University of Missouri, St. Louis

IRL @ UMSL

12-1-2018

Asymptotic properties for methods combining the minimum Hellinger distance estimate and the bayesian nonparametric density estimate

Yuefeng Wu
University of Missouri-St. Louis, Department of Mathematics and Computer Science, wuyue@umsl.edu

Giles Hooker
Cornell University

Follow this and additional works at: https://irl.umsl.edu/mathstats-faculty

December 11, 2018

Asymptotic Properties for Methods Combining the Minimum Hellinger Distance Estimate and the Bayesian Nonparametric Density Estimate

Yuefeng Wu, *University of Missouri-St. Louis*
Giles Hooker, *Cornell University*

Asymptotic Properties for Methods Combining the Minimum Hellinger Distance Estimate and the Bayesian Nonparametric Density Estimate

Yuefeng Wu [1,*,†] **and Giles Hooker** [2]

[1] Department of Mathematics and Computer Science, University of Missouri Saint Louis, St. Louis, MO 63121, USA

[2] Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA; gjh27@cornell.edu

* Correspondence: wuyue@umsl.edu; Tel.: +1-314-516-6348

† Current address: ESH 320, One University Blvd. Saint Louis, MO 63121, USA.

Abstract: In frequentist inference, minimizing the Hellinger distance between a kernel density estimate and a parametric family produces estimators that are both robust to outliers and statistically efficient when the parametric family contains the data-generating distribution. This paper seeks to extend these results to the use of nonparametric Bayesian density estimators within disparity methods. We propose two estimators: one replaces the kernel density estimator with the expected posterior density using a random histogram prior; the other transforms the posterior over densities into a posterior over parameters through minimizing the Hellinger distance for each density. We show that it is possible to adapt the mathematical machinery of efficient influence functions from semiparametric models to demonstrate that both our estimators are efficient in the sense of achieving the Cramér-Rao lower bound. We further demonstrate a Bernstein-von-Mises result for our second estimator, indicating that its posterior is asymptotically Gaussian. In addition, the robustness properties of classical minimum Hellinger distance estimators continue to hold.

Keywords: robustness; efficiency; Bayesian nonparametric; Bayesian semi-parametric; asymptotic property; minimum disparity methods; Hellinger distance; Berstein von Mises theorem

1. Introduction

This paper develops Bayesian analogs of minimum Hellinger distance methods. In particular, we aim to produce methods that enable a Bayesian analysis to be both robust to unusual values in the data and to retain their asymptotic precision when a proposed parametric model is correct.

All statistical models include assumptions which may or may not be true of the mechanisms producing a given data set. Robustness is a desired property in which a statistical procedure is relatively insensitive to deviations from these assumptions. For frequentist inference, concerns are largely associated with distributional robustness: the shape of the true underlying distribution deviates slightly from the assumed model. Usually, this deviation represents the situation where there are some outliers in the observed data set; see [1] for example. For Bayesian procedures, the deviations may come from the model, prior distribution, or utility function, or some combination thereof. Much of the literature on Bayesian robustness has been concerned with the prior distribution or utility function. By contrast, the focus of this paper is robustness with respect to outliers in a Bayesian context, a relatively understudied form of robustness for Bayesian models. For example, we know that Bayesian models with heavy tailed data distributions are robust with respect to outliers for the case of one single location parameter estimated by many observations. However, as a consequence of the Crámer–Rao lower

bound and the efficiency of the MLE, modifying likelihoods to account for outliers will usually result in a loss of precision in parameter estimates when they are not necessary. The methods we propose, and the study of their robustness properties, will provide an alternative means of making any i.i.d. data distribution robust to outliers that do not lose efficiency when no outliers are present. We speculate that they can be extended beyond i.i.d. data as in [2], but we do not pursue this here.

Suppose we are given the task of estimating $\theta_0 \in \Theta$ from independent and identically distributed univariate random variables $X_1, \ldots, X_n$, where we assume each X_i has density $f_{\theta_0} \in \mathscr{F} = \{f_\theta : \theta \in \Theta\}$. Within the frequentist literature, minimum Hellinger distance estimates proceed by first estimating a kernel density $\hat{g}_n(x)$ and then choosing θ to minimize the Hellinger distance $h(f_\theta, g_n) = [\int \{f_\theta^{1/2}(x) - \hat{g}_n^{1/2}(x)\}^2 dx]^{1/2}$. The minimum Hellinger distance estimator was shown in [3] to have the remarkable properties of being both robust to outliers and statistically efficient, in the sense of asymptotically attaining the information bound, when the data are generated from f_{θ_0}. These methods have been generalized to a class of minimum disparity estimators, based on alternative measures of the difference between a kernel density estimate and a parametric model, which have been studied since then, e.g., [4–8]. While some adaptive M-estimators can be shown to retain both robustness and efficiency, e.g., [9], minimum disparity methods are the only generic methods we are aware of that retain both properties and can also be readily employed within a Bayesian context. In this paper, we only consider Hellinger distance in order to simplify the mathematical exposition; the extension to more general disparity methods can be made following similar developments to those in [5,7].

Recent methodology proposed in [2] suggested the use of disparity-based methods within Bayesian inference via the construction of a "disparity likelihood" by replacing the likelihood function when calculating the Bayesian posterior distribution; they demonstrated that the resulting expected *a posteriori* estimators retain the frequentist properties studied above. These methods first obtain kernel density estimates from data and then calculate the disparity between the estimated density function and the corresponding density functions in the parametric family.

In this paper, we propose the use of Bayesian non-parametric methods instead of the classical kernel methods in applying the minimum Hellinger distance method. One method we proposed is just to replace the kernel density estimate used in classical minimum Hellinger distance estimate by the Bayesian nonparametric expected *a posteriori* density, which we denote by MHB (minimum Hellinger distance method using a Bayesian nonparametric density estimate). The second method combines the minimum Hellinger distance estimate with the Bayesian nonparametric posterior to give a posterior distribution of the parameter of interest. This latter method is our main focus. We show that it is more robust than usual Bayesian methods and demonstrate that it retains asymptotic efficiency, hence the precision of the estimate is maintained. So far as we are aware, this is the first Bayesian method that can be applied generically and retain both robustness and (asymptotic) efficiency. We denote it by BHM (Bayesian inference using a minimum Hellinger distance).

To study the properties of the proposed new methods, we treat both MHB and BMH as special cases of semi-parametric models. The general form of a semi-parametric model has a natural parametrization $(\theta, \eta) \mapsto P_{\theta,\eta}$, where $\theta \in \Theta$ is a Euclidean parameter and $\eta \in H$ belongs to an infinite-dimensional set. For such models, θ is the parameter of primary interest, while η is a nuisance parameter. Asymptotic properties of some of Bayesian semi-parametric models have been discussed in [10]. Our disparity based methods involve parameters in Euclidean space and Hilbert space, with the former being of most interest. However, unlike many semi-parametric models in which $P_{\theta,\eta} \in \mathcal{P}$ is specified jointly by θ and η, in our case, the finite dimensional parameter and the nonparametric density functions are parallel specifications of the data distribution. Therefore, standard methods to study asymptotic properties of semi-parametric models will not apply to the study of disparity-based methods. Nevertheless, considering the problem of estimating $\psi(P)$ of some function $\psi : \mathcal{P} \mapsto \mathbb{R}^d$, where $\mathcal{P}$ is the space of the probability models P, semi-parametric models and disparity-based methods can be unified into one framework.

The MHB and BMH methods are introduced in detail in Section 2, where we also discuss some related concepts and results, such as tangent sets, information, consistency, and the specific nonparametric prior that we employ. In Section 3, both MHB and BMH are shown to be efficient, in the sense that asymptotically the variance of the estimate achieves the lower bound of the Cramér–Rao theorem. For MHB, we show that asymptotic normality of the estimate holds, where the asymptotic variance is the inverse of the Fisher information. For BMH, we show that the Bernstein von Mises (BvM) theorem holds. The robustness property and further discussion of these two methods are given in Sections 4 and 5, respectively. A broader discussion is given in Section 6.

2. Minimum Hellinger Distance Estimates

Assume that random variables $X_1, \ldots, X_n$ are independent and identically distributed (iid) with density belonging to a specified parametric family $\mathscr{F} = \{f_\theta : \theta \in \Theta\}$, where all the f_θ in the family have the same support, denoted by $supp(f)$. For simplicity, we use $\mathbb{X}_n$ to denote the random variables $X_1, \ldots, X_n$. More flexibly, we model $\mathbb{X}_n \sim g^n$, where g is a probability density function with respect to the Lebesgue measure on $supp(f)$. Let $\mathscr{G}$ denote the collection of all such probability density functions. If the parametric family contains the data-generating distribution, then $g = f_\theta$ for some θ. Formally, we can denote the probability model of the observations in the form of a semi-parametric model $(\theta, g) \mapsto P_{\theta,g}$. We aim at estimating θ and consider g as a nuisance parameter, which is typical of semi-parametric models.

Let π denote a prior on $\mathscr{G}$, and for any measurable subset $B \subset \mathscr{G}$, the posterior probability of $g \in B$ given $\mathbb{X}_n$ is

$$\pi(B \mid \mathbb{X}_n) = \frac{\int_B \prod_{i=1}^n g(X_i)\pi(dg)}{\int_{\mathscr{G}} \prod_{i=1}^n g(X_i)\pi(dg)}.$$

Let $g_n^* = \int g\pi(dg \mid \mathbb{X}_n)$ denote the Bayesian nonparametric expected *a posteriori* estimate. Our first proposed method can be described formally as follows:

MHB: Minimum Hellinger distance estimator with Bayesian nonparametric density estimation:

$$\hat{\theta}_1 = \text{argmin}_{\theta \in \Theta}\, h\left(f_\theta, g_n^*\right). \tag{1}$$

This estimator replaces the kernel density estimate in the classical minimum Hellinger distance method introduced in [3] by the posterior expectation of the density function.

For this method, we will view $\hat{\theta}_1$ as the value at g_n^* of a functional $T : \mathscr{G} \mapsto \Theta$, which is defined via

$$\|f_{T(g)}^{1/2} - g^{1/2}\| = \min_{t \in \Theta} \|f_t^{1/2} - g^{1/2}\| \tag{2}$$

where $\|\cdot\|$ denotes the L_2 metric. We can also write $\hat{\theta}_1$ as $T(g_n^*)$.

