Bernoulli **22**(2), 2016, 857–900
DOI: 10.3150/14-BEJ678

Consistency, efficiency and robustness of conditional disparity methods

GILES HOOKER

[1]*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853-4201, USA. E-mail: Giles.hooker@cornell.edu*

This paper considers extensions of minimum-disparity estimators to the problem of estimating parameters in a regression model that is conditionally specified; that is where a parametric model describes the distribution of a response y conditional on covariates x but does not specify the distribution of x. We define these estimators by estimating a non-parametric conditional density estimates and minimizing a disparity between this estimate and the parametric model averaged over values of x. The consistency and asymptotic normality of such estimators is demonstrated for a broad class of models in which response and covariate vectors can take both discrete and continuous values and incorportates a wide set of choices for kernel-based conditional density estimation. It also establishes the robustness of these estimators for a broad class of disparities. As has been observed in Tamura and Boos (*J. Amer. Statist. Assoc.* **81** (1986) 223–229), minimum disparity estimators incorporating kernel density estimates of more than one dimension can result in an asymptotic bias that is larger that $n^{-1/2}$ and we characterize a similar bias in our results and show that in specialized cases it can be eliminated by appropriately centering the kernel density estimate. We also demonstrate empirically that bootstrap methods can be employed to reduce this bias and to provide robust confidence intervals. In order to demonstrate these results, we establish a set of L_1-consistency results for kernel-based estimates of centered conditional densities.

Keywords: bootstrap; density estimation; disparity; regression; robust inference

1. Introduction

Minimum disparity estimators (MDEs) are based on minimizing a measure of distance between a non-parametric density estimate $\hat{f}_n(y)$ and a parametric family of densities $\phi_\theta(y)$. Disparities can be written in the general form Lindsay [11]:

$$D(\hat{f}_n, \theta) = \int C\left(\frac{\hat{f}_n(y) - \phi_\theta(y)}{\phi_\theta(y)}\right)\phi_\theta(y)\,\mathrm{d}\nu(y),$$

where C is a convex function with a minimum at 0 and ν is a reference measure over the space of y. The minimum disparity estimator is defined to be

$$\hat{\theta}_n = \arg\min_\theta D(\hat{f}_n, \theta).$$

When $\hat{f}_n$ is a kernel density estimate based on univariate i.i.d. data and $C(\delta)$ behaves appropriately at 0, these estimators can be shown to be asymptotically normal and efficient in the sense of having asymptotic variance given by the inverse of the Fisher information. When C

behaves appropriately at ∞, they are also robust to outliers. This was first observed in the case of Hellinger distance ($C(\delta) = [\sqrt{\delta + 1} - 1]^2$) by Beran [3] and generalized to the broader class of disparities in Lindsay [11] for discrete data and for continuous data in Basu and Lindsay [1] and Park and Basu [13]. The particular case of $C(\delta) = \mathrm{e}^{-\delta}$ was studied in Basu, Sahadeb and Vidyashankar [2]; a choice that that is both robust to outliers and to "inliers" – regions where $\delta(\cdot) = [\hat{f}_n(x\cdot) - \phi_\theta]/\phi_\theta$ is near it's negative limit of -1 and where Hellinger distance performs poorly. Tamura and Boos [16] observed that when $\hat{f}_n(x\cdot)$ is a multivariate kernel density estimate, the MDE has an asymptotic bias that is larger than $n^{-1/2}$ and hence appears in the central limit theorem for $\hat{\theta}_n$, potentially necessitating a bias correction.

Despite the potential for both robust and efficient estimation, minimum disparity estimation has seen few extensions beyond i.i.d. data. Within this context, the use of disparity methods to estimate parameters in linear regression was treated in Pak and Basu [12] by placing a disparity on the score equations and for discrete covariates in Cheng and Vidyashankar [4], but little attention has been given to more general regression problems and we take a more direct approach here. In this paper, we consider data $(X_1, Y_1), (X_2, Y_2), \ldots$ for which we have a parameterized family of densities $\phi_\theta(y|x)$ which describe the distribution of y conditional on the value of x. We construct a non-parametric conditional density estimate $\breve{f}_n(y|x)$ based on kernel densities and define two extensions of disparities:

$$D_n(\breve{f}_n, \theta) = \frac{1}{n}\sum_{i=1}^{n} D\big(\breve{f}_n(\cdot|X_i), \phi_\theta(\cdot|X_i)\big),$$

$$\tilde{D}_n(\breve{f}_n, \theta) = \int D\big(\breve{f}_n(\cdot|x), \phi_\theta(\cdot|x)\big)\hat{h}_n(x)\,\mathrm{d}x,$$

where $\hat{h}_n(x)$ is a kernel density estimate of the density of x. We show that the parameters minimizing these disparities are consistent and asymptotically normal. Furthermore, when the data are generated from a process that corresponds to some member of the parametric model, the limiting variance is given by the information matrix. Our framework is intentionally general and designed to cover a broad range of cases in which both Y_i and X_i can be vector valued and incorporate a mix of continuous- and discrete-valued components and are designed to be as general as possible. We also consider various estimates of $\breve{f}_n(y|x)$ in which some components of y are centered by a Nadaraya–Watson estimator based on some components of x. When the parametric model is incorrect, these yield different bias and variance expressions in our central limit theorem which we interpret and describe.

To achieve these results, we first demonstrate the L_1 consistency of $\breve{f}_n(\cdot|x)$ which holds uniformly over x. We also demonstrate the robustness of these estimators to outlying values in y. The effectiveness of these techniques are then examined in simulation and with real-world data.

We will introduce the specific distributional framework and assumptions in the next subsection and our conditional density estimators in Section 1.2. Because of the notational complexity involved with working with both continuous and discrete random variables as well as a division of the components of x, Section 1.3 will detail notational shorthand that will be used in various places throughout the remainder of the paper. Section 2 will develop results on the L_1 consistency of kernel-based conditional density estimators, Section 3 will then apply these results to

demonstrate the consistency of minimum-disparity estimators in conditionally specified models. We will demonstrate the asymptotic normality of these estimators in Section 4 and their robustness will be examined in Section 5. Computational details on selecting bandwidths and using the bootstrap for bias correction and inference are given in Section 6. Simulation results and real data analysis are given in Sections 7 and 8.

We have included proofs of our results in the text where they are either enlightening or short, but have reserved many for a Supplemental Appendix (Hooker [8]) and noted where these may be found.

1.1. Framework and assumptions

Throughout the following, we assume a probability space $(\Omega, \mathcal{F}, P)$ from which we observe i.i.d. random variables $\{X_{n1}(\omega), X_{n2}(\omega), Y_{n1}(\omega), Y_{n2}(\omega), n \geq 1\}$ where we have separated discrete and continuous random variables so that $X_{n1}(\omega) \in \mathbb{R}^{d_x}$, $X_{n2}(\omega) \in S_x$, $Y_{n1}(\omega) \in \mathbb{R}^{d_y}$, $Y_{n2}(\omega) \in S_y$ for countable sets S_x and S_y with joint distribution

$$g(x_1, x_2, y_1, y_2) = P(X_2 = x_2, Y_2 = y_2) P(X_1 \in \mathrm{d}x_1, Y_1 \in \mathrm{d}y_1 | X_2 = x_2, Y_2 = y_2)$$

and define the marginal and conditional densities

$$h(x_1, x_2) = \sum_{y_2 \in S_y} \int g(x_1, x_2, y_1, y_2)\, \mathrm{d}y_1, \tag{1.1}$$

$$f(y_1, y_2 | x_1, x_2) = \frac{g(x_1, x_2, y_1, y_2)}{h(x_1, x_2)} \tag{1.2}$$

on the support of (x_1, x_2).

An important aspect of this paper is to study an approach of centering y_1 by a Nadaraya–Watson estimator before estimating g. We define this generally, so that y_1 can be centered based on some components $(X_1^{\bar{m}}, X_2^{\bar{m}})$ of (X_1, X_2) and a density for the residuals can be estimated based on a different possibly-overlapping set of components $(X_1^{\bar{g}}, X_2^{\bar{g}})$. Formally, we define $(x_1^{\bar{m}}, x_2^{\bar{m}})$ and $(x_1^{\bar{g}}, x_2^{\bar{g}})$ with densities $h^{\bar{m}}(x_1^{\bar{m}}, x_2^{\bar{m}})$ and $h^{\bar{g}}(x_1^{\bar{g}}, x_2^{\bar{g}})$, respectively, where $x_1^{\bar{m}} \in \mathbb{R}^{d_{x^{\bar{m}}}}$ and $x_1^{\bar{g}} \in \mathbb{R}^{d_{x^{\bar{g}}}}$ and $x_2^{\bar{m}} \in S_{x^{\bar{m}}}$, $x_2^{\bar{g}} \in S_{x^{\bar{g}}}$. We now define the possibly vector-valued expectation of y_1 conditional on $x^{\bar{m}}$:

$$m\big(x_1^{\bar{m}}, x_2^{\bar{m}}\big) = \sum_{y_2 \in S_y} \sum_{x_2^{\bar{g}} \in S_{x^{\bar{g}}}} \iint y_1 \frac{g(x_1, x_2, y_1, y_2)}{h^{\bar{m}}(x_1^{\bar{m}}, x_2^{\bar{m}})}\, \mathrm{d}y_1\, \mathrm{d}x_1^{\bar{g}}$$

along with the residuals

$$\varepsilon = y_1 - m\big(x_1^{\bar{m}}, x_2^{\bar{m}}\big)$$

and define the joint density of these residuals, y_2, and $x^{\bar{g}}$ by

$$g^c\left(x_1^{\bar{g}}, x_2^{\bar{g}}, \varepsilon, y_2\right) = \sum_{x_2^{\bar{m}} \in S_{x^{\bar{m}}}} \int g\left(x_1, x_2, \varepsilon + m\left(x_1^{\bar{m}}, x_2^{\bar{m}}\right), y_2\right) \mathrm{d}x_1^{\bar{m}}$$

and similarly write the conditional density

$$f^c\left(\varepsilon, y_2 | x_1^{\bar{g}}, x_2^{\bar{g}}\right) = \frac{g^c(x_1^{\bar{g}}, x_2^{\bar{g}}, \varepsilon, y_2)}{h^{\bar{g}}(x_1^{\bar{g}}, x_2^{\bar{g}})},$$

where throughout this paper we will assume that the distribution of (y_1, y_2) is such that

$$f(y_1, y_2 | x_1, x_2) = f^c\left(\varepsilon + m\left(x_1^{\bar{m}}, x_2^{\bar{m}}\right), y_2 | x_1^{\bar{g}}, x_2^{\bar{g}}\right)$$

for some function $f^c(\varepsilon, y_2 | x_1^{\bar{g}}, x_2^{\bar{g}})$ that does not depend on those components of X that are not also components of $X^{\bar{g}}$.

A useful example to keep in mind is the conditionally heteroscedastic linear regression model

$$y_i = \left(x_i^{\bar{m}}\right)^T \beta + \sigma\left(\left(x_i^{\bar{g}}\right)^T \gamma\right)\varepsilon$$

for $\varepsilon \sim f(\cdot)$ in which the residual variance depends on covariates $x^{\bar{g}}$ while the mean depends on $x^{\bar{m}}$ and these may or may not be the same variables. However, our framework is considerably more general than this model and includes all of ANOVA, multiple regression, ANCOVA, multivariate regression, tabular data and generalized linear models as well as allowing for more complex models in which dependence is assumed between categorical and continuous response variables.

To appreciate the generality of class of conditional density estimates, we observe that this covers the case (1.2) by setting the collection of variables in $(x_1^{\bar{m}}, x_2^{\bar{m}})$ to be empty and $(x_1^{\bar{g}}, x_2^{\bar{g}}) = (x_1, x_2)$; in this case we understand $m(x_1^{\bar{m}}, x_2^{\bar{m}}) \equiv 0$. It also covers the "homoscedastic" in which there is no y_2 and we assume there is a density a density $f^*(e)$ such that

$$f(y_1 | x_1, x_2) = f^*\left(y_1 - m\left(x_1^{\bar{m}}, x_2^{\bar{m}}\right)\right) \tag{1.3}$$

that is, the residuals all have the same distribution. In this case, we can set $(x_1^{\bar{m}}, x_2^{\bar{m}})$ to be all the variables and remove $(x_1^{\bar{g}}, x_2^{\bar{g}})$. If we set both $x^{\bar{g}}$ and $x^{\bar{m}}$ to be the entire set x we arrive at a centered conditional density estimate

$$f^c(\varepsilon, y_2 | x_1, x_2) = f\left(\varepsilon + m(x_1, x_2), y_2 | x_1, x_2\right).$$

This centering can improve the finite sample performance of our estimator at or near the homoscedastic case in which f^c is close to constant in x_1 and hence incurs lower bias than the uncentered version.

Here we will formalize the partition of the covariate space into components associated with centering y_1 and with conditioning. To do this, we divide $x = (x_1, x_2)$ into (x^m, x^s, x^g) where x^s are the components common to both $x^{\bar{m}} = (x^m, x^s)$ and $x^{\bar{g}} = (x^s, x^g)$ with x^m and x^g containing those components only appearing one or other of the centering and conditioning variables. We define these variables to take values on spaces $\mathcal{X}^a = \mathbb{R}^{d_{xa}} \otimes S_{xa}$ for $a \in (m, s, g)$ with $\mathcal{X} = \mathcal{X}^m \otimes \mathcal{X}^s \otimes \mathcal{X}^g$ and $\mathcal{X}^{\bar{m}} = \mathcal{X}^m \otimes \mathcal{X}^s$ and $\mathcal{X}^{\bar{g}} = \mathcal{X}^s \otimes \mathcal{X}^g$, similarly the distribution of observations on these spaces will be given by $h^a(x_1^a, x_2^a)$ for a replaced by any of $(m, s, g, \bar{m}, \bar{g})$.

We note that when y_1 is vector valued, it is not necessary to center all of its components. The results below also encompass the case where only some components are centered by interpreting $m(x_1^{\bar{m}}, x_2^{\bar{m}}) = 0$ for the non-centered components. It is also possible to include y_2 within x_2^m (but not within $x_2^{\bar{g}}$) without affecting these results.

The following regularity structures may be assumed in the theorems below:

(D1) g is bounded and continuous in x_1 and y_1.
(D2) $\int y_1^2 g(x_1, x_2, y_1, y_2)\,\mathrm{d}y_1 < \infty$ for all $x \in \mathcal{X}$.
(D3) All third derivatives of g with respect to x_1 and y_1 exist, are continuous and bounded.
(D4) The support of x, $\mathcal{X}$ is compact and $h(x_1, x_2)$ is bounded away from zero with infimum

$$h^- = \inf_{(x_1,x_2)\in\mathcal{X}} h(x_1, x_2) > 0.$$

(D5) The expected value function $m(x_1, x_2)$ is bounded, as is its gradient $\nabla_{x_1} m(x_1, x_2)$.

We note that under these conditions, continuity of h and f in x_1 and y_1 is inherited from g. We also have that $\mathcal{X}^a$ is compact for $a \in (m, s, g, \bar{m}, \bar{g})$ and similarly $h^a(x_1^a, x_2^a) > h^-$. Assumption (D4) is generally employed for models involving non-parametric smoothing and is required for the uniform convergence results that we establish; in practice it is often possible to bound the range of values that a covariate can take. This assumption is, however, more restrictive than required for general regression problems and can, in fact, be removed in special cases of the methods studied here. We have noted where this is possible below, with results provided in Supplemental Appendix E (Hooker [8]).

In the case of centered densities (i.e., $x^{\bar{m}}$ is not trivial), we also assume that f is differentiable in y_1 and has a finite second moment, uniformly over x:

(E1) $\sup_{(x_1,x_2)\in\mathcal{X}} \sum_{y_2\in S_y} \int |\nabla_{y_1} f(y_1, y_2|x_1, x_2)|\,\mathrm{d}y_1 < \infty$,
(E2) $\sup_{(x_1^m,x_2^m)\in\mathcal{X}^m} \sum_{y_2\in S_y} \int |y_1^2 f(y_1, y_2|x_1, x_2)|\,\mathrm{d}y_1 < \infty$

and note that these conditions need only apply to those components of y_1 which are centered.

1.2. Kernel estimators

In order to apply the disparity methods described above, we will need estimates of $f^c(\varepsilon, y_2|x_1^{\bar{g}}, x_2^{\bar{g}})$ which we will obtain through kernel density and Nadaraya–Watson estimators.

