Ann Inst Stat Math (2018) 70:761–788
https://doi.org/10.1007/s10463-017-0608-y

Efficient and robust tests for semiparametric models

Jingjing Wu[1] · Rohana J. Karunamuni[2]

Received: 18 March 2016 / Revised: 11 October 2016 / Published online: 11 March 2017

Abstract In this paper, we investigate a hypothesis testing problem in regular semiparametric models using the Hellinger distance approach. Specifically, given a sample from a semiparametric family of ν-densities of the form $\{f_{\theta,\eta} : \theta \in \Theta, \eta \in \Gamma\}$, we consider the problem of testing a null hypothesis $H_0 : \theta \in \Theta_0$ against an alternative hypothesis $H_1 : \theta \in \Theta_1$, where η is a nuisance parameter (possibly of infinite dimensional), ν is a σ-finite measure, Θ is a bounded open subset of $\mathbb{R}^p$, and Γ is a subset of some Banach or Hilbert space. We employ the Hellinger distance to construct a test statistic. The proposed method results in an explicit form of the test statistic. We show that the proposed test is asymptotically optimal (i.e., locally uniformly most powerful) and has some desirable robustness properties, such as resistance to deviations from the postulated model and in the presence of outliers.

Keywords Tests of hypotheses, Hellinger distance · Semiparametric models · Asymptotic optimality · Robustness · Adaptivity

---

This research was supported by research grants from the Natural Sciences and Engineering Research Council of Canada.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-017-0608-y) contains supplementary material, which is available to authorized users.

✉ Rohana J. Karunamuni
R.J.Karunamuni@ualberta.ca

[1] Department of Mathematics and Statistics, University of Calgary, Calgary, AB T2N 1N4, Canada

[2] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada

1 Introduction

In this paper, we investigate optimal and robust tests for semiparametric models. Specifically, let $(\mathcal{X}, \mathcal{A}, \nu)$ be a σ-finite measure space and f be a density with respect to (w.r.t.) the measure ν. Suppose that the observations $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d) random variables from a distribution with ν-density belong to the (regular) semiparametric family given by

$$\mathcal{F} = \{f_{\theta,\eta} : \theta \in \Theta, \eta \in \Gamma\}, \tag{1}$$

where θ is the parameter of interest and a η is a nuisance parameter, which is possibly of infinite dimensional. We assume that Θ is a bounded open subset of $\mathbb{R}^p$ and Γ is a subset of some Banach space B with a norm $\|\cdot\|_B$. We consider the problem of testing a null hypothesis $H_0 : \theta \in \Theta_0$ against an alternative hypothesis $H_1 : \theta \in \Theta_1$, where $\Theta_i \subseteq \Theta$, $i = 0, 1$.

Numerous models fall into class (1), and well-known examples include semiparametric mixture models (Lindsay 1995; van der Vaart 1996; Murphy and van der Vaart 2000), errors-in-variables models (Bickel and Ritov 1987; Murphy and van der Vaart 1996), regression models (van der Vaart 2000) and Cox model for survival analysis (Cox 1972). More references, examples and an overview of the main ideas and techniques of semiparametric inference can be found in the monographs of Bickel et al. (1998), van der Vaart (2000) and Kosorok (2008).

The goal of semiparametric inference is to construct efficient estimators and optimal test statistics for evaluating semiparametric model parameters. The most common approach to efficient estimation and optimal testing is based on modifications of the likelihood approach. These modifications are necessary due to complications resulting from the presence of an infinite-dimensional nuisance parameter in the model (1). In general, the presence of this nuisance parameter induces a loss of efficiency. An estimator/test that remains asymptotically efficient/optimal in these conditions is called *adaptive* (Bickel 1982).

One of the most useful methods of constructing test statistics for complicated models is the likelihood ratio method, mainly because it gives an explicit definition of the test statistic and an explicit form for the rejection region. It is known that most likelihood based tests in general are asymptotically optimal but not robust against outliers in the data and for model misspecification (Huber and Ronchetti 2009, Ch. 13). On the other hand, optimality of test statistics is closely related to efficient estimation; asymptotically efficient estimators generally yield asymptotically optimal tests and confidence bands (Pfanzagl and Wefelmeyer 1982). A similar assertion, however, is not certain about robustness of testing procedures, i.e., a robust test based on a "robust estimator" may not necessarily be fully robust (Huber and Ronchetti 2009, Ch. 12). It has been proved that minimum Hellinger distance (MHD) estimators for parametric models are efficient under the model and have excellent robustness properties (Beran 1977; Simpson 1987; Lindsay 1994). Recently, Wu and Karunamuni (2012, 2015) have constructed efficient MHD and profile MHD estimators for the semiparametric family $\mathcal{F}$ defined in (1). Thus, their estimators can be employed to construct a test statistic for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ for the family $\mathcal{F}$, among others. However, most MHD estimators are implicitly defined and computation of such estimators requires an iterative algorithm.

In this paper, we follow a different approach. Our idea is to treat the hypothesis testing problem as a model selection problem. By implementing a modified version of the "Hellinger information criterion" (HIC) introduced in Woo and Sriram (2006), we construct a test statistic for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. That is, we first form a test statistic using the proposed HIC for the null and local (contiguous) alternative hypotheses. Then, this approach is modified to obtain a test statistic for composite alternatives, see Sect. 2 below for specific details. The HIC of Woo and Sriram (2006, 2007) was defined specifically for mixture models and is motivated by the classical Akaike-type criterion for model selection and the third approach of Poland and Shachter (1994, Sec. 4) for model selection involving the Kullback–Leibler distance. Such a procedure is intuitively reasonable and more appropriate since a hypothesis testing problem is basically a model selection problem in a broad sense. Instead of the HIC, other information criteria can also be used, such as the L_2 information criterion of Umashanger and Sriram (2009).

Asymptotically, the likelihood ratio technique measures a certain distance between the maximum likelihood estimators under the null and full hypotheses; see, e.g., van der Vaart (2000). Advantages of the proposed method are clear though for two reasons. First, it does not require a separate construction of an MHD estimator, which is itself a formidable task in many semiparametric models. Second, the resulting test statistic is explicitly defined, a clear preference in applications. The key question then is whether this method leads to an optimal and robust test statistic for the semiparametric family $\mathcal{F}$. We investigate this problem in this paper and obtain a very positive result. Specifically, we construct a test statistic explicitly and show that it is asymptotically optimal (i.e., locally uniformly most powerful). Moreover, the proposed test statistic has some desirable robustness properties such as resistance to outliers and model misspecification. The asymptotic optimality and robustness combined with an explicit form make the proposed test statistic appealing in practice. Detailed constructions of the proposed Hellinger distance test statistic are also exhibited for a symmetric location family and a scale mixture model, as special cases of (1).

Stather (1981), Simpson (1989) and Basu et al. (2013, 2016) have developed test procedures based on the Hellinger distance for parametric models. In general, statistical inference for parametric models based on divergence measures, in which the Hellinger distance is a special case, see monographs of Pardo (2006) and Basu et al. (2011). MHD estimators for special cases of the semiparametric models (1) have been studied in Beran (1978), Wu (2007), Wu and Karunamuni (2009), Karunamuni and Wu (2009, 2011), Wu et al. (2010) and Tang and Karunamuni (2013). For the general semiparametric models (1), see Wu and Karunamuni (2012, 2015) for MHD and profile MHD estimators. However, to the best of our knowledge, Hellinger distance-based tests for general semiparametric models have not been investigated in the literature yet.

This paper is organized as follows. Section 2 presents the proposed test statistic for the semiparametric models (1). The asymptotic properties of the test statistic and some discussions are given in Sect. 3. Some robustness properties of the test statistic are discussed in Sect. 4. Detailed constructions of the proposed test statistic for two examples are given in Sect. 5. A Monte Carlo study and some concluding remarks are given in Sects. 6 and 7, respectively.

2 Test statistic

Throughout we assume that the model $\mathcal{F}$ defined by (1) is *identifiable* in the sense that $\|f_{\theta_1,\eta_1}^{1/2} - f_{\theta_2,\eta_2}^{1/2}\| = 0$ implies $\theta_1 = \theta_2$ and $\eta_1 = \eta_2$, where $\|\cdot\|$ denotes the L_2-norm and $L_2(\nu)$ denote the collection of all measurable functions that are square integrable w.r.t. measure ν, i.e., $L_2(\nu) = \{f : \int f^2 \mathrm{d}\nu < \infty\}$. We first introduce some properties of the parametric family $\mathcal{F}_g = \{f_{t,g} : t \in \Theta\}$ for each $g \in \Gamma$. We assume that $\mathcal{F}_g$ identifiable in the sense that $\|f_{\theta_1,g}^{1/2} - f_{\theta_2,g}^{1/2}\| = 0$ implies $\theta_1 = \theta_2$. We set $s_{t,g} = f_{t,g}^{1/2}$ for $t \in \Theta$. If the map $t \mapsto s_{t,g}$ is continuous in $L_2(\nu)$, then we say $\mathcal{F}_g$ is *Hellinger continuous*. We say $\mathcal{F}_g$ is *Hellinger differentiable* at an interior point θ of Θ if the map $t \mapsto s_{t,g}$ is differentiable at θ: there is a vector $\dot{s}_{\theta,g}$ with components in $L_2(\nu)$ such that

$$\left\| s_{\theta+t,g} - s_{\theta,g} - t^\top \dot{s}_{\theta,g} \right\| = o(|t|) \tag{2}$$

as $t \to 0$, where $t^\top$ is the transpose of vector $t \in \Theta$ and $|t| = \sum |t_i|$ denotes the l_1-norm of vector $t \in \mathbb{R}^p$. In this case, we call $I_{\theta,g} = 4\int \dot{s}_{\theta,g}\dot{s}_{\theta,g}^\top \,\mathrm{d}\nu$ the *information matrix at* θ and $\dot{\ell}_{\theta,g} = 2\dot{s}_{\theta,g}/s_{\theta,g}$ the *score function at* θ. Since $2\int \dot{s}_{\theta,g}s_{\theta,g}\,\mathrm{d}\nu$ is the gradient of the constant map $t \mapsto (s_{t,g}, s_{t,g}) = 1$ at θ, $\dot{s}_{\theta,g}$ and $s_{\theta,g}$ are orthogonal, i.e., $\int \dot{s}_{\theta,g}s_{\theta,g}\,\mathrm{d}\nu = 0$. We say $\mathcal{F}_g$ is *twice Hellinger differentiable at* θ if $\mathcal{F}_g$ is Hellinger differentiable in a neighborhood of θ and there is a matrix $\ddot{s}_{\theta,g}$ with entries in $L_2(\nu)$ such that

$$\left\| \dot{s}_{\theta+t,g} - \dot{s}_{\theta,g} - \ddot{s}_{\theta,g}t \right\| = o(|t|). \tag{3}$$

We say $\mathcal{F}_g$ is *Hellinger-regular* if $\mathcal{F}_g$ is identifiable, Hellinger continuous and twice Hellinger differentiable at each interior point θ of Θ with positive definite information matrix $I_{\theta,g}$.