In a more general form, what we estimate is the value $\psi(P)$ of some functional $\psi : \mathcal{P} \mapsto \mathbb{R}^d$, where the P stands for the common distribution from which data are generated, and $\mathcal{P}$ is the set of all possible values of P, which also denotes the corresponding probability model. In the setting of minimum Hellinger distance estimation, the model $\mathcal{P}$ is set as $\mathscr{F} \times \mathscr{G}$, P can be specified as $P_{\theta,g}$, and $\psi(P) = \psi(P_{\theta,g}) = \theta$. For the methods we proposed in this paper, we will focus on the functional $T : \mathscr{G} \mapsto \Theta$, for a given $\mathscr{F}$, as defined above. Note that the constraint associated with the family $\mathscr{F}$ is implicitly applied by T.

Using functional T, we can also propose a Bayesian method, which assigns nonparametric prior on the density space and gives inference on the unknown parameter θ of a parametric family as follows:

BMH: Bayesian inference with minimum Hellinger distance estimation:

$$\pi(\theta \mid \mathbb{X}_n) = \pi(T(g) \mid \mathbb{X}_n). \tag{3}$$

A nonparametric prior π on the space $\mathscr{G}$ and the observation $\mathbb{X}_n$ leads to the posterior distribution $\pi(g \mid \mathbb{X}_n)$, which can then be converted to the posterior distribution of the parameter $\theta \in \Theta$ through the functional $T : \mathscr{G} \mapsto \Theta$.

In the following subsections, we discuss properties associated with the functional T as well as the consistency of MHB and BHM, and we provide a detailed example of the random histogram prior that we will employ and its properties that will be used for the discussion of efficiency in Section 2.1.

2.1. Tangent Space and Information

In this subsection, we obtain the efficient influence function of the functional T on the linear span of the tangent set on g_0 and show that the local asymptotic normality (LAN) expansion related to the norm of the efficient influence function attains the Caramér–Rao bound. These results play important roles in showing that BvM holds for the BMH method in the next section.

Estimating the parameter by $T(g)$ under the assumption $g \in \mathscr{G}$ uses less information than estimating this parameter for $g \in \mathscr{G}^* \subset \mathscr{G}$. Hence, the lower bound of the variance of $T(g)$ for $g \in \mathscr{G}$ should be at least the supremum of the lower bounds of all parametric sub-models $\mathscr{G}^* = \{G_\lambda : \lambda \in \Lambda\} \subset \mathscr{G}$.

To use mathematical tools such as functional analysis to study the properties of the proposed methods, we introduce some notations and concepts below. Without loss of generality, we consider one-dimensional sub-models $\mathscr{G}^*$, which pass through the "true" distribution, denoted by G_0 with density function g_0. We say a sub-model indexed by t, $\{g_t : 0 < t < \epsilon\} \subset \mathscr{G}$, is differentiable in quadratic mean at $t = 0$ if we have that, for some measurable function $q : supp(g_0) \mapsto \mathbb{R}$,

$$\int \left[\frac{dG_t^{1/2} - dG_0^{1/2}}{t} - \frac{1}{2} q dG_0^{1/2} \right]^2 \to 0 \tag{4}$$

where G_t is the cumulative distribution function associated with g_t. Functions $q(x)$s are known as the score functions associated with each sub-model. The collection of these score functions, which is called a tangent set of the model $\mathscr{G}$ at g_0 and denoted by $\dot{\mathscr{G}}_{g_0}$, is induced by the collection of all sub-models that are differentiable at g_0.

We say that T is differentiable at g_0 relative to a given tangent set $\dot{\mathscr{G}}_{g_0}$, if there exists a continuous linear map $\dot{T}_{g_0} : L_2(G_0) \mapsto \mathbb{R}$ such that for every $q \in \dot{\mathscr{G}}_{g_0}$ and a sub-model $t \mapsto g_t$ with score function q, there is

$$\frac{T(g_t) - T(g_0)}{t} \to \dot{T}_{g_0} q \tag{5}$$

where $L^2(G_0) = \{q : supp(g_0) \mapsto \mathbb{R}, \int q^2(x) g_0(x) dx < \infty\}$. By the Riesz representation theorem for Hilbert spaces, the map $\dot{T}_{g_0}$ can always be written in the form of an inner product with a fixed vector-valued, measurable function $\tilde{T}_{g_0} : supp(g_0) \mapsto \mathbb{R}$,

$$\dot{T}_{g_0} q = \langle \tilde{T}_{g_0}, q \rangle_{G_0} = \int \tilde{T}_{g_0} q \, dG_0.$$

Let $\tilde{T}_{g_0}$ denote the unique function in $\overline{lin}\dot{\mathscr{G}}_{g_0}$, the closure of the linear span of the tangent set. The function $\tilde{T}_{g_0}$ is the efficient influence function and can be found as the projection of any other "influence function" onto the closed linear span of the tangent set.

For a sub-model $t \mapsto g_t$ whose score function is q, the Fisher information about t at 0 is $G_0 q^2 = \int q^2 dG_0$. In this paper, we use the notation Fg to denote $\int g \, dF$ for a general function g and distribution F. Therefore, the "optimal asymptotic variance" for estimating the functional $t \mapsto T(g_t)$, evaluated at $t = 0$, is greater than or equal to the Caramér–Rao bound

$$\frac{(dT(g_t)/dt)^2}{G_0 q^2} = \frac{\langle \tilde{T}_{g_0}, q \rangle_{G_0}^2}{\langle q, q \rangle_{G_0}}.$$

The supremum of the right-hand side (RHS) of the above expression over all elements of the tangent set is a lower bound for estimating $T(g)$ given model $\mathcal{G}$, if the true model is g_0. The supremum can be expressed in the norm of the efficient influence function $\tilde{T}_{g_0}$ by Lemma 25.19 in [11]. The lemma and its proof is quite neat, and we reproduce it here for the completeness of the argument.

Lemma 1. *Suppose that the functional $T : \mathcal{G} \mapsto \mathbb{R}$ is differentiable at g_0 relative to the tangent set $\dot{\mathcal{G}}_{g_0}$. Then*

$$\sup_{q \in lin\dot{\mathcal{G}}_{g_0}} \frac{\langle \tilde{T}_{g_0}, q \rangle_{G_0}^2}{\langle q, q \rangle_{G_0}} = G_0 \tilde{T}_{g_0}^2.$$

Proof. This is a consequence of the Cauchy–Schwarz inequality $(G_0 \tilde{T}_{g_0} q)^2 \leq G_0 \tilde{T}_{g_0}^2 G_0 q^2$ and the fact that, by definition, the efficient influence function, $\tilde{T}_{g_0}$, is contained in the closure of $lin\dot{\mathcal{G}}_{G_0}$. □

Now we show that functional T is differentiable under some mild conditions and construct its efficient influence function in the following theorem.

Theorem 1. *For the functional T defined in Equation (2), and for $t \in \Theta \subset \mathbb{R}$, let $s_t(x)$ denote $f_\theta^{1/2}(x)$ for $\theta = t$. We assume that there exist $\dot{s}_t(x)$ and $\ddot{s}_t(x)$ both in L_2, such that for α in a neighborhood of zero,*

$$\begin{align}
s_{t+\alpha}(x) &= s_t(x) + \alpha \dot{s}_t(x) + \alpha u_\alpha(x) \tag{6}\\
\dot{s}_{t+\alpha}(x) &= \dot{s}_t(x) + \alpha \ddot{s}_t(x) + \alpha v_\alpha(x), \tag{7}
\end{align}$$

where u_α and v_α converge to zero as $\alpha \to 0$. Assuming $T(g_0) \in int(\Theta)$, the efficient influence function of T is

$$\tilde{T}_{g_0} = \left(-\left[\int \ddot{s}_{T(g_0)}(x) g_0^{\frac{1}{2}}(x) dx \right]^{-1} + a_t \right) \frac{\dot{s}_{T(g_0)}(x)}{2g_0^{\frac{1}{2}}(t)} \tag{8}$$

where a_t converges to 0 as $t \to 0$. In particular, for $g_0 = f_\theta$,

$$\tilde{T}_{f_\theta} = \left(-\left[\int \ddot{s}_\theta(x) s_\theta(x) dx \right]^{-1} + a_t \right) \frac{\dot{s}_\theta(x)}{2s_\theta(x)}. \tag{9}$$

Proof. Let the t-indexed sub-model be

$$g_t := (1 + tq(x)) g_0(x)$$

where $q(x)$ satisfies $\int q(x) g_0(x) dx = 0$ and $q \in L_2(g_0)$. By direct calculation, we see that q is the score function associated with such a sub-model at $t = 0$ in the sense of Equation (4) and thus the collection of q is the maximal tangent set.

