I^D(\theta_g)\right)$.

The proofs of these theorems are deferred to the Appendix A, but the following remarks are in order:

1. If the proposed parametric model is correct and $g = f_{\theta_0}$, then $I^D(\theta_0)$ is the Fisher information and θ_n^* is efficient.

2. The proofs of these results follow the same principles as those given 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.

3. 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 (A5) can be shown hold. In particular, the results in Hooker and Vidyashankar (2010a) and Hooker and Vidyashankar (2010b) suggest that the disparity methods for regression problems detailed in Sections 5 and 7 will also yield efficient estimates.

4. We assume that that the parameter space, Θ, is compact; this result is also used in conditions that guarantee (A5). As noted in Beran (1977), as well as others, the result continues to hold if Θ can be appropriately embedded in a compact space. Alternatively, Cheng and Vidyashankar (2006) assume local compactness.

5. Assumption (A2) can be replaced by bounds on the absolute size of the third derivatives as in Ghosh et al. (2006). For Hellinger distance, L_1 continuity of the second derivatives of $D^2(g, f_\theta)$ can be obtained when $(\nabla_\theta f_\theta(x))/f_\theta(x)$ and $(\nabla^2_\theta f_\theta(x))/f_\theta(x)$ are square integrable with respect to f_θ. This can be extended to general disparities when $A(\delta)$ $A'(\delta)$ and $(1+\delta)A'(\delta)$ are all bounded; see Lindsay (1994) and Park and Basu (2004). Equivalent conditions on third derivatives require bounding further derivatives of f_θ.

6. We have assumed the data are drawn from a fixed density $g(\cdot)$ in the same manner that a fixed parameter value θ_0 is assumed to generate the data in the case of likelihood studied in Ghosh et al. (2006). A Bayesian extension of this theorem would include a non-parametric prior on possible densities $g(\cdot)$. Asymptotic results would then need to be studied over the distribution of $(X_1, \ldots, X_n, g)$ drawn from the product space.

4.2 Robustness

To describe robustness, we view our estimates as functionals $T(h)$ of a density h. We analyze the behavior of $T(h)$ under the sequence of perturbations $h_{k,\epsilon}(x) = (1-\epsilon)g(x) + \epsilon t_k(x)$ for any sequence of densities $t_k(\cdot)$ and $0 \le \epsilon \le 1$. We measure robustness via two quantities, namely the influence function:

$$IF_T(t_k) = \lim_{\epsilon \to 0} \frac{T_n((1-\epsilon)g + \epsilon t_k) - T_n(g)}{\epsilon} \tag{11}$$

and the breakdown point:

$$B(T) = \inf \left\{ \epsilon : \sup_k |T_n((1-\epsilon)g + \epsilon t_k)| \le \infty \right\}, \tag{12}$$

(see Huber (1981)). EDAP estimates turn out to be highly robust. In fact, while the most common robust estimators have breakdown points of 0.5, for most of the commonly-used robust disparities the D-posterior breakdown point is 1. As described previously this is due to the fact that these disparities are bounded above. We point out here that the Kullback-Leibler disparity is not bounded above and is not robust, both in frequentist and in Bayesian settings.

Theorem 3. *Let $D(g, f_\theta)$ be bounded for all θ and all densities g and let $\int \|\theta\|_2 \pi(\theta) d\theta < \infty$, then the breakdown point is 1.*

We remark that this result demonstrates that the disparity approximation to the likelihood is weak in its tails; it is precisely this property from which robustness is derived. However, it gains strength as n increases, becoming asymptotically equivalent to the likelihood at any θ.

The condition that $D(g, f_\theta)$ be bounded holds true if G is bounded above and below; this is assumed in Park and Basu (2004) and holds for the negative exponential disparity. It can be readily verified that $0 \leq 2HD(g, f_\theta) \leq 4$.

To examine the influence function, we assume that the limit may be taken inside all integrals in (11) and obtain

$$\mathrm{IF}(\theta; g, t_k) = E_{P_D(\theta|g)}\left[\theta C_{nk}(\theta, g)\right] - \left[E_{P_D(\theta|g)}\theta\right]\left[E_{P_D(\theta|g)}C_{nk}(\theta, g)\right].$$

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

$$\begin{aligned} C_{nk}(\theta, g) &= \frac{d}{d\epsilon} n \int G\left(\frac{h_{k,\epsilon}(x)}{f_\theta(x)} - 1\right) f_\theta(x) dx \Bigg|_{\epsilon=0} \\ &= n \int G'\left(\frac{g(x)}{f_\theta(x)} - 1\right)(g(x) - t_k(x)) dx. \end{aligned}$$

Thus, if we can establish the posterior integrability of $C_{nk}(\theta, f)$ and $\theta C_{nk}(\theta, f)$, the influence function will be everywhere finite. This is trivially true if $G'(\cdot)$ is bounded, as is the case for the negative exponential disparity. However, G' is not bounded at -1 for Hellinger distance. To handle this case, we require a more complex condition:

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

$$e_0 = \sup_x \int \left| G'\left(\frac{g(x)}{f_\theta(x)} - 1\right)\pi(\theta)\right| d\theta < \infty \tag{13}$$

and

$$e_1 = \sup_x \int \left| \theta G'\left(\frac{g(x)}{f_\theta(x)} - 1\right)\pi(\theta)\right| d\theta < \infty \tag{14}$$

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

In the case of Hellinger distance the conditions of Theorem 4 require the boundedness of

$$r(x) = \int \frac{\sqrt{f_\theta(x)}}{\sqrt{g(x)}}\pi(\theta) d\theta.$$

The proofs of Theorems 3 and 4 have been left to Appendix B. Since $t_k(x)$ can be made to concentrate on regions where $r(x)$ is large, we conjecture that the conditions in Theorem 4 are necessary. In fact, this requirement means that the H-posterior influence function will not be bounded for a large collection of parametric families. As in Beran (1977), however, we note that Theorem 3 guarantees that α-level influence functions will be bounded for each α.

4.3 Simulation Studies

To illustrate the small sample performance of D-posteriors, we undertook a simulation study for i.i.d. data from Gaussian and log-Gamma distributions and these are reported in detail in Appendix C.1. In both cases, we used the same random-walk Metropolis algorithm to sample from the posterior and the H- and N-posteriors. In general, we record an approximate 10% increase in the acceptance rate for the disparity-based procedures and a 10% increase in variance. The H-posterior, however, demonstrated higher variance for the log-Gamma distribution due to its tendency to create inliers to which Hellinger distance is sensitive. Incorporating outliers into the data strongly biased the posterior for both distributions, but the disparity-based methods were essentially unaffected. The effect of the size of the outlier is investigated in the second plot of Figure 1 where the EDAPs for both disparities smoothly down-weigh the outlying point, while the posterior is highly sensitive to it.

5. 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-stationary 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 and Vidyashankar (2010b).

Consider the classical regression framework with data $(Y_1, X_1), \ldots, (Y_n, X_n)$ is a collection 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):

$$g_n^c(y|x) = \frac{\frac{1}{nc_{n1}c_{2,n}} \sum_{i=1}^n K\left(\frac{y-Y_i}{c_{n1}}\right) K\left(\frac{\|x-X_i\|}{c_{2,n}}\right)}{\frac{1}{nc_{2,n}} \sum_{i=1}^n K\left(\frac{\|x-X_i\|}{c_{2,n}}\right)}. \tag{15}$$

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_θ as follows:

$$D^c(g_n^c, f_\theta) = \sum_{i=1}^n D\left(g_n^c(\cdot|X_i), f_\theta(\cdot|X_i)\right). \tag{16}$$

As before, for Bayesian inference we replace the log likelihood by minus the conditional disparity (16):

$$e^{l(Y_|X_i,\theta)}\pi(\theta) \approx e^{-D^c(g_n^c,f_\theta)}\pi(\theta).$$

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 ϕ_{μ,σ^2} is Gaussian density with mean μ and variance σ^2. An empirical investigation of this estimate is deferred to Section 7 where some alternative formulations of disparity are also studied.

When the X_i are discrete, (15) 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

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

When the n_i are small, or for high-dimensional covariate spaces the non-parametric estimate $g_n(y|X_i)$ can become inaccurate. We discuss techniques for reducing the dimension of the space over which a conditional density is estimated in Section 7.

6. DISPARITY METRICS AND THE PLUG-IN PRINCIPLE

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:

$$P(Y, Z, \theta) = P_1(Y|Z, \theta)P_2(Z|\theta)P_3(\theta) \tag{17}$$

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 principle*. 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 (17) 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 μ. 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

$$P(Y, Z, \mu) = \prod_{i=1}^{n}\left(\prod_{j=1}^{n_i}\phi_{0,\sigma^2}(Y_i - Z_i)\right)\prod_{i=1}^{n}\phi_{\mu,\tau^2}(Z_i)\pi(\mu, \sigma^2, \tau^2) \tag{18}$$

where ϕ_{μ,σ^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 principle described above.

If there are concerns that the distributional assumptions on the ϵ_{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:

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

and replace (18) with

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

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

$$g_n(z; Z) = \frac{1}{nc_{2,n}} \sum_{i=1}^{n} K\left(\frac{z - Z_i}{c_{2,n}}\right)$$

and use

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

as the D-posterior. For inference using this posterior, both μ 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:

$$P_{D_{12}}(Z, Y, \mu) = e^{-\sum_{i=1}^n n_i D(g_n^c(\cdot|Z_i;Z), \phi_{0,\sigma^2}(\cdot))} e^{-nD(g_n(\cdot;Z), \phi_{\mu,\tau^2}(\cdot))}\pi(\mu, \sigma^2, \tau^2).$$

A simulation study considering all these approaches with Hellinger distance chosen as the disparity is described in Appendix C.2. 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 for those that do not the estimation of both the random effect mean and variance is substantially biased.