Note that the semiparametric family (1) can be written as a union of the parametric models $\mathcal{F}_g = \{f_{t,g} : t \in \Theta\}$, $g \in \Gamma$. For $g \in \Gamma$, define

$$\dot{\Gamma}_g = \left\{ h \in B : \lim_{n\to\infty} \left\| n^{1/2}(g_n - g) - h \right\|_B = 0 \text{ for some sequence } \{g_n\} \subseteq \Gamma \right\}. \tag{4}$$

For each interior point θ of Θ, γ and g in Γ and each sequence t_n in Θ converging to θ, suppose that there exist bounded linear operators $A_{t,g} : B \mapsto L_2(\nu)$ satisfying

$$\|s_{t_n,\gamma} - s_{t_n,g} - A_{t_n,g}(\gamma - g)\| = o(\|\gamma - g\|_B) \tag{5}$$

and

$$\sup_{h\in K} \|A_{t_n,g}h - A_{\theta,g}h\|_2 \to 0$$

for any compact subset K of B. We say the family $\mathcal{F}$ is *adaptive* at (t, g) if

$$\int \dot{s}_{t,g} A_{t,g} h \, \mathrm{d}\nu = 0, \quad h \in \dot{\Gamma}_g.$$

Let $H(f, g) = \| f^{1/2} - g^{1/2} \|$ denote the Hellinger distance between two densities f and g. Then, $H^2(f, g) = 2 - 2 < f^{1/2}, g^{1/2} >$. Suppose $X_1, \ldots, X_n$ are i.i.d random variables with common density f, and let $\hat{f}$ denote a nonparametric density estimator of f based on $X_1, \ldots, X_n$. Then, following the HIC of Woo and Sriram (2006, 2007), we propose an HIC as follows:

$$HIC(f, \hat{f}) = H^2(f, \ \hat{f}) + n^{-1}\alpha(n) m_f,$$

where $\alpha(n)$ depends only on n and m_f is the number of parameters in f. If one wishes to test $H_0 : f = f_0$ against $H_1 : f = f_1$, then the difference

$$HIC(f_0, \hat{f}) - HIC(f_1, \hat{f}) \tag{6}$$

can be used as a test statistic, and H_0 is rejected for large values of the preceding difference. One can also employ the difference $H^2(f_0, \hat{f}) - H^2(f_1, \hat{f})$ for testing $H_0 : f = f_0$ against $H_1 : f = f_1$; see, e.g., Stather (1981) and Basu et al. (2013).

Now suppose that f is a member of the semiparametric family $\mathcal{F}$ defined by (1). We consider the problem of testing null hypothesis $H_0 : \theta = \theta_0$ against a local alternative hypothesis of the form $H_1 : \theta_n = \theta_0 + \beta e n^{-1/2}$ for some known value $\theta_0 \in \Theta$, where $\beta > 0$ is a fixed number, e denotes the $p \times 1$ unit length Euclidean vector and n is the sample size. Note that H_1 represents a shrinking "contiguous alternative" in a $n^{-1/2}$-neighborhood of θ_0. Then, the difference (6) now takes the form

$$\left\| f_{\theta_0,\eta}^{1/2} - \hat{f}^{1/2} \right\|^2 - \left\| f_{\theta_n,\eta}^{1/2} - \hat{f}^{1/2} \right\|^2 \tag{7}$$

with $\theta_n = \theta_0 + \beta e n^{-1/2}$. Note that the difference in (7) is equal to

$$2 \int (s_{\theta_n,\eta} - s_{\theta_0,\eta}) \hat{f}^{1/2} \mathrm{d}\nu. \tag{8}$$

If the family $\mathcal{F}_\eta = \{f_{\theta,\eta} : \theta \in \Theta\}$ is Hellinger-regular, then the map $t \to s_{t,\eta}$ is differentiable with derivative $\dot{s}_{t,\eta}$, and thus, the expression in (8) can be written as

$$2\beta n^{-1/2} e^\top \int \dot{s}_{\theta_0,\eta} \hat{f}^{1/2} \mathrm{d}\nu + o(n^{-1/2}).$$

The preceding expression suggests that the use of constant multiples of $\int \dot{s}_{\theta_0,\eta} \hat{f}^{1/2} \mathrm{d}\nu$ as an alternative to the expression (8). Let $\mathcal{G}$ denote the space of all densities w.r.t. the dominating measure ν, and define a functional $T_\eta : \mathcal{G} \to \mathbb{R}^p$ by

$$T_\eta(g) = 4 \int \dot{s}_{\theta_0,\eta} g^{1/2} \mathrm{d}\nu. \tag{9}$$

Then, $T_\eta(\hat{f})$ is equal to $4 \int \dot{s}_{\theta_0,\eta} \hat{f}^{\,1/2} \mathrm{d}\nu$. Moreover, $T_\eta(\hat{f})$ does not depend on β and, therefore, a statistic based on it can be used as a test statistic for testing $H_0 : \theta = \theta_0$ against the composite hypothesis $H_1 : \theta > \theta_0$ for θ close to θ_0, provided of course η is known, θ_0 is an interior point of Θ and Θ is an open subset of $\mathbb{R}$. This is precisely the situation where the locally most powerful (LMP) test is implemented (Lehmann 1997). In application, however, $\dot{s}_{\theta_0,\eta}$ must be replaced by an estimator since η is unknown, and this results in a test statistic of the form

$$4 \int \hat{\rho}_{\theta_0} \hat{f}^{1/2} \mathrm{d}\nu, \tag{10}$$

where $\hat{\rho}_{\theta_0}$ is an estimator of $\dot{s}_{\theta_0,\eta}$. We will show that the statistic in (10) and the LMP test statistic are asymptotically equivalent under some regularity conditions. Thus, the statistic given by (10) and the LMP statistic share some (asymptotic) optimality properties. The advantage of the test statistic (10) is that it possesses excellent robustness properties, which the LMP test statistic generally lacks.

3 Asymptotics

In this section, we establish asymptotic properties and derive the power function of the proposed test statistic. First we obtain a stochastic expansion of $4 \int \hat{\rho}_{\theta_0} \hat{f}^{1/2} \mathrm{d}\nu$ defined by (10). The proofs of the theorems stated below are given in Appendix. In what follows, asymptotic results are as $n \to \infty$.

Theorem 1 *Assume that the parametric family $\mathcal{F}_\eta = \{f_{t,\eta} : t \in \Theta\}$ is Hellinger-regular for each $\eta \in \Gamma$. Let $\{a_n\}$ be a sequence of positive numbers such that $a_n = o(n^{-1/2})$ as $n \to \infty$. Suppose that the density estimator $\hat{f}$ satisfies*

$$\int \left(\hat{s} - s_{\theta_0,\eta}\right)^2 d\nu = O_P(a_n) \tag{11}$$

and

$$\int \dot{s}_{\theta_0,\eta} \hat{s}\, d\nu = \frac{1}{n} \sum_{i=1}^{n} \frac{\dot{s}_{\theta_0,\eta}}{2 s_{\theta_0,\eta}}(X_i) + o_P(n^{-1/2}), \tag{12}$$

where $\hat{s} = \hat{f}^{1/2}$. Suppose that $\hat{\rho}_{\theta_0}$ satisfies

$$\int (\hat{\rho}_{\theta_0} - \dot{s}_{\theta_0,\eta})^2 d\nu = o_P((n a_n)^{-1}) \tag{13}$$

and

$$\int \hat{\rho}_{\theta_0} s_{\theta_0,\eta} d\nu = o_P(n^{-1/2}). \tag{14}$$

Then we have

$$4\int \hat{\rho}_{\theta_0}\hat{s}\, d\nu = \frac{1}{n}\sum_{i=1}^{n}\dot{\ell}_{\theta_0,\eta}(X_i) + o_P(n^{-1/2}). \tag{15}$$

The next theorem establishes the asymptotic normality of $4\int \hat{\rho}_{\theta_0}\hat{s}\, \mathrm{d}\nu$.

Theorem 2 *Assume that the conditions of Theorem 1 hold. Then we have*

(i) *under H_0 (i.e., $X_1, \ldots, X_n$ are i.i.d. with density $f_{\theta_0,\eta}$)*

$$4n^{1/2}\int \hat{\rho}_{\theta_0}\hat{s}\, d\nu \xrightarrow{\mathcal{D}} N(0, I_{\theta_0,\eta});$$

(ii) *under H_1 (i.e., $X_1, \ldots, X_n$ are i.i.d. with density $f_{\theta_n,\eta}$)*

$$n^{1/2}\left[4\int \hat{\rho}_{\theta_0}\hat{s}\, d\nu - T_\eta(f_{\theta_n,\eta})\right] \xrightarrow{\mathcal{D}} N(0, I_{\theta_0,\eta}),$$

where $I_{\theta_0,\eta} = 4\int \dot{s}_{\theta_0,\eta}\dot{s}_{\theta_0,\eta}^{\top}\, d\nu$, the information matrix at θ_0, and T_η is given by (9).

From part (i) of Theorem 2, it follows that the sequence $\hat{W}_n := 4n^{1/2}\hat{I}^{-1/2}\int \hat{\rho}_{\theta_0}\hat{s}\, \mathrm{d}\nu$ converges in distribution to the $N(0, I_p)$ distribution, where $\hat{I}$ denotes a consistent estimator of $I_{\theta_0,\eta}$ and I_p denotes the p-dimensional identity matrix. For univariate θ, then the null hypothesis H_0 would be rejected if $\hat{W}_n$ exceeds z_α, and such a test is asymptotically of level α, where $P(Z \geq z_\alpha) = \alpha$ with $Z \sim N(0, 1)$ and $0 < \alpha < 1$. For p-variate θ, one can use the test statistic $\hat{W}_n^T\hat{W}_n$ and rejects H_0 if $\hat{W}_n^T\hat{W}_n > \chi^2_{\alpha,p}$, where $P(\chi^2_p \geq \chi^2_{\alpha,p}) = \alpha$ with χ^2_p denoting a chi-square random variable with p degrees of freedom.

The asymptotic quality of a sequence of tests may be judged from the limit of the sequence of local power functions. The power function of the test statistic $\hat{W}_n$ in the univariate θ case can be written as

$$\begin{aligned}\pi_n(\theta_n) &= P_{\theta_n}\left\{\hat{W}_n > z_\alpha\right\}\\ &= P_{\theta_n}\left\{4n^{1/2}\hat{I}^{-1/2}\int \hat{\rho}_{\theta_0}\hat{s}\, \mathrm{d}\nu > z_\alpha\right\}\\ &= P_{\theta_n}\left\{n^{1/2}\hat{I}^{-1/2}\left[4\int \hat{\rho}_{\theta_0}\hat{s}\, \mathrm{d}\nu - T_\eta(f_{\theta_n,\eta})\right] > z_\alpha - n^{1/2}\hat{I}^{-1/2}T_\eta(f_{\theta_n,\eta})\right\}.\end{aligned} \tag{16}$$

Since $\mathcal{F}_\eta$ is Hellinger-regular, we can write

$$\begin{aligned}T_\eta(f_{\theta_n,\eta}) &= 4\int \dot{s}_{\theta_0,\eta}s_{\theta_n,\eta}\mathrm{d}\nu\\ &= 4\int \dot{s}_{\theta_0,\eta}\left[s_{\theta_0,\eta} + n^{-1/2}\beta e^{\top}\dot{s}_{\theta_0,\eta} + \Delta_n\right]\mathrm{d}\nu\\ &= n^{-1/2}\beta I_{\theta_0,\eta}e + o(n^{-1/2}),\end{aligned} \tag{17}$$

where the last equality follows from $\int \dot{s}_{\theta_0,\eta} s_{\theta_0,\eta}\, \mathrm{d}\nu = 0$ and $\left|\int \dot{s}_{\theta_0,\eta} \Delta_n \mathrm{d}\nu\right| \leq \left\|\dot{s}_{\theta_0,\eta}\right\| \|\Delta_n\| = o(n^{-1/2})$ by the Cauchy–Schwarz inequality. Then, since $\hat{I} \to I_{\theta_0,\eta}$ in probability as $n \to \infty$, we obtain from (16) and (17) that

$$\pi_n(\theta_n) \to 1 - \Phi\left(z_\alpha - \beta I_{\theta_0,\eta}^{1/2}\right), \tag{18}$$

where Φ denotes the cumulative distribution function of $N(0, 1)$ distribution.