Specifically, we first estimate the density of the centering variables $(X_1^{\bar{m}}, X_2^{\bar{m}})$:

$$\hat{h}_n^m\big(x_1^{\bar{m}}, x_2^{\bar{m}}, \omega\big) = \frac{1}{n c_{nx_2^{\bar{m}}}^{d_{x\bar{m}}}} \sum_{i=1}^n K_x^m\left(\frac{x_1^{\bar{m}} - X_{i1}^{\bar{m}}(\omega)}{c_{nx_2^{\bar{m}}}}\right) I_{x_2^{\bar{m}}}\big(X_{i2}^{\bar{m}}(\omega)\big) \tag{1.4}$$

and define a Nadaraya–Watson estimator for the continuous response variables y_1 based on them:

$$\hat{m}_n\big(x_1^{\bar{m}}, x_2^{\bar{m}}, \omega\big) = \frac{(1/n c_{nx_2^{\bar{m}}}^{d_{x\bar{m}}}) \sum_{i=1}^n Y_{i1}(\omega) K_x^m((x_1^{\bar{m}} - X_{i1}^{\bar{m}}(\omega))/c_{nx_2^{\bar{m}}}) I_{x_2^{\bar{m}}}(X_{i2}^{\bar{m}}(\omega))}{\hat{h}_n^m(x_1^{\bar{m}}, x_2^{\bar{m}})}. \tag{1.5}$$

We then obtain residuals from this estimator

$$\tilde{E}_i(\tilde{m}, \omega) = Y_i(\omega) - \tilde{m}\big(X_{i1}^{\bar{m}}(\omega), X_{i2}^{\bar{m}}(\omega)\big), \qquad i = 1, \ldots, n \tag{1.6}$$

and use these with the Y_{i2} to obtain a joint density estimate with the $(X_{i1}^{\bar{g}}, X_{i2}^{\bar{g}})$:

$$\begin{aligned} &\hat{g}_n\big(x_1^{\bar{g}}, x_2^{\bar{g}}, e, y_2, \tilde{m}, \omega\big) \\ &\quad = \frac{1}{n c_{nx_2^{\bar{g}}}^{d_{x\bar{g}}} c_{ny_2}^{d_y}} \sum_{i=1}^n K_x\left(\frac{x^{\bar{g}} - X_{i1}^{\bar{g}}(\omega)}{c_{nx_2^{\bar{g}}}}\right) K_y\left(\frac{e - \tilde{E}_i(\tilde{m}, \omega)}{c_{ny_2}}\right) I_{x_2^{\bar{g}}}\big(X_{i2}^{\bar{g}}(\omega)\big) I_{y_2}\big(Y_{i2}(\omega)\big). \end{aligned} \tag{1.7}$$

We then estimate the density of the $(X_{i1}^{\bar{g}}, X_{i2}^{\bar{g}})$ alone

$$\hat{h}_n\big(x_1^{\bar{g}}, x_2^{\bar{g}}, \omega\big) = \frac{1}{n c_{nx_2^{\bar{g}}}^{d_{x\bar{g}}}} \sum_{i=1}^n K_x\left(\frac{x_1^{\bar{g}} - X_{i1}^{\bar{g}}(\omega)}{c_{nx_2^{\bar{g}}}}\right) I_{x_2^{\bar{g}}}\big(X_{i2}^{\bar{g}}(\omega)\big) \tag{1.8}$$

and use these to obtain an estimate of the conditional distribution of the centered responses:

$$\hat{f}_n(e, y_2 | x_1, x_2, \omega) = \frac{\hat{g}_n(x_1^{\bar{g}}, x_2^{\bar{g}}, e, y_2, \hat{m}_n, \omega)}{\hat{h}_n(x_1^{\bar{g}}, x_2^{\bar{g}}, \omega)}. \tag{1.9}$$

Finally, we shift $\hat{f}_n$ by $\hat{m}_n$ to remove the centering:

$$\breve{f}_n(y_1, y_2 | x_1, x_2, \omega) = \hat{f}_n\big(y_1 - \hat{m}_n\big(x_1^{\bar{m}}, x_2^{\bar{m}}, \omega\big), y_2 | x_1^{\bar{g}}, x_2^{\bar{g}}, \omega\big). \tag{1.10}$$

Throughout the above, $I_x(X)$ is the indicator function of $X = x$ and K_x, K_x^m and K_y are densities on the spaces $\mathbb{R}^{d_{x\bar{g}}}$, $\mathbb{R}^{d_{x\bar{m}}}$ and $\mathbb{R}^{d_y}$, respectively. We have used $c_{nx_2^{\bar{m}}}$, $c_{nx_2^{\bar{g}}}$ and c_{ny_2} to distinguish the different rates which these bandwidths will need to follow. Further conditions on these are detailed below.

Here we have employed the errors $\tilde{E}_i(\tilde{m}, \omega)$ for the sake of notational compactness. We have defined centering by a generic $\tilde{m}$ in (1.6)–(1.7), which we will employ in developing its L_1

convergence below, but have replaced this with $\hat{m}_n$ in (1.9) and (1.10) to indicate real-world practice.

In the case of uncentered conditional density estimates ($x^{\bar{m}}$ trivial), these reduce to

$$\hat{g}_n^*(x_1, x_2, y_1, y_2, \omega) = \frac{1}{n c_{nx_2}^{d_x} c_{ny_2}^{d_y}} \sum_{i=1}^{n} K_x\left(\frac{x_1 - X_{i1}(\omega)}{c_{nx_2}}\right) K_y\left(\frac{y_1 - Y_{i1}(\omega)}{c_{ny_2}}\right) \times I_{x_2}\big(X_{i2}(\omega)\big) I_{y_2}\big(Y_{i2}(\omega)\big), \tag{1.11}$$

$$\begin{aligned} \hat{h}_n^*(x_1, x_2, \omega) &= \frac{1}{n c_{nx_2}^{d_x}} \sum_{i=1}^{n} K_x\left(\frac{x - X_{i1}(\omega)}{c_{nx_2}}\right) I_{x_2}\big(X_{i2}(\omega)\big) \\ &= \sum_{y_2 \in S_y} \int_{\mathbb{R}^{d_y}} \hat{g}_n^*(x_1, x_2, y_1, y_2, \omega)\, \mathrm{d}y_1, \end{aligned} \tag{1.12}$$

$$\hat{f}_n^*(y_1, y_2 | x_1, x_2, \omega) = \frac{\hat{g}_n^*(x_1, x_2, y_1, y_2, \omega)}{\hat{h}_n^*(x_1, x_2, \omega)}. \tag{1.13}$$

And for homoscedastic regression estimators ($x^{\bar{g}}$ and y_2 empty), we have

$$\hat{m}_n(x_1, x_2, \omega) = \frac{\sum_{i=1}^{n} Y_{i1}(\omega) K_x((x_1 - X_{i1}(\omega))/c_{nx_2}) I_{x_2}(X_{i2}(\omega))}{\sum_{i=1}^{n} K_x((x_1 - X_{i1}(\omega))/c_{nx_2}) I_{x_2}(X_{i2}(\omega))}, \tag{1.14}$$

$$\hat{f}_n^c(e, \omega) = \frac{1}{n c_{ny_2}^{d_y}} \sum_{i=1}^{n} K_y\left(\frac{e - (Y_i(\omega) - \hat{m}_n(X_{i1}(\omega), X_{i2}(\omega)))}{c_{ny_2}}\right), \tag{1.15}$$

$$\tilde{f}_n(y_1 | x_1, x_2, \omega) = \hat{f}_n^c\big(y_1 - \hat{m}_n(x_1, x_2, \omega), \omega\big) \tag{1.16}$$

with notation c_{ny_2} maintained as a bandwidth for the sake of consistency.

We note that while these estimates do require some extra computational work, they are not, in fact, more computationally burdensome than the methods proposed for independent, univariate data in Beran [3]. The evaluation cost of each of the density estimates and non-parametric smooths above is $\mathrm{O}(n)$ operations and $\check{f}_n(y_1, y_2 | x_1, x_2)$ can be evaluated in a few lines of code in the `R` programming language. In simulations reported in Section 7 the computing time required of our methods exceeds that of maximum likelihood methods by a factor of 10, and alternative robust methods by a factor of 5, rendering them very feasible in practical situations.

Throughout we make the following assumptions on the kernels K_x, K_x^m, and K_y. These will all conform to conditions on a general kernel $K(z)$ over a Euclidean space of appropriate dimension $\mathbb{R}^{d_z}$:

(K1) $K(z)$, is a density on $\mathbb{R}^{d_z}$.
(K2) For some finite K^+, $\sup_{z \in \mathbb{R}^{d_z}} K(z) < K^+$.
(K3) $\lim \|z\|^{2d_z} K(z) \to 0$ as $\|z\| \to \infty$.
(K4) $K(z) = K(-z)$.
(K5) $\int \|z\|^2 K(z)\, \mathrm{d}z < \infty$.
(K6) K has bounded variation and finite modulus of continuity.

We also assume that following properties of the bandwidths. These will be given in terms of the number of observations falling at each combination values of the discrete variables.

$$n\big(x_2^a\big) = \sum_{i=1}^{n} I_{x_2^a}\big(X_{2i}^a(\omega)\big), \qquad n(y_2) = \sum_{i=1}^{n} I_{y_2}\big(Y_{2i}(\omega)\big),$$

$$n\big(x_2^a, y_2\big) = \sum_{i=1}^{n} I_{x_2^a}\big(X_{2i}^a(\omega)\big) I_{y_2}\big(Y_{2i}(\omega)\big),$$

where these rates are defined for a covering any of $(m, s, g, \bar{m}, \bar{g})$ or the whole space. As $n \to \infty$:

(B1) $c_{nx_2} \to 0$, $c_{ny_2} \to 0$.
(B2) $n(x_2^a) c_{nx_2^a}^{d_{xa}} \to \infty$ for all $x_2 \in S_x$ and $n(x_2^a, y_2) c_{nx_2^a}^{d_{xa}} c_{ny_2}^{d_y} \to \infty$ for all $(x_2^a, y_2) \in S_{x^a} \otimes S_y$.
(B3) $n(x_2^a) c_{nx_2^a}^{2d_{xa}} \to \infty$.
(B4) $n(x_2^a, y_2) c_{nx_2^a}^{2d_{xa}} c_{ny_2}^{2d_y} \to \infty$.
(B5) $\sum_{n(x_2^a)=1}^{\infty} c_{nx_2^a}^{-d_{xa}} \mathrm{e}^{-\gamma n(x_2^a) c_{nx_2^a}^{d_x}} \leq \infty$ for all $\gamma > 0$.
(B6) $n(y_2) c_{ny_2}^4 \to 0$ if $d_y = 1$ and $n(x_2^a) c_{nx_2^a}^4 \to 0$ if $d_{xa} = 1$,

where the sum is taken to be over all observations in the case that X_2^a or Y_2 are singletons.

1.3. Notational conventions

Because of the complexity involved in dealing with two partitions, $x = (x^m, x^s, x^g)$ and $x = (x_1, x_2)$, along with kernel estimators and integrals, this paper will take some notational shortcuts; which ones we take will differ between sections. These will allow us to ignore notational complexities that do not affect the particular results being discussed. Here we will forecast these.

Section 2 demonstrates the consistency of kernel-based conditional density estimates. This section will require the distinction between continuous-valued and discrete-valued components of x and y and we will emphasize the division $x = (x_1, x_2)$. However the particular division between centering and conditioning variables will not be important in our calculations and we will thus suppress this notation. Formally, our results will apply to the case where both $x^{\bar{m}}$ and $x^{\bar{g}}$ contain all the components of x. However, they extend to any partition following modification of the bandwidth scaling to reflect the dimension of the real-valued components (x^m, x^s, x^g). We have kept the notation of $X_{i1}(\omega)$ depending on ω throughout this section facilitate the precise description of convergence results.

In Sections 3 and 4, the opposite case will be true. We will suppress the distinction between discrete and continuous random variables but the partition of the covariates into centering and conditioning components will have a substantial effect on our results. Here, for the sake of notational compactness we define a measure ν over $\mathbb{R}^{d_y} \otimes S_y$ and μ over $\mathbb{R}^{d_x} \otimes S_x$ given by the product of

counting and Lebesgue measure. Where needed, we will write for any function $F(x_1, x_2, y_1, y_2)$,

$$\sum_{x\in S_x, y\in S_y} \iint F(x_1, x_2, y_1, y_2)\,\mathrm{d}x_1\,\mathrm{d}y_1 = \iint F(x, y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x). \tag{1.17}$$

We will similarly define measures μ^g, μ^m, $\mu^{\bar{g}}$ and $\mu^{\bar{m}}$ over $\mathcal{X}^g$, $\mathcal{X}^m$, $\mathcal{X}^{\bar{g}}$ and $\mathcal{X}^{\bar{m}}$, respectively. In some places, we will refer to the centered $\varepsilon = y - m(x^{\bar{m}})$ where we will understand $m(x^{\bar{m}})$ to be zero on the discrete-valued components of y as well as those components of y_1 which are not being centered. In this context, we will subsume the indicator functions used above within the kernel and understand

$$K_x\left(\frac{x^{\bar{g}} - X_i^{\bar{g}}}{c_{nx_2^{\bar{g}}}}\right) = K_x\left(\frac{x_1^{\bar{g}} - X_{i1}^{\bar{g}}}{c_{n\bar{g}}}\right) I_{x_2^{\bar{g}}}\big(X_{i2}^{\bar{g}}\big).$$

Here we have changed bandwidth notation to $c_{n\bar{g}}$ in favor of $c_{nx_2^{\bar{g}}}$ and understand that c_{na} can depend on x_2^a, but we have maintained the distinction as to which of $\bar{m}$ or $\bar{g}$ a belongs to. We will also encounter a change of variables written as

$$\int F\big(x^{\bar{g}}, y\big)\frac{1}{c_{n\bar{g}}^{d_{x^{\bar{g}}}}} K_x\left(\frac{x^{\bar{g}} - X_i^{\bar{g}}}{c_{n\bar{g}}}\right)\mathrm{d}\mu^{\bar{g}}\big(x^{\bar{g}}\big) = \int F\big(X_i^{\bar{g}} + c_{n\bar{g}}u, y\big)K_x(u)\,\mathrm{d}u$$

in which we will interpret u as being a vector which is non-zero only on the continuous components of $x^{\bar{g}}$. Similar conventions will be employed for all other components of x and of y. In these sections, we will drop ω from our notation for the sake of compactness and because it will be less relevant to defining our results.

2. Consistency results for conditional densities over spaces of mixed types

In this section, we will provide a number of L_1 consistency results for kernel estimates of densities and conditional densities of multivariate random variables in which some coordinates take values in Euclidean space while others take values on a discrete set. Pointwise consistency of conditional density estimates of this form can be found in, for example, Li and Racine [10] and Hansen [7]. However, we are unaware of equivalent L_1 results which will be necessary for our development of conditional disparity-based inference. Throughout, we have assumed that both the conditioning variable x and the response y are multivariate with both types of coordinates. The specification to univariate models, or models with only discrete or only continuous variables in either x or y (and to unconditional densities) is readily seen to be covered by our results as well.

As a further generalization of the results in Li and Racine [10], we include the centered version of conditional density estimates defined by (1.7)–(1.10). We will demonstrate the consistency

of results for these estimates, from which consistency for uncentered conditional densities and results for homoscedastic conditional densities (1.3) are special cases.

Supplemental Appendix B (Hooker [8]) provides a set of intermediate results on the uniform and L_1 convergence of non-parametric regression and centered density estimates of missed types. Following these, we are able to establish the uniform (in x) L_1 (in y) convergence of multivariate densities:

Theorem 2.1. *Let* $\{(X_{n1}, X_{n2}, Y_{n1}, Y_{n2}), n \geq 1\}$ *be given as in Section* 1.1 *under assumptions* (D1)–(D4), (K1)–(K6) *and* (B1)–(B5) *then there exists a set* B *with* $P(B) = 1$ *such that for all* $\omega \in B$

$$\sup_{(x_1,x_2)\in\mathcal{X}} \sum_{y_2\in S_y} \int \left|\hat{g}_n(x_1, x_2, y_1, y_2, \hat{m}_n, \omega) - g(x_1, x_2, y_1, y_2, m)\right| \mathrm{d}y_1 \to 0. \tag{2.1}$$

The proof of this theorem is given in Supplemental Appendix C.2 (Hooker [8]). The results above can now be readily extended to equivalent L_1 results for conditional densities. We begin by considering centered densities and then proceed to uncenter them.

Theorem 2.2. *Let* $\{(X_{n1}, X_{n2}, Y_{n1}, Y_{n2}), n \geq 1\}$ *be given as in Section* 1.1 *under assumptions* (D1)–(D4), (K1)–(K6) *and* (B1)–(B5):

1. *There exists a set* B_I *with* $P(B_I) = 1$ *such that for all* $\omega \in B_I$,

$$\sum_{x_2\in S_x} \sum_{y_2\in S_y} \int h(x_1, x_2) \left|\hat{f}_n(\varepsilon, y_2|x_1, x_2, \omega) - f^c(\varepsilon, y_2|x_1, x_2)\right| \mathrm{d}\varepsilon \, \mathrm{d}x_1 \to 0. \tag{2.2}$$

2. *If further, assumptions* (D4) *and* (B5) *hold, there exists a set* B_S *with* $P(B_S) = 1$ *such that for all* $\omega \in B_S$:

$$\sup_{(x_1,x_2)\in\mathcal{X}} \sum_{y_2\in S_y} \int \left|\hat{f}_n(\varepsilon, y_2|x_1, x_2, \omega) - f^c(\varepsilon, y_2|x_1, x_2)\right| \mathrm{d}\varepsilon \to 0. \tag{2.3}$$

The proof of this theorem is given in Supplemental Appendix C.2 (Hooker [8]). From here, we can examine the behavior of $\breve{f}_n$.