By the definition of T, $T(g_0)$ maximizes $\int s_t(x) g_0^{1/2}(x) dx$. From Equation (6), we have that

$$lim_{\alpha \to 0} \alpha^{-1} \int [s_{t+\alpha}(x) - s_t(x)] g_0^{1/2}(x) dx = \int \dot{s}_t(x) g_0^{1/2}(x) dx. \tag{10}$$

Since $T(g_0) \in int(\Theta)$, we have that

$$\int \dot{s}_{T(g_0)}(x) g_0^{1/2}(x) dx = 0. \tag{11}$$

Similarly, $\int \dot{s}_{T(g_t)}(x) g_t^{1/2}(x) dx = 0$. Using Equation (7) to substitute $\dot{s}_{T(g_t)}$, we have that

$$0 = \int [\dot{s}_{T(g_0)}(x) + \ddot{s}_{T(g_0)}(x)(T(g_t) - T(g_0)) + v_t(x)(T(g_t) - T(g_0))] g_t^{1/2}(x) dx$$

where $v_t(x)$ converge in L_2 to zero as $t \to 0$ since $T(g_t) \to T(g_0)$. Thus,

$$
\begin{aligned}
\lim_{t \to 0} \frac{1}{t} & [T(g_t) - T(g_0)] \\
&= -\lim_{t \to 0} \frac{1}{t} \left[\int (\ddot{s}_{T(g_0)}(x) + v_t(x)) g_t^{\frac{1}{2}}(x) dx \right]^{-1} \int \dot{s}_{T(g_0)}(x) g_t^{1/2}(x) dx \\
&= \lim_{t \to 0} \frac{1}{t} \left(-\left[\int (\ddot{s}_{T(g_0)}(x)) g_0^{\frac{1}{2}}(x) dx \right]^{-1} + a_t \right) \int \dot{s}_{T(g_0)}(x) (g_t^{\frac{1}{2}}(x) - g_0^{\frac{1}{2}}(x)) dx \\
&= \left(-\left[\int (\ddot{s}_{T(g_0)}(x)) g_0^{\frac{1}{2}}(x) dx \right]^{-1} + a_t \right) \int \frac{\dot{s}_{T(g_0)}(x)}{2 g_0^{\frac{1}{2}}(x)} q(x) g_0(x) dx.
\end{aligned}
$$

Since by the definition of $\tilde{T}$, which requires $\int \tilde{T}_{g_0} g_0(x) dx = 0$, we have that

$$
\begin{aligned}
\tilde{T}_{g_0} &= \left(-\left[\int \ddot{s}_{T(g_0)}(x) g_0^{\frac{1}{2}}(x) dx \right]^{-1} + a_t \right) \left(\frac{\dot{s}_{T(g_0)}(x)}{2 g_0^{\frac{1}{2}}(x)} - \int \frac{\dot{s}_{T(g_0)}(x)}{2} g_0^{\frac{1}{2}}(x) dx \right) \\
&= \left(-\left[\int \ddot{s}_{T(g_0)}(x) g_0^{\frac{1}{2}}(x) dx \right]^{-1} + a_t \right) \frac{\dot{s}_{T(g_0)}(x)}{2 g_0^{\frac{1}{2}}(x)}.
\end{aligned}
$$

By the same argument we can show that, when $g_0 = f_\theta$, Equation (9) holds. $\square$

Some relatively accessible conditions under which Equations (6) and (7) hold are given by Lemmas 1 and 2 in [3]. We do not repeat them here.

Now we can expand T at g_0 as

$$
T(g) - T(g_0) = \langle \frac{g - g_0}{g_0}, \tilde{T}_{g_0} \rangle_{G_0} + \tilde{r}(g, g_0) \tag{12}
$$

where $\tilde{T}$ is given in Theorem 1 and $\tilde{r} = 0$.

2.2. Consistency of MHB and BMH

Since $T(g)$ may have more than one value, the notation $T(g)$ is used to denote any arbitrary one of the possible values. In [3], the existence, continuity in Hellinger distance, and uniqueness of functional T are ensured under the following condition:

A1 (i) Θ is compact, (ii) $\theta_1 \neq \theta_2$ implies $f_{\theta_1} \neq f_{\theta_2}$ on a set of positive Lebesgue measures, and (iii), for almost every x, $f_\theta(x)$ is continuous in θ.

When a Bayesian nonparametric density estimator is used, we assume the posterior consistency:
A2 For any given $\epsilon > 0$, $\pi\{g : h(g, f_{\theta_0}) > \epsilon \mid \mathbb{X}_n\} \to 0$ in probability.

Under Conditions A1 and A2, consistency holds for MHB and BMH.

Theorem 2. *Suppose that Conditions A1 and A2 hold, then*

1. $\|g_n^{*1/2} - f_{\theta_0}^{1/2}\|^2 \to 0$ *in probability,* $T(g_n^*) \to T(f_{\theta_0})$ *in probability, and hence* $\hat{\theta}_1 \to \theta_0$ *in probability;*
2. *For any given $\epsilon > 0$,* $\pi(|\theta - \theta_0| > \epsilon \mid \mathbb{X}_n) \to 0$ *in probability.*

Proof. Part 1: To show that $\|g_n^{*1/2} - f_{\theta_0}^{1/2}\|^2 \to 0$ in probability, which is equivalent to showing that $\int \left(\int g\pi(dg \mid \mathbb{X}_n)^{1/2} - f_{\theta_0}^{1/2} \right)^2 dx \to 0$ in probability, it is sufficient to show that $\int |\int g\pi(dg \mid \mathbb{X}_n) - f_{\theta_0}| dx \to 0$ in probability, since $h^2(f,g) \le \|f-g\|_1$. We have that

$$
\begin{aligned}
\int \left| \int g\pi(dg \mid \mathbb{X}_n) - f_{\theta_0} \right| dx &= \int \left| \int (g - f_{\theta_0})\pi(dg \mid \mathbb{X}_n) \right| dx \\
&\le \int \int |g - f_{\theta_0}| \, \pi(dg \mid \mathbb{X}_n) dx \\
&= \int \int |g - f_{\theta_0}| \, dx \pi(dg \mid \mathbb{X}_n) \\
&\le \int \sqrt{2} h(g, f_{\theta_0}) \pi(dg \mid \mathbb{X}_n).
\end{aligned}
$$

Note that the change of order of integration is due to Fubini's theorem and the last inequality is due to $\|f - g\|_1 \le \sqrt{2} h(f,g)$. Split the integral on the right-hand side of the above expression into two parts:

$$
\int_{\mathscr{A}} \sqrt{2} h(g, f_{\theta_0}) \pi(dg \mid \mathbb{X}_n) + \int_{\mathscr{A}^c} \sqrt{2} h(g, f_{\theta_0}) \pi(dg \mid \mathbb{X}_n)
$$

where $\mathscr{A} = \{g : h(g, f_{\theta_0}) \le \epsilon\}$ for any given $\epsilon > 0$. The first term is bounded by ϵ by construction. By Condition A1, the posterior of measure of $\mathscr{A}^c$ to 0 in probability as $n \to \infty$. Since Hellinger distance is bounded by 2, so does the second term above. This completes the proof for $\|g_n^{*1/2} - f_{\theta_0}^{1/2}\|^2 \to 0$ in probability.

To show $T(g_n^*) \to T(f_{\theta_0})$ and $\hat{\theta}_1 \to \theta_0$ in probability, we need that the functional T is continuous and unique at f_{θ_0}, which is proved by Theorem 1 in [3] under Condition A1.

Part 2: By Condition A1 and Theorem 1 in [3], the functional T is continuous and unique at f_{θ_0}. Hence, for any given $\epsilon > 0$, there exist $\delta > 0$ such that $|T(g) - T(f_{\theta_0})| < \epsilon$ when $h(g, f_{\theta_0}) < \delta$. By Condition A2, we have that $\pi(h(g, f_{\theta_0}) < \delta) \to 1$, which implies that $\pi(|\theta - \theta_0| < \epsilon) \to 1$ in probability. $\square$

It should be noted that, if we change the ϵ in Condition A2 to ϵ_n, a sequence converging to 0, then we can apply the results for the concentration rate of the Bayesian nonparametric density estimation here. However, such an approach cannot lead to the general "efficiency" claim, no matter in the form of rate of concentration or asymptotic normality. There are two reasons for this. First, the rate of concentration for Bayesian nonparametric posterior is about $n^{-2/5}$ for a rather general situation and $(\log n)^a \times n^{-1/2}$, where $a > 0$, for some special cases (see [12–14]). This concentration rate is not sufficient in many situations to directly imply that the concentration of the corresponding parametric estimates achieves the lower bound of the variance given in the Cramér–Rao theorem. Second, the Hellinger distances between pairs of densities as functions of parameters vary among different parametric families. Therefore, obtaining the rate of concentration in parameters from the rate of convergence in the densities cannot be generally applied to different distribution families.

It should also be noted that, although Θ is required to be compact in Condition A1, Theorem 2 is useful for a Θ that is not compact, as long as the parametric family $f_\theta : \theta \in \Theta$ can be re-parameterized where the space of new parameters can be embedded within a compact set. An example of re-parameterizing a general location-scale family with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ to a family with parameters $t_1 = tan^{-1}(\mu)$ and $t_2 = tan^{-1}(\sigma)$, where $\Theta_{(t_1,t_2)} = (-\pi/2, \pi/2) \times (0, \pi/2)$ and $\Theta \subset \bar{\Theta} = [-\pi/2, \pi/2] \times [0, \pi/2]$, is discussed in [3], and the conclusions of Theorem 1 in [3] is still valid for a location-scale family. Therefore, Theorem 2 remains valid for the same type of the families, whose parameter space may not be compact and for the same reasons; the compactness requirement stated in the theorem is mainly for mathematical simplicity.

2.3. Prior on Density Functions

We introduce a random histogram as an example for priors used in Bayesian nonparametric density estimation. It can be seen as a simplified version of a Dirichlet process mixture (DPM) prior, which is commonly used in practice. Both DPM and random histogram are mixture densities. While DPM uses a Dirichlet process to model the weights within an infinite mixture of kernels, the random histogram prior only has a finite number of components. Another difference is that, although we specify the form of the kernel function for DPM, the kernel function could be any density function in general, while the random histogram uses only the uniform density as its mixing kernel. Nevertheless, the limit on the finite number of the mixing components is not that important in practice, since the Dirichlet process will always be truncated in computation. In the next section, we will verify that the random histogram satisfies the conditions that are needed for our proposed methods to be efficient. On the other hand, although we believe that DPM should also lead to efficiency, the authors are unaware of the theoretical results or tools required to prove it. This is mostly due to the flexibility of DPM, which in turn significantly increases the mathematical complexity of the analysis.

For any $k \in \mathbb{N}$, denote the set of all regular k bin histograms on $[0, 1]$ by $\mathscr{H}_k = \{f \in L^2([0,1]) : m(x) = \sum_{j=1}^{k} f_j \mathbb{1}_{I_j}(x), f_j \in \mathbb{R}, j = 1, \ldots, k\}$, where $I_j = [(j-1)/k, j/k)$. Denote the unit simplex in $\mathbb{R}^k$ by $\mathscr{S}_k = \{\omega \in [0,1]^k : \sum_{j=1}^{k} \omega_j = 1\}$. The subset of $\mathscr{H}_k$, $\mathscr{H}_k^1 = \{f \in L^2(\mathbb{R}), f(x) = f_{\omega,k} = k \cdot \sum_{j=1}^{k} \omega_j \mathbb{1}_{I_j}(x), (\omega_1, \ldots, \omega_k) \in \mathscr{S}_k\}$, denotes the collection of densities on $[0, 1]$ in the form of a histogram.