When η is known, the LMP test statistic $(n\hat{I})^{-1/2} \sum \dot{\ell}_{\theta_0,\eta}(X_i)$ in the univariate case has the same asymptotic power function as the right-hand side of (18) for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta_n = \theta_0 + \beta n^{-1/2}$, where $\dot{\ell}_{\theta_0,\eta}$ denotes the score function for θ_0; see, e.g., Lehmann (1999). Thus, the LMP statistic and the proposed test statistic $\hat{W}_n$ are asymptotically equally efficient. It is known that the LMP test is asymptotically (locally) optimal if η is known, i.e., the LMP test is asymptotically (locally) uniformly most powerful. In this sense, $\hat{W}_n$ is *adaptive* since $\hat{W}_n$ operates under the assumption that η is unknown. The concept of adaptivity is generally reserved for estimators in semiparametric families. It means that one can estimate the parameter θ as well asymptotically not knowing the nuisance parameter η as knowing η (Bickel 1982). Due to the fact that η is unknown in the present setup, the LMP test statistic is not available in practice. Nevertheless, tests based on the likelihood ratio and profile likelihood ratio statistics have been developed in this context that are asymptotically optimal in the "semiparametric sense," see Murphy and van der Vaart (1996, 1997, 2000) and Banerjee (2000, 2005), among others.

Using a Monte Carlo study, we will show that the statistic $\hat{W}_n$ performs better than the likelihood ratio test statistic in the presence of outliers. Theoretically, we will also observe that the statistic $\hat{W}_n$ is not affected when the chosen model is only approximately correct; this would be the case, for example, if a few observations are not consistent with the chosen model. In other words, $\hat{W}_n$ would be robust against deviations from the postulated model as well as in the presence of outliers, which the likelihood ratio and LMP test statistics generally lack. Thus, $\hat{W}_n$ would be an attractive alternative to the likelihood ratio and LMP test statistics in practice when the postulated model may not be totally correct and when the outliers seem to be present in the data.

As Lehmann (1999), van der Vaart (2000) and others have argued, it is enough to consider an alternative hypothesis in a small neighborhood of the null hypothesis since any reasonable test can discriminate well between the null hypothesis and a "distant" alternative, particularly if the number of observations is large. In other words, interest tends to focus on distinguishing the hypothetical value θ_0 from nearby values of θ when dealing with large samples. If the true value is some distance from θ_0, a large sample will typically reveal this so strikingly that a formal test may be unnecessary. A good test proves itself having power in discriminating "close" alternatives. Thus, such a local optimality property is of considerable importance (Lehmann 1999). Furthermore, it is well known in the literature that to make an informative comparison between sequences of tests, one should study the performance of tests in problems that becomes harder as more observations become available. One way of making a testing problem harder is to choose null and alternative hypotheses closer to each other. See Chapter 14 of van der Vaart (2000) for more discussions on above points.

It is appropriate to make a few comments here about the estimators $\hat{s}$ and $\hat{\rho}_{\theta_0}$ used in Theorems 1 and 2 above. For $\hat{s}$, one can employ an estimator based on the empirical density and a kernel-type density estimator in the discrete and continuous cases, respectively. For $\hat{\rho}_{\theta_0}$, two types of estimators can be constructed: "plug-in"- and "direct"-type estimators. In the former case, we first obtain an estimate of the nuisance parameter η from the data and then plug it into the expression of $\dot{s}_{\theta_0,\eta}$. In the latter case, one must obtain an estimator directly by examining the expression of $\dot{s}_{\theta_0,\eta}$. Recall that if η is known, then $\dot{s}_{\theta_0,\eta}$ is typically just the usual parametric score function $\dot{\ell}_{\theta,\eta}$ for θ times $\frac{1}{2} f_{\theta,\eta}^{1/2}$. Thus, the problem reduces to estimation of the score function $\dot{\ell}_{\theta,\eta}$ and the (discrete or continuous) density $f_{\theta,\eta}$. There are a number of methods available in the literature for estimation of the the score function directly; see, e.g., Bickel et al. (1998) and van der Vaart (2000). Alternatively, one can employ readily available nonparametric density estimation techniques to construct an estimator of $\dot{s}_{\theta_0,\eta}$ as a ratio estimator in the continuous case.

We now discuss the assumptions and conclusions made in Theorem 1. In some sense, conclusion (15) obtained on the test statistic, $4\int \hat{\rho}_{\theta_0}\hat{s}\,\mathrm{d}\nu$, is similar to the "asymptotic linearity" condition generally assumed on estimators. Asymptotic linearity is satisfied by many estimators in parametric and semiparametric contexts; see, e.g., Bickel et al. (1998) and van der Vaart (2000). Conditions (11)–(14) give sufficient conditions for conclusion (15). Conditions (11) and (13) give required rates of the mean square errors (MSEs) of the estimators $\hat{s}$ and $\hat{\rho}_{\theta_0}$; (13) might be harder to verify compared to (11) since $\hat{\rho}_{\theta_0}$ estimates a ratio, $\dot{s}_{\theta_0,\eta}$. Condition (12) is not hard to verify in practice for smoothed estimators $\hat{s}$, especially those estimators based on kernel density estimators; see Sect. 5 for more details. In fact, condition (12) can be further simplified, see Remark 3 below. Finally, condition (14) is the "no-bias" condition. This condition is similar to no-bias conditions used in the maximum likelihood estimation for semiparametric models, see van der Vaart (1996) and van der Vaart (2000, Chapter 25). Since $\int \dot{s}_{\theta_0,\eta}s_{\theta_0,\eta}\mathrm{d}\nu = 0$, the condition $\int \hat{\rho}_{\theta_0}s_{\theta_0,\eta}\mathrm{d}\nu = o_P(n^{-1/2})$ means that the "bias" of the estimator $\hat{\rho}_{\theta_0}$, due to estimating the quantity $\dot{s}_{\theta_0,\eta}$, converges to zero slightly faster than $n^{-1/2}$. Such a condition comes out naturally in the proof.

Remark 1 If $\hat{\rho}_{\theta_0} = \dot{s}_{\theta_0,\hat{\eta}}$ for some estimator $\hat{\eta}$ of η, then condition (14) reduces to $\int \dot{s}_{\theta_0,\hat{\eta}}s_{\theta_0,\eta}\mathrm{d}\nu = o_P(n^{-1/2})$. If the model $\mathcal{F}$ is adaptive at (t, g), then one can write

$$\begin{aligned}
-\int \dot{s}_{\theta_0,\hat{\eta}}s_{\theta_0,\eta}\mathrm{d}\nu &= \int \dot{s}_{\theta_0,\hat{\eta}}(s_{\theta_0,\hat{\eta}} - s_{\theta_0,\eta})\mathrm{d}\nu \\
&= \int \dot{s}_{\theta_0,\eta}(s_{\theta_0,\hat{\eta}} - s_{\theta_0,\eta} - A_{\theta_0,\eta}(\hat{\eta}-\eta))\mathrm{d}\nu + o_p(n^{-1/2}) \\
&\quad + \int (\dot{s}_{\theta_0,\hat{\eta}} - \dot{s}_{\theta_0,\eta})(s_{\theta_0,\hat{\eta}} - s_{\theta_0,\eta})\mathrm{d}\nu. \qquad (19)
\end{aligned}$$

From (5) it follows that

$$\sup_{t\in\Theta,|t-\theta|\le Ca_n^{1/2}} \|s_{t,\hat{\eta}} - s_{t,\eta} - A_{t,\eta}(\hat{\eta}-\eta)\| = o(\|\hat{\eta}-\eta\|_B).$$

Thus, if $\|\hat{\eta} - \eta\|_B = O_P(n^{-1/2})$ then using the Cauchy–Schwarz inequality the first term of (19) is of order $o_P(n^{-1/2})$. The second term of (19) is of order $o_P(n^{-1/2})$ from (13) and if $\int (s_{\theta_0,\hat{\eta}} - s_{\theta_0,\eta})^2 = O_P(a_n)$. In cases in which nuisance parameter η is not estimable at $\sqrt{n}$-rate then the Taylor expansion must be carried out into its second-order term. Then, it may be sufficient to have $\|\hat{\eta}-\eta\|_B = o_P(n^{-1/4})$, provided the first term of (19) is bounded by $\|\hat{\eta} - \eta\|_B^2$.

Remark 2 Suppose $\hat{\rho}_{\theta_0} = \dot{s}_{\theta_0,\hat{\eta}}$ for some estimator $\hat{\eta}$ of η. Then, condition (14) holds if $\|\dot{s}_{\theta_0,\eta}\| < \infty$ and $\|s_{\theta_0,\hat{\eta}} - s_{\theta_0,\eta}\| = o_P(n^{-1/2})$. To see this clearly, note that $\int \dot{s}_{\theta_0,\hat{\eta}} s_{\theta_0,\hat{\eta}} \mathrm{d}\nu = \frac{1}{2}\frac{\partial}{\partial\theta_0}\int f_{\theta_0,\hat{\eta}}\mathrm{d}\nu = 0$ and then by the Cauchy–Schwarz inequality and (13),

$$\begin{aligned}\left|\int \dot{s}_{\theta_0,\hat{\eta}} s_{\theta_0,\eta}\mathrm{d}\nu\right| &= \left|\int \dot{s}_{\theta_0,\hat{\eta}}(s_{\theta_0,\hat{\eta}} - s_{\theta_0,\eta})\mathrm{d}\nu\right| \\ &\le \|\dot{s}_{\theta_0,\hat{\eta}}\| \cdot \|s_{\theta_0,\hat{\eta}} - s_{\theta_0,\eta}\| \\ &\le \left(\|\dot{s}_{\theta_0,\eta}\| + \|\dot{s}_{\theta_0,\hat{\eta}} - \dot{s}_{\theta_0,\eta}\|\right)\|s_{\theta_0,\hat{\eta}} - s_{\theta_0,\eta}\| \\ &= o_P(n^{-1/2}).\end{aligned}$$

Remark 3 Sufficient conditions for (12) are given by

$$\sqrt{n}\left[\int \frac{\dot{s}_{\theta_0,\eta}}{s_{\theta_0,\eta}}\hat{f}\,\mathrm{d}\nu - \frac{1}{n}\sum_{i=1}^{n}\frac{\dot{s}_{\theta_0,\eta}}{s_{\theta_0,\eta}}(X_i)\right] = o_P(1), \tag{20}$$

$$\sqrt{n}\int \frac{|\dot{s}_{\theta_0,\eta}|}{s_{\theta_0,\eta}}\left(\hat{f}^{1/2} - f_{\theta_0,\eta}^{1/2}\right)^2 \mathrm{d}\nu = o_P(1). \tag{21}$$

To see this more clearly, apply the algebraic identity

$$b^{1/2} - a^{1/2} = (b-a)/\left(2a^{1/2}\right) - \left(b^{1/2} - a^{1/2}\right)^2/2a^{1/2}$$

for $b \ge 0$ and $a > 0$. Since $\int \dot{s}_{\theta_0,\eta} s_{\theta_0,\eta}\mathrm{d}\nu = 0$, then using (20) and (21) we obtain

$$\begin{aligned}\sqrt{n}\int \dot{s}_{\theta_0,\eta}\hat{s}\,\mathrm{d}\nu &= \sqrt{n}\int \dot{s}_{\theta_0,\eta}(\hat{s} - s_{\theta_0,\eta})\,\mathrm{d}\nu \\ &= \sqrt{n}\int \frac{\dot{s}_{\theta_0,\eta}}{2s_{\theta_0,\eta}}(\hat{f} - f_{\theta_0,\eta})\,\mathrm{d}\nu - n^{1/2}\int \frac{\dot{s}_{\theta_0,\eta}}{2s_{\theta_0,\eta}}\left(\hat{f}^{1/2} - f_{\theta_0,\eta}^{1/2}\right)^2 \mathrm{d}\nu \\ &= \sqrt{n}\int \frac{\dot{s}_{\theta_0,\eta}}{2s_{\theta_0,\eta}}(\hat{f} - f_{\theta_0,\eta})\,\mathrm{d}\nu + R_n \\ &= \sqrt{n}\int \frac{\dot{s}_{\theta_0,\eta}}{2s_{\theta_0,\eta}}\hat{f}\,\mathrm{d}\nu + R_n \\ &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\dot{s}_{\theta_0,\eta}}{2s_{\theta_0,\eta}}(X_i) + o_P(1) + R_n,\end{aligned}$$

with $|R_n| \le \sqrt{n}\int \frac{|\dot{s}_{\theta_0,\eta}|}{s_{\theta_0,\eta}}(\hat{f}^{1/2} - f_{\theta_0,\eta}^{1/2})^2\,\mathrm{d}\nu = o_P(1)$, and hence (12) holds.