Theorem 2.3. *Let* $\{(X_{n1}, X_{n2}, Y_{n1}, Y_{n2}), n \geq 1\}$ *be given as in Section* 1.1 *under assumptions* (E1)–(E2), (D1)–(D4), (K1)–(K6) *and* (B1)–(B5):

1. *There exists a set* B_I *with* $P(B_I) = 1$ *such that for all* $\omega \in B_I$,

$$\sum_{x_2\in S_x} \sum_{y_2\in S_y} \int h(x_1, x_2) \left|\breve{f}_n(y_1, y_2|x_1, x_2, \omega) - f(y_1, y_2|x_1, x_2)\right| \mathrm{d}y_1 \, \mathrm{d}x_1 \to 0. \tag{2.4}$$

2. *If further, assumptions* (D4) *and* (B5) *hold, there exists a set* B_S *with* $P(B_S) = 1$ *such that for all* $\omega \in B_S$:

$$\sup_{(x_1,x_2)\in\mathcal{X}} \sum_{y_2\in S_y} \int \left| \breve{f}_n(y_1, y_2|x_1, x_2, \omega) - f(y_1, y_2|x_1, x_2) \right| \mathrm{d}y_1 \to 0. \tag{2.5}$$

Proof. We begin by writing

$$\begin{aligned}
&\sum_{y_2\in S_y} \int \left| \breve{f}_n(y_1, y_2|x_1, x_2, \omega) - f(y_1, y_2|x_1, x_2) \right| \mathrm{d}y_1 \\
&\quad \le \sum_{y_2\in S_y} \int \left| \breve{f}_n(y_1, y_2|x_1, x_2, \omega) - f^c\big(y_1 - \hat{m}_n(x_1, x_2), y_2|x_1, x_2\big) \right| \mathrm{d}y_1 \\
&\qquad + \sum_{y_2\in S_y} \int \left| f^c\big(y_1 - \hat{m}_n(x_1, x_2), y_2|x_1, x_2\big) - f^c\big(y_1 - m(x_1, x_2), y_2|x_1, x_2\big) \right| \mathrm{d}y_1 \\
&\quad \le \sum_{y_2\in S_y} \int \left| \breve{f}_n(y_1, y_2|x_1, x_2, \omega) - f^c\big(y_1 - \hat{m}_n(x_1, x_2), y_2|x_1, x_2\big) \right| \mathrm{d}y_1 \\
&\qquad + \sup_{(x_1,x_2)\in\mathcal{X}} \left| \hat{m}_n(x_1, x_2) - m(x_1, x_2) \right| \sum_{y_2\in S_y} \int \left| \nabla_{y_1} f^c(y_1, y_2|x_1, x_2) \right| \mathrm{d}y_1.
\end{aligned}$$

The first term of the last line converges almost surely from Theorem 2.2 applied either marginalized over (x_1, x_2) to obtain (2.4) or after taking a supremum to obtain (2.5). The second term follows from Theorem B.2 in the Supplemental Appendix (Hooker [8]) and assumption (E1). □

These results can now be applied to the more regular conditional density estimates (1.11)–(1.13) and homoscedastic conditional density estimates (1.14–1.16). For the sake of completeness, we state these directly as corollaries without proof.

Corollary 2.1. *Let* $\{(X_{n1}, X_{n2}, Y_{n1}, Y_{n2}), n \ge 1\}$ *be given as in Section* 1.1 *under assumptions* (D1)–(D3), (K1)–(K6) *and* (B1)–(B2) *then*:

1. *For almost all* $x = (x_1, x_2) \in \mathbb{R}^{d_x} \otimes S_x$ *there exists a set* B_x *with* $P(B_x) = 1$ *such that for all* $\omega \in B_x$

$$\sum_{y_2\in S_y} \int \left| \hat{f}_n^*(y_1, y_2|x_1, x_2, \omega) - f(y_1, y_2|x_1, x_2) \right| \mathrm{d}y_1 \to 0. \tag{2.6}$$

2. *There exists a set* B_I *with* $P(B_I) = 1$ *such that for all* $\omega \in B_I$,

$$\sum_{x_2\in S_x} \sum_{y_2\in S_y} \int h(x_1, x_2) \left| \hat{f}_n^*(y_1, y_2|x_1, x_2, \omega) - f(y_1, y_2|x_1, x_2) \right| \mathrm{d}y_1 \, \mathrm{d}x_1 \to 0. \tag{2.7}$$

3. *If further, assumptions* (D4) *and* (B5) *hold, there exists a set* B_S *with* $P(B_S) = 1$ *such that for all* $\omega \in B_S$

$$\sup_{(x_1,x_2)\in\mathcal{X}} \sum_{y_2\in S_y} \int \left|\hat{g}_n^*(x_1, x_2, y_1, y_2, \omega) - g(x_1, x_2, y_1, y_2)\right| \mathrm{d}y_1 \to 0 \tag{2.8}$$

and

$$\sup_{(x_1,x_2)\in\mathcal{X}} \sum_{y_2\in S_y} \int \left|\hat{f}_n^*(y_1, y_2|x_1, x_2, \omega) - f(y_1, y_2|x_1, x_2)\right| \mathrm{d}y_1 \to 0. \tag{2.9}$$

Corollary 2.2. *Let* $\{(X_{n1}, X_{n2}, Y_{n1}), n \geq 1\}$ *be given as in Section* 1.1 *with the restriction* (1.3)*, under assumptions* (D1)–(D4), (E1)–(E2), (K1)–(K6), (B1)–(B2) *and* (B5) *there exists a set* B *with* $P(B) = 1$ *such that for all* $\omega \in B$

$$\int \left|\hat{f}_n^c(e, \omega) - f^c(e)\right| \mathrm{d}e \to 0 \tag{2.10}$$

and

$$\sup_{(x_1,x_2)\in\mathcal{X}} \int \left|\tilde{f}_n(y_1|x_1, x_2, \omega) - f^c\big(y_1 - m(x_1, x_2)\big)\right| \mathrm{d}y_1 \to 0. \tag{2.11}$$

The above theorems rely on the compactness of $\mathcal{X}$ (assumption (D4)), this is necessary due to the estimate $\hat{m}_n(x_1^{\bar{m}}, x_2^{\bar{m}})$, and is necessary for uniform convergence in $\mathcal{X}$. However, a weaker version can be given for non-centered densities which does not require a compact support:

Theorem 2.4. *Let* $\{(X_{n1}, Y_{n1}), n \geq 1\}$ *be given as in Section* 1.1 *under assumptions* (D1)–(D3), (K1)–(K6) *and* (B1)–(B2) *then for almost all* $x = (x_1, x_2)$ *there exists a set* B_x *with* $P(B_x) = 1$ *such that for all* $\omega \in B_x$

$$\sum_{y_2\in S_y} \int \left|\hat{g}_n(x_1, x_2, y_1, y_2, \omega) - g(x_1, x_2, y_1, y_2)\right| \mathrm{d}y_1 \to 0. \tag{2.12}$$

Proof. For (2.12), we observe that

$$\sum_{x_2\in S_x} \sum_{y_2\in S_y} \iint \left|\hat{g}_n(x_1, x_2, y_1, y_2, \omega) - g(x_1, x_2, y_1, y_2)\right| \mathrm{d}y_1 \, \mathrm{d}x_1 = \sum_{x_2\in S_x} \int T_n(x_1, x_2) \, \mathrm{d}x_1 \to 0$$

almost surely with $T_n(x_1, x_2) > 0$, see [5], Chapter 3, Theorem 1. Thus $T_n(x_1, x_2) \to 0$ for almost all (x_1, x_2). □

In particular, we can rely on this theorem to remove assumption (D4) from the minimum disparity methods studied below in special cases that employ $\hat{g}_n$ as a density estimate. Relevant further results are given in Supplemental Appendix E (Hooker [8]).

3. Consistency of minimum disparity estimators for conditional models

In this section, we define minimum disparity estimators for the conditionally specified models based on distributions and data defined in Section 1.1. For the purposes of notational simplicity, we will ignore the distinction between continuous and discrete random variables X_1, X_2 and Y_1, Y_2, but we will make use of the division $x = (x^m, x^s, x^g)$ into those covariates x^m used to center the estimated density, those used to condition, x^g, and those in both, x^s. We assume that a parametric model has been proposed for these data of the form

$$f(y|x) = \phi(y|x, \theta),$$

where we assume that the X_i are independently drawn from a distribution $h(x)$ which is not parametrically specified. For this model, the maximum likelihood estimator for θ given observations (Y_i, X_i), $i = 1, \ldots, n$ is

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max \sum_{i=1}^{n} \log \phi(Y_i|X_i, \theta)$$

with attendant asymptotic variance

$$I(\theta_0) = n \iint \nabla_\theta^2 \big[\log \phi(y|x, \theta_0)\big] \phi(y|x, \theta_0) h(x)\, \mathrm{d}\nu(y)\, \mathrm{d}\mu(x)$$

when the specified parametric model is correct at $\theta = \theta_0$.

In the context of disparity estimation, for every value x we define the conditional disparity between f and ϕ as

$$D(f, \phi|x, \theta) = \int C\left(\frac{f(y|x)}{\phi(y|x, \theta)} - 1\right) \phi(y|x, \theta)\, \mathrm{d}\nu(y)$$

in which C is a strictly convex function from $\mathbb{R}$ to $[-1\ \infty)$ with a unique minimum at 0. Classical choices of C include $\mathrm{e}^{-x} - 1$, resulting in the negative exponential disparity (NED) and $[\sqrt{x+1} - 1]^2 - 1$, which corresponds to Hellinger distance (HD).

These disparities are combined over observed X_i by averaging the disparity between f and ϕ evaluated at each X_i

$$D_n(f, \theta) = \frac{1}{n} \sum_{i=1}^{n} D(f, \phi|X_i, \theta)$$

(note that the Y_i only appear here when f is replaced by an estimate $\breve{f}_n$) or by integrating over the estimated density of $x^{\bar{g}}$:

$$\tilde{D}_n(f, \theta) = \frac{1}{n} \sum_{i=1}^{n} \int D\big(f, \phi|X_i^m, x^{\bar{g}}, \theta\big) \hat{h}_n\big(x^{\bar{g}}\big)\, \mathrm{d}\mu^{\bar{g}}\big(x^{\bar{g}}\big)$$

with limiting cases

$$D_\infty(f,\theta) = \int D(f,\phi|x,\theta)h(x_1,x_2)\,\mathrm{d}\mu(x)$$

and

$$\tilde{D}_\infty(f,\theta) = \iint D\big(f,\phi|x^m,x^{\bar{g}},\theta\big)h^m\big(x^m\big)h^{\bar{g}}\big(x^{\bar{g}}\big)\,\mathrm{d}\mu^m\big(x^m\big)\,\mathrm{d}\mu^{\bar{g}}\big(x^{\bar{g}}\big).$$

We now define the corresponding conditional minimum disparity estimators:

$$\hat{\theta}_n^D = \arg\min_{\theta\in\Theta} D_n(\breve{f}_n,\theta), \qquad \tilde{\theta}_n^D = \arg\min_{\theta\in\Theta} \tilde{D}_n(\breve{f}_n,\theta).$$

Here we note that when the model is correct – that is $f(y|x) = \phi(y|x,\theta_0)$ – we have that θ_0 minimizes both $D_\infty(f,\theta)$ and $\tilde{D}_\infty(f,\theta)$.

Under this definition, we first establish the existence and consistency of $\hat{\theta}_n^D$. To do so, we note that disparity results all rely on the boundedness of $D(f,\phi|X_i,\theta)$ over θ and f and a condition of the form that for any conditional densities f_1 and f_2,

$$\sup_{\theta\in\Theta}\big|D(f_1,\phi|x,\theta) - D_n(f_2,\phi|x,\theta)\big| \le K\int\big|f_1(y|x) - f_2(y|x)\big|\,\mathrm{d}\nu(y) \tag{3.1}$$

for some $K > 0$. In the case of Hellinger distance (Beran [3]), $D(g,\theta) < 2$ and (3.1) follows from Minkowski's inequality. For the alternate class of divergences studied in Park and Basu [13], boundedness of D is established from assuming that $\sup_{t\in[-1,\infty)}|C'(t)| \le C^* < \infty$ which also provides

$$\begin{aligned}
&\left|\int\left[C\left(\frac{f_1(y|x)}{\phi(y|x,\theta)} - 1\right) - C\left(\frac{f_2(y|x)}{\phi(y|x,\theta)} - 1\right)\right]\phi(y|x,\theta)\,\mathrm{d}\nu(y)\right| \\
&\quad\le C^*\int\left|\frac{f_1(y|x)}{\phi(y|x,\theta)} - \frac{f_2(y|x)}{\phi(y|x,\theta)}\right|\phi(y|x,\theta)\,\mathrm{d}\nu(y) \\
&\quad= C^*\int\big|f_1(y|x) - f_2(y|x)\big|\,\mathrm{d}\nu(y).
\end{aligned}$$

For simplicity, we therefore use (3.1) as a condition below.

In general, we will require the following assumptions:

- (P1) There exists N such that $\max_{i\in1,\ldots,n}|\sum_{i=1}^n \phi(y|X_i,\theta_1) - \phi(y_i|X_i,\theta_2)| > 0$ with probability 1 on a nonzero set of dominating measure in y whenever $n > N$ and $\theta_1 \ne \theta_2$.
- (P2) $\phi(y|x,\theta)$ is continuous in θ for almost every (x,y).
- (P3) $D_n(f,\phi|x,\theta)$ is uniformly bounded over f in the space of conditional densities, $(x_1,x_2)\in\mathcal{X}$ and $\theta\in\Theta$ and (3.1) holds.
- (P4) For every f, there exists a compact set $S_f \subset \Theta$ and N such that for $n \ge N$,

$$\inf_{\theta\in S_f^c} D_n(f,\theta) > \inf_{\theta\in S_f} D_n(f,\theta).$$

These assumptions combine those of Park and Basu [13] for a general class of disparities with the identifiability condition (P4) which appears in [15], equation (3.3), which relaxes the assumption of compactness of Θ; see also Cheng and Vidyashankar [4]. Together, these provide the following results.

Theorem 3.1. *Under assumptions* (P1)–(P4), *define*

$$T_n(f) = \arg\min_{\theta\in\Theta} D_n(f,\theta), \tag{3.2}$$

for $n = 1,\dots,\infty$ *inclusive*, *then*:

(i) *For any* $f \in \mathcal{F}$ *there exists* $\theta \in \Theta$ *such that* $T_n(f) = \theta$.
(ii) *For* $n \geq N$, *for any* θ, $\theta = T_n(\phi(\cdot|\cdot,\theta))$ *is unique.*
(iii) *If* $T_n(f)$ *is unique and* $f_m \to f$ *in* L_1 *for each* x, *then* $T_n(f_m) \to T_n(f)$.

The same results hold for

$$\tilde{T}_n(f) = \arg\min_{\theta\in\Theta} \tilde{D}_n(f,\theta).$$

Proof. (i) Existence. We first observe that it is sufficient to restrict the infimum in (3.2) to S_f. Let $\{\theta_m\colon \theta_m \in S_f\}$ be a sequence such that $\theta_m \to \theta$ as $m \to \infty$. Since

$$C\left(\frac{f(y|x)}{\phi(y|x,\theta_m)} - 1\right)\phi(y|x,\theta_m) \to C\left(\frac{f(y|x)}{\phi(y|x,\theta)} - 1\right)\phi(y|x,\theta)$$

by assumption (P2), using the bound on $D(f,\phi,\theta)$ from assumption (P3) we have $D_n(f,\theta_m) \to D_n(f,\theta)$ by the dominated convergence theorem. Hence $D_n(f,t)$ is continuous in t and achieves its minimum for $t \in S_f$ since S_f is compact.

(ii) Uniqueness. This is a consequence of assumption (P1) and the unique minimum of C at 0.

(iii) Continuity in f. For any sequence $f_m(\cdot|x) \to f(\cdot|x)$ in L_1 for every x as $m \to \infty$, we have

$$\sup_{\theta\in\Theta}\left|D_n(f_m,\theta) - D_n(f,\theta)\right| \to 0 \tag{3.3}$$

from assumption (P3).

Now consider $\theta_m = T_n(f_m)$. We first observe that there exists M such that for $m \geq M$, $\theta_m \in S_f$ otherwise from (3.3) and assumption (P4)

$$D_n(f_m,\theta_m) > \inf_{\theta\in S_f} D_n(f_m,\theta)$$

contradicting the definition of θ_m.

Now suppose that θ_m does not converge to θ_0. By the compactness of S_f we can find a subsequence $\theta_{m'} \to \theta^* \neq \theta_0$ implying $D_n(f,\theta_{m'}) \to D_n(f,\theta^*)$ from assumption (P2). Combining this with (3.3) implies $D_n(f,\theta^*) = D_n(f,\theta_0)$, contradicting the assumption of the uniqueness of $T_n(f)$. □

Theorem 3.2. *Let* $\{(X_{n1}, X_{n2}, Y_{n1}, Y_{n2}), n \geq 1\}$ *be given as in Section* 1.1 *and define*

$$\theta_n^0 = \arg\min_{\theta \in \Theta} D_n(f, \theta)$$

for every n *including* ∞. *Further, assume that* θ_∞^0 *is unique in the sense that for every* ε *there exists* δ *such that*

$$\left\|\theta - \theta_\infty^0\right\| > \varepsilon \quad \Rightarrow \quad D_\infty(f, \theta) > D_\infty\left(f, \theta_\infty^0\right) + \delta$$

then under assumptions (D1)–(D4), (K1)–(K6), (B1)–(B2) *and* (P1)–(P4):

$$\hat{\theta}_n = T_n(\breve{f}_n) \to \theta_\infty^0 \qquad \text{as } n \to \infty \text{ almost surely.}$$

Similarly,

$$\tilde{T}_n(\breve{f}_n) = \arg\min_{\theta \in \Theta} \tilde{D}_n(\breve{f}_n, \theta) \to \tilde{\theta}_\infty^0 \qquad \text{as } n \to \infty \text{ almost surely.}$$

Proof. First, we observe that for every f, it is sufficient to restrict attention to S_f and that

$$\sup_{\theta \in S_f} \left|D_n(f, \theta) - D_\infty(f, \theta)\right| \to 0 \qquad \text{almost surely} \tag{3.4}$$

from the strong law of large numbers, the compactness of S_f and the assumed continuity of C and of ϕ with respect to θ.

Further,

$$\begin{aligned} \sup_{m \in \mathbb{N}, \theta \in \Theta} \left|D_m(\breve{f}_n, \theta) - D_m(f, \theta)\right| &\leq C^* \sup_{x \in \mathcal{X}} \int \left|\breve{f}_n(y|x) - f(y|x)\right| \mathrm{d}\nu(y) \\ &\to 0 \qquad \text{almost surely,} \end{aligned} \tag{3.5}$$

where the convergence is obtained from Theorem 2.3.