7. DIMENSION REDUCTION METHODS

The conditional disparity formulation outlined above requires the estimation of the density of a response conditioned on a potentially high-dimensional set of covariates. While this dimensionality

does not affect the asymptotic or computational performance of disparity-based Bayesian inference, it may result in poor performance in small samples. In this section, we explore two methods for reducing the dimension of the conditioning spaces. The first is referred to as the "marginal formulation" and requires only a univariate, unconditional, density estimate. This is a Bayesian extension of the approach suggested in Hooker and Vidyashankar (2010a). It is more stable and computationally efficient than schemes based on nonparametric estimates of conditional densities. However, in a linear-Gaussian model with Gaussian covariates, it requires an external specification of variance parameters for identifiability. For this reason, we propose a two-step Bayesian estimate. The asymptotic analysis for i.i.d. data can be extended to this approach by using the technical ideas in Hooker and Vidyashankar (2010a).

The second method produces a conditional formulation that relies on the structure of a homoscedastic location-scale family $P(Y_i|X_i, \theta, \sigma) = f_\sigma(y - \eta(X_i, \theta))$ and we refer to it as the "conditional-homoscedastic" approach. This method provides a full conditional estimate by replacing a non-parametric conditional density estimate with a two-step procedure as proposed in Hansen (2004). The method involves first estimating the mean function non-parametrically and then estimating a density from the resulting residuals.

7.1 Marginal Formulation

Hooker and Vidyashankar (2010a) proposed basing inference on a marginal estimation of residual density in a nonlinear regression problem. A model of the form

$$Y_i = \eta(X_i, \theta) + \epsilon_i$$

is assumed for independent ϵ_i from a scale family with mean zero and variance σ^2. θ is an unknown parameter vector of interest. A disparity method was proposed based on a density estimate of the residuals

$$e_i(\theta) = Y_i - \eta(X_i, \theta)$$

yielding the kernel estimate

$$g_n^m(e, \theta, \sigma) = \frac{1}{nc_n} \sum K\left(\frac{e - e_i(\theta)/\sigma}{c_n}\right) \tag{19}$$

and θ was estimated by minimizing $D(\phi_{0,1}(\cdot), g_n^m(\cdot, \theta, \sigma))$ where $\phi_{0,1}$ is is the postulated density. As described above, in a Bayesian context we replace the log likelihood by $-nD(\phi_{0,1}(\cdot), g_n^m(\cdot, \theta, \sigma))$.

This formulation has the advantage of only requiring the estimate of a univariate, unconditional density $g_n^m(\cdot, \theta, \sigma)$. This reduces the computational cost considerably as well as providing a density estimate that is more stable in small samples.

Hooker and Vidyashankar (2010a) proposed a two-step procedure to avoid identifiability problems in a frequentist context. This involves replacing σ by a robust external estimate $\tilde{\sigma}$. It was observed that estimates of θ were insensitive to the choice of $\tilde{\sigma}$. After an estimate $\hat{\theta}$ was obtained by minimizing $D(\phi_{0,1}(\cdot), g_n^m(\cdot, \theta, \tilde{\sigma}))$, an efficient estimate of σ was obtained by re-estimating σ based on a disparity for the residuals $e_i(\hat{\theta})$. A similar process can be undertaken here.

In a Bayesian context a plug-in estimate for σ^2 also allows the use of the marginal formulation: an MCMC scheme is undertaken with the plug-in value of σ^2 held fixed. A pseudo-posterior distribution for σ can then be obtained by plugging in an estimate for θ to a Disparity-Posterior for σ. More explicitly, the following scheme can be undertaken:

1. Perform an MCMC sampling scheme for θ using a plug-in estimate for σ^2.

2. Approximate the posterior distribution for σ^2 with an MCMC scheme to sample from the D-posterior $P_D(\sigma^2|y) = e^{-nD(g_n(\cdot,\hat{\theta},\sigma),\phi_{0,1}(\cdot))}\pi(\sigma^2)$ where $\hat{\theta}$ is the EDAP estimate calculated above.

This scheme is not fully Bayesian in the sense that fixed estimators of σ and θ are used in each step above. However, the ideas in Hooker and Vidyashankar (2010a) can be employed to demonstrate that under these schemes the two-step procedure will result in statistically efficient estimates and asymptotically correct credible regions. We note that while we have discussed this formulation with respect to regression problems, it can also be employed with the plug-in principle for random-effects models and we use this in Section 8.2, below.

The formulation presented here resembles the methods proposed in Pak and Basu (1998) based on a sum of disparities between weighted density estimates of the residuals and their expectations assuming the parametric model. For particular combinations of kernels and densities, these estimates are efficient, and the sum of disparities, appropriately scaled, should also be substitutable

for the likelihood in order to achieve an alternative D-posterior.

7.2 Nonparametric Conditional Densities for Regression Models in Location-Scale Families

Under a homoscedastic location-scale model $p(Y_i|X_i,\theta,\sigma) = f_\sigma(Y_i - \eta(X_i,\theta))$ where f_σ is a distribution with zero mean, an alternative density estimate may be used. We first define a non-parametric estimate of the mean function

$$m_n(x) = \frac{\sum Y_i K\left(\frac{x-X_i}{c_{2,n}}\right)}{\sum K\left(\frac{x-X_i}{c_{2,n}}\right)}$$

and then a non-parametric estimate of the residual density

$$g_n^{c2}(e) = \frac{1}{nc_{1,n}} \sum K\left(\frac{e - y_i + m_n(X_i)}{c_{1,n}}\right).$$

We then consider the disparity between the proposed $f_{\theta,\sigma}$ and g_n:

$$D^{c2}(g_n,\theta,\sigma) = \sum D(g_n^{c2}(\cdot + m(X_i)), f_\sigma(\cdot + \eta(X_i,\theta))).$$

As before, $-D^{c2}(g_n, f)$ can be substituted for the log likelihood in an MCMC scheme.

Hansen (2004) remarks that in the case of a homoscedastic conditional density, g_n^{c2} has smaller bias than g_n^c. This formulation does not avoid the need to estimate the high-dimensional function $m_n(x)$. However, the shift in mean does allow the method to escape the identification problems of the marginal formulation while retaining some of its stabilization.

Appendix C.3 gives details of a simulation study of both conditional formulations and the marginal formulation above 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 Section 5 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; negative exponential disparity is less sensitive to this problem (Basu et al., 1997).

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; the full study is presented in Nielsen et al. (2010). The data used in this example are reported in Appendix D.

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

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

where N_i are the pre-treatment egg counts and k_i are the post-treatment egg counts. The log normal was chosen due to the very small empirical probabilities in the data. We observe the data (k_i, N_i) and desire an estimate of μ and σ. The likelihood for these data are

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

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 distribution by use of a negative exponential disparity:

$$g_n(p; p_1, \ldots, p_n) = \frac{1}{nc_n} \sum K\left(\frac{p - \log(p_i)}{c_n}\right)$$

$$l^N(\mu, \sigma | k, N) = -\sum_{i=1}^{n} \left[k_i \log p_i + (N_i - k_i) \log(1 - p_i)\right] - nD(g_n(\cdot; p_1, \ldots, p_n), \phi_{\mu,\sigma^2}(\cdot))$$

In order to perform a Bayesian analysis, μ was given a $N(0, 5)$ prior and σ^2 an inverse Gamma prior with shape parameter 3 and scale parameter 0.5. A random walk Metropolis algorithm was run for this scheme with parameterization $(\mu, \log(\sigma), \log(p_1), \ldots, \log(p_n))$ 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 Figure 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 Figure 2 suggests a lower-outlier with $\log(p_i)$ 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.

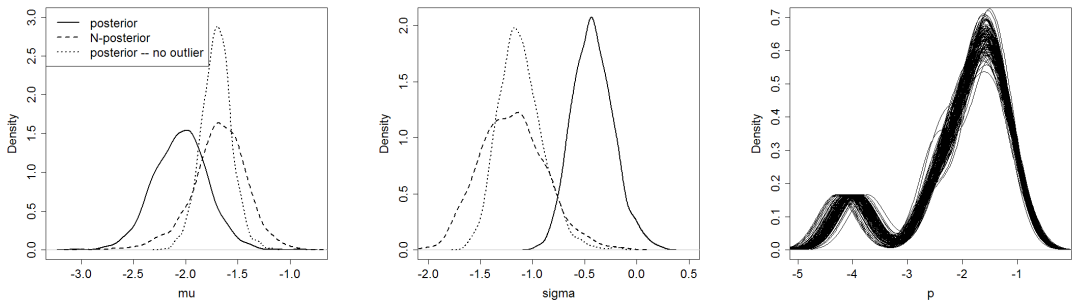

Figure 2: Posterior distributions for the parasite data. Left: posteriors for μ with and without an outlier and the N-posterior. Middle: posteriors for σ. Right: samples of g_n based on draws from the posterior for $p_1, \ldots, p_n$, demonstrating an outlier at -3.