4 Robustness properties

To see the robustness properties of our test statistic, we define a functional T based on (7) as

$$T(f) = \left\| f_{\theta_0,\eta}^{1/2} - f^{1/2} \right\|^2 - \left\| f_{\theta_n,\eta}^{1/2} - f^{1/2} \right\|^2 = 2\int (s_{\theta_n,\eta} - s_{\theta_0,\eta}) f^{1/2} \mathrm{d}\nu.$$

If $\mathcal{F}_\eta$ is Hellinger-regular, we further consider the functional T_η defined in (9):

$$T_\eta(f) = 4\int \dot{s}_{\theta_0,\eta} f^{1/2} \mathrm{d}\nu.$$

For densities f and g, using the Cauchy–Schwarz inequality and noting that $\int (g^{1/2} - f^{1/2})^2 \mathrm{d}\nu \le \int |g - f| \mathrm{d}\nu = \|g - f\|_1$, we have

$$|T_\eta(g) - T_\eta(f)| \le 4\|\dot{s}_{\theta_0,\eta}\| \cdot \|g^{1/2} - f^{1/2}\| \le 4\|\dot{s}_{\theta_0,\eta}\|(\|g - f\|_1)^{1/2}. \tag{22}$$

Suppose $\dot{s}_{\theta_0,\eta} \in L_2$, then we see from (22) that $|T_\eta(g) - T_\eta(f)| \to 0$ whenever $\|g - f\|_1 \to 0$. This shows that small distortions away from f do not affect the value of T_η very much. In other words, the test statistic $4\int \hat{\rho}_{\theta_0} \hat{f}^{1/2} \mathrm{d}\nu$ defined by (10) will not be much affected by small departures from the true model.

To see how small distortions away from $\eta \in \Gamma$ affect the value of our test statistic, we define a functional $\tilde{T} : \Gamma \to \mathbb{R}^p$ as

$$\tilde{T}(g) = 4\int \dot{s}_{\theta_0,g} f^{1/2} \mathrm{d}\nu.$$

Suppose that there exist bounded linear operators $A^*_{\theta,g} : B \mapsto L_2(\nu)$ such that for each interior point θ of Θ and $\gamma, g \in \Gamma$, we have

$$\|\dot{s}_{\theta,\gamma} - \dot{s}_{\theta,g} - A^*_{\theta,g}(\gamma - g)\| = o(\|\gamma - g\|_B).$$

Again using the Cauchy–Schwarz inequality, we obtain

$$|\tilde{T}(\gamma) - \tilde{T}(g)| \le 4\|\dot{s}_{\theta_0,\gamma} - \dot{s}_{\theta_0,g}\| \le 4\|A^*_{\theta_0,g}\| \cdot \|\gamma - g\|_B + o(\|\gamma - g\|_B). \tag{23}$$

We see from (23) that small distortions away from g do not affect the value of the functional $\tilde{T}$ very much. Since

$$\left| 4\int \varphi f^{1/2} \mathrm{d}\nu - 4\int \dot{s}_{\theta_0,\eta} f^{1/2} \mathrm{d}\nu \right| \le 4\int \left|\varphi - \dot{s}_{\theta_0,\eta}\right| f^{1/2} \mathrm{d}\nu \le 4\|\varphi - \dot{s}_{\theta_0,\eta}\|,$$

we see that small distortions away from $\dot{s}_{\theta_0,\eta}$ also do not affect the value of the test statistic very much.

5 Examples

In this section, we consider two examples of the semiparametric models (1), namely a symmetric location model and a scale mixture model. In each case, we will construct the proposed test statistic $4\int \hat{\rho}_{\theta_0}\hat{s}\,\mathrm{d}\nu$ defined by (10). Further, we will also show that the conditions of Theorem 3.1 can be verified for suitable estimators $\hat{\rho}_{\theta_0}$ and $\hat{s} = \hat{f}^{1/2}$.

5.1 Symmetric location model

Assume that the random variables $X_1, \ldots, X_n$ are i.i.d from a member of the semiparametric family defined by

$$\mathcal{F} = \{f_{\theta,\eta}(x) = \eta(x-\theta) :\ \theta \in \Theta,\ \eta \in \Gamma\},$$

where η is unknown and is assumed belongs to the class

$$\Gamma = \left\{\eta :\ \eta > 0,\ \int \eta(x)\mathrm{d}x = 1,\ \eta(-x) = \eta(x),\ \eta \text{ is absolutely continuous a.e. with } \int \frac{(\eta^{(1)}(x))^2}{\eta(x)}\mathrm{d}x < \infty\right\}. \tag{24}$$

Further, assume that Θ is a bounded open interval of $\mathbb{R}$. For instance, one can set $\Theta = (-M, M)$ for some large positive number M. We will assume that $\mathcal{F}_\eta = \{f_{\theta,\eta} :\ \theta \in \Theta\}$ is Hellinger-regular for each $\eta \in \Gamma$. Let $\hat{f}(x)$ denote a kernel density estimator based on $X_1, \ldots, X_n$:

$$\hat{f}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right), \tag{25}$$

where kernel K is a nonvanishing bounded density, symmetric about zero, twice continuously differentiable and satisfies $\int u^i K(u)du < \infty$ for $i = 2, 4$, and bandwidth sequence $\{h_n\}$ satisfies $h_n > 0$ and $h_n \to 0$ as $n \to \infty$. For convenience, we assume that h_n is not random. However, in practice, h_n may be chosen by a data-driven method, such as the cross-validation method.

We now show a plug-in-type estimator of the statistic $4\int \hat{\rho}_{\theta_0}\hat{s}\,\mathrm{d}\nu$ by constructing an estimator for the nuisance parameter η. First, let $\tilde{\theta} = \tilde{\theta}(X_1, \ldots, X_n)$ denote a preliminary location estimator of θ possessing the property that $n^{1/2}(\tilde{\theta} - \theta) = O_P(1)$; that is, $\tilde{\theta}$ is a $\sqrt{n}$-consistent estimator of θ. Location estimators that satisfy this property can be easily found; see, e.g., Bickel et al. (1998). Since $\eta(x) = f(x+\theta)$, intuitively we can construct an estimator of η by $\hat{f}(x+\tilde{\theta})$, where $\hat{f}$ is given by (25). Following an idea of Beran (1978), we define a symmetric truncated version

$$\hat{\eta}(x) = \frac{1}{2}\left[\hat{f}(x+\tilde{\theta}) + \hat{f}(-x+\tilde{\theta})\right]b_n^2(x) \tag{26}$$

as our estimator of η, where b_n is given by

$$b_n(x) = \begin{cases} 1 & \text{if } |x| \leq c_n \\ b(x-c_n) & \text{if } c_n \leq x \leq c_n+1 \\ b(x+c_n) & \text{if } -c_n-1 \leq x \leq -c_n \\ 0 & \text{otherwise,} \end{cases} \tag{27}$$

where function $b(.)$ has range [0, 1], is symmetric about zero with $b(0) = 1$, vanishes outside $[-1, 1]$, and is twice absolutely continuous with derivatives $b^{(i)}$ $(i = 1, 2)$ bounded on the real line, and $\{c_n\}$ is a sequence of positive numbers such that $c_n \to \infty$ as $n \to \infty$. It is easy to show that $\hat{\eta}$ is a consistent estimator of η. Furthermore, $\hat{\eta}(x) = \hat{\eta}(-x)$ and $\hat{\eta}^{(1)}(-x) = -\hat{\eta}^{(1)}(x)$, where $\hat{\eta}^{(1)}$ denotes the first derivative of $\hat{\eta}$. The statistic $4\int \hat{\rho}_{\theta_0}\hat{s}\,\mathrm{d}\nu$ now takes the form $4\int \dot{s}_{\theta_0,\hat{\eta}}\hat{s}\,\mathrm{d}x$, where $\hat{s} = \hat{f}^{1/2}$ with $\hat{f}$ given by (25). Moreover, note that,

$$\begin{aligned} 2\int \hat{\rho}_{\theta_0} s_{\theta_0,\eta}\mathrm{d}x = 2\int \dot{s}_{\theta_0,\hat{\eta}} s_{\theta_0,\eta}\mathrm{d}x &= -\int \frac{\hat{\eta}^{(1)}(x-\theta_0)}{\sqrt{\hat{\eta}(x-\theta_0)}}\sqrt{\eta(x-\theta_0)}\mathrm{d}x \\ &= -\int \frac{\hat{\eta}^{(1)}(x)}{\sqrt{\hat{\eta}(x)}}\sqrt{\eta(x)}\mathrm{d}x \\ &= 0, \end{aligned}$$

since $\hat{\eta}$ and η are symmetric and $\hat{\eta}^{(1)}$ is antisymmetric about 0. Thus, the no-bias condition (14) is trivially satisfied in this case.

As in previous sections, denote $\dot{f}_{t,\eta}(\cdot) = \frac{\partial}{\partial t} f_{t,\eta}(\cdot) = \frac{\partial}{\partial t}\eta(\cdot - t)$, $s_{t,\eta}(\cdot) = f_{t,\eta}^{1/2}(\cdot) = \eta^{1/2}(\cdot - t)$, $\dot{s}_{t,\eta}(\cdot) = \frac{\partial}{\partial t} s_{t,\eta}(\cdot) = \frac{\partial}{\partial t}\eta^{1/2}(\cdot - t)$ and $\ddot{s}_{t,\eta}(\cdot) = \frac{\partial^2}{\partial t^2} s_{t,\eta}(\cdot) = \frac{\partial^2}{\partial t^2}\eta^{1/2}(\cdot - t)$. For notational convenience, we will denote $f_{\theta,\eta}$ and h_n by f and h, respectively, in what follows. Let $B_n = \{x \in \mathbb{R} : |x-\theta| \leq c_n\}$ and $w_n(x) = \sup_{t\in\Theta, |t-\theta|\leq Ca_n^{1/2}} \frac{|\eta^{(1)}(x-t)|}{\sqrt{\eta(x-t)}}$ for any constant $C > 0$ and a sequence of positive numbers $\{a_n\}$ such that $a_n = o(n^{-1/2})$ as $n \to \infty$. We also denote $\bar{f}(x) = E[\hat{f}(x)] = \int f(x-hu)K(u)\,du$, $C_{n1} = \int_{B_n^c} f^{1/2}(x)\mathrm{d}x$, $C_{n2} = \int_{B_n^c} |f^{(2)}(x)|^{1/2}\mathrm{d}x$, $C_{n3} = \int_{B_n^c} \left|\frac{f^{(1)}}{f}\right|(x) f^{1/2}(x)\mathrm{d}x$, $C_{n4} = \int_{B_n} \left|\frac{f^{(1)}}{f}\right|(x)\mathrm{d}x$, $C_{n5} = \int_{B_n^c} f(x)\mathrm{d}x$ and $C_{n6} = \int_{B_n^c} \frac{(f^{(1)})^2}{4f}(x)\mathrm{d}x$.

To prove Theorem 1 for the statistic $4\int \widehat{\rho}_{\theta_0}\hat{s}\mathrm{d}x$ under above specifications, we now state four lemmas. The proofs of these lemmas follow from routine algebra and, therefore, are relegated to supplementary material.