Suppose that $\hat{\theta}_n$ does not converge to θ_∞^0, then we can find $\varepsilon > 0$ and a subsequence $\hat{\theta}_{n'}$ such that $\|\hat{\theta}_{n'} - \theta_\infty^0\| > \varepsilon$ for all n'. However, on this subsequence

$$\begin{aligned} D_{n'}(\breve{f}_{n'}, \hat{\theta}_{n'}) &= D_{n'}(\breve{f}_{n'}, \theta_0) + \left(D_{n'}(f, \theta_0) - D_{n'}(\breve{f}_{n'}, \theta_0)\right) + \left(D_\infty(f, \theta_0) - D_{n'}(f, \theta_0)\right) \\ &\quad + \left(D_\infty(f, \hat{\theta}_{n'}) - D_\infty(f, \theta_0)\right) \\ &\quad + \left(D_{n'}(f, \hat{\theta}_{n'}) - D_\infty(f, \hat{\theta}_{n'})\right) + \left(D_{n'}(\breve{f}_{n'}, \hat{\theta}_{n'}) - D_{n'}(f, \hat{\theta}_{n'})\right) \\ &\leq D_{n'}(\breve{f}_{n'}, \theta_0) + \delta \\ &\quad - 2 \sup_{\theta \in \Theta} \left|D_{n'}(f, \theta) - D(f, \theta)\right| - 2 \sup_{\theta \in \Theta} \left|D_{n'}(\breve{f}_{n'}, \theta) - D_{n'}(f, \theta)\right| \end{aligned}$$

but from (3.4) and (3.5) we can find N so that for $n' \geq N$

$$\sup_{\theta \in \Theta} \left|D_{n'}(f, \theta) - D(f, \theta)\right| \leq \frac{\delta}{6}$$

and

$$\sup_{\theta\in\Theta}\left|D_{n'}(\breve{f}_{n'},\theta)-D_{n'}(f,\theta)\right|\leq\frac{\delta}{6}$$

contradicting the optimality of $\hat{\theta}_{n'}$. The proof for $\tilde{T}_n(\breve{f}_n)$ follows analogously. □

The compactness assumption (D4) used above can be removed for the special case of an uncentered density employed with our second estimator: $\tilde{T}(\hat{f}_n^*)$. This is stated in Theorem E.1 in Supplemental Appendix E (Hooker [8]).

4. Asymptotic normality and efficiency of minimum disparity estimators for conditional models

In this section, we demonstrate the asymptotic normality and efficiency of minimum conditional disparity estimators. In order to simplify some of our expressions, we introduce the following notation, that for a column vector A we define the matrix

$$A^{TT}=AA^T.$$

This will be particularly useful in defining information matrices.

We will also frequently use the notation $y=(y_1,y_2)$ and $x=(x_1,x_2)$, ignoring the distinction between real and discrete valued variables. It will be particularly relevant to distinguish $x^{\bar{g}}$ and $x^{\bar{m}}$ along with their subsets x^g and x^m that are solely in $x^{\bar{g}}$ or $x^{\bar{m}}$, respectively, along with the shared dimensions x^s. Because our notation would otherwise become unwieldy, we will subsume indicator functions within kernels, and, for example, understand

$$K_x\left(\frac{x^{\bar{g}}-X_i^{\bar{g}}}{c_{n\bar{g}}}\right)=K_x\left(\frac{x_1^{\bar{g}}-X_{i1}^{\bar{g}}}{c_{nx_2^{\bar{g}}}}\right)I_{x_2^{\bar{g}}}\left(X_{i2}^{\bar{g}}\right),$$

where we have also suppressed the x_2 indicator in the bandwidth $c_{n\bar{g}}$. Within this context, we will also occasionally abuse notation when changing variables and write $x^{\bar{g}}+c_{n\bar{g}}v$ in which we understand that the additive term only corresponds to the continuous-valued entries in $x^{\bar{g}}$. We will also express integration with respect to the distribution $\mu(x)$ and $\nu(y)$ and denote μ^g, μ^m, μ^s, $\mu^{\bar{g}}$ and $\mu^{\bar{m}}$ the measures marginalized to the corresponding dimensions of $\mathcal{X}$.

The proof techniques employed here are an extension of those developed in i.i.d. settings in Beran [3]; Tamura and Boos [16]; Lindsay [11]; Park and Basu [13]. In particular we will require the following assumptions:

(N1) Define

$$\Psi_\theta(x,y)=\frac{\nabla_\theta\phi(y|x,\theta)}{\phi(y|x,\theta)}$$

then

$$\sup_{x\in\mathcal{X}}\int\Psi_\theta(x,y)\Psi_\theta(x,y)^T f(y|x)\,\mathrm{d}\nu(y)<\infty$$

elementwise. Further, there exists $a_y > 0$ such that

$$\sup_{x\in\mathcal{X}} \sup_{\|t\|\le a_y} \sup_{\|s\|\le a_y} \int \Psi_\theta(x_1+s, x_2, y_1+t, y_2)^2 f(y|x)\,\mathrm{d}\nu(y) < \infty$$

and

$$\sup_{x\in\mathcal{X}} \sup_{\|t\|\le a_y} \sup_{\|s\|\le a_y} \int \big(\nabla_{y_1}\Psi_\theta(x_1+t, x_2, y_1+s, y_2)\big)^2 f(y|x)\,\mathrm{d}\nu(y) < \infty,$$

and

$$\sup_{x\in\mathcal{X}} \sup_{\|t\|\le a_y} \sup_{\|s\|\le a_y} \int \big(\nabla_{x}\Psi_\theta(x_1+t, x_2, y_1+s, y_2)\big)^2 f(y|x)\,\mathrm{d}\nu(y) < \infty.$$

(N2) There exists sequences b_n and α_n diverging to infinity along with a constant $c > 0$ such that:

(i) $nK_x(b_n/c_{nx}) \to 0$, $nK_y(b_n/c_{ny}) \to 0$ and

$$n \sup_{x\in\mathcal{X}} \sup_{\|u\|>b_n} \iint_{\|v\|>b_n} \Psi_\theta^2(x + c_{nx}u, y + c_{ny}v)K_y^2(u)K_x^2(v)g(x,y)\,\mathrm{d}v\,\mathrm{d}\nu(y) \to 0$$

elementwise.

(ii) $\sup_{x\in\mathcal{X}} nP(\|Y_1 - c_{ny}b_n\| > \alpha_n - c) \to 0$.

(iii)

$$\sup_{x\in\mathcal{X}} \frac{1}{\sqrt{n}c_{nx}^{d_x}c_{ny}^{d_y}} \int_{\|y_1\|\le\alpha_n+c} \big|\Psi_\theta(x,y)\big|\,\mathrm{d}\nu(y) \to 0.$$

(iv)

$$\sup_{x\in\mathcal{X}} \sup_{\|t\|\le b_n} \sup_{\|s\|\le b_n} \sup_{\|y_1\|<\alpha_n} \frac{g(x + c_{nx}s, y + c_{ny}t)}{g(x,y)} = \mathrm{O}(1).$$

(N3) $\sup_{y,x} \sqrt{\phi(y|x,\theta)}\nabla_\theta\Psi_\theta(y,x) = S < \infty$.

(N4) C is either given by Hellinger distance $C(x) = [\sqrt{x+1} - 1]^2 - 1$ or

$$A_1(r) = -C''(r-1)r, \qquad A_2(r) = C(r-1) - C'(r-1)r,$$
$$A_3(r) = C''(r-1)r^2$$

are all bounded in absolute value as is $r^2C^{(3)}(r)$.

Assumption (N1) ensures that the likelihood score function is well controlled including for small location changes of x_1 and y_1. Assumption (N2) requires Ψ_θ and y_1 to have well-behaved tails relative to K_y. In particular, assumption (N2)(i) allows us to truncate the kernels at b_n

which will prove mathematically convenient throughout the remainder of the section. Assumption (N3) concerns the regularity of the parametric model and in particular ensures that the second derivative of Hellinger distance with respect to parameters is well behaved. Assumption (N4) is a restatement of conditions on the residual adjustment function in Lindsay [11] and Park and Basu [13]; a wide class of disparities satisfy these conditions including NED, we refer the reader to Lindsay [11] for a more complete discussion. As was the case for assumption (P3), we treat Hellinger distance separately in assumption (N4) as it does not conform to the general assumptions on C, but the relevant bounds can be demonstrated by other means in the proof of Theorem 4.1 below.

The demonstration of a central limit theorem involves bounding the score function for a general disparity in terms of that for Hellinger distance and then taking Taylor expansion of this score. For this we need two lemmas. The first is that the weighted Hellinger distance between $\hat{f}_n$ and its expectation is smaller than $\sqrt{n}$. This is used in Theorem 4.1 to remove terms involving $\sqrt{\hat{f}_n}$.

Lemma 4.1. *Let* $\{(X_n, Y_n), n \geq 1\}$ *be given as in Section* 1.1, *under assumptions* (D1)–(D4), (K1)–(K6), (B1)–(B4), *and* (N1)–(N2)(iv) *for any function* $J(y, x)$ *satisfying the conditions on* Ψ *in assumptions* (N1)–(N2)(iv)

$$\sqrt{n} \sup_{x \in \mathcal{X}} \iint J\bigl(e + \hat{m}_n\bigl(x^{\bar{m}}\bigr), x\bigr)\left(\sqrt{\hat{f}_n(e|x)} - \sqrt{\frac{E\hat{g}_n(x, e, \hat{m}_n)|\hat{m}_n}{E\hat{h}_n(x)}}\right)^2 \mathrm{d}\nu(e) \to 0 \tag{4.1}$$

in probability and

$$\begin{aligned}\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \iint &\left(\sqrt{\hat{f}_n\bigl(e|X_i^m, x^g\bigr)} - \sqrt{\frac{E\hat{g}_n(X_i^m, x^g, e, \hat{m}_n)|\hat{m}_n}{E\hat{h}_n(x^g)}}\right)^2 \\ &\times J\bigl(e + \hat{m}_n\bigl(X_i^m\bigr), X_i^m, x^g\bigr)\hat{h}_n\bigl(x^g\bigr)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu^g\bigl(x^g\bigr) \to 0\end{aligned} \tag{4.2}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \iint J\bigl(e + \hat{m}_n\bigl(X_i^m\bigr), X_i\bigr)\left(\sqrt{\hat{f}_n(e|X_i)} - \sqrt{\frac{E\hat{g}_n(X_i, e, \hat{m}_n)|\hat{m}_n}{\hat{h}_n(X_i)}}\right)^2 \mathrm{d}\nu(e) \to 0. \tag{4.3}$$

The proof of this lemma is given in Supplemental Appendix D.2 (Hooker [8]).

A second lemma states that integrating a function $J(y, x)$ with respect to $\hat{f}_n$ yields a central limit theorem. In the below, we have used a subscript x_* to help differentiate which components are being integrated with respect to which measure.

Lemma 4.2. *Let* $\{(X_n, Y_n), n \geq 1\}$ *be given as in Section* 1.1, *under assumptions* (D1)–(D4), (E1)–(E2), (K1)–(K6), (B1)–(B2) *and* (P1)–(P4) *for any for any function* $J(y, x)$ *satisfying the conditions on* Ψ *in assumptions* (N1)–(N2)(iv) *and*

$$V_J = \iint U_J^{TT}(\varepsilon, x) g^c(x, \varepsilon)\,\mathrm{d}\nu(\varepsilon)\,\mathrm{d}\mu(x) < \infty,$$

where

$$\begin{aligned}U_J(\varepsilon,x) &= \int J\bigl(\varepsilon + m\bigl(x_*^m, x^s\bigr), x_*^m, x^{\bar g}\bigr)h^m\bigl(x_*^m\bigr)\,\mathrm{d}\mu^m\bigl(x_*^m\bigr)\\ &\quad + \iint J\bigl(e_* + m\bigl(x^m, x_*^s\bigr), x^m, x_*^{\bar g}\bigr)g^c(x_*, e_*)\,\mathrm{d}\nu(e_*)\,\mathrm{d}\mu(x_*)\\ &\quad + \varepsilon\iint \nabla_y J\bigl(e_* + m\bigl(x^{\bar m}\bigr), x^{\bar m}, x_*^g\bigr)\\ &\qquad\qquad \times \frac{g^c(x_*^m, x^s, x_*^g, e_*)h^m(x^m)}{h^{\bar m}(x^{\bar m})}\,\mathrm{d}\nu(e_*)\,\mathrm{d}\mu^g\bigl(x_*^g\bigr)\,\mathrm{d}\mu^m\bigl(x_*^m\bigr)\\ &\quad - \varepsilon\iint \nabla_y J\bigl(e_* + m\bigl(x_*^m, x^s\bigr), x_*^m, x^s, x_*^g\bigr)\\ &\qquad\qquad \times \frac{g^c(x^{\bar m}, x_*^g, e_*)h^m(x_*^m)}{h^{\bar m}(x^{\bar m})}\,\mathrm{d}\nu(e_*)\,\mathrm{d}\mu^g\bigl(x_*^g\bigr)\,\mathrm{d}\mu^m\bigl(x_*^m\bigr)\end{aligned}$$

elementwise, then

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^n \iint J\bigl(y, X_i^m, x^{\bar g}\bigr)\hat g_n(y,x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x) - B_n\right] \to N(0, V_J) \tag{4.4}$$

in distribution where

$$B_n = 2\iint J\bigl(e + m\bigl(x_*^m, x^{\bar g}\bigr), x_*^s, x^g\bigr)E\hat g_n\bigl(x^g, e, m\bigr)h^m\bigl(x_*^m\bigr)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x).$$

Similarly, if

$$\tilde V_J = \iint \tilde U_J^{TT}(\varepsilon,x)g^c(x,\varepsilon)\,\mathrm{d}\nu(\varepsilon)\,\mathrm{d}\mu(x) < \infty,$$

where

$$\begin{aligned}\tilde U_J(\varepsilon,x) &= \iint J\bigl(e_* + m\bigl(x_*^m, x^s\bigr), x\bigr)\frac{g(x_*^m, x^{\bar g}, e_*)}{h^{\bar g}(x^{\bar g})}\,\mathrm{d}\mu^m\bigl(x_*^m\bigr)\,\mathrm{d}\nu(e)\\ &\quad + \iint J\bigl(\varepsilon + m\bigl(x_*^m, x^s\bigr), x_*^m, x^{\bar g}\bigr)\frac{h(x_*^m, x^{\bar g})}{h^{\bar g}(x^{\bar g})}\,\mathrm{d}\mu^m\bigl(x_*^m\bigr)\\ &\quad + \varepsilon\iint \nabla_y J\bigl(e_* + m\bigl(x^{\bar m}\bigr), x^{\bar m}, x_*^g\bigr)\frac{g(x_*^m, x^s, x_*^g, e_*)h(x^m, x_*^{\bar g})}{h^{\bar g}(x^s, x_*^g)h^{\bar m}(x^{\bar m})}\,\mathrm{d}\nu(e_*)\,\mathrm{d}\mu(x_*)\\ &\quad - \varepsilon\iiint \nabla_y J\bigl(e_* + m\bigl(x_*^m, x^s\bigr), x_*^m, x^s, x_*^g\bigr)\\ &\qquad\qquad \times \frac{g^c(x^{\bar m}, x_*^g, e_*)h(x_*^m, x^s, x_*^g)}{h^{\bar g}(x^s, x_*^g)h^m(x^{\bar m})}\,\mathrm{d}\nu(e_*)\,\mathrm{d}\mu^m\bigl(x_*^m\bigr)\,\mathrm{d}\mu^g\bigl(x_*^g\bigr)\\ &\quad + \iint J\bigl(e_* + m\bigl(x_*^m, x^s\bigr), x_*^m, x^g\bigr)\frac{g^c(x_*^m, x^{\bar g}, e_*)}{h^{\bar g}(x^{\bar g})}\,\mathrm{d}\nu(e_*)\,\mathrm{d}\mu^m\bigl(x_*^m\bigr)\end{aligned}$$

elementwise, then

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\int J(y, X_i)\,\breve{f}_n(y|X_i)\,\mathrm{d}\nu(y) - \tilde{B}_n\right] \to N(0, \tilde{V}_J) \tag{4.5}$$

in distribution with

$$\tilde{B}_n = 2\iint J\big(e + m\big(x^{\bar{m}}\big), x\big)\frac{E\,\hat{g}_n(x^g, e, m)}{E\hat{h}_n(x^{\bar{g}})}h(x)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x).$$

The proof of this lemma is reserved to Supplemental Appendix D.1 (Hooker [8]).

The bias and variance terms found in this lemma are rather complex due to their generality and it will be helpful here to note the resulting expressions for four simplifying cases and the consequence of these. Further, in Theorem 4.1 we will investigate

$$\Psi_\theta(y|x) = \frac{\nabla_\theta \phi_\theta(y|x)}{\phi_\theta(y|x)}, \tag{4.6}$$

where if $\phi_\theta(y|x)$ has the form $\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)$ we have that

$$\Psi_\theta(y|x) = -\frac{\partial_\theta m(x^{\bar{m}}, \theta)\,\partial_y \phi(y - m(x^{\bar{m}}; \theta_1)|x^{\bar{g}}; \theta)}{\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)} + \frac{\partial_\theta \phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)}{\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)},$$

where ∂ is used to represent a partial gradient and ∇ the total gradient. We also have that

$$\begin{aligned}\nabla_y J(y|x) = &-\frac{\partial_\theta m(x^{\bar{m}}, \theta) D_y^2 \phi(y - m(x^{\bar{m}}; \theta_1)|x^{\bar{g}}; \theta)}{\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)} \\ &+ \frac{D_{\theta y}^2 \phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)}{\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)} \\ &+ \frac{\partial_\theta m(x^{\bar{m}}, \theta)\,\partial_y \phi(y - m(x^{\bar{m}}; \theta_1)|x^{\bar{g}}; \theta)}{\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)} \frac{\partial_y \phi(y - m(x^{\bar{m}}; \theta_1)|x^{\bar{g}}; \theta)^T}{\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)} \\ &- \frac{\partial_\theta \phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)}{\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)} \frac{\partial_y \phi(y - m(x^{\bar{m}}; \theta_1)|x^{\bar{g}}; \theta)^T}{\phi(y - m(x^{\bar{m}}; \theta)|x^{\bar{g}}; \theta)},\end{aligned}$$

where we take $\partial_y^2 \phi$ to be the Hessian with respect to y and $\partial_{\theta y}^2 \phi$ to be the corresponding matrix of cross derivatives. In each of these cases, we demonstrate that substituting in $f(y|x) = \phi_\theta(y|x)$ results in variance terms given by the Fisher information

$$I(\theta) = \iint \frac{\nabla_\theta \phi_\theta(e|x) \nabla_\theta \phi_\theta(e|x)^T}{\phi_\theta(e|x)} h(x)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x)$$

or the equivalent based on centering by $m(x, \theta)$ above.