The set $\mathscr{H}_k$ is a closed subset of $L_2[0, 1]$. For any function $f \in L_2[0, 1]$, denote its projection in the L_2 sense on $\mathscr{H}_k$ by $f_{[k]}$, where $f_{[k]} = k \sum_{j=1}^{k} \mathbb{1}_{I_j} \int_{I_j} f$.

We assign priors on $\mathscr{H}_k^1$ via k and $(\omega_1, \ldots, \omega_k)$ for each k. A degenerate case is to let $k = K_n = o(n)$. Otherwise, let p_k be a distribution on positive integers, where

$$k \sim p_k, e^{-b_1 k \log(k)} \leq p_k(k) \leq e^{-b_2 k \log(k)} \tag{13}$$

for all k large enough and some $0 < b_1 < b_2 < \infty$. For example, Condition (13) is satisfied by the Poisson distribution, which is commonly used in Bayesian nonparametric models.

Conditionally on k, we consider a Dirichlet prior on $\omega = \{\omega_1, \ldots, \omega_k\}$:

$$\omega \sim \mathscr{D}(\alpha_{1,k}, \ldots, \alpha_{k,k}), \qquad c_1 k^{-a} \leq \alpha_{j,k} \leq c_2 \tag{14}$$

for some fixed constants $a, c_1, c_2 > 0$ and any $1 \leq j \leq k$. For posterior consistency, we need the following condition:

$$\sup_{k \in \mathscr{K}_n} \sum_{j=1}^{k} \alpha_{j,k} = o(\sqrt{n}) \tag{15}$$

where $\mathscr{K}_n \subset \{1, 2, \ldots, \lfloor n/(\log n)^2 \rfloor\}$.

The consistency result of this prior is given by Proposition 1 in the supplement to [15]. For $n \geq 2, k \geq 1, M > 0$, let

$$A_{n,k}(M) = \{g \in \mathscr{H}_k^1, h(g, g_{0,[k]})\} < M\epsilon_{n,k} \tag{16}$$

where $\epsilon_{n,k}^2 = k \log n / n$ denote a neighborhood of $g_{0,[k]}$, and we have that

- (a) there exist $c, M > 0$ such that

$$P_0 \left[\exists k \leq \frac{n}{\log n}; \pi[g \notin A_{n,k}(M) \mid \mathbb{X}_n, k] > e^{-ck \log n} \right] = o(1). \tag{17}$$

- (b) Suppose $g_0 \in \mathscr{C}^\beta$ with $0 < \beta \leq 1$, if $k_n(\beta) = (n \log n)^{1/(2\beta+1)}$ and $\epsilon_n(\beta) = k_n(\beta)^{-\beta}$, then, for k_1 and a sufficiently large M,

$$\pi[h(g_0, g) \leq M\epsilon_n(\beta); k \leq k_1 k_n(\beta) \mid \mathbb{X}_n] = 1 + o_p(1), \tag{18}$$

  where $\mathscr{C}^\beta$ denotes the class of β-Hölder functions on $[0, 1]$.

This means that the posterior of the density function concentrates around the projection $g_{0[k]}$ of g_0 and around g_0 itself in terms of the Hellinger distance. We can easily conclude that $\pi(\mathscr{K}_n \mid \mathbb{X}_n) = 1 + o(1)$ from Equation (18) for $g_0 \in \mathscr{C}^\beta$.

It should be noted that, although the priors we defined above are on the densities on $[0, 1]$, this is for mathematical simplicity, which could easily be extended to the space of probability densities on any given compact set. Further, transformations of $\mathbb{X}_n$, similar to those discussed at the end of Section 2.2, can extend the analysis to the real line (refer to [3,16] for more example and details).

3. Efficiency

We say that both MHB and BMH methods are efficient if the lower bound of the variance of the estimate, in the sense of Cramèr and Rao's theorem, is achieved.

3.1. Asymptotic Normality of MHB

Consider the maximal tangent set at g_0, which is defined as $\mathcal{H}_T = \{q \in L^2(g_0), \int qg_0 = 0\}$. Denote the inner product on $\mathcal{H}_T$ by $\langle q_1, q_2 \rangle_L = \int q_1 q_2 g_0$, which induces the L-norm as

$$\|g\|_L^2 = \int_0^1 (g - G_0 g)^2 g_0. \tag{19}$$

Note that the inner product $\langle \cdot, \cdot \rangle_L$ is equivalent to the inner product introduced in Section 2.1, and the induced L-norm corresponds to the local asymptotic normality (LAN) expansion. Refer to [17] and Theorem 25.14 in [11] for more details.

With functional T and priors on g defined in the previous section, Theorem 3 shows that the MHB method is efficient when the parametric family contains the true model.

Theorem 3. *Let two priors π_1 and π_2 be defined by Equations (13)–(14) and let a prior on k be either a Dirac mass at $k = K_n = n^{1/2}(\log n)^{-2}$ for π_1 or $k \sim \pi_k$ given by Equation (13) for π_2. Then the limit distribution of $n^{1/2}[T(g_n^*) - T(g_0)]$ under g_0 as $n \to \infty$ is $Norm(0, \|\tilde{T}_{g_0}\|_L^2)$, where $\|\tilde{T}_{g_0}\|_L^2 = I(\theta_0)^{-1}$ when $g_0 = f_{\theta_0}$.*

Proof. To prove this result, we verify Lemma 25.23 in [11], which is equivalent to showing that

$$\sqrt{n}(T(g_n^*) - T(g_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{T}_{g_0}(X_i) + o_p(1).$$

By the consistency result provided for priors π_1 and π_2 in the previous section, we consider only $g_n^* \in A_{n,k}$ for an n that is sufficiently large. Then by Equation (12) we have that

$$\sqrt{n}(T(g_n^*) - T(g_0)) = \sqrt{n} \left\langle \frac{g_n^* - g_0}{g_0}, \tilde{T}_{g_0} \right\rangle_L + o_p(1).$$

Therefore, showing

$$\sqrt{n} \int_0^1 (g_n^*(x) - g_0(x))\tilde{T}_{g_0}(x)dx = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{T}_{g_0}(X_i) + o_p(1)$$

will complete the proof. Due to $\int_0^1 g_0(x)\tilde{T}_{g_0}(x)dx = 0$, we now need to show that $\int_0^1 g_n^*(x)\tilde{T}_{g_0}(x)dx = (1/n)\sum_{i=1}^n \tilde{T}_{g_0}(X_i) + o_p(1)$. By the law of large numbers, we have that $\frac{1}{n}\sum_{i=1}^n \tilde{T}_{g_0}(X_i) - G_0\tilde{T}_{g_0} = o_p(1)$, and $\int_0^1 g_n^*(x)\tilde{T}_{g_0}(x)dx - G_0\tilde{T}_{g_0} = o_p(1)$ due to the posterior consistency demonstrated above. Therefore, we have that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \tilde{T}_{g_0}(X_i) - \int_0^1 g_n^*(x)\tilde{T}_{g_0}(x)dx \\
&= \frac{1}{n}\sum_{i=1}^n \tilde{T}_{g_0}(X_i) - \int_0^1 g_n^*(x)\tilde{T}_{g_0}(x)dx + \int_0^1 g_n^*(x)\tilde{T}_{g_0}(x)dx - G_0\tilde{T}_{g_0} \\
&= o_p(1).
\end{aligned}
$$

$\square$

3.2. The Bernstein von Mises Theorem for BMH

Theorem 2.1 in [15] yielded a general result and approach to show that the BvM Theorem holds for smooth functionals in some semi-parametric models. The theorem shows that, under the continuity and consistency condition, the moment generating function (MGF) of the parameter endowed with a posterior distribution can be calculated approximately through the local asymptotic normal (LAN) expansion, and its convergence to an MGF of some normal random variable can then be shown under some assumptions on the statistical model.

We will show that the BvM theorem holds for the BMH Method via Theorem 4. The result also shows that the approach given in [15] can be applied not only to simple examples but also to relatively complicated frameworks. To prove it, we introduced Lemma 2, which is modified from Proposition 1 in [15], the proof of which was not given explicitly in the original paper.

For mathematical simplicity, we assume that the true density f_{θ_0} belongs to the set $\mathscr{F}$, which is restricted to the space of all densities that are bounded away from 0 and ∞ on $[0, 1]$. As noted above, the compactness of the domain can be relaxed by considering transformations of the parameters and random variables.

To state the lemma, we need several more notations. Assume that the functional T satisfies Equation (12) with bounded efficient influence function $\tilde{T}_{g_0} \neq 0$. We denote $\tilde{T}_{g_0}$ by $\tilde{T}$, where $\tilde{T}_{[k]}$ denotes the projection of $\tilde{T}$ on $\mathscr{H}_k$. For $k \geq 1$, let

$$
\begin{aligned}
\hat{T}_k &= T(g_{0[k]}) + \frac{\mathbb{G}_n \tilde{T}_{[k]}}{\sqrt{n}}, & V_k &= \|\tilde{T}_{[k]}\|_L^2 \\
\hat{T} &= T(g_0) + \frac{\mathbb{G}_n \tilde{T}}{\sqrt{n}}, & V &= \|\tilde{T}\|_L^2
\end{aligned}
\tag{20}
$$

and denote

$$
\mathbb{G}_n(g) = W_n(g) = \frac{1}{\sqrt{n}}\sum_{i=1}^n [g(x_i) - G_0(g)].
\tag{21}
$$

Lemma 2. *Let g_0 belong to $\mathscr{G}$, let the prior π be defined as in Section 2.3, and let Conditions (13, 14, 15) be satisfied. Consider estimating a functional $T(g)$, differentiable with respect to the tangent set $\mathscr{H}_T := \{q \in L^2(g_0), \int_{[0,1]} qg_0 = 0\} \subset \mathscr{H} = L^2(g_0)$, with efficient influence function $\tilde{T}_{g_0}$ bounded on $[0, 1]$, and with $\tilde{r}$ defined in Equation (12), for $\mathscr{K}_n$ as introduced in Equation (15). If*

$$
\max_{k \in \mathscr{K}_n} \left| \|\tilde{T}_{[k]}\|_L^2 - \|\tilde{T}\|_L^2 \right| = o_p(1)
\tag{22}
$$

$$
\max_{k \in \mathscr{K}_n} \mathbb{G}_n(\tilde{T}_{[k]} - \tilde{T}) = o_p(1)
\tag{23}
$$

$$\sup_{k \in \mathcal{K}_n} \sup_{g \in A_{n,k}(M)} \sqrt{n}\tilde{r}(g, g_0) = o_p(1) \tag{24}$$

for any $M > 0$ and $A_{n,k}(M)$ defined as in (16), as $n \to \infty$, and

$$\max_{k \in \mathcal{K}_n} \sqrt{n} \left| \int (\tilde{T} - \tilde{T}_{[k]})(g - g_0) \right| = o(1), \tag{25}$$

then the BvM theorem for the functional T holds.