Lemma 1 *Let $\hat{f}$ be defined by (25). Suppose the density f has absolutely continuous derivatives $f^{(i)}$, $i = 1, 2$, and $\int \psi(x)dx < \infty$, where $\psi(x) = \frac{(f^{(2)})^2}{2f}(x) + \frac{(f^{(1)})^4}{8f^3}(x)$.*

Then

$$\int \left[\hat{f}^{1/2}(x) - f^{1/2}(x)\right]^2 dx = O_P\left(h^4 + c_n(nh)^{-1} + C_{n1}(nh)^{-1/2} + C_{n2}(nh^{-1})^{-1/2} + n^{-1/2}h\right). \tag{28}$$

Further, if $\int |\frac{f^{(1)}}{f}|(x)\psi(x)dx < \infty$ *then*

$$\int \left|\frac{f^{(1)}}{f}\right|(x)\left[\hat{f}(x) - f(x)\right]^2 dx = O_P\left(h^4 + C_{n4}(nh)^{-1} + C_{n3}(nh)^{-1/2} + (nh^{-1})^{-1/2}\right). \tag{29}$$

Lemma 2 *Assume that the conditions of Lemma 1 hold. Further assume that* $f^{(4)}$ *exists and is bounded, and* $f^{(i)}$, $i = 1, 2, 4$, *satisfy following conditions:*

$$\int_{-\infty}^{\infty} f^{-1}(x)f^{(1)}(x)f^{(2)}(x)\,dx = 0, \tag{30}$$

$$\int_{-\infty}^{\infty} f^{-1}(x)\left|f^{(1)}f^{(4)}\right|(x)\,dx < \infty,$$

$$\int_{-\infty}^{\infty} f(x)\left(\frac{f^{(1)}}{f}\right)^2(x)\,dx < \infty. \tag{31}$$

Then, we have

$$\int_{-\infty}^{\infty} \dot{s}_{\theta,\eta}(x)\hat{f}^{1/2}(x)dx = \frac{1}{n}\sum_{i=1}^{n}\frac{\dot{s}_{\theta,\eta}}{s_{\theta,\eta}}(X_i) + O_P\left(h^4 + C_{n4}(nh)^{-1} + C_{n3}(nh)^{-1/2} + (nh^{-1})^{-1/2}\right).$$

Lemma 3 *Assume that the conditions of Lemma 1 hold. Further assume that* η *satisfies*

$$\int \eta^{-1}(x)\left(\eta^{(1)}\right)^2(x)\,dx < \infty. \tag{32}$$

Then, with $\hat{\eta}$ *defined by* (26), *we have*

$$\int \left[s_{t,\hat{\eta}}(x) - s_{t,\eta}(x)\right]^2 dx = O_P\left(h^4 + c_n(nh)^{-1} + C_{n1}(nh)^{-1/2} + C_{n2}(nh^{-1})^{-1/2} + n^{-1/2}h + C_{n5} + n^{-1}\right). \tag{33}$$

Lemma 4 *Let $\hat{f}$ be defined by (25) with kernel K further satisfying $K^{(1)}/K$ bounded. Assume that the conditions of Lemma 1 hold. Assume also that $f^{(2)}$ and $f^{(3)}$ exist and are bounded, and that η satisfies*

$$\int (g^{(2)})^2(x)\, dx < \infty, \tag{34}$$

where $g = \eta^{1/2}$. Further suppose that $\psi(x) = \frac{(f^{(2)})^2}{2f}(x) + \frac{(f^{(1)}(x))^4}{8f^3(x)}$ is bounded,

$$\int \left(\frac{f^{(1)}}{f}\right)^2 (x)\psi(x)dx < \infty, \tag{35}$$

$$\int \left(\frac{f^{(1)} f^{(2)}}{f^{3/2}}\right)^2 (x)dx < \infty, \tag{36}$$

and

$$\int f^{-1}(x)(f^{(3)})^2(x)dx < \infty. \tag{37}$$

Then, we have

$$\int \left[\dot{s}_{t,\hat{\eta}}(x) - \dot{s}_{t,\eta}(x)\right]^2 dx = O_P\left(h^4 + c_n(nh^3)^{-1} + C_{n5} + C_{n6} + n^{-1}\right). \tag{38}$$

Theorem 3 *Assume that the conditions in Lemmas 1 to 4 hold at $\theta = \theta_0$. Further assume that the bandwidth $h = h_n$ in (25) is of the form $h = O(n^{-1/5})$. Suppose that the sequence $\{c_n\}$ satisfies the conditions $c_n = o(n^{1/10})$, $C_{n1} = o(n^{-3/10})$, $C_{n2} = o(n^{-1/10})$, $C_{n3}n^{-2/5} = o(n^{-1/2})$, $C_{n4}n^{-4/5} = o(n^{-1/2})$, $C_{n5} = o(n^{-2/5})$ and $C_{n6} = o(n^{-2/5})$. Then the conclusion of Theorem 1 holds.*

The proof of Theorem 3 follows from Lemmas 1 to 4. For the normal location family, i.e., $f_{\theta,\eta}(x) = \eta(x-\theta)$ with $\eta(x) = (2\pi)^{-1/2}e^{-x^2/2}$, $-\infty < x < \infty$, it is easy to show that there exists a sequence $\{a_n\}$ satisfying the conditions of Theorem 3 when one chooses $B_n = \{x \in \mathbb{R} : |x-\theta| \le c_n\}$ with $c_n = (2\log n)^{1/2}$ and bounded Θ. For the double-exponential family, i.e., $f_{\theta,\eta}(x) = \eta(x-\theta)$ with $\eta(x) = 2^{-1}e^{-|x|}$, $-\infty < x < \infty$, the choices of $c_n = \log n$ and bounded Θ would be appropriate to verify the conditions in Theorem 3. In fact, the sequence $\{a_n\}$ used in Theorem 3 has the form $\{n^{-7/10}\}$ in both cases with the preceding choices of c_n. Furthermore, conditions (31) to (32) and (34) to (37) are easily satisfied for both of these families.

5.2 Scale mixture model

Let ϕ denote a probability density that is symmetric about zero and consider the mixture model

$$f_{\theta,\eta}(x) = \int_0^\infty \frac{1}{z}\phi\left(\frac{x-\theta}{z}\right) \mathrm{d}\eta(z), \quad -\infty < x < \infty. \tag{39}$$

We will also assume that the parameter space Θ is a bounded open interval of $\mathbb{R}$. Further assume that the class $\mathcal{F}_\eta = \{f_{\theta,\eta} : \theta \in \Theta\}$ is Hellinger-regular and that the information matrix $I_{\theta,\eta}$ is finite for each η. Then, the mixture density $f_{\theta,\eta}$ is symmetric about θ, and θ can be estimated asymptotically efficiently with a fully adaptive estimator (Bickel 1982; van der Vaart 1996). For simplicity, it is assumed that the unknown mixing distribution η (nuisance parameter) is supported on a fixed interval $[m, M] \subset (0, \infty)$. Here we are interested in showing that the conditions of Theorems 1 can be verified for a suitable estimator $\hat{\rho}_{\theta_0}$ of $\dot{s}_{\theta_0,\eta}$.

Suppose that $X_1, \ldots, X_n$ is a random sample from (39). Then by symmetry, for fixed θ, the variables $Y_i = |X_i - \theta|$ are sampled from the density $g(s) = 2\varphi(s)I\{s \geq 0\}$, where $\varphi(s) = \int_0^\infty \frac{1}{z}\phi(\frac{s}{z})\mathrm{d}\eta(z)$. We will use the Y_i's to construct an estimator $\hat{\rho}_{\theta_0}$. From the fact that ϕ is symmetric about zero, we have $f_{\theta,\eta}(x) = \frac{1}{2}g(|x-\theta|)$. By straightforward calculations, we obtain

$$\dot{s}_{t,\eta}(x) = \frac{g^{(1)}(|x-t|)\mathrm{sign}(x-t)}{2^{3/2}g^{1/2}(|x-t|)}, \tag{40}$$

where $g^{(1)}$ is the first derivative of g. Define

$$\hat{g}_n(s) = \frac{1}{nh_n\beta_0(s)}\sum_{i=1}^{n} K\left(\frac{s-Y_i}{h_n}\right), \tag{41}$$

as the boundary kernel estimator of the density $g(s)$, where the kernel K and the bandwidth h_n are defined as in (25) and $\beta_0(s) = \int_{-\infty}^{s/h} K(u)du$. Furthermore, define

$$\hat{g}_n^{(1)}(s) = \frac{1}{nh_n^2}\sum_{i=1}^{n} \frac{\beta_0(s)K^{(1)}\left(\frac{s-Y_i}{h_n}\right) - \delta_0(s)K\left(\frac{s-Y_i}{h_n}\right)}{\beta_1(s)\delta_0(s) - \beta_0(s)\delta_1(s)}$$

as the boundary kernel estimator of $g^{(1)}(s)$, where $\beta_1(s) = \int_{-\infty}^{s/h} uK(u)du$ and $\delta_k(s) = \int_{-\infty}^{s/h} u^k K^{(1)}(u)du$ for $k = 0, 1$. We now define

$$\hat{\rho}_t(x) = \frac{\hat{g}_n^{(1)}(|x-t|)\mathrm{sign}(x-t)}{2^{3/2}\hat{g}_n^{1/2}(|x-t|)} I_{\{\gamma_n \leq |x-t| \leq \alpha_n\}} \tag{42}$$

as our proposed estimator of $\rho_{t,\eta} = \dot{s}_{t,\eta}$, where $\alpha_n \to +\infty$ and $\gamma_n \to 0$ as $n \to \infty$. Then $\hat{\rho}_{\theta_0}(x)$, with $Y_i = |X_i - \theta_0|$, is the estimator of $\dot{s}_{\theta_0,\eta}$.

Observe that the no-bias condition (14) is trivially satisfied for $\hat{\rho}_t$ defined by (42), since $\int \hat{\rho}_t s_{t,\eta}\mathrm{d}x = 0$ for any $t \in \mathrm{int}(\Theta)$, due to the fact that $\hat{\rho}_t$ and $s_{t,\eta}$ are antisymmetric and symmetric, respectively, about 0.

Denote $C_{n7} = \int_{\alpha_n}^{+\infty} \frac{(g^{(1)})^2}{4g}(s)ds$ and $C_{n8} = \int_0^{\gamma_n} \frac{(g^{(1)})^2}{4g}(s)ds$. To prove Theorem 1 for the statistic $4\int \widehat{\rho}_{\theta_0}\hat{s}\,\mathrm{d}x$ under the scale mixture model (39), we first state a lemma, and its proof is again given in supplementary material.

Lemma 5 *Suppose that the ratio* $K^{(1)}(s)/K(s)$ *is bounded and* $\int_{\gamma_n}^{+\infty}\beta_1^2(s)ds = O(h^2)$, $\inf_{s\in[\gamma_n,+\infty)}|\beta_1(s)\delta_0(s)-\beta_0(s)\delta_1(s)|\geq\gamma_0$ *for some positive constant* γ_0. *Assume that* $\frac{[g^{(1)}(x)]^2}{g(x)}$ *is bounded and* $\int_0^{+\infty}\frac{[g^{(2)}(x)]^2}{g(x)}dx<\infty$. *Then we have*

$$\int(\widehat{\rho_t}(x)-\dot{s}_{t,\eta}(x))^2dx = O_P\left((nh^3)^{-1}\alpha_n+h_n^2+C_{n7}+C_{n8}\right). \tag{43}$$

Theorem 4 *Assume that the conditions in Lemmas 1, 2 and 5 hold at* $\theta=\theta_0$. *Further assume that the bandwidths* h_n *in (25) and (41) are of the form* $h_n=O(n^{-1/5})$. *Suppose that the sequences* $\{c_n\}$, $\{\alpha_n\}$ *and* $\{\gamma_n\}$ *satisfy the conditions* $c_n=o(n^{1/10})$, $\alpha_n=o(n^{1/10})$, $C_{n1}=o(n^{-3/10})$, $C_{n2}=o(n^{-1/10})$, $n^{-2/5}C_{n3}=o(n^{-1/2})$, $n^{-4/5}C_{n4}=o(n^{-1/2})$, $C_{n7}=o(n^{-2/5})$ *and* $C_{n8}=o(n^{-2/5})$. *Then Theorem 1 holds.*

The proof of Theorem 4 follows from Lemmas 1, 2 and 5. For $\phi(s)=(2\pi)^{-1/2}e^{-s^2/2}$ in (39), it is easy to show that the conditions of Theorem 4 are satisfied with $K(s)=2^{-1}e^{-|s|}$, $c_n=c_0\log n$ for some positive constant c_0, $\alpha_n=\alpha_0\log n$ for some positive constant α_0 and $\gamma_n=h\log(h^{-1})$. In fact, since $\int_{-\infty}^{+\infty}uK(u)du=0$, we have $\beta_1(s)=-\int_{s/h}^{+\infty}uK(u)du$. For $\phi(s)=(2\pi)^{-1/2}e^{-s^2/2}$ and $\gamma_n=h\log(h^{-1})$, by simple calculation, we obtain

$$\int_{\gamma_n}^{+\infty}\beta_1^2(s)ds=\frac{h}{16}\left[2\left(1+\frac{\gamma_n}{h}\right)^2+2\left(1+\frac{\gamma_n}{h}\right)+1\right]e^{-\frac{2\gamma_n}{h}}=O(h^2)$$

and

$$\inf_{s\in[\gamma_n,+\infty)}|\beta_1(s)\delta_0(s)-\beta_0(s)\delta_1(s)|=\inf_{s\in[\gamma_n,+\infty)}\left(1-e^{-\frac{\gamma_n}{h}}-\frac{\gamma_n}{2h}e^{-\frac{2\gamma_n}{h}}\right)\to 1.$$

The other conditions of Theorem 4 can easily be verified using routine algebra.