Non-centered: $x^{\bar{m}} = \phi$. This corresponds to the simplest case of a conditional density estimate. Here we have

$$U_J(y,x) = J(y,x),$$
$$B_n = 2\iint J(y,x)E\hat{g}_n(x,y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x).$$

We remark here that the bias B_n corresponds to the bias found in Tamura and Boos [16] for multivariate observations. As observed there, the bias in the estimate $\hat{g}_n$ is $O(c_{n\bar{g}}^2 + c_{ny}^2)$ and that of $\hat{h}_n$ is $O(c_{n\bar{g}}^2)$, regardless of the dimension of $x_1^{\bar{g}}$ and y_1. However the variance is of order $n^{-1}c_{n\bar{g}}^{d_{x\bar{g}}}c_{ny}^{d_y}$ (corresponding to assumption (B2)), meaning that for $d_x + d_y > 3$, the asymptotic bias in the Central Limit theorem is $\sqrt{n}c_{n\bar{g}}^2 c_{ny}^2 \to \infty$ and will not become zero when the variance is controlled. We will further need to restrict to $nc_{n\bar{g}}^{2d_{x\bar{g}}}c_{ny}^{2d_y} \to \infty$, effectively reducing the unbiased central limit theorem to the cases where there is only one continuous variable, although it can be either in y or x. As in Tamura and Boos [16] we also note that this bias is often small in practice; Section 6 demonstrates that a bootstrap method can remove it. We also note that in this case, the assumption of a compact domain for the covariates x can be relaxed.

In the case of (4.5), we have

$$\tilde{U}_J(y,x) = J(y,x) + 2\int J(y_*,x)f(y_*|x)\,\mathrm{d}\nu(y_*),$$
$$\tilde{B}_n = 2\iint J(y,x)\frac{E\hat{g}_n(y,x)}{E\hat{h}_n(x)}h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x),$$

where we note the additional variance due to the summation over X_i values. In this case, the assignment (4.6) with $f(y|x) = \phi_\theta(y|x)$ gives us that the variance is the information matrix directly. For $\tilde{U}_J$, we observe that

$$\int J(y_*,x)f(y_*|x)\,\mathrm{d}\nu(y_*) = \int \nabla_\theta \phi_\theta(y|x)\,\mathrm{d}\nu(y_*) = 0$$

since $\phi_\theta(y|x)$ integrates to 1 for each x and each θ, yielding the same variance term as above. The bias here is of the same order as above.

Homoscedastic: $x^{\bar{g}} = \phi$. Here the density estimate assumes that y has a location-scale family with $y - m(x)$ independent of x. In this case,

$$U_J(\varepsilon,x) = \int J\big(\varepsilon + m(x), x_*\big)h(x_*)\,\mathrm{d}\mu(x_*)$$
$$+ \iint J\big(e_* + m(x), x\big)g^c(x_*,e_*)\,\mathrm{d}\mu(x_*)\,\mathrm{d}\nu(e_*)$$
$$+ \varepsilon\iint \nabla_y J\big(e_* + m(x), x\big)g^c(x_*,e_*)\,\mathrm{d}\mu(x_*)\,\mathrm{d}\nu(e_*)$$

$$-\varepsilon \iint \nabla_y J\big(e_* + m(x_*), x_*\big) g^c(x_*, e_*)\, \mathrm{d}\nu(e)\, \mathrm{d}\mu(x_*),$$

$$B_n = 2 \iint J\big(e + m(x), x\big) E\,\hat{g}_n(x, e, m) h(x)\, \mathrm{d}\nu(e)\, \mathrm{d}\mu(x).$$

Here we observe that the bias is again of order c_{nx}^2. However, for e and x^m both univariate it is possible to make $\sqrt{n}B_n \to 0$ while retaining consistency of $\hat{g}_n(e, m)$ and $\hat{m}_n(x^m)$.

We also have

$$\tilde{U}_J(\varepsilon, x) = U_J(\varepsilon, x), \qquad \tilde{B}_n = B_n$$

since in this case, both estimators are equal.

When we make the replacement (4.6), we assume that the assumed residual density $\phi(e; \theta)$ is parameterized so that

$$\phi(e; \theta) = \phi^*(S_\theta e; \theta)$$

with

$$\int e e^T \phi^*(e, \theta)\, \mathrm{d}\nu(e) = \int \frac{\nabla_e \phi^*(e, \theta)^{TT}}{\phi^*(e; \theta)}\, \mathrm{d}e = I \quad \text{and} \quad \int e \phi^*(e; \theta) = 0$$

for all θ where I is the $d_y \times d_y$ identity matrix. The second equality can always be achieved by re-parameterizing so that $\phi^*(e; \theta) = \phi(I(\theta)^{1/2} e; \theta)$ along with appropriate centering. The first equality requires that the variance in $\phi^*(e; \theta)$ be equal to the Fisher information for the location family $\phi^*(e + \mu; \theta)$; this condition is satisfied, for example, for the multivariate normal density. We now have that the total gradient is

$$\nabla_e \phi^*(S_\theta e; \theta) = S_\theta \partial_e \phi^*(S_\theta e; \theta)$$

and hence

$$U_J(\varepsilon, x) = \overline{\partial_\theta m} S_\theta \frac{\partial_y \phi^*(S_\theta \varepsilon; \theta)}{\phi^*(S_\theta \varepsilon; \theta)} + \frac{\partial_\theta \phi(\varepsilon; \theta)}{\phi(\varepsilon; \theta)} + \varepsilon\big(\partial_\theta m(x, \theta) - \overline{\partial_\theta m}\big) S_\theta S_\theta^T,$$

where we have used the shorthand

$$\overline{\partial_\theta m} = \int_{\mathcal{X}} \partial_\theta m(x, \theta) h(x)\, \mathrm{d}\mu(x)$$

along with the observation that

$$\int \partial_y \phi(e; \theta)\, \mathrm{d}\nu(e) = \int \partial_\theta \phi(e; \theta)\, \mathrm{d}\nu(e) = \int \partial_y^2 \phi(e; \theta)\, \mathrm{d}\nu(e) = \int \partial_{y\theta}^2 \phi(e; \theta)\, \mathrm{d}\nu(e) = 0$$

and some cancelation. We have retained ϕ instead of ϕ^* in terms involving ∂_θ for the sake of notational compactness.

We now have that

$$\begin{aligned}
&\int U_J(e,x)^{TT}\phi(e;\theta)h(x)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x)\\
&\quad=(\overline{\partial_\theta m}S_\theta)^{TT}+\int\frac{\partial_\theta\phi(e;\theta)^{TT}}{\phi(e;\theta)}\,\mathrm{d}\nu(e)\\
&\qquad-\overline{\partial_\theta m}S_\theta\int\frac{\partial_y\phi(e;\theta)\,\partial_\theta\phi(e;\theta)^T}{\phi(e;\theta)}\,\mathrm{d}\nu(e)-\int\frac{\partial_\theta\phi(e;\theta)\,\partial_y\phi(e;\theta)^T}{\phi(e;\theta)}\,\mathrm{d}\nu(e)S_\theta^T\overline{\partial_\theta m}^T\\
&\qquad+\iint\bigl(\partial_\theta m(x;\theta)-\overline{\partial_\theta m}\bigr)eS_\theta S_\theta^T S_\theta S_\theta^T\varepsilon^T\bigl(\partial_\theta m(x;\theta)-\overline{\partial_\theta m}\bigr)\phi(e;\theta)h(x)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x)
\end{aligned}$$

by making a change of variables $\varepsilon=S_\theta^{-1}e$ in the last line and some cancelation we have that

$$\begin{aligned}
&\int U_J(e,x)^{TT}\phi(e;\theta)h(x)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x)\\
&\quad=\iint\partial_\theta m(x;\theta)\frac{\partial_y\phi(e;\theta)^{TT}}{\phi(e;\theta)}\partial_\theta m(x;\theta)^T\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)+\int\frac{\partial_\theta\phi(e;\theta)^{TT}}{\phi(e;\theta)}\,\mathrm{d}\nu(e)\\
&\qquad-\overline{\partial_\theta m}\int\frac{\partial_y\phi(e;\theta)\,\partial_\theta\phi(e;\theta)^T}{\phi(e;\theta)}\,\mathrm{d}\nu(e)-\int\frac{\partial_\theta\phi(e;\theta)\,\partial_y\phi(e;\theta)^T}{\phi(e;\theta)}\,\mathrm{d}\nu(e)\overline{\partial_\theta m}^T
\end{aligned}$$

which is readily verified to be the Fisher information for this model.

Where $\theta=(\theta_1,\theta_2)$ can be partitioned into parameters θ_1 that appear only in $m(x;\theta_1)$ and parameters θ_2 that appear only in $\phi(e;\theta_2)$ the terms on the second line above are zero and the resulting information matrix is diagonal. In the classical case of nonlinear regression with homoscedastic normal errors, we have

$$y_i=m(x_i,\theta)+\varepsilon_i,\qquad \varepsilon_i\sim N\bigl(0,\sigma^2\bigr)$$

the score covariance for (θ,σ) reduces to

$$\int U_J(e,x)^{TT}\phi(e;\theta)h(x)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x)=\begin{bmatrix}\dfrac{1}{\sigma^2}\displaystyle\int\nabla_\theta m(x;\theta)^{TT}h(x)\,\mathrm{d}x & 0\\ 0 & \dfrac{1}{2\sigma^4}\end{bmatrix}.$$

Joint centering and conditioning: $x^s=x$. Here we center and condition on the entire set of x. In this case our results are those of the uncentered case:

$$\begin{aligned}
U_J(e,x)&=J\bigl(e+m(x),x\bigr),\\
B_n&=2\iint J\bigl(e+m(x),x\bigr)E\hat{g}_n(x,e,m)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x).
\end{aligned}$$

For (4.5):

$$\tilde{U}_J(e,x) = U_J(e,x) + 2\int J\big(e_* + m(x),x\big) f(e_*|x)\,\mathrm{d}\nu(e_*),$$

$$\tilde{B}_n = 2\iint J\big(e + m(x),x\big)\frac{E\hat{g}_n(x,e)}{E\hat{h}_n(x)} h(x)\,\mathrm{d}\nu(e)\,\mathrm{d}\mu(x).$$

In this case, $U_J(x,y)$ and $\tilde{U}_J(x,y)$ are exactly the same as the non-centered case, yielding the information matrix with the replacement (4.6).

We note that while the non-centered and the jointly centered and conditioned cases always yield the Fisher information under the substitution (4.6), the case of centering by some variables and conditioning on others need not. Even in the homoscedastic case, efficiency is only gained when the variance of the model for the residuals is equal to the Fisher information for its mean. However, under these conditions, we can gain efficiency while reducing the bias in the central limit theorem above.

Employing these lemmas, we can demonstrate a central limit theorem for minimum conditional disparity estimates:

Theorem 4.1. *Let* $\{(X_{n1}, X_{n2}, Y_{n1}, Y_{n2}), n \geq 1\}$ *be given as in Section* 1.1, *under assumptions* (D1)–(D4), (E1)–(E2), (K1)–(K6), (B1)–(B4), (P1)–(P4) *and* (N1)–(N4) *define*

$$\theta_f = \arg\min_{\theta\in\Theta} D_\infty(f,\theta)$$

and

$$\begin{aligned}
H^D(\theta) &= \nabla_\theta^2 D_\infty(f,\theta)\\
&= \int A_2\left(\frac{f(y|x)}{\phi(y|x,\theta)}\right)\nabla_\theta^2\phi(y|x,\theta)h(x)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)\\
&\quad + \int A_3\left(\frac{f(y|x)}{\phi(y|x,\theta)}\right)\frac{\nabla_\theta\phi(y|x,\theta)^{TT}}{\phi(y|x,\theta)}h(x)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y),\\
I^D(\theta) &= H^D(\theta)V^D(\theta)^{-1}H^D(\theta),\\
\tilde{I}^D(\theta) &= H^D(\theta)\tilde{V}^D(\theta)^{-1}H^D(\theta)
\end{aligned}$$

then

$$\sqrt{n}\big[T_n(\breve{f}_n) - \theta_f - B_n\big] \to N\big(0, I^D(\theta_f)^{-1}\big)$$

and

$$\sqrt{n}\big[\tilde{T}_n(\breve{f}_n) - \theta_f - \tilde{B}_n\big] \to N\big(0, \tilde{I}^D(\theta_f)^{-1}\big)$$

in distribution where B_n, $\tilde{B}_n$, $V^D(\theta)$ *and* $\tilde{V}^D(\theta)$ *are obtained by substituting*

$$J(y,x) = A_1\left(\frac{f(y|x)}{\phi(y|x,\theta_f)}\right)\frac{\nabla_\theta\phi(y|x,\theta_f)}{\phi(y|x,\theta_f)} \tag{4.7}$$

into the expressions for B_n, $\tilde{B}_n$, V_J and $\tilde{V}_J$ in Lemma 4.2.

Here we note that in the case that $f = \phi_{\theta_0}$ for some θ_0, that $\theta_f = \theta_0$ and further since $A_1(1) = A_2(1) = A_3(1) = 1$ we have that $H^D(\theta_f)$ is given by the Fisher information for ϕ_{θ_0}. Since we have demonstrated above that $V^D(\theta_f)$ and $\tilde{V}^D(\theta_f)$ also correspond to the Fisher information in particular cases above, when this holds $I^D(\theta_f)$ and $I^{\tilde{D}}(\theta_f)$ also give us the Fisher information and hence efficiency.

Proof of Theorem 4.1. We will define $\bar{T}_n$, and $f_K(y|x)$ to be either the pair $(T_n, E[\hat{g}_n(x,y)]/\hat{h}_n(x))$ or $(\tilde{T}_n, E\hat{g}_n(x,y)/E\hat{h}_n(x))$. Our arguments now follow those in Tamura and Boos [16] and Park and Basu [13].

Since $\bar{T}_n(f)$ satisfies

$$\nabla_\theta D_n\big(f, \bar{T}_n(f)\big) = 0$$

we can write

$$\sqrt{n}\big(\bar{T}_n(\breve{f}_n) - \theta_0\big) = -\big[\nabla_\theta^2 D_n\big(\breve{f}_n, \theta^+\big)\big]^{-1}\sqrt{n}\nabla_\theta D_n(\breve{f}_n, \theta_0)$$

for some θ^+ between $\bar{T}_n(\breve{f}_n)$ and θ_f. It is therefore sufficient to demonstrate:

(i) $\nabla_\theta^2 D_n(\breve{f}_n, \theta^+) \to H^D(\theta_f)$ in probability.
(ii) $\sqrt{n}[\nabla_\theta D_n(\breve{f}_n, \theta_f) - \bar{B}_n] \to N(0, V^D(\theta_f)^{-1})$ in distribution

with $\bar{B}_n$ given by B_n or $\tilde{B}_n$ as appropriate.

We begin with (i) where we observe that by assumption (N4), $A_2(r)$ and $A_3(r)$ are bounded and the result follows from Theorems 2.3 and 3.2 and the dominated convergence theorem. In the case of Hellinger distance

$$\begin{aligned}\nabla_\theta^2 D(\breve{f}_n, \phi|x, \theta) &= \int\left[\frac{\nabla_\theta^2\phi(y,x,\theta)}{\sqrt{\phi(y,x,\theta)}} - \frac{\nabla_\theta\phi(y,x,\theta)^{TT}}{\phi(y,x,\theta)^{3/2}}\right]\sqrt{\breve{f}_n(y|x)}\,\mathrm{d}\nu(y)\\ &= \int\sqrt{\phi(y,x,\theta)}\nabla_\theta\Psi_\theta(y,x,\theta)\sqrt{\breve{f}_n(y|x)}\,\mathrm{d}\nu(y)\end{aligned}$$

so that $|\nabla_\theta^2 D_n(\breve{f}_n, \phi|x, \theta^+) - H^D(\theta_f)|$ can be expressed as

$$\begin{aligned}&\iint\sqrt{\phi(y,x,\theta)}\nabla_\theta\psi(y,x,\theta)\big(\sqrt{\breve{f}_n(y|x)} - \sqrt{f(y|x)}\big)\,\mathrm{d}\nu(y)h(x)\,\mathrm{d}\mu(x)\\ &\quad+\int\left(\frac{\nabla_\theta^2\phi(y,x,\theta^+)}{\sqrt{\phi(y,x,\theta^+)}} - \frac{\nabla_\theta^2\phi(y,x,\theta_f)}{\sqrt{\phi(y,x,\theta_f)}}\right)\sqrt{f(y|x)}\,\mathrm{d}\nu(y)h(x)\,\mathrm{d}\mu(x)\\ &\le \sup_{x\in\mathcal{X}} S\left(\int\big|\breve{f}_n(y|x) - f(y|x)\big|\,\mathrm{d}\nu(y)\right)^{1/2} + \mathrm{o}_p(1)\\ &= \mathrm{o}_p(1).\end{aligned}$$

Where the calculations above follow from assumption (N3), bounding (squared) Hellinger distance by L_1 distance, the uniform L_1 convergence of $\breve{f}_n$ (Theorem 2.1) and the consistency of θ (Theorem 3.2).