While this example has demonstrated the practical use of disparity methods in Bayesian inference, we feel it important to note that in the context of generalized linear mixed models the distinction between outlying random effects and mis-specification of a link function is not always clear. In particular, an alternative transformation of the p_i such as the square-root transform would not spread out low values to the same extent. The use of disparities therefore controls for outliers only in the context of the particular probability model being employed. The square-root transform would restrict the distribution of the $\sqrt{p}_i$ to the half-line and while asymptotic results continue to hold for kernel densities, this suggest a density estimate which is truncated at 0 and the properties of disparity-based methods for such estimates have not been studied.

8.2 Class Survey Data

Our second data set are from an in-class survey in an introductory statistics course held at Cornell University in 2009. Students were asked to specify 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 Appendix D. Our object is to examine the expected rate of increase in income and any differences in this rate or in the over-all salary level between American and foreign students. From the plot of these data in Figure 8 in Appendix D some potential outliers in both over-all 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

$$Y_{ijk} = b_{0ij} + b_{1j}t_k + \epsilon_{ijk} \tag{20}$$

where Y_{ijk} is log income for the ith 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

$$l(Y, \beta, \sigma^2, \tau_0^2) = -\sum_{i=1}^{n} \sum_{j\in\{a,f\}} \sum_{k=1}^{4} \frac{1}{2\sigma^2} \left(Y_{ijk} - b_{0ij} - \beta_{1j}t_k\right)^2 - \sum_{i=1}^{n} \sum_{j\in\{a,f\}} \frac{1}{2\tau_0^2} \left(b_{0ij} - \beta_{0j}\right)^2 \tag{21}$$

to which we attach Gaussian priors centered at zero with standard deviations 150 and 0.5 for the β_{0j} and β_{1j} respectively and Gamma priors with shape parameter 3 and scale 0.5 and 0.05 for τ_0^2 and σ^2. These are chosen to correspond to the approximate orders of magnitude observed in the maximum likelihood estimates of the b_{0ij}, β_{1j} and residuals.

As in Section 6 we can robustify this likelihood in two different ways: either against the distributional assumptions on the ϵ_{ijk} or on the b_{0ij}. In the latter case we form the density estimate

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

and replace the second term in (21) with $-2nD(g_n(\cdot; \boldsymbol{\beta}), \phi_{0,\tau_0^2}(\cdot))$. Here we have used

$$\boldsymbol{\beta} = (\beta_{0a}, \beta_{0f}, \beta_{1a}, \beta_{1f}, b_{01a}, b_{01f}, \ldots, b_{0na}, b_{0nf})$$

as an argument to g_n to indicate its dependence on the estimated parameters. We have chosen to combine the b_{0ia} and the b_{0if} together in order 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 ϵ_{ijk} since the b_{0ij} cannot be identified marginally. Instead, we estimate a density at each ij:

$$g_{ij,n}^c(e; \boldsymbol{\beta}) = \frac{1}{4nc_n} \sum_{k=1}^{4} K\left(\frac{e - (Y_{ijk} - b_{0ij} - \beta_{1j}t_k)}{c_n}\right)$$

and replace the first term with $\sum_{i=1}^{n}\sum_{j\in\{a,f\}} 4D(g^c_{ij,n}(\cdot;\beta),\phi_{0,\sigma^2}(\cdot))$. 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:

$$\begin{aligned}
\hat{\beta}_{0j} &= \frac{1}{n}\hat{b}_{0i} \\
e_{ijk} &= Y_{ijk} - \hat{b}_{0ij} - \hat{\beta}_{1j}t_k \\
\hat{\sigma}^2 &= \frac{1}{8n-1}\sum_{ijk} e^2_{ijk} \\
\hat{\tau}^2_0 &= \frac{1}{2n-1}\sum_{ij}(\hat{b}_{0ij} - \hat{\beta}_{0j})^2.
\end{aligned}$$

The bandwidth selector in Sheather and Jones (1991) was applied to the $\hat{b}_{0ij} - \hat{b}_{0j}$ to obtain a bandwidth for $g_n(b;\boldsymbol{\beta})$. The bandwidth for $g^c_{ij,n}(e;\boldsymbol{\beta})$ 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 Figure 3. Here we have plotted only the differences $\beta_{0f} - \beta_{0a}$ and $\beta_{1f} - \beta_{1a}$ 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 ($P(\beta_{1f} < \beta_{1a}) < 0.02$ 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 ($P(\beta_{1f} < \beta_{1a}) > 0.145$ in both cases). We also observe that the observation variance is significantly reduced for posteriors in which the observation likelihood is replaced by Hellinger distance, but that uncertainty in the difference $\beta_{0f} - \beta_{0a}$ is increased.

Investigating these differences, there were two foreign students who's over-all 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 $\beta_{1a} > \beta_{1f}$ to 0.11 and decreased the estimate

of σ from 0.4 to 0.3. Removing the evident high outlier with a considerable departure from trend at age 45 in Figure 8 in Appendix D further reduced the EAP of σ to 0.185, in the same range as those obtained from robustifying the observation distribution.

A further model exploration allows a random slope for each student in addition to the random offset. The model now becomes

$$Y_{ijk} = b_{0ij} + b_{1ij}t_k + \epsilon_{ijk} \tag{22}$$

with additional distributional assumptions

$$b_{1ij} \sim N(\beta_{1j}, \tau_1^2)$$

and an additional term

$$-\frac{1}{2\tau_1^2}\sum_{i=1}^{n}\sum_{j\in\{a,f\}}(b_{1ij}-\beta_{1j})^2$$

added to (21). Here, this term can be robustified in a similar manner to the robustification of the b_{0ij}. However, we note that a robustification of the error terms would require the estimation of a conditional density for each ij – based on only four data points. We viewed this as being too little to achieve reasonable results and therefore employed the marginal formulation described in Section 7. Specifically, we first obtained residuals e_{ijk} for the random slope model from the maximum likelihood estimates for each subject-specific effect and estimated

$$\hat{\sigma}^2 = \frac{1}{0.674\sqrt{2}}|e_{ijk} - \text{median}(e)|.$$

Following this, we estimated a combined density for all residuals, conditional on the random effects

$$g_n^m(e;\boldsymbol{\beta}) = \frac{1}{8nc_n}\sum_{i=1}^{n}\sum_{j\in\{a,f\}}\sum_{k=1}^{4}K\left(\frac{e-(Y_{ijk}-b_{0ij}-b_{1ij}t_k)}{c_n}\right)$$

and replaced the first term in (21) with $-8nD(g_n^m(\cdot;\boldsymbol{\beta}),\phi_{0,\hat{\sigma}^2}(\cdot))$. Following the estimation of all other parameters, we obtained new residuals $\tilde{e}_{ijk} = Y_{ijk} - \tilde{b}_{0ij} - b_{1ij}t_k$ where the $\tilde{b}_{0ij}$ and $\tilde{b}_{1ij}$ are the EDAP estimators. We then re estimated σ^2 based on its H-posterior using the $\tilde{e}_{ijk}$ as data. In this particular case a large number of outliers from a concentrated peak (see Figure 4) meant that the use of Gauss-Hermite quadrature in the evaluation of

$$HD(g_n^m(\cdot,\tilde{\boldsymbol{\beta}}),\phi_{0,\sigma^2}) = 2 - 2\int\left(\sqrt{g_n^m(e;\tilde{\boldsymbol{\beta}})}/\sqrt{\phi_{0,\sigma^2}(e)}\right)\phi_{0,\sigma^2}(e)de$$

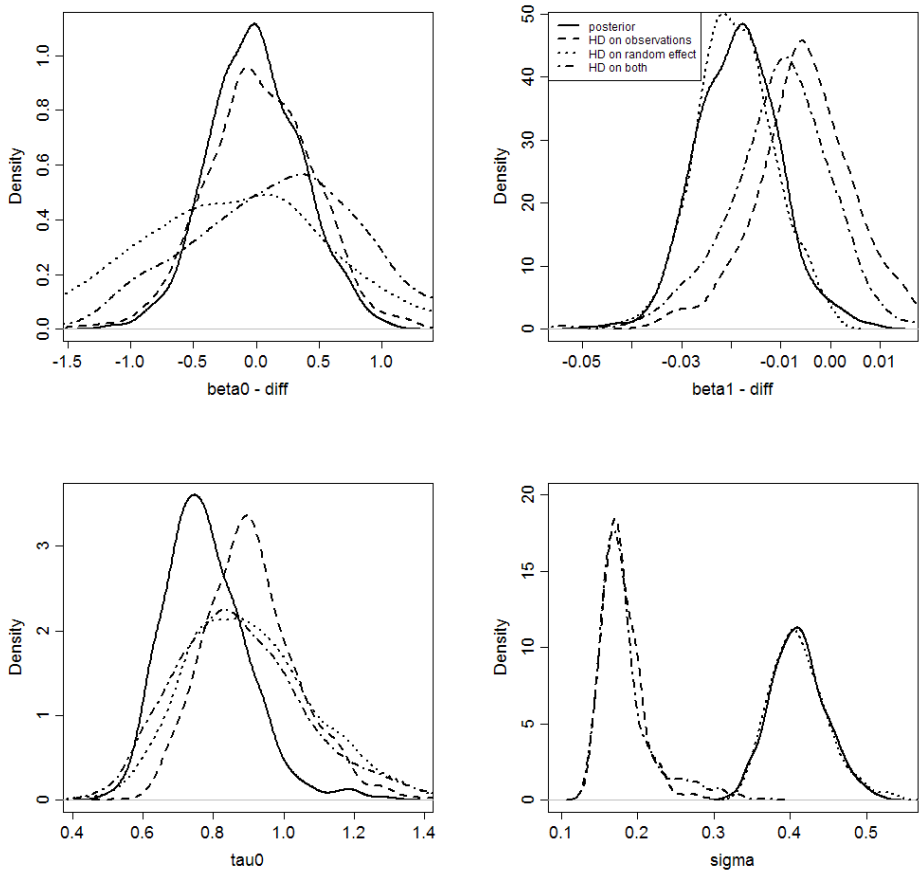

Figure 3: Analysis of the class survey data using a random intercept model with Hellinger distance replacing the observation likelihood, the random effect likelihood or both. Top: differences in intercepts between foreign and American students (left) and differences in slopes (right). Bottom: random 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.

suffered from large numerical errors and we therefore employed a Monte Carlo integral based on 400 data points drawn from g_n^m instead, using the estimate (5) which required approximately 30 times the computing time as compared to using Gauss-Hermite quadrature. To estimate both σ^2 and the other parameters we used a Metropolis random walk algorithm which was again run for 200,000 iterations with estimates based on every 100th sample in the second half of the chain.