Proof. To show that BvM holds is to show that the posterior distribution converges to a normal distribution. If we have that

$$
\begin{aligned}
\pi[\sqrt{n}(T - \hat{T}_k) \le z \mid \mathbb{X}_n] &= \sum_{k \in \mathcal{K}_n} \pi[k \mid \mathbb{X}_n]\pi\left[\sqrt{n}(T - \hat{T}) \le z + \sqrt{n}(\hat{T} - \hat{T}_k) \mid \mathbb{X}_n, k\right] + o_p(1) \\
&= \sum_{k \in \mathcal{K}_n} \pi[k \mid \mathbb{X}_n]\Phi\left(\frac{z + \sqrt{n}(\hat{T} - \hat{T}_k)}{\sqrt{V_k}}\right) + o_p(1),
\end{aligned}
\tag{26}
$$

then the proof will be completed by showing that the RHS of Equation (26) reduces from the mixture of normal to the target law $N(0, V)$.

By Condition (22), we have that V_k goes to V uniformly for $k \in \mathcal{K}_n$. Due to the definition of $\tilde{T}$ and the Lemma 4 result (iii) in the supplement of [15], we have that

$$
\begin{aligned}
\sqrt{n}(\hat{T} - \hat{T}_k) &= \sqrt{n}\left(T(g_0) - T(g_{[k]})\right) + \mathbb{G}_n(\tilde{T} - \tilde{T}_{[k]}) \\
&= \sqrt{n}\int \tilde{T}(g_{0[k]} - g_0) + \mathbb{G}_n(\tilde{T}_{[k]} - \tilde{T}) + o_p(1) \\
&= \sqrt{n}\int (\tilde{T} - \tilde{T}_{[k]})(g_{0[k]} - g_0) + \mathbb{G}_n(\tilde{T}_{[k]} - \tilde{T}) + o_p(1).
\end{aligned}
$$

By Conditions (25) and (23), the last line converges to 0 uniformly for $k \in \mathcal{K}_n$.

Therefore, showing that for any given k, Equation (26) holds will complete the proof. We prove this by showing that the MGF (Laplace transformation) of the posterior distribution of the parameter of interest converges to the MGF of some normal distribution, which implies that the posterior converges to the normal distribution weakly by Lemmas 1 and 2 in supplement to [15] or Theorem 2.2 in [18].

First, consider the deterministic $k = K_n$ case. We calculate the MGF as

$$
E[e^{t\sqrt{n}(T(g) - \hat{T}(g_{0[k]}))} \mid \mathbb{X}_n, A_n] = \frac{\int_{A_n} e^{t\sqrt{n}(T(g) - \hat{T}(g_{0[k]})) + l_n(g) - l_n(g_{0[k]})} d\pi(g)}{\int_{A_n} e^{l_n(g) - l_n(g_{0[k]})} d\pi(g)}
\tag{27}
$$

where $l_n(g)$ is the log-likelihood for given g and $\mathbb{X}_n$. Based on the LAN expansion of the log-likelihood and the smoothness of the functional, the exponent in the numerator on the RHS of the equation can be transformed with respect to $\bar{T}_{(k)} = \tilde{T}_{[k]} - \int \tilde{T}_{[k]} g_{0[k]}$

$$
\begin{aligned}
&t\sqrt{n}(T(g) - \hat{T}_k) + l_n(g) - l_n(g_{0[k]}) \\
&= t\sqrt{n}\left(T(g) - T(g_{0[k]}) - \frac{\mathbb{G}_n\tilde{T}_{[k]}}{\sqrt{n}}\right) + l_n(g) - l_n(g_{0[k]}) \\
&= t\sqrt{n}\left(\left\langle \log\frac{g}{g_{0[k]}} - \int \log\frac{g}{g_{0[k]}}g_{0[k]}, \bar{T}_{[k]}\right\rangle_L + \mathscr{B}(g, g_{0[k]}) + \tilde{r}(g, g_{0[k]}) - \frac{\mathbb{G}_n\tilde{T}_{[k]}}{\sqrt{n}}\right) \\
&\quad - \frac{1}{2}\left\|\sqrt{n}\log\frac{g}{g_{0[k]}}\right\|_L^2 + W_n\left(\sqrt{n}\log\frac{g}{g_{0[k]}}\right) + R_{n,k}(g, g_{0[k]})
\end{aligned}
$$

where $\mathscr{B}(g, g_0) = \int_0^1 [\log(g/g_0) - (g - g_0)/g_0](x)\tilde{T}_{g_0}(x)g_0(x)dx$. Note that $\mathbb{G}_n = W_n$ and add a term of $(t^2/2)\|\bar{T}_{(k)}\|_L^2$. Re-arranging the RHS expression above, we have

$$
t\sqrt{n}(T(g) - \hat{T}_k) + l_n(g) - l_n(g_{0,[k]})
$$

$$
= -\frac{n}{2}\left\|\log\frac{g}{g_{0[k]}} - \frac{t}{\sqrt{n}}\bar{T}_{(k)}\right\|_{L,k}^2 + \sqrt{n}W_n\left(\log\frac{g}{g_{0[k]}} - \frac{t}{\sqrt{n}}\bar{T}_{(k)}\right)
$$

$$
+ \frac{t^2}{2}\|\bar{T}_{(k)}\|_{L,k}^2 + t\sqrt{n}\mathscr{B}_{n,k} + R_{n,k}(g, g_{0[k]}) + \tilde{r}(g, g_{0[k]})
$$

$$
= -\frac{n}{2}\left\|\log\frac{ge^{-\frac{t}{\sqrt{n}}\bar{T}_{(k)}}}{g_{0[k]}}\right\|_{L,k}^2 + \sqrt{n}W_n\left(\log\frac{ge^{-\frac{t}{\sqrt{n}}\bar{T}_{(k)}}}{g_{0[k]}}\right) + \frac{t^2}{2}\|\bar{T}_{(k)}\|_{L,k}^2
$$

$$
+ t\sqrt{n}\mathscr{B}_{n,k} + R_{n,k}(g, g_{0[k]}) + \tilde{r}(g, g_{0[k]}).
$$

This is because the cross term in calculating the first term in the second line above is equal to the inner product term in the equation above it.

Let $g_{t,k} = ge^{-\frac{t}{\sqrt{n}}\bar{T}_{(k)}}/Ge^{-\frac{t}{\sqrt{n}}\bar{T}_{(k)}}$, the RHS of the above equation can be written as

$$
\frac{t^2}{2}\|\bar{T}_{(k)}\|_{L,k}^2 + l_n(g_{t,k}) - l_n(g_{0[k]}) + o(1). \tag{28}
$$

Substituting the corresponding terms on the RHS of Equation (27) by (28), we have that

$$
E[e^{t\sqrt{n}(T(g) - \hat{T}(g_{0[k]}))} \mid \mathbb{X}_n, A_n] = e^{(t^2/2)\|\bar{T}_{(k)}\|_{L,k}^2 + o(1)} \times \frac{\int_{A_{n,k}} e^{l_n(g_{t,k}) - l_n(g_{0[k]})}d\pi_k(g)}{\int_{A_{n,k}} e^{l_n(g) - l_n(g_{0[k]})}d\pi_k(g)}. \tag{29}
$$

Notice that the integration in the denominator of the second term is an expectation based on a Dirichlet distribution on ω as described in Equation (14) and that $g_{t,k} = k\sum_{j=1}^k \zeta_j \mathbb{1}_{I_j}$, where

$$
\zeta_j = \frac{\omega_j \gamma_j^{-1}}{\sum_{j=1}^k \omega_j \gamma_j^{-1}} \tag{30}
$$

with $\gamma_j = e^{t\bar{T}_j/\sqrt{n}}$ and $\bar{T}_j := k\int_{I_j} \bar{T}_{(k)}$. Let $S_{\gamma^{-1}(\omega)} = \sum_{j=1}^k \omega_j \gamma_j^{-1}$, by (30). We then have $S_\gamma^{-1}(\zeta) = S_{\gamma^{-1}}(\omega)$. Now using these notations,

$$
\frac{\int_{A_{n,k}} e^{l_n(g_{t,k}) - l_n(g_{0[k]})}d\pi_k(g)}{\int_{A_{n,k}} e^{l_n(g) - l_n(g_{0[k]})}d\pi_k(g)} = \frac{\int_{A_{n,k}} e^{l_n(g_{t,k}) - l_n(g_{0[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k} - 1}/B(\alpha)d\omega}{\int_{A_{n,k}} e^{l_n(g) - l_n(g_{0[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k} - 1}/B(\alpha)d\omega}
$$

$$
= \frac{\int_{A_{n,k}} e^{l_n\left(k\sum_{j=1}^k \frac{\omega_j \gamma_j^{-1}}{\sum_{j=1}^k \omega_j \gamma_j^{-1}}\mathbb{1}_{I_j}\right) - l_n(g_{0[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k} - 1}d\omega}{\int_{A_{n,k}} e^{l_n\left(k\sum_{j=1}^k \omega_j \mathbb{1}_{I_j}\right) - l_n(g_{0[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k} - 1}d\omega}
$$

$$
= \frac{\int_{A_{n,k}} e^{l_n\left(k\sum_{j=1}^k \zeta_j \mathbb{1}_{I_j}\right) - l_n(g_{0[k]})} \Delta_\zeta \prod_{j=1}^k [\gamma_j \zeta_j S_\gamma^{-1}(\zeta)]^{\alpha_{j,k} - 1}d\zeta}{\int_{A_{n,k}} e^{l_n\left(k\sum_{j=1}^k \omega_j \mathbb{1}_{I_j}\right) - l_n(g_{0[k]})} \prod_{j=1}^k \omega_j^{\alpha_{j,k} - 1}d\omega} \tag{31}
$$

where $\Delta_\zeta = S_\gamma^{-k}(\zeta) \prod_{j=1}^k \gamma_j$ is the Jacobian of the change of variable, $(\omega_1, \ldots, \omega_{k-1}) \to (\zeta_1, \ldots, \zeta_{k-1})$, which is given in Lemma 5 in supplement of [15], and $B(\alpha) = \prod_{i=1}^k \Gamma(\alpha_i)/\Gamma(\sum_{i=1}^k \alpha_i)$ is the constant for normalizing Dirichlet distribution.