Turning to (ii) where we observe that by the boundedness of C and the dominated convergence theorem, we can write $\nabla_\theta D_n(\breve{f}_n, \phi|x, \theta) - \bar{B}_n$ as

$$\begin{aligned}
&\int A_2\left(\frac{\breve{f}_n(y|x)}{\phi(y|x,\theta)}\right)\nabla_\theta\phi(y|x,\theta)\,\mathrm{d}\nu(y) - \bar{B}_n \\
&\quad= \int A_1\left(\frac{f_K(y|x)}{\phi(y|x,\theta)}\right)\frac{\nabla_\theta\phi(y|x,\theta)}{\phi(y|x,\theta)}\big[\breve{f}_n(y|x) - f_K(y|x)\big]\,\mathrm{d}\nu(y) \\
&\qquad+ \int\left[A_2\left(\frac{\breve{f}_n(y|x)}{\phi(y|x,\theta)}\right) - A_2\left(\frac{f_K(y|x)}{\phi(y|x,\theta)}\right)\right]\nabla_\theta\phi(y|x,\theta)\,\mathrm{d}\nu(y) \\
&\qquad- \int A_1\left(\frac{f_K(y|x)}{\phi(y|x,\theta)}\right)\left(\frac{\breve{f}_n(y|x)}{\phi(y|x,\theta)} - \frac{f_K(y|x)}{\phi(y|x,\theta)}\right)\nabla_\theta\phi(y|x,\theta)\,\mathrm{d}\nu(y)
\end{aligned}$$

from a minor modification Lemma 25 of Lindsay [11] we have that by the boundedness of A_1 and A_2 there is a constant B such that

$$\big|A_2\big(r^2\big) - A_2\big(s^2\big) - \big(r^2 - s^2\big)A_1\big(s^2\big)\big| \le \big(r^2 - s^2\big)B$$

substituting

$$r = \sqrt{\frac{\breve{f}_n(y|x)}{\phi(y|x,\theta)}}, \qquad s = \sqrt{\frac{f_K(y|x)}{\phi(y|x,\theta)}}$$

we obtain

$$\begin{aligned}
&\int A_2\left(\frac{\breve{f}_n(y|x)}{\phi(y|x,\theta)}\right)\nabla_\theta\phi(y|x,\theta)\,\mathrm{d}\nu(y) - \bar{B}_n \\
&\quad= \int A_1\left(\frac{f_K(y|x)}{\phi(y|x,\theta)}\right)\frac{\nabla_\theta\phi(y|x,\theta)}{\phi(y|x,\theta)}\big[\breve{f}_n(y|x) - f_K(y|x)\big]\,\mathrm{d}\nu(y) \\
&\qquad+ B\int\frac{\nabla_\theta\phi(y|x,\theta)}{\phi(y|x,\theta)}\big(\sqrt{\breve{f}_n(y|x)} - \sqrt{f_K(y|x)}\big)^2\,\mathrm{d}\nu(y).
\end{aligned}$$

The result now follows from Lemmas 4.1 and 4.2.

For the special case of Hellinger distance, we observe that

$$\nabla_\theta D_n(\breve{f}_n, \phi|x, \theta) = \int\frac{\nabla_\theta\phi(y, x, \theta)}{\sqrt{\phi(y, x, \theta)}}\sqrt{\breve{f}_n(y|x)}\,\mathrm{d}\nu(y)$$

and applying the identity $\sqrt{a}-\sqrt{b}=(a-b)/2\sqrt{a}+(\sqrt{b}-\sqrt{a})^2/2\sqrt{a}$ with $a=f_K(y|x)$ and $b=\breve{f}_n(y|x)$, we obtain

$$\begin{aligned}
&\sqrt{n}\int \frac{\nabla_\theta \phi(y,x,\theta)}{\sqrt{\phi(y,x,\theta)}}\big(\sqrt{\breve{f}_n(y|x)}-\sqrt{f_K(y|x)}\big)\,\mathrm{d}\nu(y)\\
&\quad=\sqrt{n}\int \frac{\nabla_\theta \phi(y,x,\theta)}{2\sqrt{\phi(y,x,\theta)f_K(y|x)}}\big(\breve{f}_n(y|x)-f_K(y|x)\big)\,\mathrm{d}\nu(y)\\
&\qquad-\sqrt{n}\int \frac{\nabla_\theta \phi(y,x,\theta)}{2\sqrt{\phi(y,x,\theta)f_K(y|x)}}\big(\sqrt{\breve{f}_n(y|x)}-\sqrt{f_K(y|x)}\big)^2\,\mathrm{d}\nu(y)\\
&\quad=\sqrt{n}\left(\int \frac{\nabla_\theta \phi(y,x,\theta)}{2\sqrt{\phi(y,x,\theta)f_K(y|x)}}\breve{f}_n(y|x)-B_n\right)+\mathrm{o}_p(1),
\end{aligned}$$

where we have applied Lemma 4.1 to the second term in the expression above, and can now obtain the result from Lemma 4.2 and the convergence of $f_K(y|x)$ to $f(y|x)$. □

We note here that Theorem 4.1 relies on assumption (D4) only through the consistency of $\bar{T}_n(\breve{f}_n)$ and Lemmas 4.1 and 4.2. In the case of $\tilde{T}_n(\hat{f}_n^*)$ (uncentered densities with the integral form of the disparity), we can remove this condition by employing Theorem E.1, and Lemmas E.1 and E.2 from Supplemental Appendix E (Hooker [8]).

5. Robustness properties

An important motivator for the study of disparity methods is that in addition to providing statistical efficiency as demonstrated above, they are also robust to contamination from outlying observations. Here we investigate the robustness of our estimates through their breakdown points. These have been studied for i.i.d. data in Beran [3]; Park and Basu [13]; Lindsay [11] and the extension to conditional models follows similar lines.

In particular, we examine two models for contamination:

1. To mimic the "homoscedastic" case, we contaminate $g(x_1,x_2,y_1,y_2)$ with outliers independent of (x_1,x_2). That is, we define the contaminating density

$$g_{\varepsilon,z}(x_1,x_2,y_1,y_2)=(1-\varepsilon)g(x_1,x_2,y_1,y_2)+\varepsilon\delta_z(y_1,y_2)h(x_1,x_2), \tag{5.1}$$

where δ_z is a contamination density parameterized by z such that δ_z becomes "outlying" as $z\to\infty$. Typically, we think of δ_z as having small support centered around z. This results in the conditional density

$$f_{\varepsilon,z}(y_1,y_2|x_1,x_2)=(1-\varepsilon)f(y_1,y_2|x_1,x_2)+\varepsilon\delta_z(y_1,y_2)$$

which we think of as the result of smoothing a contaminated residual density. We note that we have not changed the marginal distribution of (x_1,x_2) via this contamination. This particularly applies to the case where only y_1 is present and the estimate (1.14)–(1.16) is employed.

2. In the more general setting, we set

$$g_{\varepsilon,z}(x_1, x_2, y_1, y_2) = (1-\varepsilon)g(x_1, x_2, y_1, y_2) + \varepsilon\delta_z(y_1, y_2)J_U(x_1, x_2)h(x_1, x_2), \quad (5.2)$$

where $J_U(x_1, x_2)$ is the indicator of $(x_1, x_2) \in U$ scaled so that $h(x_1, x_2)J_U(x_1, x_2)$ is a distribution. This translates to the conditional density

$$f_{\varepsilon,z}(y_1, y_2|x_1, x_2) = \begin{cases} f(y_1, y_2|x_1, x_2), & (x_1, x_2) \notin U, \\ (1-\varepsilon)f(y_1, y_2|x_1, x_2) + \varepsilon\delta_z(y_1, y_2), & (x_1, x_2) \in U \end{cases}$$

which localizes contamination in covariate space. Note that the marginal distribution is now scaled differently in U.

Naturally, this characterization (5.1) does not account for the effect of outliers on the Nadaraya–Watson estimator (1.14). If these are localized in covariate space, however, we can think of (1.16) as being approximately a mixture of the two cases above. As we will see the distinction between these two will not affect the basic properties below. Throughout we will write $\delta_z(y_1, y_2|x_1, x_2)$ in place of $\delta_z(y_1, y_2)$ or $\delta_z(y_1, y_2)J_U(x_1, x_2)$ as appropriate. $h(x_1, x_2)$ will be taken to be modified according to (5.2) if appropriate.

We must first place some conditions on δ_z:

C1. δ_z is orthogonal in the limit to f. That is

$$\lim_{z\to\infty} \sum_{y_2 \in S_y} \int \delta_z(y_1, y_2|x_1, x_2) f(y_1, y_2|x_1, x_2)\, \mathrm{d}y_1 = 0 \qquad \forall (x_1, x_2).$$

C2. δ_z is orthogonal in the limit to ϕ:

$$\lim_{z\to\infty} \sum_{y_2 \in S_y} \int \delta_z(y_1, y_2|x_1, x_2)\phi(y_1, y_2|x_1, x_2, \theta)\, \mathrm{d}y_1 = 0 \qquad \forall (x_1, x_2).$$

C3. ϕ becomes orthogonal to f for large θ:

$$\lim_{\|\theta\|\to\infty} \sum_{y_2 \in S_y} \int f(y_1, y_2|x_1, x_2)\phi(y_1, y_2|x_1, x_2, \theta) = 0 \qquad \forall (x_1, x_2).$$

C4. $C(-1)$ and $C'(\infty)$ are both finite or the disparity is Hellinger distance.

In the following result with use $T[f] = \arg\min D_\infty(f, \theta)$ for any f in place of our estimate $\hat{\theta}$.

Theorem 5.1. *Under assumptions* C1–C4 *under both contamination models* (5.1) *and* (5.2) *define* ε^* *to satisfy*

$$\big(1-2\varepsilon^*\big)C'(\infty) = \inf_{\theta\in\Theta} D\big(\big(1-\varepsilon^*\big)f, \theta\big) - \lim_{z\to\infty} \inf_{\theta\in\Theta} D\big(\varepsilon^*\delta_z, \theta\big) \quad (5.3)$$

with $C'(\infty)$ *replaced by* 1 *in the case of Hellinger distance then for* $\varepsilon < \varepsilon^*$

$$\lim_{z\to\infty} T[f_{\varepsilon,z}] = T\big[(1-\varepsilon)f\big]$$

and in particular the breakdown point is at least ε^*: for $\varepsilon < \varepsilon^*$,

$$\sup_z \big\| T[f_{\varepsilon,z}] - T\big[(1-\varepsilon)f\big] \big\| < \infty.$$

Proof. We begin by observing that by assumption C1, for any fixed θ,

$$\begin{aligned} D(f_{\varepsilon,z},\theta) &= \iint_{A_z(x)} C\left(\frac{f_{\varepsilon,z}(y|x)}{\phi(y|x,\theta)} - 1\right)\phi(y|x,\theta)h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x) \\ &\quad + \iint_{A_z^c(x)} C\left(\frac{f_{\varepsilon,z}(y|x)}{\phi(y|x,\theta)} - 1\right)\phi(y|x,\theta)h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x) \\ &= D_{A_z}(f_{\varepsilon,z},\theta) + D_{A_z^c}(f_{\varepsilon,z},\theta), \end{aligned}$$

where $A_z(x) = \{y\colon \max(f(y|x), \phi(y|x,\theta)) > \delta_z(y|x)\}$. We note that for any η with z sufficiently large that

$$\sup_{x\in\mathcal{X}} \sup_{y\in A_z(x)} \delta_z(y|x) < \eta \quad \text{and} \quad \sup_{(x)\in\mathcal{X}} \sup_{y\in A_z^c(x)} f(y|x) < \eta$$

and thus for sufficiently large z,

$$\begin{aligned} &\big|D(f_{\varepsilon,z},\theta) - \big(D_{A_z}\big((1-\varepsilon)f,\theta\big) + D_{A_z^c}(\varepsilon\delta_z,\theta)\big)\big| \\ &\quad \le \iint C\left(\frac{\eta}{\phi(y|x,\theta)} - 1\right)\phi(y|x,\theta)h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x) \\ &\quad \le \eta \sup_t \big|C'(t)\big| \end{aligned}$$

hence

$$\sup_\theta \big|D(f_{\varepsilon,z},\theta) - \big(D_{A_z}\big((1-\varepsilon)f,\theta\big) + D_{A_z^c}(\varepsilon\delta_z,\theta)\big)\big| \to 0. \tag{5.4}$$

We also observe that for any fixed θ,

$$\begin{aligned} D_{A_z^c}(\varepsilon\delta_z,\theta) &= \iint_{A_z^c(x)} C\left(\frac{2\varepsilon\delta_z(y)}{\phi(y|x,\theta)} - 1\right)\phi(y|x,\theta)h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x) \\ &\quad + \iint_{A_z^c(x)} \varepsilon\delta_z(y|x)C'\big(t(y,x)\big)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x) \\ &\to \varepsilon C'(\infty) \end{aligned}$$

for $t(y,x)$ between $\varepsilon\delta_z(y|x)/\phi(y|x,\theta)$ and $2\varepsilon\delta_z(y|x)/\phi(y|x,\theta)$ since $t(y,x)\to\infty$, $C(\cdot)$ and $C'(\cdot)$ are bounded and $\int_{A_z^c(x)} \phi(y|x,\theta)\,\mathrm{d}\nu(y) \to 0$.

Similarly,

$$D_{A_z^c}\big((1-\varepsilon)f,\theta\big) \to D\big((1-\varepsilon)f,\theta\big)$$

and thus

$$D(f_{\varepsilon,z},\theta)\to D\big((1-\varepsilon)f,\theta\big)+\varepsilon C'(\infty)$$

which is minimized at $\theta=T[f_{\varepsilon,z}]$.

It remains to rule out divergent sequences $\|\theta_z\|\to\infty$. In this case, we define $B_z(x)=\{y\colon\ f(y|x)>\max(\varepsilon\delta_z(y|x),\phi(y|x,\theta_z))\}$ and note that from the arguments above

$$D_{B_z}\big((1-\varepsilon)f,\theta_z\big)\to(1-\varepsilon)C'(\infty)$$

and

$$D_{B_z^c}(\varepsilon\delta,\theta_z)\to D(\varepsilon\delta,\theta_z)$$

and hence

$$\lim_{z\to\infty}D(f_{\varepsilon,z},\theta_z)>\lim_{z\to\infty}\inf_{\theta\in\Theta}D(\varepsilon\delta_z,\theta)+(1-\varepsilon)C'(\infty)>D\big(f_{\varepsilon,z},T\big[(1-\varepsilon)f\big]\big)$$

from (5.3), yielding a contradiction.

In the case of Hellinger distance, we observe

$$\begin{aligned}
&\big|D(f_{\varepsilon,z},\theta)-\big(D\big((1-\varepsilon)f,\theta\big)+D(\varepsilon\delta_z,\theta)\big)\big|\\
&\quad=\iint\sqrt{\phi(y|x,\theta)}\big(\sqrt{f_{\varepsilon,z}(y|x)}-\sqrt{(1-\varepsilon)f(y|x)}-\sqrt{\varepsilon\delta_z(y|x)}\big)h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x)\\
&\quad\le\iint\big(\sqrt{f_{\varepsilon,z}(y|x)}-\sqrt{(1-\varepsilon)f(y|x)}-\sqrt{\varepsilon\delta_z(y|x)}\big)^2h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x)\\
&\quad=\iint\big[2(1-\varepsilon)f(y|x)+2\varepsilon\delta(y|x)\big]h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x)\\
&\qquad-2\iint\big(\sqrt{f_{\varepsilon,z}(y|x)}\big(\sqrt{(1-\varepsilon)f(y|x)}+\sqrt{\varepsilon\delta_z(y|x)}\big)\big)h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x),
\end{aligned}$$

where, by dividing the range of y into $A_z(x)$ and $A_z^c(x)$ as above, we find that on $A_z(x)$, for any $\eta>0$ and z sufficiently large,

$$\big|(1-\varepsilon)f(y|x)-\sqrt{f_{\varepsilon,z}(y|x)}\big(\sqrt{(1-\varepsilon)f(y|x)}+\sqrt{\varepsilon\delta_z(y|x)}\big)\big|\le\sqrt{\varepsilon\eta f_{\varepsilon,z}(y|x)}+\varepsilon\eta$$

which with the corresponding arguments on $A_z^c(x)$ yields (5.4). We further observe that for fixed θ

$$D(\varepsilon\delta_z,\theta)=1+\varepsilon-\sqrt{\varepsilon}\int\sqrt{\delta_z(y|x)\phi(y|x,\theta)}h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x)\to1+\varepsilon$$

and for $\|\theta_z\|\to\infty$,

$$D\big((1-\varepsilon)f,\theta_z\big)=2-\varepsilon-\sqrt{1-\varepsilon}\int\sqrt{f(y|x)\phi(y|x,\theta_z)}h(x)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x)\to2-\varepsilon$$

from which the result follows from the same arguments as above. □

These results extend on Park and Basu [13] and Beran [3] and a number of ways and a few remarks are warranted:

1. In Beran [3], Θ is assumed to be compact, allowing θ_z to converge at least on a subsequence. This removes the $\|\theta_z\| \to \infty$ case and the result can be shown for $\varepsilon \in [0, 1)$.
2. We have not assumed that the uncontaminated density f is a member of the parametric class ϕ_θ. If $f = \phi_{\theta_0}$ for some θ_0, then we observe that by Jensen's inequality

$$D\big((1-\varepsilon)\phi_{\theta_0}, \theta\big) > C(-\varepsilon) = D\big((1-\varepsilon)\phi_{\theta_0}, \theta_0\big)$$

hence $T[(1-\varepsilon)\phi_{\theta_0}] = \theta_0$. We can further bound $D(\varepsilon\delta_z, \theta) > C(\varepsilon - 1)$ in which case (5.3) can be bounded by

$$(1-2\varepsilon)C'(\infty) \geq C(-\varepsilon) - C(\varepsilon - 1)$$

which is satisfied for $\varepsilon = 1/2$. We note that in the more general condition, if $(1-\varepsilon)f$ is closer to the family ϕ_θ than $\varepsilon\delta_z$ at $\varepsilon = 1/2$, the breakdown point will be greater than $1/2$; in the reverse situation it will be smaller.