Some results from this analysis are displayed in Figure 4. The residual distribution of the $\tilde{e}_{ijk}$ show a very strong peak and a number of isolated outliers. The estimated standard deviation of the residual distribution is therefore very different between those methods that are robust to outliers and those that are not; the mean posterior σ was increased by a factor of four between those methods using a Hellinger disparity and those using the random effect log likelihood. The random slope variance was estimated to be small by all methods – we speculate that the distinction between random effect log likelihoods and Hellinger methods is bias due to bandwidth size – but this was not enough to overcome the differences between the methods concerning the distinction between β_{1f} and β_{1a}.

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* in order to demonstrate the immediate applicability of our methods in combination with existing computational tools. We expect that a more careful study of sampling 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

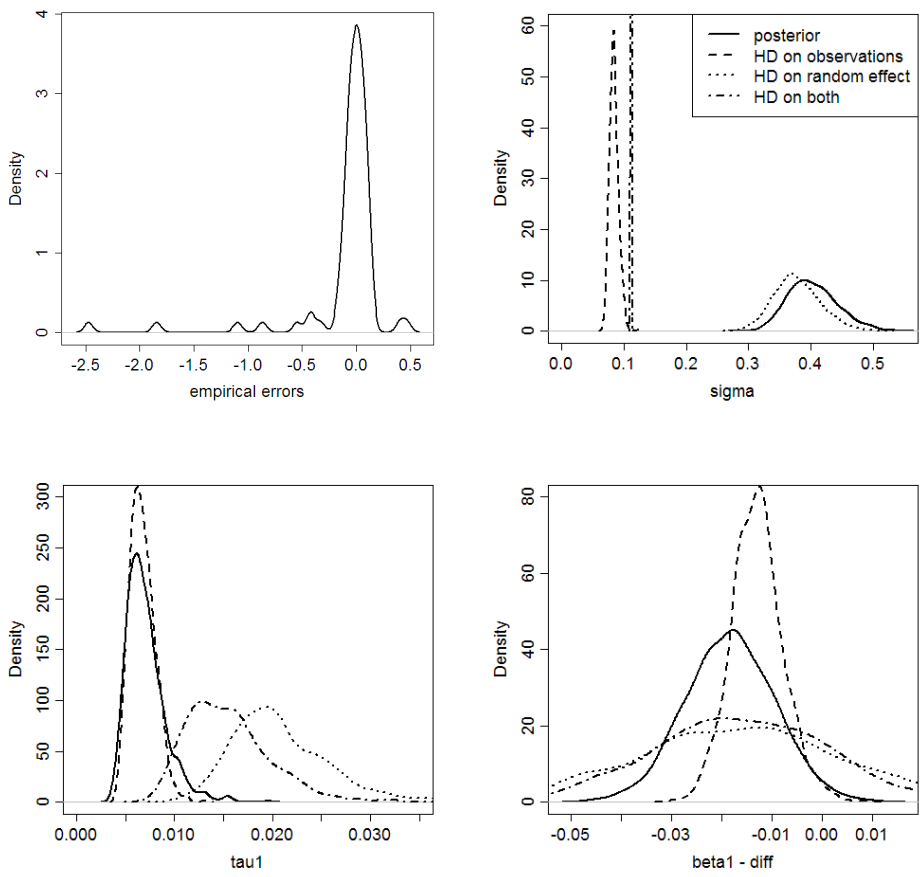

Figure 4: Analysis of a random-slope random-intercept model for the class survey data. Top left: a density estimate of the errors following a two-step procedure with the error variance held constant. This shows numerous isolated outliers than create an ill-conditioned problem for Gauss-Hermite quadrature methods. Top right: estimates of residual standard deviation replacing various terms in the log likelihood with Hellinger distance. The effect of outliers is clearly apparent in producing an over-estimate of variance. Bottom left: estimated variance of the random slope. Bottom right: estimated difference in mean slope between American and foreign students.

a hierarchical model. Further, estimated densities can indicate how the current model may be improved. 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 and Vidyashankar (2010a) can be used to extend these to the regression framework. Our proposal of hierarchical models remains under mathematical investigation, but we expect that similar results can be established in this case. The methodology can also be applied within a frequentist context to define an alternative marginal likelihood for random effects models, although the numerical estimation of such models is likely to be problematic.

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 breakdown points of 1, indicating hyper-insensitivity to the data. Theoretically, some form of boundedness has been used within proofs of the efficiency of minimum disparity estimators. However these results suggest an investigation of the necessity of this assumption and the development of new disparities which diverge at a rate slow enough to retain robustness.

The use of a kernel density estimate may also be regarded as inconsistent with a Bayesian context and it may therefore be desirable to employ non-parametric Bayesian density estimates as an alternative. Results for disparity estimation are heavily dependent on properties of kernel density estimates and this extension will require significant mathematical development.

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 though 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.

REFERENCES

Albert, J. (2008). *LearnBayes: Functions for Learning Bayesian Inference*. R package version 2.0.

Albert, J. (2009). *Bayesian Computation with R*. New York: Springer.

Andrade, J. A. A. and A. O'Hagan (2006). Bayesian robustness modeling using regularly varying distributions. *Bayesian Analysis 1*(1), 169–188.

Basu, A., S. Sarkar, and A. N. Vidyashankar (1997). Minimum negative exponential disparity estimation in parametric models. *Journal of Statistical Planning and Inference 58*, 349–370.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Annals of Statistics 5*, 445–463.

Berger, J., E. Moreno, L. Pericchi, M. Bayarri, J. Bernardo, J. Cano, J. De la Horra, J. Martn, D. Ros-Insa, B. Betr, A. Dasgupta, P. Gustafson, L. Wasserman, J. Kadane, C. Srinivasan, M. Lavine, A. OHagan, W. Polasek, C. Robert, C. Goutis, F. Ruggeri, G. Salinetti, and S. Sivaganesan (1994). An overview of robust bayesian analysis. *TEST 3*, 5–124. 10.1007/BF02562676.

Cheng, A.-L. and A. N. Vidyashankar (2006). Minimum Hellinger distance estimation for randomized play the winner design. *Journal of Statistical Planning and Inference 136*, 1875–1910.

Devroye, L. and G. Györfi (1985). *Nonparametric Density Estimation: The L1 View*. New York: Wiley.

Dey, D. K. and L. R. Birmiwal (1994). Robust bayesian analysis using divergence measures. *Statistics and Probability Letters 20*, 287–294.

Engel, J., E. Herrmann, and T. Gasser (1994). An iterative bandwidth selector for kernel estimation of densities and their derivatives. *Journal of Nonparametric Statistics 4*, 2134.

Ghosh, J. K., M. Delampady, and T. Samanta (2006). *An Introduction to Bayesian Analysis*. New York: Springer.

Hansen, B. E. (2004). Nonparametric conditional density estimation.

Hooker, G. and A. N. Vidyashankar (2010a). Minimum disparity methods for nonlinear regression – marginal approach. *in preparation*.

Hooker, G. and A. N. Vidyashankar (2010b). Minimum disparity methods for nonlinear regression – conditional approach. *in preparation*.

Huber, P. (1981). *Robust Statistics*. New York: Wiley.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics 22*, 1081–1114.

Nielsen, M., Vidyashankar, A.N., B. Hanlon, S. Petersen, and R. Kaplan (2010). Hierarchical models for evaluating anthelmintic resistance in livestock parasites using observational data from multiple farms. *under reivew*.

original by Matt Wand. R port by Brian Ripley., S. (2009). *KernSmooth: Functions for kernel smoothing*. R package version 2.23-3.

Pak, R. J. and A. Basu (1998). Minimum disparity estimation in linear regression models: Distribution and efficiency. *Annals of the Institute of Statistical Mathematics 50*, 503–521.

Park, C. and A. Basu (2004). Minimum disparity estimation: Asymptotic normality and breakdown point results. *Bulletin of Informatics and Cybernetics 36*.

Peng, F. and D. K. Dey (1995). Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics 23*, 199–213.

Sheather, S. J. and M. C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B 53*, 683690.

Silverman, B. W. (1982). *Density Estimation*. Chapman and Hall.

Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American statistical Association 82*, 802–807.

Simpson, D. G. (1989). Hellinger deviance test: efficiency, breakdown points and examples. *Journal of the American Statistical Association 84*, 107–113.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predicive class probabilities. *Machine Learning 46*, 21–52.

Stigler, S. M. (1973). The asymptotic distribution of the trimmed mean. *Annals of Statistics 1*, 427–477.