6 Monte Carlo studies

In this section, we examine the finite sample performance of the test statistic given by (10). Specifically, we present numerical studies of the test statistic for the symmetric location model described in Sect. 5.1.

Suppose $X_1,\ldots,X_n$ are i.i.d from $f_{\theta,\eta}(x)=\eta(x-\theta)$ with η belonging to the class Γ defined by (24). Without loss of generality, we test the true value $\theta_0=0$, i.e., $H_0:\theta=0$ against $H_1:\theta_n=\beta n^{-1/2}$ for some $\beta>0$. In this simulation study, we take η to be the standard normal distribution.

As discussed in Sect. 5.1, the statistic of (10) reduces to

$$4\int\hat{\rho}_{\theta_0}\hat{f}^{1/2}\mathrm{d}x=4\int\dot{s}_{\theta_0,\hat{\eta}}\hat{f}^{1/2}\mathrm{d}x=-2\int\frac{\hat{\eta}^{(1)}}{\hat{\eta}^{1/2}}(x-\theta_0)\hat{f}^{1/2}(x)\mathrm{d}x$$
$$=-2\int\frac{\hat{\eta}^{(1)}}{\hat{\eta}^{1/2}}(x)\hat{f}^{1/2}(x+\theta_0)\mathrm{d}x,$$

where $\hat{f}$ and $\hat{\eta}$ are defined by (25) and (26), respectively. By Theorem 2, under H_0, this statistic has asymptotic variance $n^{-1}I_{\theta_0,\eta} = 4n^{-1}\int \dot{s}^2_{\theta_0,\eta}\mathrm{d}x = n^{-1}\int \frac{(\eta^{(1)})^2}{\eta}(x - \theta_0)\mathrm{d}x$, which can be estimated by $n^{-1}\int \frac{(\hat{f}^{(1)})^2}{\hat{f}}(x)\mathrm{d}x$ or $n^{-1}\int \frac{(\hat{\eta}^{(1)})^2}{\hat{\eta}}(x)\mathrm{d}x$. As a result, under H_0,

$$T_1 = -2n^{1/2}\left[\int \frac{(\hat{f}^{(1)})^2}{\hat{f}}(x)\mathrm{d}x\right]^{-1/2} \int \frac{\hat{\eta}^{(1)}}{\hat{\eta}^{1/2}}(x)\hat{f}^{1/2}(x+\theta_0)\mathrm{d}x \tag{44}$$

and

$$T_2 = -2n^{1/2}\left[\int \frac{(\hat{\eta}^{(1)})^2}{\hat{\eta}}(x)\mathrm{d}x\right]^{-1/2} \int \frac{\hat{\eta}^{(1)}}{\hat{\eta}^{1/2}}(x)\hat{f}^{1/2}(x+\theta_0)\mathrm{d}x \tag{45}$$

follow approximately the standard normal distribution when the sample size n is large. Thus, we would reject H_0 if $T_1 > 1.645$ (or $T_2 > 1.645$) when, for example, $\alpha = 0.05$ is used as the significance level.

For comparison purposes, we will also give simulation results for the likelihood ratio test statistic in this case. Note that for this symmetric location model, with probability one, the midpoint of any pair of distinct X_i's yields the maximum likelihood estimator (MLE). As a result, the likelihood ratio test statistic is $2\log 4$, i.e., a constant. In other words, the likelihood ratio test statistic gives us no guidance in testing the parameter θ. To avoid this problem, we used a modified version with a plug-in density estimator of η. With location parameter θ, an estimate of $f_{\theta,\eta}(x)$ is

$$\tilde{f}_\theta(x) = \frac{1}{2}\left[\hat{f}(x) + \hat{f}(2\theta - x)\right],$$

where $\hat{f}$ is given by (25). Then, the "plug-in" likelihood ratio test statistic is

$$T_{\mathrm{lrt}} = 2\left[\sup_{\theta\in\mathbb{R}} \sum_{i=1}^n \log \tilde{f}_\theta(X_i) - \sum_{i=1}^n \log \tilde{f}_0(X_i)\right]. \tag{46}$$

When η is the standard normal distribution, the likelihood ratio test statistic is equal to

$$\begin{aligned} T_{\mathrm{lrt}0} &= 2\left[\sup_{\theta\in\mathbb{R}} \sum_{i=1}^n \log \eta(X_i - \theta) - \sum_{i=1}^n \log \eta(X_i)\right] \\ &= 2\left[\sum_{i=1}^n \log \eta(X_i - \bar{X}) - \sum_{i=1}^n \log \eta(X_i)\right], \end{aligned} \tag{47}$$

where $\bar{X}$ is the sample mean. The statistic $T_{\mathrm{lrt}0}$ is considered here solely for comparison purposes, and note that it is not available in practice, as η is unknown in the semiparametric setup. We reject H_0 when $T_{\mathrm{lrt}} > \chi^2_{1,0.05}$ (or $T_{\mathrm{lrt}0} > \chi^2_{1,0.05}$) if $\alpha = 0.05$.

For the statistics T_1 and T_2 defined by (44) and (45), the estimator $\hat{\eta}$ is chosen as the one defined by (26) with $b(x) = [\frac{1}{2} + \frac{1}{2}\cos(\pi x)]I_{[-1,1]}$ and $c_n = \log n$, which satisfy the conditions on $b(x)$ and c_n discussed in Sect. 5.1. Here $I_{[-1,1]}$ denotes the indicator function over the interval $[-1, 1]$. Even though one can employ $\hat{\eta}$ defined by (26) as an estimator of η, the truncation term b_n used in (26) has been introduced there solely for technical purposes, and as we will observe later that $\hat{\eta}$ works equally well even without the truncation term b_n. For the preliminary location estimator $\tilde{\theta}$ in (26), we used the sample median. For the kernel estimator $\hat{f}$ defined in (25), we employed the logistic kernel $K(x) = (e^x + e^{-x} + 2)^{-1}$ and bandwidth $h_n = 0.5 S_n n^{-1/5}$, where $S_n = S_n(X_1, \ldots, X_n)$ is a robust scale statistic (generally estimate the scale parameter of the distribution). Here we take $S_n = (.674)^{-1}\text{median}\{|X_i - \tilde{\theta}|\}$. Note that this choice of K satisfies the conditions of Lemma 4. For comparison, we also give results when the quartic kernel function $K(x) = \frac{15}{16}(1 - x^2)^2 I_{[-1,1]}(x)$ is used, for which $K^{(1)}/K$ is not bounded, so it does not satisfy the conditions of Lemma 4. We have taken $N = 10000$ repetitions and different sample sizes $n = 50, 100, 250, 500, 1000$ in our simulation. Tests were carried out with a normal approximation and the significance level $\alpha = 0.05$. The calculated empirical type I errors are presented in Table 1. By definition, T_{lrt0} does not depend on the choice of kernel function.

In Table 1, we observe that when the proposed logistic kernel is used, all empirical values of the probability of type I error for both T_1 and T_2 are very close to 0.05, while T_1 performs slightly better than T_2 and both are much better than T_{lrt}. When the quartic kernel is used, T_1 tends to give smaller probability of type I error than 0.05, while both T_2 and T_{lrt} give much higher probability of type I error than the level 0.05. This demonstrates that the logistic and quartic kernels perform quite differently, with the former generally showing a chance of much closer values to 0.05 for the type I error, regardless of whether T_1, T_2 or T_{lrt} is considered. Note that the logistic kernel is the recommended one here, whereas the quartic kernel does not satisfy the conditions of Lemma 4. The results for the quartic kernel are given here for comparison purposes only. The kernel density estimator $\hat{f}$ given by (25) performs equally well with or without the truncation term b_n. We also observe that T_{lrt0} performs consistently best over different sample sizes with probability of type I error being close to 0.05 over T_1, T_2 and T_{lrt}. Note again that T_{lrt0} is considered here solely for comparison purposes, and it is not available in practice.

Empirical powers of T_1 and T_2 were also calculated and compared to those of T_{lrt} and T_{lrt0} based on $N = 10000$ repetitions and the significance level $\alpha = 0.05$. We also considered several values for β, namely $\beta = 1, 2, 4, 10$. For the reasons discussed above, we only report here the results for the logistic kernel estimator without the truncation term, and the results are similar when truncation is applied. Simulation results for the empirical powers of the statistics under study are given in Table 2. In Table 2, we observe that the power increases first and then stabilizes when the sample size increases. It even decreases slightly for some test statistics considered. For a fixed sample size, the power should increase when the β value increases, intuitively speaking. This is the case for all the test statistics considered.

To investigate if the proposed test statistics T_1 and T_2 have retained any robustness properties that are generally exhibit in MHD estimation/testing context, we examined their behavior in the presence of outliers. In particular, we looked at how outliers affect

Table 1 Probability of type I error of test statistics T_1, T_2, T_{lrt} and T_{lrt0}

n	Logistic kernel			Quartic kernel			T_{lrt0}
	With truncation $T_1(T_2)$	No truncation $T_1(T_2)$	T_{lrt}	With truncation $T_1(T_2)$	No truncation $T_1(T_2)$	T_{lrt}	
50	0.0436 (0.0544)	0.0434 (0.0544)	0.0593	0.0126 (0.1791)	0.0126 (0.1791)	0.1864	0.0487
100	0.0440 (0.0527)	0.0440 (0.0525)	0.0721	0.0176 (0.2237)	0.0176 (0.2237)	0.1904	0.0463
250	0.0467 (0.0523)	0.0464 (0.0523)	0.0662	0.0249 (0.2584)	0.0249 (0.2584)	0.1934	0.0496
500	0.0476 (0.0536)	0.0476 (0.0535)	0.0663	0.0234 (0.2579)	0.0234 (0.2579)	0.1946	0.0491
1000	0.0517 (0.0562)	0.0517 (0.0563)	0.0662	0.0200 (0.2482)	0.0200 (0.2482)	0.2021	0.0517

Table 2 Power of test statistics T_1, T_2, T_{lrt} and T_{lrt0} for $\beta = 1, 2, 4, 10$

n	$\beta = 1$			$\beta = 2$			$\beta = 4$			$\beta = 10$		
	$T_1(T_2)$	T_{lrt}	T_{lrt0}	$T_1(T_2)$	T_{lrt}	T_{lrt0}	$T_1(T_2)$	T_{lrt}	T_{lrt0}	$T_1(T_2)$	T_{lrt}	T_{lrt0}
50	0.2186 (0.2549)	0.1514	0.1660	0.5757 (0.6188)	0.4409	0.5109	0.9835 (0.9878)	0.9469	0.9805	0.9999 (1)	1	1
100	0.2327 (0.2623)	0.1826	0.1697	0.5888 (0.6263)	0.4836	0.5082	0.9896 (0.9896)	0.9305	0.9797	1 (1)	0.9596	1
250	0.2448 (0.2678)	0.1890	0.1709	0.6058 (0.6308)	0.5064	0.5090	0.9875 (0.9900)	0.9648	0.9796	1 (1)	0.9942	1
500	0.2471 (0.2656)	0.1890	0.1700	0.6252 (0.6469)	0.5211	0.5229	0.9873 (0.9901)	0.9711	0.9792	1 (1)	0.9994	1
1000	0.2530 (0.2667)	0.1830	0.1711	0.6208 (0.6398)	0.5145	0.5117	0.9885 (0.9902)	0.9716	0.9794	1 (1)	1	1

the chance of making type I error. In general, the presence of an outlier in the same direction as the alternative hypothesis would increase the chance of making type I error and the power simultaneously, no matter which test statistic is considered. Therefore, we are more concerned about the probability of type I error rather than the power of a test statistic.