We emphasize here that we consider robustness here in the sense of having outliers in the response variables Y_i. Outliers in the X_i result in points of high leverage, to which our methods are not robust. Robustness in this sense would require a weighted combination of the $D_n(f, \phi|x, \theta)$ as an objective and the resulting efficiency properties of the model are not clear.

6. Bandwidth selection, bootstrapping, bias correction and inference

The results in the previous sections indicate that minimum disparity estimates based on nonparametric conditional density estimates are efficient in the sense that their asymptotic variance is identical to the Fisher information when the model is correct. They are also robust to outliers. This comes at a price, however, of a bias that is asymptotically non-negligible. Here, we propose to correct this bias with a bootstrap based on the estimated conditional densities. This will also provide a means of inference that does not assume the parametric model. We also provide details of the bandwidth selection methods used in our empirical studies. The details in this section are heuristic choices applied to the simulation studies in Section 7 and real data analysis in Section 8.

6.1. Bandwidth selection

Bandwidth selection is not particularly well studied for multivariate or conditional density estimates and software implementing existing methods is not readily available. Here, we employed a naïve cross-validation approach designed to be methodologically straightforward. In particular:

1. We chose bandwidths $c_{n\tilde{m}}$ for $\hat{m}_n$ by cross-validating squared error.

2. We chose bandwidths $c_{n\bar{g}}$ associated with $x^{\bar{g}}$ in $\hat{h}_n$ by cross-validating the non-parametric log likelihood:

$$c_{n\bar{g}} = \arg\max \sum_{i=1}^{n} \log \hat{h}_n^{-i}\left(X_i^{\bar{g}}\right),$$

   where $\hat{h}_n^{-i}$ is the estimate $\hat{h}_n$ based on the data set with $X_i^{\bar{g}}$ removed.
3. We fixed $\hat{m}_n$ and $\hat{h}_n$ and their bandwidths and chose c_{ny} based on cross-validating the non-parametric conditional log likelihood:

$$c_{ny} = \arg\max \sum_{i=1}^{n} \log \hat{g}_n^{-i}\left(Y_i - \hat{m}_n\left(X_i^{\bar{m}}\right), X_i^{\bar{g}}\right).$$

   Noting that the denominator in the conditional density becomes an additive term after taking logs and does not change with c_{ny}.

Where we also used discrete values X_2, these bandwidths were estimated for each value of X_2 separately at each step. The resulting bandwidths were then averaged in order to improve the stability of bandwidth selection.

6.2. Bootstrapping

We have two aims in bootstrapping: bias correction and inference. Nominally, we can base inference on the asymptotic normality results established in Theorem 4.1 using the inverse of the Fisher information as the variance for the estimated parameters. However, the coverage probabilities of confidence intervals based on these results will be poor due to the non-negligible bias in the theorem; it will also not provide correct coverage when the assumed parametric model is incorrect.

As an alternative, we propose a bootstrap based on the estimated non-parametric conditional densities. That is, to create each bootstrap sample, we simulate a new response Y_i^* from $\check{f}_n(\cdot|X_{i1}, X_{i2})$ for $i = 1, \ldots, n$ and use these to re-estimate parameters $\hat{\theta}$. For continuous Y_{i1}, simulating from this density can be achieved by choosing Y_{j1} with weights $K([X_i - X_j]/c_{n\bar{g}})$ and then simulating from the density $c_{ny}^{-\mathrm{dy}} K((y - Y_i)/c_{ny})$. For discrete Y_{i2}, simulating from the non-parametric multinomial model is straightforward.

In the simulation experiments below, we examine a number of different choices of $x^{\bar{m}}$ and $x^{\bar{g}}$ and each is bootstrapped separately. For maximum likelihood and other robust estimators, we employ a residual bootstrap for continuous responses and a parametric bootstrap for discrete responses.

We also examine a hybrid method proposed in Hooker and Vidyashankar [9] in which we replace $\hat{m}_n$ with a parametric regression model $m(x, \theta)$. We then minimize the disparity between the estimated density of residuals (which varies with parameters) and a parametric residual den-

sity. Specifically, we set

$$E_i(\theta) = Y_i - m(X_i, \theta),$$
$$\tilde{f}_n(e, \theta) = \frac{1}{nc_n} \sum_{i=1}^{n} K\left(\frac{e - E_i(\theta)}{c_n}\right),$$
$$\tilde{\theta}_n = \arg\min_{\theta \in \Theta} \int C\left(\frac{\tilde{f}_n(e, \theta)}{\phi(e)} - 1\right)\phi(e)\,\mathrm{d}e.$$

This formulation avoids conditional density estimation (and hence asymptotic bias) at the expense of a parameter-dependent kernel density estimate for the residuals. In this formulation $\phi(e)$ is a reference residual density in which a scale parameter has been robustly estimated. In the simulations below, the scale parameter is re-estimated via a disparity method with the remaining θ held fixed. For this case, we employ a parametric bootstrap at the estimated parameters, but sample from the estimated non-parametric residual density. Throughout, we keep the estimated bandwidths fixed.

6.3. Inference

Given a bootstrap sample θ_b^*, $b = 1, \ldots, B$ along with our original estimate $\hat{\theta}$, we conduct inference along well established lines:

- Obtain a bias corrected estimate

$$\hat{\theta}^c = 2\hat{\theta} - \frac{1}{B}\sum_{b=1}^{B} \theta_b^*.$$

- Estimate a bootstrap standard error, $\widehat{\mathrm{se}}(\theta)$, from the sample standard deviation of θ_b.
- Construct confidence intervals $[\hat{\theta}^c - 1.96\widehat{\mathrm{se}}(\theta), \hat{\theta}^c + 1.96\widehat{\mathrm{se}}(\theta)]$.

The performance of these confidence intervals will be examined in the simulation studies below, but we make a couple of remarks on this:

1. Our bootstrap scheme amounts to simulation under the model $\breve{f}_n$. Given the convergence of $\breve{f}_n$ to f in Theorem 2.1 and the continuity of $I^D(\theta)$ and $\tilde{I}^D(\theta)$ in f, the bootstrap standard error can be readily shown to be consistent for the sampling standard error of $\hat{\theta}$. Similarly, since density estimates with bandwidths c_{ny} and $2c_{ny}$ converge, the bias correction incurs no additional variance.
2. The bias correction for the proposed bootstrap approximates considering the difference between estimating $\breve{f}_n$ with bandwidths c_{ny} and $2c_{ny}$; this is exactly true when employing a Gaussian kernel. The bias terms in Lemma 4.2 are readily shown to be $\mathrm{O}(c_{ny}^2)$ which would suggest a corrected estimate of the form $(4\hat{\theta} - 1/B\sum\theta_b^*)/3$ instead of the linear correction proposed above. However the estimate is also biassed due to the nonlinear dependence of $\hat{\theta}$ on $\breve{f}_n$ regardless of the value of c_{ny}. This bias is asymptotically negligible, but we

have found the proposed correction to provide better performance at realistic sample sizes. A combined bias correction associated with explicitly obtaining an estimate at $2c_{ny}$ to correct for smoothing bias with a bootstrap estimate to correct for intrinsic bias may improve performance further, but this is beyond the scope of this paper.

7. Simulation studies

Here we report simulation experiments designed to evaluate the methods analyzed above. Our examples are all based on conditionally-specified regression models. In all of these, we generate a three-dimensional set of covariates in the following manner:

1. Generate $n \times 3$ matrix X from a Uniform random variable on $[-1, 1]$.
2. Post-multiply this matrix by a $\sqrt{8}/3$ times a matrix with unit diagonal and 0.25 in all off-diagonal entries to create correlation.
3. Replace the third column of X with the indicator of the corresponding entry being greater than zero.

This gives us two continuous valued covariates and a categorical covariate all of which are correlated. The values of these covariates were regenerated in each simulation.

Using these covariates, we simulated data from two models:

- A linear regression with Gaussian errors and all coefficients equal to 1:

$$Y_i = 1 + \sum_{j=1}^{3} X_{ij} + \varepsilon_i \tag{7.1}$$

  with $\varepsilon_i \sim N(0, 1)$, This yields a signal to noise ratio of 1.62. In this model, we estimate the intercept and all regression parameters as well as the noise variance, yielding true values of $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma) = (1, 1, 1, 1, 1)$. We optimize over $\log \sigma$ to avoid boundary problems and have reported estimate and standard errors for $\log \sigma$ below.
- A logistic regression with zero intercept and all other coefficients 0.5:

$$P(X_i = 1|X_i) = \frac{e^{\sum_{j=1}^{3} 0.5X_{ij}}}{1 + e^{\sum_{j=1}^{3} 0.5X_{ij}}} \tag{7.2}$$

  in order to evaluate a categorical response model. Here only the four regression parameters were estimated.

In each model we also examined the addition of outliers. In (7.1), we changed either 1, 3, 5 or 10 of the ε_i to take values 3, 5, 10 and 15. These covariate values X_i corresponding the modified ε_i where held constant within each simulation study, but were selected in two different ways:

1. At random from among all the data.
2. Based on the points with X_{i1} closest to -0.5.

These mimic the contamination scenarios above.

In binary response data in (7.2), we require a model in which an "outlier" distribution can become orthogonal to the model distribution. For binary data this can occur only if the parametric model has $P(Y = 1|X) \approx 0$ or $P(Y = 1|X) \approx 1$ which for logistic regression can occur only at values of X that have high leverage; a robustness problem not considered in this paper. Instead, we examine a logistic binomial model based on successes out of 8 trials. For this, we have employed an exact distribution which is contaminated with $\alpha\%$ of a distribution in which points take the value 8, either uniformly as in scenario (5.1) or at the single X_i with X_{i1} closest to -0.5 as in scenario (5.2). In this case, reasonable estimates of conditional distributions would require very large sample sizes and we have based all our estimates on exact distributions.

7.1. Linear regression

For the linear regression simulations, we employed 31 points generated as above. We considered three types of density estimates corresponding to no centering (labeled HD and NED for Hellinger distance and negative exponential disparity), jointly centering and conditioning on all variables (HD.c and NED.c) and the homoscedastic model: centering by all variables but assuming a constant residual density (HD.h and NED.h). We also included the marginal method of Hooker and Vidyashankar [9] which involves only fitting a kernel density estimate to the residuals of a linear regression. Bandwidths where chosen by cross-validated log likelihood for uncontaminated data. We conducted all estimates by minimizing $D_n(\check{f}, \theta)$ with $D(\check{f}, \phi|X_i, \theta)$ approximated a Monte Carlo integral based on 101 points drawn from $\check{f}(\cdots|X_i)$.

We also included a standard linear regression (Lik) and Gervini and Yohai's estimates (Gervini and Yohai [6]) based on a Huberized estimate with an adaptively-chosen threshold (G–Y). Table 1

Table 1. Simulation results for a linear regression simulation. `Lik` are the maximum likelihood estimates, `G-Y` correspond to Gervini and Yohai's adaptive truncation estimator, `HD` is minimum Hellinger distance, `NED` is minimum negative exponential disparity based on uncentered kernel density estimates, `HD.c` and `NED.c` are centered by a Nadaraya–Watson estimator, `HD.h` and `NED.h` are based on homoscedastic conditional density estimates and `HD.m` and `NED.m` are the marginal estimators in Hooker and Vidyashankar [9]. We report the mean value over 5000 simulations as well as the standard deviation (sd) between simulations

	$\log\sigma$	sd	β_0	sd	β_1	sd	β_2	sd	β_3	sd	Time
Lik	−0.10	0.14	1.00	0.28	1.00	0.40	0.99	0.40	0.99	0.43	0.0049
G–Y	−0.10	0.19	1.00	0.30	1.00	0.43	0.99	0.42	0.99	0.46	0.0144
HD.c	0.13	0.44	0.97	0.34	0.94	0.60	0.94	0.50	1.05	0.60	0.0588
NED.c	0.12	0.23	0.98	0.30	0.95	0.40	0.94	0.40	1.04	0.45	0.0751
HD	0.26	0.37	0.94	0.40	0.87	0.39	0.86	0.51	1.11	0.58	0.0604
NED	0.25	0.16	0.94	0.30	0.87	0.35	0.87	0.36	1.11	0.45	0.0776
HD.h	−0.18	0.32	0.95	0.50	0.88	0.34	0.88	0.34	1.10	0.75	0.0616
NED.h	−0.17	0.21	0.95	0.34	0.88	0.34	0.88	0.34	1.09	0.50	0.0764
HD.m	0.05	0.17	1.00	0.29	1.00	0.43	1.00	0.42	1.00	0.45	0.0328
NED.m	0.06	0.16	1.00	0.30	1.00	0.44	1.00	0.44	1.00	0.47	0.0292

reports the means and standard deviations of the parameters in this model calculated from 5000 simulations before bootstrap methods are applied. We present computation times here as well; bootstrapping results in multiplying these times by 100 for all estimators.

As can be observed from these results, the use of multivariate density estimation creates significant biases, particularly in β_2 and β_3. This is mitigated in the centered density estimates, although not for the homoscedastic estimators. We speculate that this is because the conditional density estimate can correct for biasses from the Nadaraya–Watson estimator which the homoscedastic restriction does not allow for. The marginal methods perform considerably better and achieve similar performance to those of Gervini and Yohai [6]. We also observe that Hellinger distance estimators have large variances in some cases, mostly due to occasional outlying parameter estimates. By contrast, negative exponential disparity estimators were much more stable.

In addition to the simulations above, for each simulated data set we performed 100 bootstrap replicates as described in Section 6 and used this to both provide a bias correction and confidence intervals. The resulting point estimates and coverage probabilities are reported in Table 2. Here we see that much of the bias has been removed for all estimators except for the homoscedastic models. Coverage probabilities are at least as close to nominal values as minimum squared error estimators.

To examine results when the data are contaminated, we plot the mean estimate for β_0 under the contamination model 1 in Figure 1 as the position of the contamination increases; this mimics the bias plots of Lindsay [11], Figures 1 and 2. We have reported plots at each level of the number of contaminated observations. Here, we observe that the least squares estimator is strongly affected although most robust estimators are not. At 10 (30%) contaminated observations, the Gervini–Yohai estimator exhibits greater distortion of all except the homoscedastic and maximum likelihood estimators, although it remains robust and the tendency to ignore large outliers is evident. We speculate that the breakdown in the homoscedastic methods is because the underlying Nadaraya–Watson estimator is locally influenced strongly by these values and the

Table 2. Statistical properties (estimate, standard deviation (sd) and coverage (cov)) of inference following bootstrap bias correction and using bootstrap confidence intervals. Labels for estimators are the same as in Table 1

	β_0^c	sd	cov	β_1^c	sd	cov	β_2^c	sd	cov	β_3^c	sd	cov
Lik	1.00	0.28	0.92	1.00	0.4	0.92	1.00	0.4	0.91	0.99	0.41	0.92
Hub	1.00	0.29	0.91	1.00	0.41	0.91	1.00	0.41	0.91	0.99	0.43	0.91
G–Y	1.01	0.31	0.91	1.00	0.43	0.92	1.00	0.43	0.91	0.99	0.45	0.92
HD.c	1.00	0.31	0.95	0.99	0.43	0.93	0.99	0.43	0.93	1.00	0.46	0.95
NED.c	1.00	0.31	0.96	0.99	0.42	0.94	0.99	0.43	0.94	1.00	0.46	0.95
HD	0.99	0.42	0.98	0.98	0.59	0.95	0.97	0.48	0.94	1.02	0.62	0.98
NED	0.99	0.3	0.98	0.97	0.4	0.96	0.97	0.4	0.96	1.02	0.44	0.98
HD.h	0.99	0.58	0.81	0.97	0.38	0.86	0.98	0.39	0.84	1.02	0.76	0.81
NED.h	1.00	0.36	0.86	0.97	0.39	0.86	0.98	0.4	0.86	1.01	0.52	0.87
HD.m	1.00	0.31	0.9	0.99	0.44	0.95	1.00	0.46	0.95	1.00	0.47	0.95
NED.m	1.00	0.32	0.9	0.99	0.46	0.94	1.00	0.47	0.94	1.00	0.48	0.95

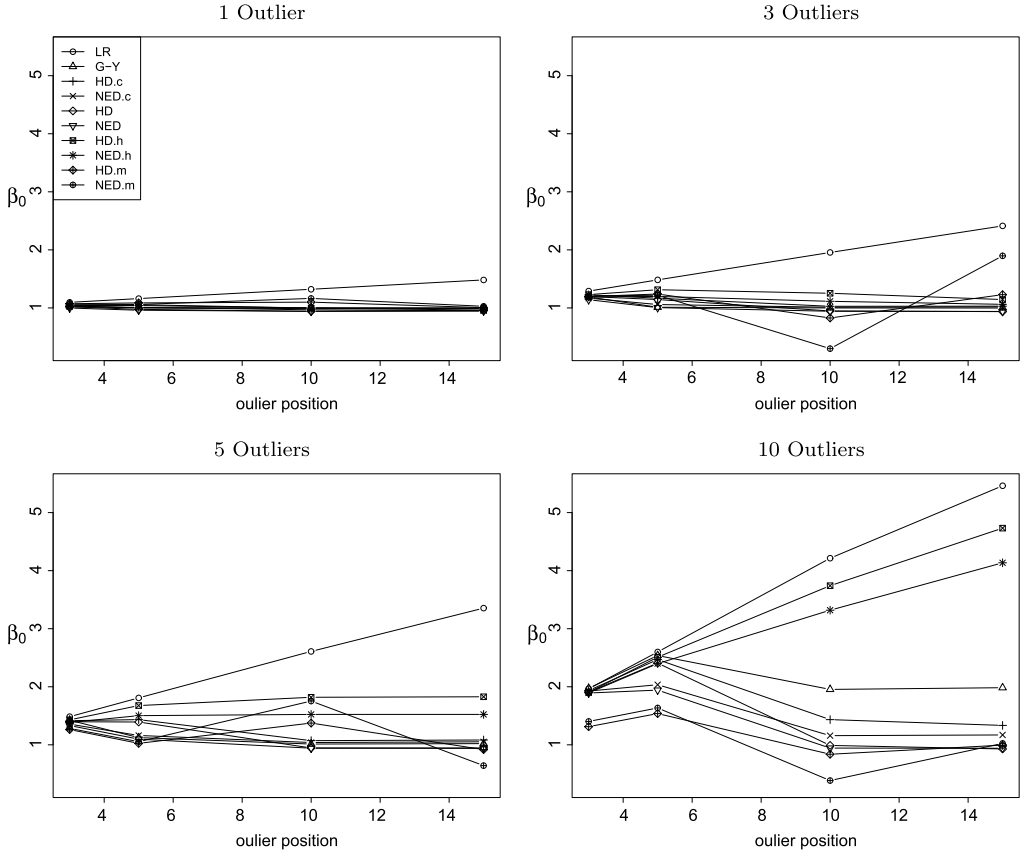

Figure 1. Mean estimates $\hat{\beta}_0$ with different levels of contamination uniformly distributed over covariate values. Each line corresponds do a different estimation method as given in the key.

homoscedastic restriction does not allow it to compensate for this. Estimates for the variance σ were similarly affected but the other regression parameters were not influenced by outliers since they were uniformly distributed over the range of covariates. A complete set of graphs is given in Figure 1 in Supplemental Appendix A (Hooker [8]).