Notice that, over the set $A_{n,k}$,

$$
\prod_{j=1}^{k}[\gamma_j S_\gamma^{-1}(\zeta)]^{\alpha_{j,k}-1}\Delta_\zeta \;=\; S_\gamma(\zeta)^{-\sum_{j=1}^{k}\alpha_{j,k}}\gamma_j^{\sum_{j=1}^{k}\alpha_{j,k}}
$$

$$
=\; S_\gamma(\zeta)^{-\sum_{j=1}^{k}\alpha_{j,k}}e^{t\sum_{j=1}^{k}a_{j,k}\bar{T}_j/\sqrt{n}} \tag{32}
$$

$$
=\; e^{t\sum_{j=1}^{k}\alpha_{j,k}\bar{T}_j/\sqrt{n}}\left(1-\frac{t}{\sqrt{n}}\int_0^1 \bar{T}_{(k)}(g-g_0)+O(n^{-1})\right)^{\sum_{j=1}^{k}\alpha_{j,k}},
$$

since

$$
S_{\gamma^{-1}}(\omega)=\int_0^1 e^{-t\bar{T}_{(k)}(x)/\sqrt{n}}g_{[k]}(x)dx=1-\frac{t}{\sqrt{n}}\int_0^1 \bar{T}_{(k)}(g_{[k]}-g_0)+O(n^{-1})
$$

by Taylor's expansion. Expression (32) converges to 1 under Condition (15), so Expression (31) converges to

$$
\frac{\int_{A_{n,k}}e^{ln(k\sum_{j=1}^{k}\zeta_j\mathbb{1}_{I_j})-ln(g_{0[k]})}\prod_{j=1}^{k}\zeta_j^{\alpha_{j,k}-1}/B(\alpha_k)d\zeta}{\int_{A_{n,k}}e^{ln(k\sum_{j=1}^{k}\omega_j\mathbb{1}_{I_j})-ln(g_{0[k]})}\prod_{j=1}^{k}\omega_j^{\alpha_{j,k}-1}/B(\alpha_k)d\omega} \tag{33}
$$

since, when $\|\omega-\omega_0\|_1 \le M\sqrt{k\log n}/\sqrt{n}$,

$$
\|\zeta-\omega_0\|_1 \le \|\omega-\omega_0\|_1+\|\omega-\zeta\|_1=\frac{M\sqrt{k\log n}+2|t|\|\tilde{T}\|_\infty}{\sqrt{n}}\le (M+1)\frac{\sqrt{k\log n}}{\sqrt{n}}
$$

and, vice versa, when $\|\zeta-\omega_0\|_1 \le M\sqrt{k\log n}/\sqrt{n}$,

$$
\|\omega-\omega_0\|_1 \le \|\omega-\zeta\|_1+\|\omega_0-\zeta\|_1=\frac{M\sqrt{k\log n}+2|t|\|\tilde{T}\|_\infty}{\sqrt{n}}\le (M+1)\frac{\sqrt{k\log n}}{\sqrt{n}}.
$$

Choosing M, such that

$$
\pi\left[\|\omega-\omega_0\|_1 \le (M+1)\sqrt{k\log n}\mid \mathbb{X}_n,k\right]=1+o_p(1), \tag{34}
$$

Expression (33) equals $1+o_p(1)$. Notice that $\|\bar{T}_{(k)}\|_{L,k}=\|\tilde{T}_{[k]}\|_L$. We then have that

$$
E^\pi\left[e^{t\sqrt{n}(T(g)-\hat{T}_k)}\mid \mathbb{X}_n,A_{n,k}\right]=e^{t^2\|\tilde{T}_{[k]}\|_L^2}\left(1+o_p(1)\right), \tag{35}
$$

which completes the proof for a fixed k case.

For a random k case, the proof will follow the same steps as the corresponding part in the proof for Theorem 4.2 in [15]. For completeness, we briefly sketch the proof here. Since k is not fixed, we will calculate $E^\pi[e^{t\sqrt{n}(T(f)-\hat{T}_k)}\mid \mathbb{X}_n]$ on $B_n=\bigcup_{1\le k\le n}A_{n,k}\bigcap\{f=f_{\omega,k},k\in\mathscr{K}_n\}$. Consider $\mathscr{K}_n$ a subset of $\{1,2,\ldots,n/\log^2 n\}$ such that $\pi(\mathscr{K}_n\mid \mathbb{X}_n)=1+o_p(1)$ by the concentration property (a) of the random histogram, we have that $\pi[B_n\mid \mathbb{X}_n]=1+o_p(1)$. We rewrite the left-hand side (LHS) of Equation (35) as $E^\pi[e^{t\sqrt{n}(T(f)-\hat{T}_k)}\mid \mathbb{X}_n,B_{n,k}]$, which is also equal to $e^{t^2\|\tilde{T}_{[k]}\|_L^2}(1+o_p(1))$. Notice that $o(1)$ in this expression is uniform in k. This is because it holds in the proof for a deterministic case for any given $k<n$. Therefore,

$$
E^\pi\left[e^{t\sqrt{n}(T(f)-\hat{T})}\mid \mathbb{X}_n,B_n\right]=\sum_{k\in\mathscr{K}_n}E^\pi\left[e^{t\sqrt{n}(T(f)-\hat{T}_k)+\hat{T}_k)-\hat{T})}\mid \mathbb{X}_n,A_{n,k},k\right]\pi[k\mid \mathbb{X}_n]
$$

$$
=(1+o(1))\sum_{k\in\mathscr{K}_n}e^{t^2 V_k/2+t\sqrt{n}(\hat{T}_k-\hat{T})}\pi[k\mid \mathbb{X}_n].
$$

Using Equations (23) and (25) together with the continuous mapping theorem for the exponential function yields that the last display converges in probability to $e^{t^2 V/2}$ as $n \to \infty$, which completes the proof. □

The following theorem shows that Method 2 is efficient, the proof of which consists in verifying that the conditions in the above lemma are satisfied.

Theorem 4. *Suppose $g_0 \in \mathscr{C}^{\beta}$ with $\beta > 0$. Let the prior on k be either a Dirac mass at $k = K_n = n^{1/2}(\log n)^{-2}$ or $k \sim \pi_k$ given by (13), and let two priors π_1 and π_2 be defined by (14) and satisfy (13). Then, for all $\beta > 1/2$, the BvM holds for $T(f)$ for both π_1 and π_2.*

Proof. For $T(f)$ such that Equation (12) is satisfied, Condition (24) is satisfied obviously.

For Equation (23), the empirical process $\mathbb{G}_n(\tilde{T}_{[k]} - \tilde{T})$ is controlled and will converge to 0 by applying Lemma 19.33 in [11].

Condition (25) is satisfied by Lemma 3 below.

Now we show that Equation (22) holds:

$$
\begin{aligned}
\|\tilde{T}_f\|_L^2 - \|\tilde{T}_{[k]}\|_L^2 &\leq \left| \int \dot{s}_{T(f)}(x)dx - \int \frac{\dot{s}_{T(f_{[k]}(x))}f(x)}{f_{[k]}(x)}dx \right| \\
&\lesssim \left| \int \dot{s}_{T(f)}(x)f_{[k]}(x) - \dot{s}_{T(f_{[k]})}(x)f(x) \right| \\
&= \left| \int \dot{s}_{T(f_{[k]})}(x)[f_{[k]}(x) - f(x)] \right| \\
&\lesssim \int |f_{[k]}(x) - f(x)|dx.
\end{aligned}
$$

The last equality is based on Conclusion (3) in Lemma 4 in [19], and the last inequality is due to the assumption that $\tilde{T}$ is bounded. Then the last term is controlled by $h(f, f_n)$, which completes the proof. □

Lemma 3. *Under the same conditions as in Theorem 4, Equation (25) holds.*

Proof. Since $\tilde{T} = \left(- \left[\int \ddot{s}_{T(g_0)}(x)g_0^{\frac{1}{2}}(x)dx \right]^{-1} + a_t \right) \frac{\dot{s}_{T(g_0)}(x)}{2g_0^{\frac{1}{2}}(x)}$, under the deterministic k-prior with $k = K_n = n^{1/2}(\log n)^{-2}$ and $\beta > 1/2$,

$$
\left| \int (\tilde{T} - \tilde{T}_{[k]})(g_0 - g_{0[k]}) \right| \lesssim h^2(g_0, g_{0[k]}) = o(1/\sqrt{n}).
$$

For the random k-prior, since we restrict g to be bounded from above and below, so the Hellinger and L^2-distances considered are comparable. For a given $k \in \mathcal{K}_n$, by definition, there exists $g_k^* \in \mathcal{H}_k^1$ with $h(g_0, g_k^*) \leq M\epsilon_n(\beta)$, so

$$
h^2(g_0, g_{0[k]}) \lesssim \int (g_0 - g_{0[k]})^2 \leq \int (g_0 - g_k^*)^2 \lesssim h^2(g_0, g_k^*) \lesssim \epsilon_n^2(\beta),
$$

which completes the proof. □

4. Robustness Properties

In frequentist analysis, robustness is usually measured by the influence function and breakdown point of estimators. These have been used to study robustness in minimum Hellinger distance estimators in [3] and in more general minimum disparity estimators in [2,7].

In Bayesian inference, robustness is labeled "outlier rejection" and is studied under the framework of the "theory of conflict resolution". There is a large literature on this topic, e.g., [20–22]. While the results of [22] are only about symmetric distributions, [23] provides corresponding results covering a wider class of distributions with tails in the general exponential power family. These results provide a complete theory for the case of many observations and a single location parameter.