After drawing data X_i's from the standard normal distribution with a sample size of $n = 50$, we contaminated it by replacing the last one or two observations by one or two outliers of the same value. Any integer between -15 and 15 is considered an outlier. Simulation results for the type I error probability of T_1, T_2, T_{lrt} and T_{lrt0} are plotted in Fig. 1 with (a) for contaminated data with one outlier, and (b) for contaminated data with two outliers of the same value. Again the logistic kernel and without the truncation was implemented for T_1 and T_2.

When compare (a) and (b) in Fig. 1, we observe a similar behavior for the three statistics T_1, T_2 and T_{lrt}, while T_{lrt0} performs somewhat differently for the two cases. When the outlier value is beyond the range of the interval $[-5, 5]$, the chances of making type I error for T_1, T_2 and T_{lrt} are almost constants around 0.05, 0.06 and 0.07, respectively. When the outlier value is within the range of $[-5, 5]$, the chance of making type I error for T_{lrt} is quite stable, while those for T_1 and T_2 fluctuate somewhat but still within the range of the interval $[0.01, 0.1]$. On the other hand, T_{lrt0} exhibits a completely different pattern for the probability of type I error; it is approximately symmetric about zero and increases dramatically in both directions and reaches one when two outliers are present. These observations indicates that T_1, T_2 and T_{lrt} possess some good robustness properties, while T_{lrt0} appears to be lacking in this aspect. In general, kernel estimators place higher weight over the range where the data are condensed. As a result, T_1, T_2 and T_{lrt} essentially treat outlier values as outliers and ignore them, and this may be the reason why they are robust against outliers, whereas, even with more information used (i.e., normality is assumed known), T_{lrt0} treats the presence of outlier observations as an evidence of wrong hypothesized parameter value and shows incorrectly a very high probability of type I error. This indicates that it is not robust against outliers.

We also investigated how the proposed test statistic performs when the percentage of contamination increases. For this purpose, we deliberately contaminated the sample of size $n = 50$ with a percentage of p outliers from a normal distribution $N(m, 0.5^2)$. We considered several p values: $p = 4\%, 10\%, 20\%, 30\%$, i.e., 2, 5, 10 and 15 outliers, respectively; and varying m values ranging from -15 to 15. In other words, the data are from a two-component normal mixture, and one component is the contamination data with varying contamination rate p and varying mean m. Simulation results are displayed in Fig. 2. In Fig. 2, the x- and y-axes represent the mean of the normal contamination component and the probability of rejecting H_0, respectively. Note that the y-axis of Fig. 2 is labeled as "Probability of Rejecting H_0." This is because when p is large, such as 20%, it is hard to distinguish if those 20% of data are outliers from other populations or they are just part of the population of our interest. Thus, it is not quite appropriate to conclude that rejecting H_0 means making a type I error. In Fig. 2, we observe again that the performance of T_1 and T_2 are quite similar, with T_1 having a little less chance of rejecting H_0 than T_2 . When the contamination data are on the left side of the correct data, there is no indication that H_1 (positive θ) is correct, and thus,

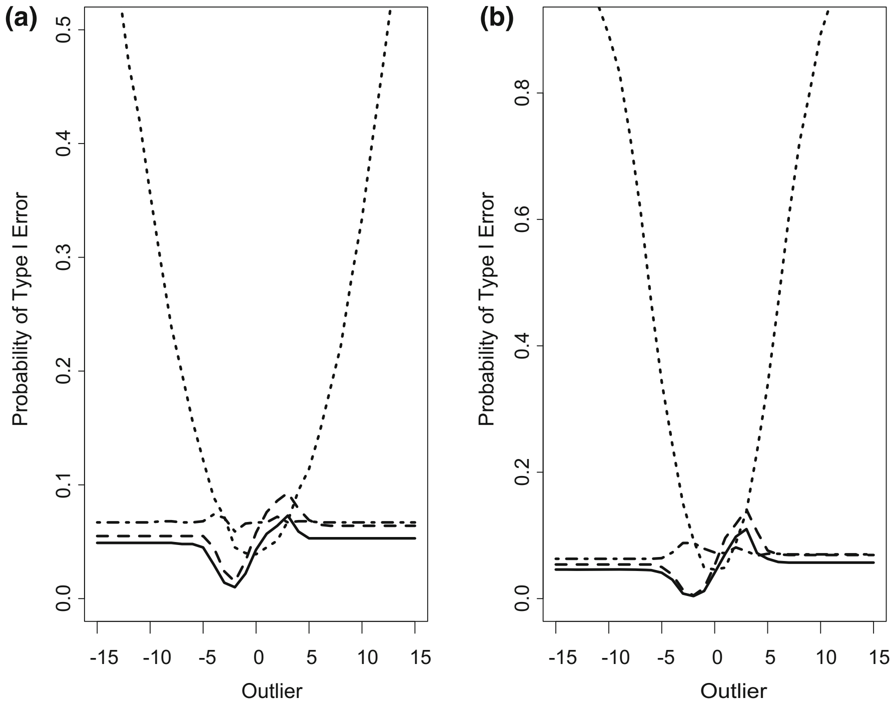

Fig. 1 Probability of type I error for T_1 (*solid*), T_2 (*dashed*), T_{lrt} (*dot dashed*) and T_{lrt0} (*dotted*). **a** One outlier. **b** Two outliers

T_1 and T_2 show a very little chance of rejecting H_0. When the contamination data are close and on the immediate right side of the correct data, then T_1 and T_2 treat them as correct data, which make the population mean positive, and as a result, T_1 and T_2 reject H_0 with high probability. When the contamination data keep moving to the right we observe the following: if p is small then T_1 and T_2 treat them as outliers and are less likely to reject H_0, and if p is large then T_1 and T_2 are not treating them as outliers completely, which results in a moderate chance of rejecting H_0. On the other hand, T_{lrt0} treats the contamination data as non-outliers whenever they are close to zero and thus rejects H_0 with high probability. Furthermore, it treats them as outliers whenever they are away from zero and thus is less likely to reject H_0. The performance of T_{lrt0} is very different from others: When the contamination data are not close to zero, T_{lrt0} has a very high probability, even close to one, of rejecting H_0.

7 Concluding remarks

Hellinger distance-based methods have been applied with great success in the literature in a variety of inference problems, especially in estimation problems of parametric and semiparametric models. There is very little research available on testing of hypotheses problems with the use of Hellinger distance (Basu et al. 2016). In fact, there is no reported work we are aware of on testing of hypotheses in semiparametric models with Hellinger distance methods. This paper is an attempt to fill in this gap in the literature.

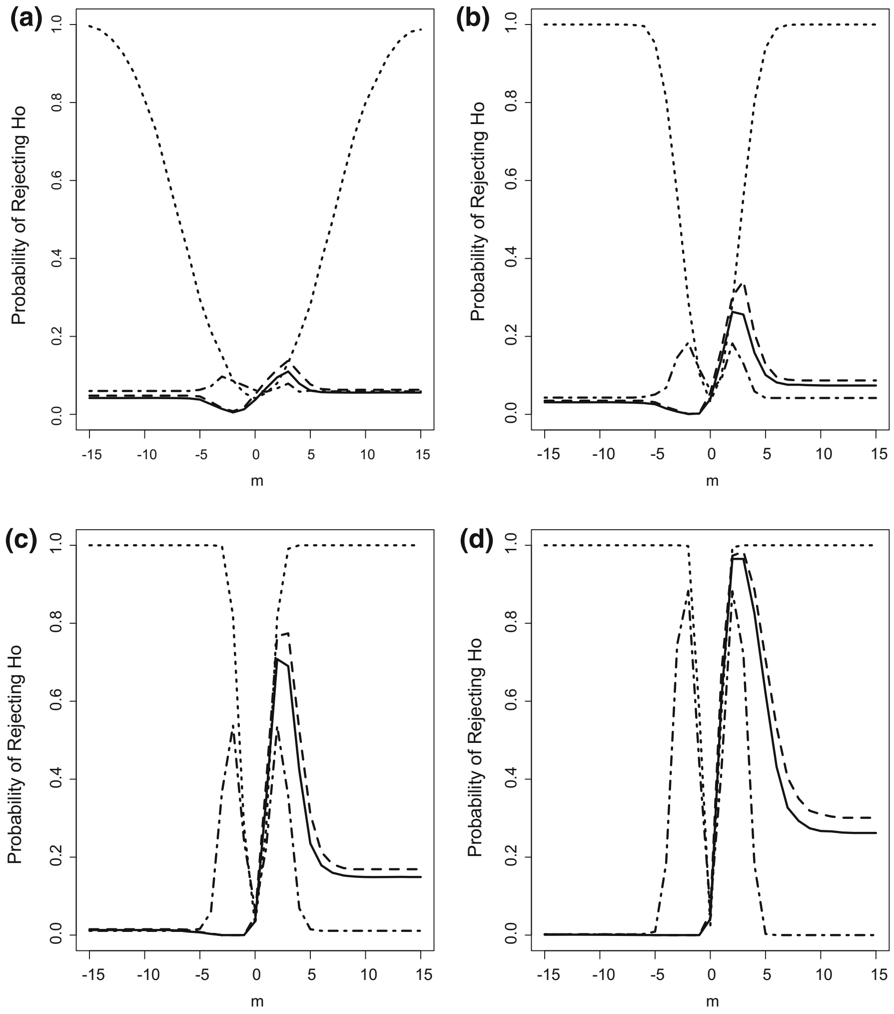

Fig. 2 Probability of type I error for T_1 (*solid*), T_2 (*dashed*), T_{lrt} (*dotdashed*) and T_{lrt0} (*dotted*) with different contamination rate p. **a** $p = 4\%$, **b** $p = 10\%$, **c** $p = 20\%$ and **d** $p = 30\%$

Treating the testing problem as a model selection problem, we have constructed a test statistic for regular semiparametric models. The proposed test statistic has an explicit expression. Furthermore, it possesses desirable properties such as asymptotic normality under both null and alternative hypotheses and asymptotic optimality (locally uniformly most powerful). We have also observed that the proposed test statistic has some desirable robustness properties such as resistance to outliers and model misspecification. Optimality combined with excellent robustness properties make the proposed testing procedure appealing in practical applications.

Although we have concentrated on the continuous data case in the development of the proposed test statistic, it is easy to modify the test statistic for the discrete case.

If we have a sample from a discrete distribution, then we use the empirical density as our density estimator $\hat{f}$ in the test statistic $4\int \hat{\rho}_{\theta_0}\hat{f}^{1/2}\mathrm{d}\nu$, where $\hat{f}$ is now given by

$$\hat{f}(x) = n^{-1}\sum_{i=1}^{n} I\{X_i = x\}, \quad x = 0, 1, 2, \ldots,$$

where $I\{\cdot\}$ is the indicator function. In this case, the test statistic $4\int \hat{\rho}_{\theta_0}\hat{f}^{1/2}\mathrm{d}\nu$ with $\hat{\rho}_{\theta_0} = \dot{s}_{\theta_0,\hat{\eta}}$ is a simple modification of the classical optimal test statistic. The optimality and robustness of the statistic are almost clear upon inspection of the form of this modification.