By contrast, under contamination model 5.2, all least-squares parameter estimates were affected by outliers. We have plotted the average estimates for each parameter for 10 outliers in Figure 2 using the same key as in Figure 1. Here we observe that most estimators were robust, although the Gervini–Yohai as well as the homoscedastic models were affected. Investigating this more closely, at this level of contamination, sampling distribution the Gervini–Yohai estimator appears multi-modal which we speculate is associated with the adaptive choice of the Huber threshold failing to reject some of the outliers. It should be noted that this behavior was not evident at smaller contamination percentages. Examining Figure 2 in Supplemental Appendix A (Hooker [8]), we observe that this breakdown in robustness occurs most dramatically only at 10

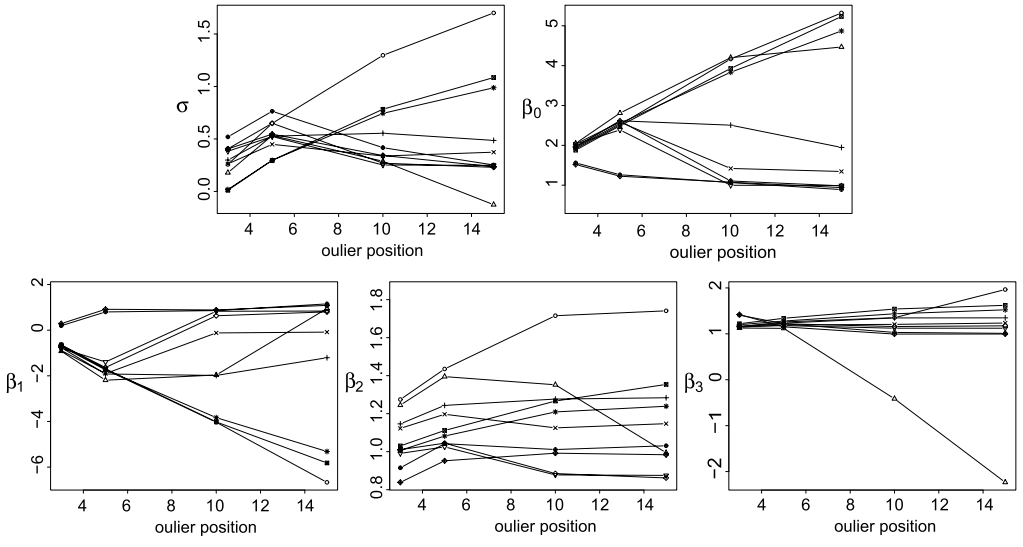

Figure 2. Mean parameter estimates with 10 outliers with values x_1 close to -0.5.

outliers, although the homoscedastic estimators (but not Gervini–Yohai) show some evidence for this at 5 outliers as well.

7.2. Logistic regression

For logistic regression there is no option to center the response before producing a conditional density estimate. We therefore examine only the logistic regression (Lik), Hellinger distance (HD) and negative exponential disparity (NED) estimators. Because logistic regression estimates are less stable than linear regression, we used 121 points generated as described above. We also note that Monte Carlo estimates are not required to evaluate the disparity in this case since it is defined as a sum over a discrete set of points. Simulation results are reported in Table 3.

There is again a noticeable bias in these estimates and we employed the bootstrapping methods outlined above both to remove the bias in the estimates and to estimate confidence intervals. For each data set, we simulated 100 bootstrap samples and used these to estimate the bias and

Table 3. Simulation results for logistic regression using maximum likelihood (`LR`), Hellinger distance (`HD`) and negative exponential disparity (NED) estimates

	β_0	sd	β_1	sd	β_2	sd	β_3	sd	Time
LR	0.00	0.29	0.53	0.42	0.52	0.42	0.52	0.44	0.01
HD	−0.01	0.33	0.57	0.44	0.56	0.44	0.58	0.49	0.01
NED	−0.01	0.29	0.51	0.39	0.5	0.39	0.54	0.44	0.01

Table 4. Simulation results for logistic regression following a bootstrap to correct for bias and construct confidence intervals using maximum likelihood LR, Hellinger distance HD and negative exponential disparity (NED) estimates with mean estimate, standard deviation across simulations (sd) and coverage of bootstrap confidence intervals (cov)

	β_0^c	sd	cov	β_1^c	sd	cov	β_2^c	sd	cov	β_3^c	sd	cov
LR	−0.01	0.28	0.97	0.51	0.4	0.97	0.49	0.4	0.97	0.5	0.42	0.97
HD	0.00	0.31	0.96	0.55	0.43	0.95	0.53	0.44	0.94	0.52	0.46	0.96
NED	−0.01	0.28	0.95	0.5	0.4	0.94	0.49	0.4	0.94	0.5	0.42	0.95

standard deviation of the estimators. In addition to removing bias, we examined the coverage of a parametric bootstrap interval based on the bias corrected estimate plus or minus 1.96 the bootstrap standard deviation. The results of these experiments are reported in Table 4 where we observe that the bias has effectively been removed, the standard deviations between the corrected estimators are very similar between the disparity methods and standard logistic regression estimates and we retain appropriate coverage levels.

The robustness of these estimates for binomial data from 8 trials at each X_i is examined in Figure 3. Here we observe that adding outliers at a single point generate classical robust behavior – the maximum likelihood estimate (calculating by minimizing the Kullback–Leibler divergence) is highly non-robust while Hellinger distance and negative exponential disparity are largely unchanged. When outliers are added uniformly, we observe more distortion of our estimates, particularly NED. This is both due to the large over-all amount of contamination (at all points rather than just one) and because we cannot achieve exact orthogonality between the generating and contaminating distributions. At $\alpha = 0.5$, there is, as expected, a significant change and both NED and HD exhibit increased distortion.

8. Real data

We demonstrate these methods with the analysis of the phosphorus content data in [14] in which plant phosphorus in corn is related to organic and non-organic phosphorus in the soil in which it is grown. In these data there is a distinct outlier that significantly affects least squares estimates. However robust procedures all produced estimates of approximately the same magnitude. We also conducted a bootstrap analysis, as described in Section 7 based on 100 bootstrap samples. The results of these are reported in Table 5.

9. Discussion

Conditionally specified models make up a large subset of the models most commonly used in applied statistics, including regression, generalized linear models and tabular data. In this paper, we investigate the use of disparity methods to perform parameter estimation across a range of such models. Our treatment is general in covering multivariate response and covariate variables

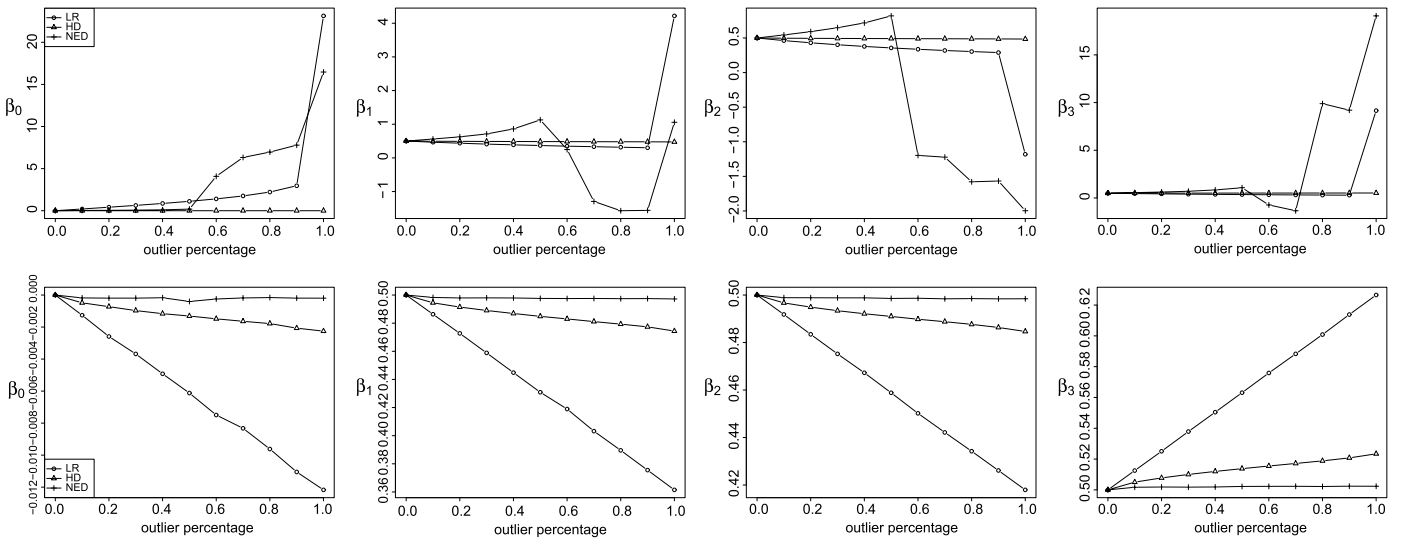

Figure 3. Mean estimates of parameters in a logistic regression as the outlier percentage increases. Top row: outliers occur uniformly over X. Bottom: outliers at a single value of X.

Table 5. Results on phosphorous data. Estimates with superscripts (β^c) incorporate a bootstrap bias correction, standard deviations are also estimated via a bootstrap

	$\log\sigma$	$\log\sigma^c$	sd	β_0	β_0^c	sd	β_1	β_1^c	sd	β_2	β_2^c	sd
LR	20.68	17.01	7.89	56.25	35.98	19.52	1.79	1.8	0.65	0.09	0.08	0.5
Hub	2.14	2.14	0.57	59.08	59.99	10.87	1.36	1.4	0.39	0.09	0.06	0.28
G–Y	2.51	2.8	0.38	66.47	63.02	8.86	1.29	1.28	0.33	−0.11	−0.05	0.23
HD.c	2.26	2.23	0.12	54.27	53.84	5.39	1.3	1.22	0.33	0.24	0.25	0.12
NED.c	2.16	2.14	0.12	53.19	53.08	6.78	1.23	1.15	0.32	0.27	0.27	0.15
HD	2.44	2.39	0.13	61.39	59.57	10.95	1.01	1.12	0.27	0.09	0.1	0.21
NED	2.4	2.4	0.16	56.78	52.45	14.08	1.03	1.15	0.3	0.19	0.25	0.26
HD.h	2.42	2.45	0.2	50.8	44.02	10.33	1.47	1.53	0.32	0.2	0.22	0.25
NED.h	2.33	2.32	0.18	52.77	49.08	10.29	1.35	1.31	0.3	0.21	0.24	0.26
HD.m	2.35	2.17	0.33	74.71	70.99	13.69	1.58	1.08	1.09	−0.42	−0.22	0.45
NED.m	2.36	2.28	0.32	60.33	57.2	11.46	1.21	1.08	0.71	0.1	0.22	0.35

and allowing for both discrete and continuous elements of each and almost any probabilistic relationship between them. We have also investigated the use of centering continuous responses by a Nadaraya–Watson estimator based on a subset of the covariates and presented a complete theory covering all ways to divide covariates into centering and conditioning variables. Along the way we have established uniform L_1 convergence results for a class of non-parametric conditional density estimates as well as the consistency and a central limit theorem for disparity-based models. These theoretical results highlight the consequences of different choices of density estimate and disparity when the model is incorrectly specified and demonstrate the limitations of centering densities within this methodology unless the same covariates are used within both the centering estimate and to condition. We have also established a bootstrap bias correction and inference methodology that has sound theoretical backing.

There are many direction for future study, starting from these methods. As is the case for disparity estimators for multivariate data, the use of conditional kernel densities results in a bias in parameter estimates that cannot be ignored in our central limit theorem, except in special cases. Empirically, our bootstrap methods reduce this bias, but more sophisticated alternatives are possible. We have not investigated using alternatives to Nadaraya–Watson estimators, but conjecture that doing so may also reduce bias. In a linear regression model, for example, the use of a local linear smoother should completely remove the bias from $\hat{m}_n$ when the model is true. More generally, centering based on a localized version of the assumed parametric model may be helpful. An alternative method of removing the bias follows the marginal approaches explored in Hooker and Vidyashankar [9]. In this approach, the non-parametric density estimate becomes dependent on a parametric transformation of the data that is chosen in such a way that at the true parameters the transformed data have independent dimensions. This would allow the use of univariate density estimates, thereby removing the asymptotic bias.

In our examples, we have employed cross-validated log likelihood to choose bandwidths and the robustness of this choice has not been investigated. We speculate that a form of weighted cross-validation may produce more robust bandwidth selection. We have also focussed solely on

kernel-based methods; little is known about the use of alternative density estimates in disparity measures, although see Wu and Hooker [17] for an exploration of non-parametric Bayesian methods combined with disparities.

Empirically, our methods perform very well in both the precision and robustness of our estimators. Within our experiments, NED generally improved upon HD methods; we speculate this is due to Hellinger distance's sensitivity to inliers (see Lindsay [11]) and hence added variability if the non-parametric estimate is sometimes multi-modal. Moreover, in distinction to alternatives, our methods provide a generic means of obtaining both robustness and efficiency across a very wide range of applicable regression models.

The need for kernel density estimates for responses and covariates at each level of the combined categorical variables limits the set of situations in which our estimates are feasible at realistic sample sizes. They are nonetheless relevant for non-trivial practical problems in data analysis; the marginal approaches in Hooker and Vidyashankar [9] also represent a means of approaching higher-dimensional covariate spaces. These results open the way for the application of minimum disparity estimates to a wide range of real-world data analysis problems.

Supplementary Material

Proofs and simulations for consistency, efficiency and robustness of conditional disparity methods (DOI: 10.3150/14-BEJ678SUPP; .pdf). We provide additional supporting simulations of the efficiency and robustness of the conditional disparity methods along with proofs of the results stated above.

Acknowledgements

Research supported in part by NSF Grants DEB-0813743, CMG-0934735 and DMS-1053252. The author thanks Anand Vidyashankar for many helpful discussions.

References

[1] Basu, A. and Lindsay, B.G. (1994). Minimum disparity estimation for continuous models: Efficiency, distributions and robustness. *Ann. Inst. Statist. Math.* **46** 683–705. MR1325990

[2] Basu, A., Sarkar, S. and Vidyashankar, A.N. (1997). Minimum negative exponential disparity estimation in parametric models. *J. Statist. Plann. Inference* **58** 349–370. MR1450021

[3] Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463. MR0448700

[4] Cheng, A.-L. and Vidyashankar, A.N. (2006). Minimum Hellinger distance estimation for randomized play the winner design. *J. Statist. Plann. Inference* **136** 1875–1910. MR2255602

[5] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. New York: Wiley. MR0780746

[6] Gervini, D. and Yohai, V.J. (2002). A class of robust and fully efficient regression estimators. *Ann. Statist.* **30** 583–616. MR1902900

[7] Hansen, B.E. (2004). Nonparametric conditional density estimation. Available at http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf.
[8] Hooker, G. (2014). Supplement to "Consistency, efficiency and robustness of conditional disparity methods." DOI:10.3150/14-BEJ678SUPP.
[9] Hooker, G. and Vidyashankar, A.N. (2014). Bayesian model robustness via disparities. *TEST* **23** 556–584. MR3252095
[10] Li, Q. and Racine, J.S. (2007). *Nonparametric Econometrics*: *Theory and Practice*. Princeton, NJ: Princeton Univ. Press. MR2283034
[11] Lindsay, B.G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.* **22** 1081–1114. MR1292557
[12] Pak, R.J. and Basu, A. (1998). Minimum disparity estimation in linear regression models: Distribution and efficiency. *Ann. Inst. Statist. Math.* **50** 503–521. MR1664536
[13] Park, C. and Basu, A. (2004). Minimum disparity estimation: Asymptotic normality and breakdown point results. *Bull. Inform. Cybernet.* **36** 19–33. MR2139489
[14] Rousseeuw, P.J. and Leroy, A.M. (2005). *Robust Regression and Outlier Detection*. New York: Wiley.
[15] Simpson, D.G. (1987). Minimum Hellinger distance estimation for the analysis of count data. *J. Amer. Statist. Assoc.* **82** 802–807. MR0909985
[16] Tamura, R.N. and Boos, D.D. (1986). Minimum Hellinger distance estimation for multivariate location and covariance. *J. Amer. Statist. Assoc.* **81** 223–229. MR0830585
[17] Wu, Y. and Hooker, G. (2013). Hellinger disance and Bayesian non-parametrics: Hierarchical models for robust and efficient Bayesian inference. Under review.

Received April 2014 and revised August 2014