We examine the behavior of the methods MHB and BMH under a mixture model for gross errors. Let δ_z denote the uniform density of the interval $(z - \epsilon, z + \epsilon)$, where $\epsilon > 0$ is small, and let $f_{\theta,\alpha,z} = (1 - \alpha)f_\theta + \alpha\delta_z$, where $\theta \in \Theta$ and $\alpha \in [0,1)]$ and z is a real number. The density $f_{\theta,\alpha,z}$ models a situation, where $100(1 - \alpha)\%$ observations are distributed from f_θ, and $100\alpha\%$ of the observations are the gross errors located near z.

Theorem 5. *For every $\alpha \in (0,1)$ and every $\theta \in \Theta$, denote the mixture model for gross errors by $f_{\theta,\alpha,z}$. We then have that $lim_{z\to\infty}lim_{n\to\infty}T(g_n^*) = \theta$, under the assumptions of Theorem 3 and that, for the BMH method, $\pi(T(g) \mid \mathbb{X}_n) \to \phi(\theta, \|\tilde{T}_{f_{\theta,\alpha,z}}\|_L^2)$ in the distribution as $n \to \infty$ and $z \to \infty$, where ϕ denotes the probability function of the normal distribution, when conditions in Theorem 4 are satisfied.*

Proof. By Theorem 7 in [3], for functional T, as we defined and under the conditions in this theorem, we have that

$$lim_{z\to\infty}T(f_{\theta,\alpha,z}) = \theta.$$

We also have that, for MHB, under conditions of Theorem 3, $lim_{n\to\infty}T(g_n^*) \to T(f_{\theta,\alpha,z})$ in probability. Combining the two results, $lim_{z\to\infty}lim_{n\to\infty}T(g_n^*) = \theta$, when the data is generated from a contaminated distribution as $f_{\theta,\alpha,z}$. Similarly, by Theorem 4, we have that $\pi(T(g) \mid \mathbb{X}_n) \to \phi(T(f_{\theta,\alpha,a}), \|\tilde{T}_{f_{\theta,\alpha,z}}\|_L^2)$ in distribution as $n \to \infty$, and which converges to $\phi(\theta, \|\tilde{T}_{f_{\theta,\alpha,z}}\|_L^2)$, as $z \to \infty$. □

5. Demonstration

We provide a demonstration of both BMH and MHB methods on two data sets: the classical Newcomb light speed data (see [24,25]), in which 2 out of 66 values are clearly negative oultiers, and a bivariate simulation containing 10% contamination in two asymmetric locations.

We have implemented the BMH and MHB methods using two Bayesian nonparametric priors:

1. the random histogram prior studied in this paper based on a fixed $k = 100$ with the range naturally extended to the range of the observed data (this is applied only to our first univariate example).
2. the popular Dirichlet Process (DP) kernel mixture of the form

$$
\begin{aligned}
y_i \mid \mu_i, \Sigma_i &\sim N(\mu_i, \Sigma_i) \\
(\mu_i, \Sigma_i) \mid G &\sim G \\
G \mid \alpha, G_0 &\sim DP(\alpha G_0)
\end{aligned}
$$

where the baseline distribution is the conjugate normal-inverted Wishart,

$$G_0 = N(\mu \mid m_1, (1/k_0)\Sigma)IW(\Sigma \mid \nu_1, \psi_1).$$

Note that, when y_i values are univariate observations, the inverse Wishart (IW) distribution reverts to an inverse Gamma distribution. To complete the model specification, independent hyperpriors are assumed

$$\alpha \mid a_0, b_0 \sim Gamma(a_0, b_0)$$
$$m_1 \mid m_2, s_2 \sim N(m_2, s_2)$$
$$k_0 \mid \tau_1, \tau_2 \sim Gamma(\tau_1/2, \tau_2/2)$$
$$\psi_1 \mid \nu_2, \psi_2 \sim IW(\nu_2, \psi_2).$$

We obtain posteriors for both using BUGS. We have elected to use BUGS here as opposed to the package `DPpackage` within R despite the latter's rather efficient MCMC algorithms because our BMH method requires direct access to samples from the posterior distribution as opposed to the expected *a posteriori* estimate. The R package `distrEx` is then used to construct the sampled density functions and calculated the Hellinger distance between the sampled densities from the nonparametric model and the assumed normal distribution. The R package `optimx` is also used to find the minima of the Hellinger distances. The time cost of our methods are dominated by the optimization step rather than by the obtaining of samples from the posterior density.

We first apply BMH and MHB on the Simon Newcomb's measurements to measure the speed of light. The data contains 66 observations. For this example, we specify the parameters and hyper-parameters of the DPM as $\alpha = 1$, $m_2 = 0$, $s_2 = 1000$, $\tau_1 = 1$, $\tau_2 = 100$, and $\nu_2 = 2$, $\psi_2 = 1$. We plot the data and a bivariate contour of the BMH posterior for both the mean and variance of the assumed normal in Figure 1, where, despite outliers, the BvM result is readily apparent.

Table 1 summarizes these estimates. We report the estimated mean and variance with and without the obvious outliers as well as the same quantities estimated using both MHB and BMH methods with the last of these being the expected *a posteriori* estimates. Quantities in parentheses given the "natural" standard error for each quantity: likelihood estimates correspond to standard normal theory—dividing the estimated standard error by $\sqrt{n}$, and BMH standard errors are obtained from the posterior distribution. For MHB, we used a bootstrap and note that, while the computational cost involved in estimating MHB is significantly lower than BMH when obtaining a point estimate, the standard errors require and MCMC chain for each bootstrap, significantly raising the cost of obtaining these estimates. We observe that both prior specifications result in parameter estimates that are identical to two decimal places and very close to those obtained after removing outliers.
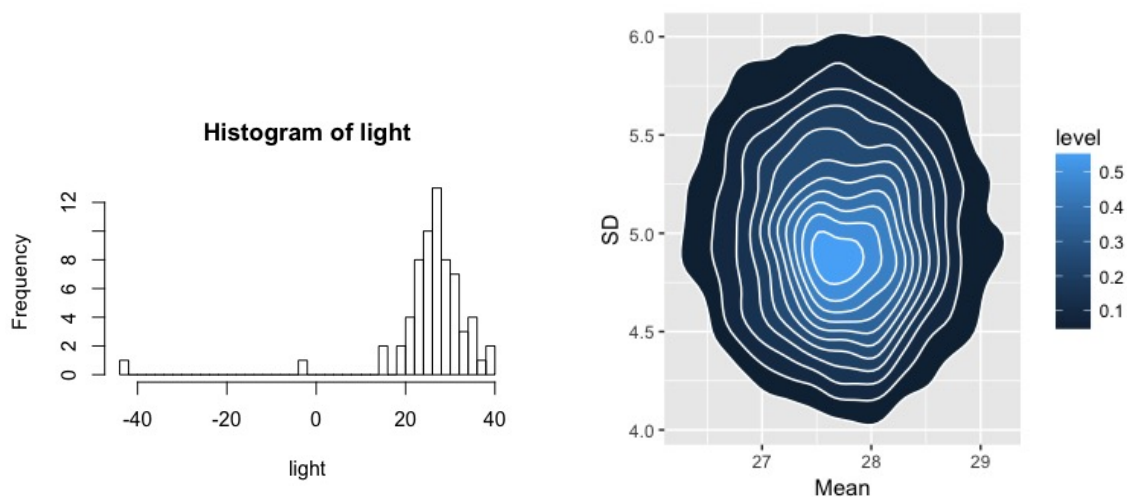

Figure 1. Left: Histogram of the light speed data; Right: bivariate contour plots of the posterior for the mean and variance of these data from the BMH method.

Table 1. Estimation results for Newcomb's light speed data. Direct Estimate refers to the standard mean and variance estimates, and Without Outliers indicates the same estimates with outliers removed. The first row for each parameter gives the estimate under a Dirichlet process prior and the second using a random histogram. Standard errors for each estimate are given in parentheses: these are from the normal theory for the first two columns via a bootstrap for MHB and from posterior samples for BMH.

	Direct Estimate	Without Outliers	MHB	BMH
$\hat{\mu}$	26.21 (1.32)	27.75 (0.64)	27.72 (0.64)	27.73 (0.63)
			27.72 (0.64)	27.73 (0.63)
$\hat{\sigma}$	10.75 (3.40)	5.08 (0.46)	5.07 (0.46)	5.00 (0.47)
			5.07 (0.46)	5.00 (0.47)

To examine the practical implementation of methods that go beyond our theoretical results, we applied these methods to a simulated two-dimensional data set of 100 data points generated from a standard normal with two contamination distributions. Specifically, our data distribution comes from

$$\frac{9}{10} N \left(\begin{pmatrix} 10 \\ 5 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 5 \end{pmatrix} \right) + \frac{1}{20} N \left(\begin{pmatrix} -2 \\ 5 \end{pmatrix}, \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix} \right) + \frac{1}{20} N \left(\begin{pmatrix} 10 \\ 14 \end{pmatrix}, \begin{pmatrix} 0.4 & -0.1 \\ -0.1 & 0.4 \end{pmatrix} \right)$$

where exactly five points were generated from each of the second-two Gaussians. Our DP prior used the same hyper-parameters as above with the exception that Ψ_1 was obtained from the empirical variance of the (contaminated) data, and (m_2, S_2) were extended to their 2-dimensional form as $((0,0)^T, diag(1000, 1000))$. Figure 2 plots these data along with the posterior for the two means. Figure 3 provides posterior distributions for the components of the variance matrix. Table 2 presents estimation results for the full data and those with the contaminating distributions removed as well as from the BMH method. Here we again observe that BMH yields results that are very close to those obtained using the uncontaminated data. There is some more irregularity in our estimates, particularly in Figure 3, which we speculate is due to poor optimization. There is considerable scope to improve the numerics of minimum Hellinger distance methods more generally, but this is beyond the scope of this paper.
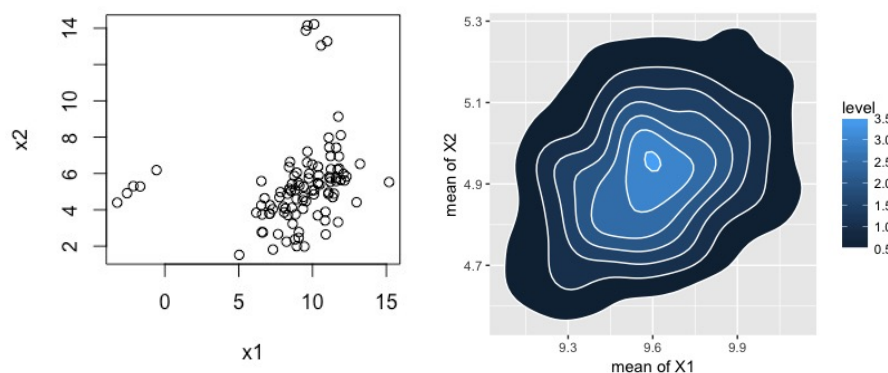

Figure 2. Left: simulated two-dimensional normal example with two contamination components; Right: BMH posterior for the mean vector (μ_1, μ_2).