Tamura, R. N. and D. D. Boos (1986). Minimum Hellinger distances estimation for multivariate location and and covariance. *Journal of the American Statistical Association 81*, 223–229.

A. PROOFS OF EFFICIENCY

A.1 Efficiency Conditions for Minimum Disparity Estimators

Here we provide conditions that ensure the consistency and asymptotic normailty of the minimum-disparity estimator $\hat{\theta}_n$. There is a slight variation through the literature in conditions required for efficiency (see Beran (1977), Basu et al. (1997), Park and Basu (2004) and Cheng and Vidyashankar (2006)). The conditions given below are adapted from Cheng and Vidyashankar (2006) for the specific case of Hellinger distance. A small modification of these conditions will also provide the consistency and asymptotic normality for more general disparities under appropriate conditions on $G(\cdot)$.

We first require conditions on the data and the proposed parametric family:

(D1) $X_1, \ldots, X_n$ are i.i.d. with distribution given by the density function $g(\cdot)$.

(D2) $f_\theta(\cdot)$ is twice continuously differentiable with respect to θ.

(D3) $\left\| \nabla_\theta \left(\sqrt{f_\theta(\cdot)} \right) \right\|_2$ is continuous and bounded.

(D4) $f_\theta^{-1}(x) \left[\nabla_\theta f_\theta(x) \right]$ is continuous and bounded in L_2 at $\theta = \theta_g$.

(D5) $f_\theta^{-1/2}(x) \left[\nabla_\theta^2 f_\theta(x) \right] - f_\theta^{-3/2}(x) \left[\nabla_\theta f_\theta(x) \right] \left[\nabla_\theta f_\theta(x) \right]^T$ is continuous and bounded in L_2 at $\theta = \theta_g$.

(D6) $f_\theta^{-1/2}(x) \left[\nabla_\theta f_\theta(x) \right] \left[\nabla_\theta f_\theta(x) \right]^T$ is continuous and bounded in L_2 at $\theta = \theta_g$.

We also require conditions on the kernel function K in the kernel density estimate and its relationship to the parametric density family:

(K1) $K(\cdot)$ is symmetric about 0 and $\int K(t)dt = 1$.

(K2) The bandwidth is chosen so that $c_n \to 0$, $nc_n^2 \to 0$, $nc_n \to \infty$.

(K3) There is a sequence a_n, $n \geq 0$ diverging to infinity such that

(a) For X a random variable with density $f_{\theta_g}(\cdot)$

$$\lim_{n \to \infty} nP\left(|X| > a_n \right) = 0,$$

(b)

$$\sup_{n\geq 1}\sup_{|x|<a_n}\sup_{t\in R}\left|K(t)\frac{f_{\theta_g}(x+tc_n)}{f_{\theta_g}(x)}\right|<\infty,$$

(c) The parametric score functions have regular central behavior relative to the bandwidth:

$$\lim_{n\to\infty}\frac{1}{n^{1/2}c_n}\int_{-a_n}^{a_n}\frac{\nabla_\theta f_{\theta_g}(x)}{f_{\theta_g}(x)}dx=0$$

and

$$\lim_{n\to\infty}\frac{1}{n^{1/2}c_n^4}\int_{-a_n}^{a_n}\frac{\nabla_\theta f_{\theta_g}(x)}{f_{\theta_g}(x)}dx=0,$$

(d) The score functions are smooth with respect to K in an L_2 sense:

$$\lim_{n\to\infty}\sup_{t\in R}\int_{\mathbb{R}}K(t)\left[\frac{\nabla_\theta f_{\theta_g}(x+c_nt)}{f_{\theta_g}(x+c_nt)}-\frac{\nabla_\theta f_{\theta_g}(x)}{f_{\theta_g}(x)}\right]^2 f_{\theta_g}(x)dx=0.$$

This statement of assumptions in particular remove the condition that $K(t)$ has compact support, which was assumed in Beran (1977); Basu et al. (1997); Park and Basu (2004). These assumptions significantly expand the class of kernels available for use and hence expands the applicability of Theorem 1 (see Hooker and Vidyashankar (2010a) for formal details). In practice, it is numerically more stable to use a Gaussian kernel or some other distribution with support on the whole real line and we have used Gaussian kernels throughout our numerical experiments.

A.2 Proof of Theorem 1

We begin with the following Lemma:

Lemma 1. *Let*

$$w_n(t)=\pi(\hat{\theta}_n+t/\sqrt{n})e^{-nD\left(g_n,f_{\hat{\theta}_n+t\sqrt{n}}\right)+nD\left(g_n,f_{\hat{\theta}_n}\right)}-\pi(\theta_g)e^{-\frac{1}{2}t'I^D(\theta_g)t}$$

then under (A1)-(A5)

$$\int|w_n(t)|dt\xrightarrow{a.s.}0 \text{ and } \int\|t\|_2|w_n(t)|dt\xrightarrow{a.s.}0.$$

Proof. We divide the integral into $A_1=\{\|t\|_2>\delta\sqrt{n}\}$ and $A_2=\{\|t\|_2\leq\delta\sqrt{n}\}$:

$$\int|w_n(t)|dt=\int_{A_1}|w_n(t)|dt+\int_{A_2}|w_n(t)|dt \tag{A.1}$$

and show that each vanishes in turn. First, since

$$\sup_{\theta\in\Theta}\left|D\left(g_n,f_\theta\right)-D\left(g,f_\theta\right)\right|\xrightarrow{a.s.}0$$

by Assumption (A3), for some $\epsilon>0$ with probability 1 it follows that

$$\exists N:\forall n\geq N,\ \sup_{\|t\|_2>\delta}D(g_n,f_{\hat{\theta}_n+t\sqrt{n}})-D\left(g_n,f_{\hat{\theta}_n}\right)>-\epsilon.$$

This now allows us to demonstrate the convergence of the first term in (A.1):

$$\begin{aligned}\int_{A_1}|w_n(t)|dt &\leq \int_{A_1}\pi(\hat{\theta}_n+t/\sqrt{n})e^{-nD\left(g_n,f_{\hat{\theta}_n+t\sqrt{n}}\right)+nD\left(g_n,f_{\hat{\theta}_n}\right)}dt\\ &\quad+\int_{A_1}\pi(\theta_g)e^{-\frac{1}{2}t'I^D(\theta_g)t}dt\\ &\leq e^{-n\epsilon}+\pi(\theta_g)\left(\frac{|I^D(\theta_g)|}{2\pi}\right)^{p/2}P(\|Z\|_2>\sqrt{n}\delta)\\ &\xrightarrow{a.s.}0\end{aligned}$$

where Z is a $N(0,I^D(\theta_g))$ random variable.

We now deal with the second term in (A.1). Notice that

$$nD\left(g_n,f_{\hat{\theta}_n+t/\sqrt{n}}\right)-nD\left(g_n,f_{\hat{\theta}_n}\right)=\frac{1}{2}t'I_n^D(\theta_n')$$

for $\theta_n'=\hat{\theta}_n+\alpha t/\sqrt{n}$ with $0\leq\alpha\leq1$ and therefore

$$\begin{aligned}w_n(t)&=\pi(\hat{\theta}_n+t/\sqrt{n})e^{-\frac{1}{2}t'I_n^D(\theta_n')}-\pi(\theta_g)e^{-\frac{1}{2}t'I^D(\theta_g)t}\\ &\to 0\end{aligned}$$

for every t.

By Assumption (A2) we can choose δ so that $I^D(\theta)\succ 2M$ if $\|\theta-\theta_g\|_2\leq 2\delta$ for some positive definite matrix M where $A\succ B$ indicates $t'At>t'Bt$ for all t. Since $\|\theta_n'-\hat{\theta}_n\|\leq\delta$ with probability 1 for all n sufficiently large

$$e^{-nD\left(g_n,f_{\hat{\theta}_n+t\sqrt{n}}\right)+nD\left(g_n,f_{\hat{\theta}_n}\right)}\leq e^{-\frac{1}{2}t'Mt}.$$

Therefore

$$\int_{A_2}|w_n(t)|dt\leq\int_{A_2}\pi(\hat{\theta}_n+t/\sqrt{n})e^{-\frac{1}{2}t'Mt}+\pi(\theta_g)\int_{A_2}e^{-\frac{1}{2}tI^D(\theta_g)t}dt<\infty.$$

and the result follows from the pointwise convergence of $w(t)$ and the dominated convergence theorem.

We can prove

$$\int \|t\|_2 |w_n(t)| dt \xrightarrow{a.s.} 0$$

in an analogous manner by observing that on A_1

$$\begin{aligned}\int_{A_1} \|t\|_2 |w_n(t)| dt &\leq \int_{A_1} \|t\|_2 \pi(\hat{\theta}_n + t/\sqrt{n}) e^{-nD\left(g_n, f_{\hat{\theta}_n + t\sqrt{n}}\right) + nD\left(g_n, f_{\hat{\theta}_n}\right)} dt \\ &\quad + \int_{A_1} \pi(\theta_g) \|t\|_2 e^{-\frac{1}{2} t' I^D(\theta_g) t} dt \\ &\xrightarrow{a.s.} 0\end{aligned}$$

and on A_2, $\|t\|_2 |w_n(t)| \xrightarrow{a.s.} 0$ and

$$\int_{A_2} \|t\|_2 |w_n(t)| dt \leq \int_{A_2} \|t\|_2 \pi(\hat{\theta}_n + t/\sqrt{n}) e^{-\frac{1}{2} t' M t} + \pi(\theta_g) \int_{A_2} \|t\|_2 e^{-\frac{1}{2} t I^D(\theta_g) t} dt < \infty.$$

□

Following this lemma, we prove Theorem 1.