The proposed testing method may be applied to various applications. In particular, the proposed approach can be used to test the parameters in the two-sample semiparametric density ratio model examined in Wu et al. (2010) and Wu and Karunamuni (2009). This type of statistical model naturally arises in case-control studies and logistic discriminant analyses. After a delicate reparametrization, this semiparametric model is equivalent to a prospective logistic regression model, which is widely used in health-related applications and in analysis of case-control studies. Furthermore, with appropriate modifications, the proposed testing procedure can also be studied for testing the parameters in dose-response studies models (see Karunamuni et al. 2015, and the references therein). In a typical dose-response studies model, at a given dose x, one assumes that the response Y is a Bernoulli random variable with probability of "success" being $g(x)$, i.e., $\Pr(Y = 1|x) = g(x)$. The statistical problem concerns the estimation of "effective dose" levels defined as $ED_p = g^{-1}(p)$ with $0 < p < 1$, where $g^{-1}(\cdot)$ is the inverse function of $g(\cdot)$. Note that ED_p can be interpreted as the dose at which the probability of response is p. For example, if $p = 0.5$ then $ED_{0.5}$ is the dose that produces a desired effect in half of the test population. Pharmacology studies typically focus on estimating $ED_{0.5}$, whereas in toxicology studies the main interest is estimating ED_p for smaller values of p. The importance of estimating extreme percentage points, such as $ED_{0.90}$ and $ED_{0.95}$, is also well known. It is commonly assumed a parametric model for the dose–response curve: $g_\theta(x) = F(\mathbf{z}^T\theta)$, where $\theta = (\alpha, \beta)^T$ are unknown parameters, $\mathbf{z}(x) = (1, x)^T$, and F is some known cumulative distribution function, also known as the *link function*. Then, the data $\{Y_1, \ldots, Y_n\}$ follow a semiparametric Bernoulli model: $f_{\theta,\eta}(y) = \Pr(Y = y) = (p_{\theta,\eta})^y(1 - p_{\theta,\eta})^{1-y}$, $y = 0, 1$, where $p_{\theta,\eta} = \int g_\theta(x)\mathrm{d}\eta(x)$ with $\eta(.)$ being an unknown distribution of the X's. Hence, the basic problem reduces to estimation/testing of unknown parameter vector $\theta = (\alpha, \beta)^T$ based on a sample $\{Y_1, \ldots, Y_n\}$. These results will be discussed in a separate article.

Acknowledgements We wish to thank the Chief Editor, Professor Kenji Fukumizu, an Associate Editor and two reviewers for their helpful comments and suggestions that led to a substantial improvement in this paper. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

Appendix

Proof of Theorem 1 Write

$$\begin{aligned}\int \hat{\rho}_{\theta_0}\hat{s}\,\mathrm{d}\nu &= \int (\hat{\rho}_{\theta_0} - \dot{s}_{\theta_0,\eta})\hat{s}\,\mathrm{d}\nu + \int \dot{s}_{\theta_0,\eta}\hat{s}\,\mathrm{d}\nu \\ &= \int (\hat{\rho}_{\theta_0} - \dot{s}_{\theta_0,\eta})(\hat{s} - s_{\theta_0,\eta})\mathrm{d}\nu + \int (\hat{\rho}_{\theta_0} - \dot{s}_{\theta_0,\eta})s_{\theta_0,\eta}\mathrm{d}\nu + \int \dot{s}_{\theta_0,\eta}\hat{s}\,\mathrm{d}\nu \\ &= I_1 + I_2 + I_3. \end{aligned} \tag{48}$$

Since $\mathcal{F}_\eta$ is Hellinger-regular, we have $\int \dot{s}_{t,\eta}s_{t,\eta}d\nu = 0$ for $t \in \Theta$. Using this result together with (14), we obtain

$$\begin{aligned} I_2 &= \int (\hat{\rho}_{\theta_0} - \dot{s}_{\theta_0,\eta})s_{\theta_0,\eta}\mathrm{d}\nu \\ &= o_P(n^{-1/2}). \end{aligned} \tag{49}$$

From (12), we also have

$$\begin{aligned} I_3 &= \int \dot{s}_{\theta_0,\eta}\hat{s}\,\mathrm{d}\nu \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{s}_{\theta_0,\eta}}{2s_{\theta_0,\eta}}(X_i) + o_P(n^{-1/2}) \\ &= \frac{1}{4n}\sum_{i=1}^{n} \dot{\ell}_{\theta_0,\eta}(X_i) + o_P(n^{-1/2}). \end{aligned} \tag{50}$$

Finally, from the Cauchy–Schwarz inequality, (11) and (13) it follows that

$$\begin{aligned} |I_1| &\le \int \left|\hat{\rho}_{\theta_0} - \dot{s}_{\theta_0,\eta}\right| \left|\hat{s} - s_{\theta_0,\eta}\right| \mathrm{d}\nu \\ &\le \left\|\hat{\rho}_{\theta_0} - \dot{s}_{\theta_0,\eta}\right\| \left\|\hat{s} - s_{\theta_0,\eta}\right\| \\ &= o_P((na_n)^{-1/2})O_P(a_n^{1/2}) \\ &= o_P(n^{-1/2}). \end{aligned} \tag{51}$$

The proof of (15) is now completed by combining (48) to (51). □

Proof of Theorem 2 Part (i) follows from (15) and the central limit theorem, since $E\dot{\ell}_{\theta_0,\eta}(X_i) = 0$ and $Var(\dot{\ell}_{\theta_0,\eta}(X_i)) = I_{\theta_0,\eta}$. Let $P_{\theta,\eta}$ denotes the probability distribution of the density $f_{\theta,\eta}$, where $f_{\theta,\eta}$ belongs to the family (1). By Le Cam's first lemma, the sequences of joint probability measures $\{P^n_{\theta_0,\eta}\}$ (the n-fold product of $P_{\theta_0,\eta}$) and $\{P^n_{\theta_0+\beta n^{-1/2},\eta}\}$ of the null and alternative hypotheses, respectively, are contiguous; see, e.g., Proposition 2.1.3 of Bickel et al. (1998, p. 395). Consequently, the statement of (15) continues to hold when $X_1, \ldots, X_n$ are i.i.d with density $f_{\theta_n,\eta}$. Furthermore, since $\mathcal{F}_\eta$ is Hellinger-regular we have

$$\begin{aligned} f_{\theta_n,\eta}^{1/2} - f_{\theta_0,\eta}^{1/2} &= s_{\theta_n,\eta} - s_{\theta_0,\eta} \\ &= \frac{1}{2} n^{-1/2} \beta \dot{s}_{\theta_0,\eta} + R_n, \end{aligned} \tag{52}$$

where $\|R_n\|^2 = o(n^{-1/2})$ from (2). Part (ii) now follows from (15) and (52). □

The proofs of Lemmas 1 to 5 are given in supplementary material.

References

Banerjee, M. (2000). Likelihood ratio inference in regular and nonregular problems. Ph.D. dissertation, University of Washington.

Banerjee, M. (2005). Likelihood ratio tests under local alternatives in regular semiparametric models. *Statistica Sinica*, *15*, 635–644.

Basu, A., Shioya, H., Park, C. (2011). *Statistical inference*: *the minimum distance approach*. Boca Raton, Florida: CRC Press.

Basu, A., Mandal, A., Martin, N., Pardo, L. (2013). Testing statistical hypotheses based on the density power divergence. *Annals of the Institute of Statistical Mathematics*, *65*, 319–348.

Basu, A., Mandal, A., Martin, N., Pardo, L. (2016). Generalized Wald-type tests based on minimum density power divergence estimators. *Statistics*, *50*, 1–26.

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

Beran, R. (1978). An efficient and robust adaptive estimator of location. *The Annals of Statistiscs*, *6*, 292–313.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistiscs*, *10*, 647–671.

Bickel, P. J., Ritov, Y. (1987). Efficient estimation in the errors in variables model. *The Annals of Statistiscs*, *15*, 513–540.

Bickel, P. J., Klaassen, C., Ritov, Y., Wellner, J. (1998). *Efficient and adaptive estimation for semiparametric models*. New York: Springer.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B*, *34*, 187–202.

Huber, P., Ronchetti, E. M. (2009). *Robust statistics* (2nd ed.). New York: Wiley.

Karunamuni, R. J., Wu, J. (2009). Minimum Hellinger distance estimation in a nonparametric mixture model. *Journal of Statistical Planning and Inference*, *139*, 1118–1133.

Karunamuni, R. J., Wu, J. (2011). One-step minimum Hellinger distance estimation. *Compututational Statistics and Data Analysis*, *55*, 3148–3164.

Karunamuni, R. J., Tang, Q., Zhao, B. (2015). Robust and efficient estimation of effective dose. *Computational Statistics and Data Analysis*, *90*, 47–60.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. New York: Springer.

Lehmann, E. L. (1997). *Testing statistical hypotheses*. New York: Springer.

Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.

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

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, vol. 5. Institute for Mathematical Statistics, Hayward, California.

Murphy, S. A., van der Vaart, A. W. (1996). Likelihood inference in the errors-in-variables model. *Journal of Multivariate Analysis*, *59*, 81–108.

Murphy, S. A., van der Vaart, A. W. (1997). Semiparametric likelihood ratio inference. *The Annals of Statistics*, *25*, 1471–1509.

Murphy, S. A., van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, *95*, 449–485.

Pardo, L. (2006). *Statistical inference based on divergence measures*. Boca Raton, FL: Chapman & Hall/CRC.

Pfanzagl, J., Wefelmeyer, W. (1982). *Contributions to a general asymptotic statistical theory*. New York: Springer.

Poland, W. B., Shachter, R. D. (1994). Three approaches to probability model selection. *Uncertainty in Artificial Intelligence*: *Proceedings of the Tenth Conference, San Mateo* (CA: Morgan Kaufmann), pp. 478–483.

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 tests: Efficiency, breakdown points and examples. *Journal of the American Statistical Association*, *84*, 107–113.

Stather, C. R. (1981). Robust statistical inference using Hellinger distance methods. Ph.D. dissertation, LaTrobe University, Australia.

Tang, Q., Karunamuni, R. J. (2013). Minimum distance estimation in a finite mixture regression mixture regression model. *Journal of Multivariate Analysis*, *120*, 185–204.

Umashanger, T., Sriram, T. N. (2009). L_2-estimation of mixture complexity for count data. *Computational Statistics and Data Analysis*, *53*, 4243–4254.

van der Vaart, A. W. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *The Annals of Statistics*, *24*, 862–878.

van der Vaart, A. W. (2000). *Asymptotic statistics*. First paperback edition: Cambridge University Press, UK.

Woo, M.-J., Sriram, T. N. (2006). Robust estimation of mixture complexity. *Journal of the American Statistical Association*, *101*, 1475–1486.

Woo, M.-J., Sriram, T. N. (2007). Robust estimation of mixture complexity for count data. *Computational Statistics and Data Analysis*, *51*, 4379–4392.

Wu, J. (2007). Minimum Hellinger distance estimation in semiparametric models. Ph.D. dissertation, University of Alberta.

Wu, J., Karunamuni, R. J. (2009). On minimum Hellinger distance estimation. *Canadian Journal of Statistics*, *37*, 514–533.

Wu, J., Karunamuni, R. J. (2012). Efficient Hellinger distance estimates for semiparametric models. *Journal of Multivariate Analysis*, *107*, 1–23.

Wu, J., Karunamuni, R. J. (2015). Profile Hellinger distance estimation. *Statistics*, *49*, 711–740.

Wu, J., Karunamuni, R. J., Zhang, B. (2010). Minimum Hellinger distance estimation in a two-sample semiparametric model. *Journal of Multivariate Analysis*, *101*, 1102–1122.