Table 2. Estimation results for a contaminated bivariate normal. We provide generating estimates, the natural maximum likelihood estimates with and without outliers and the BMH estimates. Reported BMH estimates are expected *a posteriori* estimates with posterior standard errors given in parentheses.

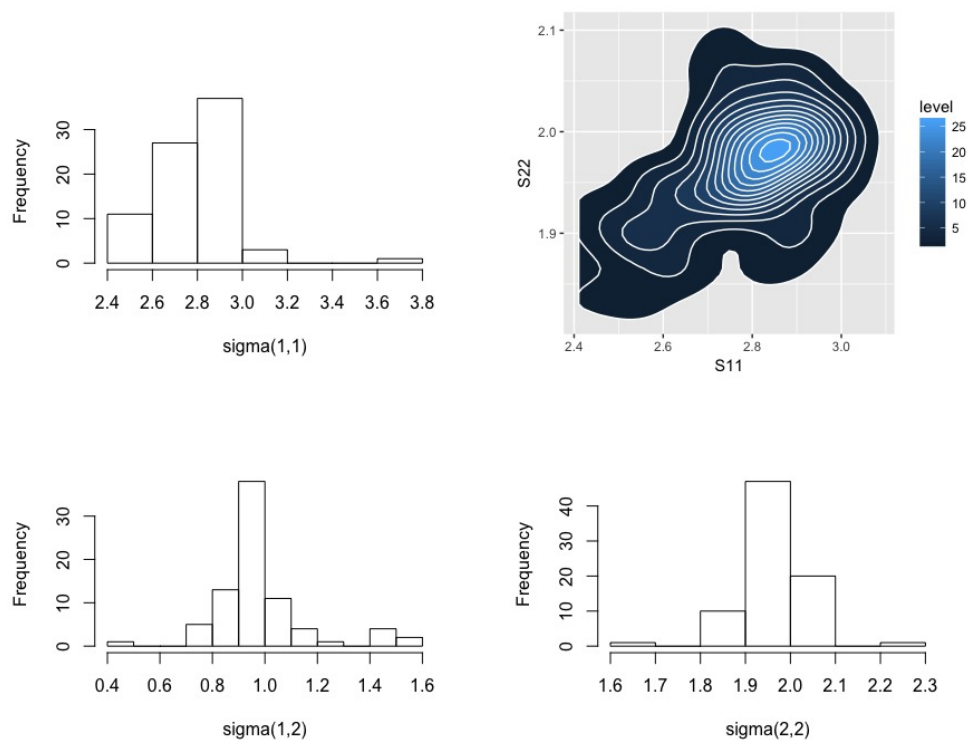|	μ_{01}	μ_{02}	Σ_{11}	Σ_{12}	Σ_{22}
True	10	5	3	1	2
Contaminated data	9.07	5.36	9.76	1.67	5.80
Data with outliers removed	9.62 (0.13)	4.91 (0.11)	3.45 (0.13)	1.49 (0.13)	2.29 (0.11)
Estimated by BMH	9.59 (0.27)	4.93 (0.19)	2.79 (0.18)	0.98 (0.18)	1.97 (0.076)

Figure 3. Posterior distributions for the elements of Σ in the simulated bivariate normal example.

6. Discussion

This paper investigates the use of minimum Hellinger distance methods that replace kernel density estimates with Bayesian nonparametric models. We show that simply substituting the expected *a posteriori* estimator will reproduce the efficiency and robustness properties of the classical disparity methods first derived in [3]. Further, inducing a posterior distribution on θ through the posterior for *g* results in a Bernstein von Mises theorem and a distributional robustness result.

There are multiple potential extensions of this work. While we have focused on the specific pairing of Hellinger distance and random histogram priors, both of these can be generalized. A more general class of disparities was examined in [7], and we believe the extension of our methods to this class are straightforward. More general Bayesian nonparametric priors are discussed in [14], where the Dirichlet process prior has been particularly popular. Extensions to each of these priors will require separate analysis (e.g. [26]). Extensions of disparities to regression models were examined in [27] using a conditional density estimate, where equivalent Bayesian nonparametrics are not as well developed. Other modeling domains such as time series may require multivariate density estimates, resulting in further challenges.

Our results are a counterpoint to the Bayesian extensions of Hellinger distance methods in [2] where the kernel density was retained for g_n but a prior was given for θ and the disparity treated as a log likelihood. Combining both these approaches represents a fully Bayesian implementation of disparity methods and is an important direction of future research.

Author Contributions: Conceptualization: Y.W. and G.H; methodology: Y.W. and G.H.; formal analysis: Y.W.; investigation: Y.W. and G.H.; writing—original draft preparation: Y.W.; writing—review and editing: Y.W. and G.H.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BMH   Bayesian Minimum Hellinger Method
MHB   Minimum Hellinger Method with Bayesian Density estimation
BvM   Bernstein von Mises
MGF   Moment Generating Function
DP   Dirichlet Process
DPM   Dirichlet Process Mixture
LAN   local asymptotic normality
BUGS   Bayesian inference Using Gibbs Sampling
MCMC   Markov Chain Monte Carlo
IW   inverse Wishart
RHS   Right hand side
LHS   Left hand side

References

1. Huber, P.J. *Robust Statistics*; Wiley: Hoboken, NJ, USA, 2004.
2. Hooker, G.; Vidyashankar, A.N. Bayesian model robustness via disparities. *Test* **2014**, *23*, 556–584. [CrossRef]
3. Beran, R. Minimum Hellinger Distance Estimates for Parametric Models. *Ann. Stat.* **1977**, *5*, 445–463. [CrossRef]
4. Basu, A.; Lindsay, B.G. Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **1994**, *46*, 683–705. [CrossRef]
5. Basu, A.; Sarkar, S.; Vidyashankar, A.N. Minimum Negative Exponential Disparity Estimation in Parametric Models. *J. Stat. Plan. Inference* **1997**, *58*, 349–370. [CrossRef]
6. Pak, R.J.; Basu, A. Minimum Disparity Estimation in Linear Regression Models: Distribution and Efficiency. *Ann. Inst. Stat. Math.* **1998**, *50*, 503–521. [CrossRef]
7. Park, C.; Basu, A. Minimum Disparity Estimation: Asymptotic Normality and Breakdown Point Results. *Bull. Inform. Cybern.* **2004** *38*, 19–33.
8. Lindsay, B.G. Efficiency versus Robustness: The case for minimum Hellinger distance and related methods. *Ann. Stat.* **1994**, *22*, 1081–1114. [CrossRef]
9. Gervini, D.; Yohai, V.J. A class of robust and fully efficient regression estimators. *Ann. Stat.* **2002**, *30*, 583–616. [CrossRef]
10. Wu, Y.; Ghosal, S. Posterior consistency for some semi-parametric problems. *Sankhyā Ser. A* **2008**, *70*, 267–313.
11. Van der Vaart, A. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, **2000**.
12. Ghosal, S.; Ghosh, J.K.; van der Vaart, A. Convergence rates of posterior distributions. *Ann. Stat.* **2000**, *28*, 500–531. [CrossRef]
13. Ghosal, S.; van der Vaart, A. Convergence rates of posterior distributions for noniid observations. *Ann. Stat.* **2007**, *35*, 192–223. [CrossRef]
14. Ghosh, J.K.; Ramamoorthi, R.V. *Bayesian Nonparametrics;* Springer: New York, NY, USA, 2003.
15. Castillo, I.; Rousseau, J.A. Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Stat.* **2015**, *43*, 2353–2383. [CrossRef]
16. Amewou-Atisso, M.; Ghosal, S.; Ghosh, J.; Ramamoorthi, R. Posterior consistency for semi-parametric regression problems. *Bernoulli* **2003**, *9*, 291–312. [CrossRef]
17. Rivoirard, V.; Rousseau, J. Bernstein-von Mises theorem for linear functionals of the density. *Ann. Stat.* **2012**, *40*, 1489–1523. [CrossRef]
18. Bagui, S.C.; Mehra, K.L. Convergence of Binomial, Poisson, Negative-Binomial, and Gamma to normal distribution: Moment generating functions technique. *Am. J. Math. Stat.* **2016**, *6*, 115–121.
19. Castillo, I.; Nickl, R. Nonparametric Bernstein-von Mises Theorems in Gaussian White Noise. *Ann. Stat.* **2013**, *41*, 1999–2028. [CrossRef]
20. De Finetti, B. The Bayesian approach to the rejection of outliers. In Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1960; University of California Press: Berkeley, CA, USA, 1961; Volume 1, pp. 199–210.
21. O'Hagan, A. On outlier rejection phenomena in Bayes inference. *J. R. Stat. Soc. B* **1979**, *41*, 358–367. [CrossRef]

22. O'Hagan, A. Outliers and credence for location parameter inference. *J. Am. Stat. Assoc.* **1990**, *85*, 172–176. [CrossRef]
23. Desgagnè, A.; Angers, J.-F. Confilicting information and location parameter inference. *Metron* **2007**, *67*, 67–97.
24. Stigler, S.M. Do Robust Estimators Work with Real Data? *Ann. Stat.* **1977**, *5*, 1055–1098. [CrossRef]
25. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; Chapman and Hall: Boca Raton, FL, USA, 2011.
26. Wu, Y.; Ghosal, S. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.* **2008**, *3*, 298–331. [CrossRef]
27. Hooker, G. Consistency, Efficiency and Robustness of Conditional Disparity Methods. *Bernoulli* **2016**, *22*, 857–900. [CrossRef]