Proof. First, from Assumption (A5),

$$\sqrt{n}\left(\hat{\theta}_n - \theta_g\right) \xrightarrow{d} N(0, I^D(\theta_g)),$$

using that $\int |g_n(t) - f_{\theta_g}(t)| dt \xrightarrow{a.s.} 0$, the continuity of G and the compactness of Θ, it follows that

$$\sup_{\theta \in \Theta} \left| D\left(g_n, f_\theta\right) - D\left(g, f_\theta\right) \right| \xrightarrow{a.s.} 0$$

and

$$D\left(g_n, f_{\hat{\theta}_n}\right) \xrightarrow{a.s.} D\left(g, f_{\theta_g}\right), \ \nabla_\theta D\left(g_n, f_{\hat{\theta}_n}\right) \xrightarrow{a.s.} \nabla_\theta D\left(g, f_{\theta_g}\right), \ \nabla^2_\theta D\left(g_n, f_{\hat{\theta}_n}\right) \xrightarrow{a.s.} \nabla^2_\theta D\left(g, f_{\theta_g}\right)$$

Now, we write that

$$\pi_n^{*D}(t) = \kappa_n^{-1} \pi(\hat{\theta}_n + t/\sqrt{n}) e^{-nD\left(g_n, f_{\hat{\theta}_n + t\sqrt{n}}\right) + nD(g_n, f_{\hat{\theta}})}$$

where κ_n is chosen so that $\int \pi_n^{*D}(t) dt = 1$. Let

$$w_n(t) = \pi(\hat{\theta}_n + t/\sqrt{n}) e^{-nD\left(g_n, f_{\hat{\theta}_n + t\sqrt{n}}\right) + nD\left(g_n, f_{\hat{\theta}_n}\right)} - \pi(\theta_g) e^{-\frac{1}{2} t' I^D(\theta_g) t}$$

from Lemma 1, we have

$$\int |w_n(t)|dt \xrightarrow{a.s.} 0$$

from which

$$\kappa_n = \int \pi(\hat{\theta}_n + t/\sqrt{n}) e^{-nD\left(g_n, f_{\hat{\theta}_n + t\sqrt{n}}\right) + nD\left(g_n, f_{\hat{\theta}_n}\right)} dt \xrightarrow{a.s.} \pi(\theta_g) \int e^{-\frac{1}{2}t' I^D(\theta_g) t} dt = \pi(\theta_g) \left(\frac{2\pi}{|I^D(\theta_g)|}\right)^{p/2}$$

and

$$\begin{aligned}
\lim_{n\to\infty} \int \left| \pi_n^{*D}(t) - \frac{\sqrt{I^D(\theta_g)}}{\sqrt{2\pi}} e^{-\frac{1}{2}t' I^D(\theta_g) t} \right| dt &= \int \left| \kappa_n^{-1} w_n(t) + \left(\kappa_n^{-1} \pi(\theta_g) - \left(\frac{|I^D(\theta_g)|}{2\pi} \right)^{p/2} \right) e^{-\frac{1}{2}t' I^D(\theta_g) t} \right| dt \\
&\leq \kappa_n^{-1} \int |w_n(t)| dt + \left(\frac{2\pi}{|I^D(\theta_g)|} \right)^{p/2} \left| \kappa_n^{-1} \pi(\theta_g) - \left(\frac{|I^D(\theta_g)|}{2\pi} \right)^{p/2} \right| \\
&\xrightarrow{a.s.} 0.
\end{aligned}$$

That the result holds for $I^D(\theta_g)$ replaced with $\hat{I}_n^D(\hat{\theta}_n)$ follows from the almost sure convergence of the latter to the former. □

A.3 Proof of Theorem 2

Proof. Let $t = (t_1, \ldots, t_p)$, from Theorem 1

$$\int t_i \pi^{*D}(t|x_1, \ldots, x_n) \xrightarrow{a.s.} \left(\frac{2\pi}{|I^D(\theta_g)|} \right)^{p/2} \int t_i e^{-\frac{1}{2}t' I^D(\theta_g) t} dt = 0.$$

Since

$$\theta_n^* = E(\hat{\theta}_n + t/\sqrt{n} | X_1, \ldots, X_n)$$

we have

$$\sqrt{n}\left(\theta_n^* - \hat{\theta}_n\right) \xrightarrow{a.s.} \left(\frac{2\pi}{|I^D(\theta_g)|} \right)^{p/2} \int t e^{-\frac{1}{2}t' I^D(\theta_g) t} dt = 0.$$

Since $\sqrt{n}\left(\hat{\theta}_n - \theta_g\right) \xrightarrow{d} N\left(0, I^D(\theta_g)\right)$, it follows that $\sqrt{n}\left(\theta_n^* - \theta_g\right) \xrightarrow{d} N\left(0, I^D(\theta_g)\right)$; hence θ_n^* is asymptotically normal, efficient as well as robust. □

B. PROOFS OF ROBUSTNESS

B.1 Proof of Theorem 3

Proof. Under the assumptions, $\sup_{\theta,g} D(g, f_\theta) = R < \infty$ and $\inf_{\theta,g} D(g, f_\theta) = r > -\infty$. Let $h_{k,\epsilon} = (1-\epsilon)g + \epsilon t_k$, then for all θ,

$$e^{-nR} \leq e^{-nD(h_{k,\epsilon}, f_\theta)} < e^{-nr}, \ \ \forall k \in 1, 2, \ldots, \ \ \forall \epsilon \in [0\ 1]$$

and therefore

$$e^{n(r-R)} E_{\pi(\theta)}\theta = \frac{\int \theta e^{-nR}\pi(\theta)d\theta}{\int e^{-nr}\pi(\theta)d\theta} \leq E_{P_D(\theta|h_{k,\epsilon})}\theta \leq \frac{\int \theta e^{-nr}\pi(\theta)d\theta}{\int e^{-nR}\pi(\theta)d\theta} = e^{n(R-r)} E_{\pi(\theta)}\theta.$$

□

B.2 Proof of Theorem 4

Proof. It is sufficient to show that

$$\left| E_{P_D(\theta|g)} C_{nk}(\theta, g) \right| < \infty \text{ and } \left| E_{P_D(\theta|g)} \left[\theta C_{nk}(\theta, g)\right] \right| < \infty.$$

We will prove the first of these, the second follows analogously.

$$\begin{aligned}
\left| E_{P_D(\theta|g)} C_{nk}(\theta, g) \right| &\leq e^{n(R-r)} \left| \int C_{nk}(\theta, g)\pi(\theta)d\theta \right| \\
&\leq e^{n(R-r)} \int \left| (g(x) - t_k(x)) \int G'\left(\frac{g(x)}{f_\theta(x)} - 1\right)\pi(\theta)d\theta \right| dx \\
&\leq e^{n(R-r)} e_0 \int |g(x) - t_k(x)| dx \qquad \text{(A.2)} \\
&< 2e^{n(R-r)} e_0.
\end{aligned}$$

where $\sup_{\theta,g} D(g, f_\theta) = R < \infty$ and $\inf_{\theta,g} D(g, f_\theta) = r > -\infty$ and (A.2) follows from the assumption (13). □

C. SIMULATION STUDIES

C.1 Gaussian and Gamma Distributions – The i.i.d. Case

We undertook a simulation study for the normal mean example in Figure 1 to examine the efficiency and robustness of Hellinger and Negative-Exponential posterior samples. 2,500 sample data sets of size 20 from a $N(1, 1)$ population were generated. For each sample data set, a random

walk Metropolis algorithm was run for 100,000 steps using a $N(0, 0.5)$ proposal distribution and a $N(0, 1)$ prior. 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` (original by Matt Wand. R port by Brian Ripley., 2009) and `LearnBayes` (Albert, 2008) packages in `R`. Expected *a posteriori* estimates for the sample mean were obtained along with 95% credible intervals from every 10th sample in the second half of the MCMC chain. An outlier was then added to each data set taking the value 20 and the estimate re-computed. The analytic posterior without the outlier is normal with mean 0.9524 (equivalently, bias of -0.048) and variance 0.0476. The results of this simulation are summarized in Table 1. As can be expected, the standard Bayesian posterior suffers from sensitivity to large values; it's theoretical bias when the outlier is added is 0.8182. However, both the Hellinger and the negative exponential posterior estimates remained nearly unchanged by the additional data point. For a data set of this size, there is approximately a 10% increase in variance for both robust estimates relative to the standard posterior.

The problem of estimating a Gaussian mean is made relatively straightforward by the symmetry of the distribution. We therefore conducted a further study, estimating both shape and scale parameters in a log-gamma distribution. In this case, the shape parameter was chosen at 5 and the scale parameter at 0.25 and these were given χ^2 priors with degrees of freedom 3 and 0.3 respectively. 5,000 data sets were simulated of 20 points each and the D-posteriors were calculated as above both with and without an outlier placed at ln(20). For this chain a random walk Metropolis algorithm was again conducted with the random walk on the log shape and log scale parameters again using the `LearnBayes` package. Table 2 reports the tabulated results. We note in particular that the efficiency of the H-posterior has been considerably reduced, as have its coverage properties; additionally two simulation runs were removed due to poor convergence. This is explained as being due to the inlier effect; the skewness of the gamma distribution produces density estimates g_n that tend to have "holes" and the Hellinger disparity is sensitive to these; an example is give in Figure 5. By contrast, the negative exponential disparity is much less sensitive.

C.2 One-Way Random Effects Models

Figure 6 demonstrates the differences resulting from robustifying different distributional assumptions in the one-way random effects model described in Section 6. We simulated a set of five

No Outliers			
	Bias	Variance	Coverage
Posterior	-0.047	0.045	0.9506
Hellinger	-0.052	0.050	0.9434
Negative Exponential	-0.066	0.048	0.9684
Outlier at 20			
	Bias	Variance	Coverage
Posterior	0.818	0.041	0.0276
Hellinger	-0.050	0.050	0.9362
Negative Exponential	-0.063	0.049	0.9634

Table 1: A simulation study for a normal mean using the usual posterior, the Hellinger posterior and the Negative Exponential posterior. Columns give the bias and variance of the posterior mean and the coverage of the central 95% credible interval based on 2,500 simulations. Note that the same data sets are used in for both tables, an outlier being added to the data sets when calculating the quantities in the lower table.

No Outliers						
	Shape			Scale		
	Bias	Variance	Coverage	Bias	Variance	Coverage
Posterior	-0.088	0.571	0.9516	0.004	0.0005	0.9562
Hellinger	-0.081	1.005	0.850	0.010	0.0011	0.839
Negative Exponential	-0.182	0.739	0.9436	0.008	0.0007	0.9556
Outlier at 20						
	Shape			Scale		
	Bias	Variance	Coverage	Bias	Variance	Coverage
Posterior	-3.068	0.001	0	0.988	0.00006	0
Hellinger	-0.013	1.046	0.8508	0.010	0.0011	0.8440
Negative Exponential	-0.210	0.725	0.9496	0.010	0.0008	0.9586

Table 2: A simulation study for a log gamma using the usual posterior, the Hellinger posterior and the Negative Exponential posterior. Columns give the bias and variance of the posterior mean and the coverage of the central 95% credible interval based on 5000 simulations. Note that the same data sets are used in for both tables, an outlier being added to the data sets when calculating the quantities in the bottom table. The shape parameter is given in the first column and the scale parameter in the second in each entry.

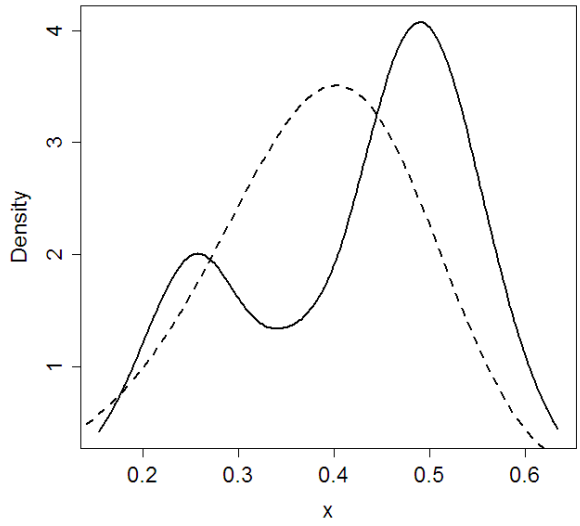

Figure 5: Comparisons of $g_n(x)$ (dashed) and the true $\log \Gamma(5, 0.25)$ density generating the data (solid). Hellinger distance estimators are sensitive to valleys in the density that are due to a density estimate naively applied to skewed data. Negative exponential disparity are robust towards these effects as well.

groups of twenty observations. Each group had variance 0.2 with a mean drawn from a $N(\mu, 1)$ population. This produces a random-effects model and the goal is to estimate μ. The plots in Figure 6 show that the use of disparity methods in either the random effect or on the residual process or on both provide very similar distributions to the correct posterior. When an outlying group is added with mean 40, those methods that replace the random effects distribution with a disparity are unaffected while those that do not are substantially biased.

In order to verify the apparent success of this method, we conducted a simulation study of a one-way random effects model with ten random effects and five observations per random effect. The random effects were simulated from a standard normal distribution, while the observations were Gaussian, centered on the random effect and with standard deviation 0.2. We simulated 1,000 versions of these data. For each version a random-walk Metropolis scheme was run to sample from the posterior, the posterior with the observation likelihood replaced by H-posterior, the posterior with the random effect likelihood replaced by H-posterior and the posterior with both replacements. All MCMC schemes were run for 10,000 steps with posterior distributions calculated based on every 5th sample from the second half of the chain. We additionally added a further random effect with five observations distributed around the value 40 with standard deviation 0.2. Bandwidth parameters were chosen by the selection criterion of Sheather and Jones (1991) based on maximum-likelihood estimates of random effects and residual errors. The results of this simulation are summarized in Table 3. Here we see that the estimation of σ is biased downwards by the estimation of a conditional density for each random effect, based on only five observations and there is more uncertainty in the estimate of τ when Hellinger distance is used in place of the random effect log likelihood. The disparity-based methods otherwise perform very similarly to the true likelihood. When an outlying random effect is added, replacing the random effect likelihood with Hellinger distance robustified inference, where those posteriors without this replacement were severely biased.

C.3 Linear Regression Models

Figure 7 provides example D-posterior distributions of all regression disparities described in Sections 5 and 7 along with the posterior for an example 3-dimensional linear regression based on 30 points. Both Hellinger and negative exponential disparities were used. Covariates were generated from a standard normal distribution with errors also generated from a standard normal

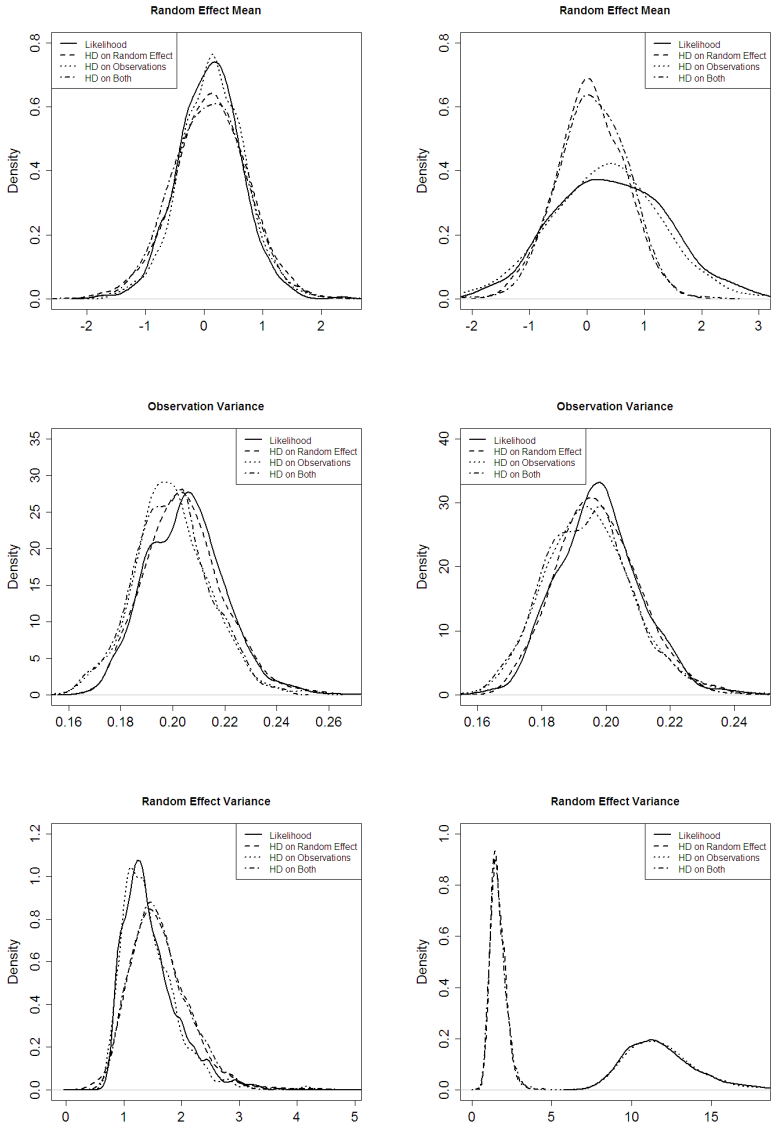

Figure 6: Posterior distributions for a random-effects model. 5 groups of 20 observations each with mean drawn from $N(\mu, 1)$ and variance 0.2. μ was given a $N(0, 1)$ prior. Posterior densities for μ were estimated from every 50th observation in the last half of a random-walk Markov chain run for 100,000 steps. Hellinger distance was used to replace the component of the likelihood representing $Y_{ij} - X_i$, representing X_i and both. Left: posterior densities for μ from the original data when the model is correct. Right: posterior densities after adding a further group of 5 observation generated from $N(40, 0.2)$. First row: estimates for μ, second: σ^2, third: τ^2.

No Outliers									
	μ	sd(μ)	coverage	σ	sd(σ)	coverage	τ	sd(τ)	coverage
Likelihood	0.0013	0.286	0.947	0.200	0.0230	0.940	0.979	0.234	0.919
HD - obs	0.0035	0.285	0.937	0.200	0.0231	0.939	1.070	0.342	0.900
HD - rand	0.0021	0.286	0.936	0.191	0.0259	0.686	0.982	0.231	0.928
HD - both	0.0024	0.291	0.933	0.191	0.0258	0.692	1.068	0.342	0.899
Outlying Random Effect									
	μ	sd(μ)	coverage	σ	sd(σ)	coverage	τ	sd(τ)	coverage
Likelihood	0.365	0.197	1.000	0.200	0.0222	0.930	10.11	0.163	0.000
HD - obs	0.002	0.288	0.926	0.200	0.0223	0.922	1.088	0.389	0.887
HD - rand	0.353	0.252	1.000	0.191	0.0249	0.684	10.13	0.196	0.000
HD - both	0.002	0.295	0.917	0.191	0.0249	0.674	1.066	0.325	0.885

Table 3: A simulation study for a one-way random effect model from using the posterior, replacing the observation likelihood with a conditional Hellinger distance, replacing the random effect likelihood with Hellinger distance and making both replacements. The columns give the mean, standard deviation and coverage of μ, σ^2 and τ^2 based on 1,000 simulations. The lower table indicates the effect of adding an outlying random effect at 40. A total of 8 simulations were excluded due to poor convergence of the MCMC chain.

distribution. The likelihood is noticeably more concentrated than the disparity-based posteriors, but all exhibit broadly similar shapes.

We have supplemented this experiment with a simulation. 1,000 data sets were simulated from a linear regression model on three covariates. The covariates were chosen from an independent standard normal distribution and were held fixed across all simulations. The parameters in the model were chosen as $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2) = (1, 1, 1, 1, 1)$. Bandwidths were chosen using the criterion in Sheather and Jones (1991) based on the observed covariates and maximum likelihood estimates of the residuals. Both Hellinger distance and negative exponential disparity were considered, and the conditional formulation, marginal formulation and conditional-homoscedastic form were used to estimate the five parameters in the model. This resulted in seven estimators including a Gaussian. For each estimate a random-walk Metropolis algorithm was run for 10,000 steps and EDAP estimates were calculated from every fifth sample in the second half of the chain. The results from this study are given in Table 4. Here the conditional form of the Hellinger distance performs poorly and this can be attributed to an under-smoothed conditional density. Other that this anomaly, we observe good agreement between all disparity-based methods and the likelihood. The choice between these will therefore depend on the amount of data available and the dimension of the covariate space.

D. DATA

Table 5 provides the values of the parasitology data set used in Section 8.1. The data used for the class survey data in Section 8.2 are provided in Table 6; they are graphed in Figure 8.

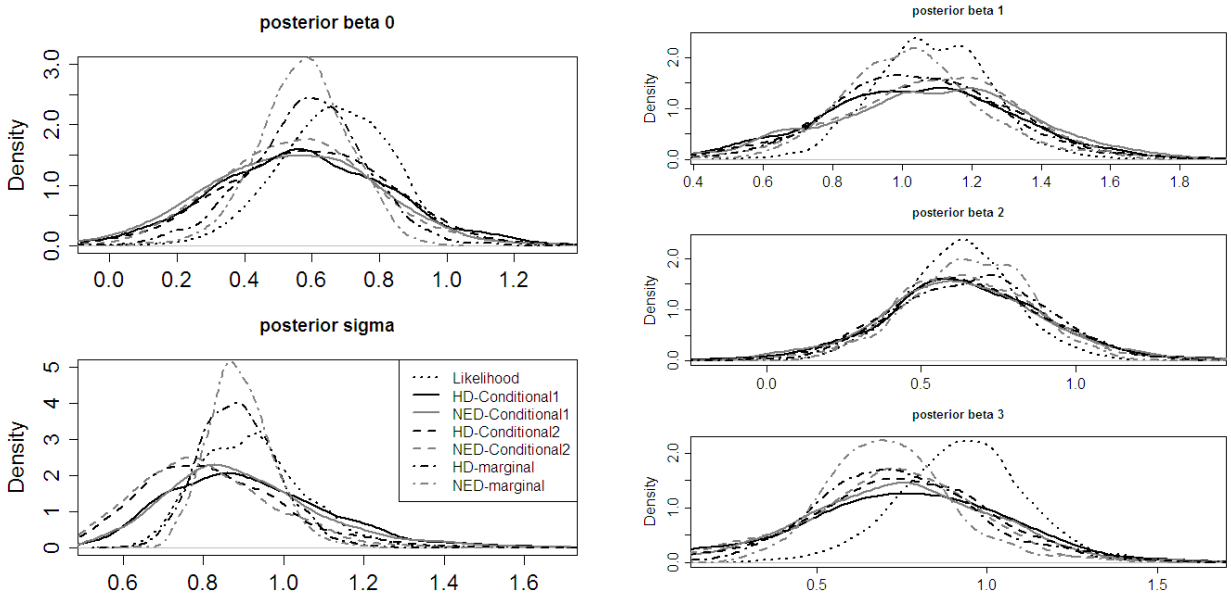

Figure 7: D-posterior inference and linear regression. Top left: posterior for β_0, bottom left: posterior for σ. Right: posterior for each β_i, $i = 1, \ldots, 3$. Thick lines: posterior based on likelihood. Solid lines: based on g_n^c, dashed: based on g_n^{c2}, dotted: marginal formulation. Black: Hellinger distance, grey: negative exponential disparity. Here 'Conditional 1' indicates the conditional density estimate in Section 5, 'Conditional 2' refers to the conditional-homoscedastic approach.

	σ		β_0		β_1		β_1		β_3	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Likelihood	1.035	0.141	0.996	0.193	0.996	0.176	0.988	0.215	0.996	0.183
HD cond. 1	1.986	2.046	0.966	0.703	1.046	1.091	0.872	1.527	0.852	1.613
NED cond. 1	1.127	0.174	0.989	0.211	1.002	0.206	0.972	0.255	0.976	0.244
HD cond. 2	0.980	0.135	0.969	0.194	0.955	0.164	0.855	0.194	0.883	0.163
NED cond. 2	1.010	0.138	0.964	0.193	0.953	0.167	0.857	0.198	0.884	0.166
HD marg.	1.074	0.137	0.993	0.199	0.999	0.189	0.981	0.229	0.989	0.199
NED marg.	1.083	0.139	0.994	0.205	0.997	0.195	0.986	0.237	0.990	0.208

Table 4: Simulation results for linear regression. Columns give mean and standard deviation of EDAP estimates of parameters based on 1,000 simulated data sets, rows correspond to posteriors using likelihood, using conditional density estimates with Hellinger distance and negative exponential disparity (cond .1), using conditional-homoscedastic density estimates with Hellinger distance and negative exponential disparity (cond. 2) and using the marginal formulation of Hellinger distance and negative exponential disparity.

Horse	1	2	3	4	5	6	7
Pre-treatment	2440	1000	1900	1820	3260	300	660
Post-treatment	580	320	400	160	60	40	120

Table 5: Data used in Section 8.1: pre- and post-treatment fecal egg count for seven horses on one farm.

Status	35	45	55	65
a	11.51293	11.775290	11.98293	12.10071
a	11.91839	12.206073	12.61154	12.20607
a	11.69525	12.100712	12.42922	12.42922
a	11.69525	12.100712	12.38839	12.79386
a	10.71442	10.819778	10.81978	10.81978
a	11.15625	11.512925	11.84940	11.98293
a	10.51867	10.596635	10.71442	11.00210
a	12.20607	12.206073	12.20607	12.20607
a	9.21034	9.903488	10.12663	10.30895
a	10.59663	11.002100	11.28978	11.40756
a	14.22098	12.100712	14.91412	15.60727
f	11.51293	11.918391	11.51293	10.81978
f	11.00210	11.002100	11.15625	11.15625
f	11.91839	12.206073	12.42922	12.42922
f	11.40756	11.695247	11.73607	11.77529
f	11.91839	13.122363	13.30468	13.30468
f	11.51293	11.695247	11.91839	12.20607
f	11.15625	11.512925	11.51293	11.69525
f	12.20607	12.611538	12.76569	11.00210
f	10.81978	11.002100	11.00210	11.00210
f	10.46310	11.002100	11.08214	11.08214

Table 6: Class survey data used in Section 8.2; columns give American (a) or foreign (f) status, and log expected salary at ages 35, 45, 55, and 65.

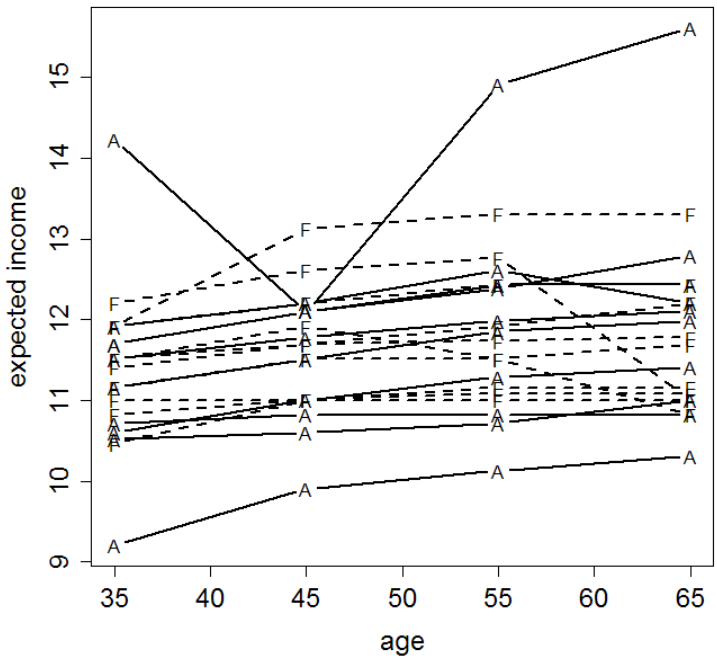

Figure 8: Responses to an in-class survey on expected income at ages 35, 45, 55 and 65. Students were either foreign born (dashed) or American (solid).