Robust Inference for Convex Pairwise Difference Estimators*

Matias D. Cattaneo† Michael Jansson‡ Kenichi Nagasawa§

June 16, 2026

Abstract

This paper develops distribution theory and bootstrap-based inference methods for a broad class of convex pairwise difference estimators. These estimators minimize a kernel-weighted convex-in-parameter function over observation pairs with similar covariates, where the similarity is governed by a localization (bandwidth) parameter. While classical results establish asymptotic normality under restrictive bandwidth conditions, we show that valid Gaussian and bootstrap-based inference remains possible under substantially weaker assumptions. First, we extend the theory of small bandwidth asymptotics to convex pairwise difference estimation settings, deriving robust Gaussian approximations even when a smaller-than-standard bandwidth is used. Second, we employ a debiasing procedure based on generalized jackknifing to enable inference with larger bandwidths, while preserving convexity of the objective function. Third, we construct a novel bootstrap method that adjusts for bandwidth-induced variance distortions, yielding valid inference across a wide range of bandwidth choices. Our proposed inference method enjoys demonstrably greater robustness, while retaining the practical appeal of convex pairwise difference estimators.

Keywords: small bandwidth asymptotics, generalized jackknife, bootstrap, U-process, pairwise comparisons, robust distribution theory.

*This paper was prepared for the Econometric Theory Lecture delivered at the 2025 International Symposium on Econometric Theory and Applications (SETA), University of Macau (China), June 1–3, 2025. It was also presented at the Econometrics Journal Lecture of the 2024 $(EC)^2$ Conference (Amsterdam), and the 2025 Conference in Honor of Bo Honoré's 65th Birthday (Princeton University). We thank the participants at these conferences for their feedback. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1947805, DMS-2210561, and SES-2241575. Jansson gratefully acknowledges financial support from the National Science Foundation through grant SES-1947662 and from the Aarhus Center for Econometrics (ACE) funded by the Danish National Research Foundation grant number DNRF186. Nagasawa gratefully acknowledges financial support from the British Academy through grant SRG24\241614.

†Department of Operations Research and Financial Engineering, Princeton University.

‡Department of Economics, UC Berkeley and ACE.

§Department of Economics, University of Warwick.

1 Introduction

Suppose $\mathbf{z}_1, \ldots, \mathbf{z}_n$ is a random sample from the distribution of a random vector $\mathbf{z}$. This paper studies the large-sample properties of the following *convex* pairwise difference estimator:

$$\widehat{\boldsymbol{\theta}}_n \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) K_{h_n}(\mathbf{w}_i - \mathbf{w}_j), \qquad K_h(\mathbf{u}) = \frac{1}{h^d} K\left(\frac{\mathbf{u}}{h}\right), \tag{1.1}$$

where $\Theta \subseteq \mathbb{R}^k$ is a parameter space, $\sum_{i<j}$ denotes $\sum_{j=2}^{n}\sum_{i=1}^{j-1}$, $(\mathbf{z}, \bar{\mathbf{z}}) \mapsto m(\mathbf{z}, \bar{\mathbf{z}}; \boldsymbol{\theta})$ is a permutation symmetric function, K is a symmetric, non-negative kernel, h_n is a positive bandwidth (or localization) parameter sequence, $\mathbf{w}$ is a continuously distributed d-dimensional subvector of $\mathbf{z}$, and where $\boldsymbol{\theta} \mapsto m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ is a *convex* function. Pairwise difference estimation, which relies on local comparisons between observation pairs, has been used to address heterogeneity in nonlinear models. See Powell (1994), Honoré and Powell (2005), and Aradillas-Lopez, Honoré, and Powell (2007) for overviews, and Section 2 for three motivating examples.

In contrast to classical extremum estimators, $\widehat{\boldsymbol{\theta}}_n$ is a local M-estimator that employs observation pairs (i, j) for which $\mathbf{w}_i$ and $\mathbf{w}_j$ are similar. The bandwidth h_n governs the degree of similarity: When $h_n \to 0$ (as $n \to \infty$), the estimator increasingly focuses on nearly-identical-in-$\mathbf{w}$ pairs. In turn, focusing on such pairs is natural in settings where identification can be based on the condition $\mathbf{w}_i \approx \mathbf{w}_j$ (combined with smoothness assumptions). The localization introduces a familiar trade-off for estimation and inference: A smaller h_n reduces bias from dissimilarity between $\mathbf{w}_i$ and $\mathbf{w}_j$, but increases variance due to fewer usable pairs. As a consequence, the large-sample behavior of $\widehat{\boldsymbol{\theta}}_n$ depends critically on a delicate bias-variance trade-off determined by h_n. This paper develops novel inference methods for convex pairwise difference estimators that are demonstrably more robust to bandwidth choice than existing methods.

Under regularity conditions and assuming that

$$nh_n^d \to \infty \qquad \text{and} \qquad nh_n^4 \to 0, \tag{1.2}$$

the pairwise difference estimator $\widehat{\boldsymbol{\theta}}_n$ is known to be asymptotically linear:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\psi}_0(\mathbf{z}_i) + o_{\mathbb{P}}(1) \rightsquigarrow \mathsf{N}(\mathbf{0}, \mathbb{E}[\boldsymbol{\psi}_0(\mathbf{z})\boldsymbol{\psi}_0(\mathbf{z})']), \tag{1.3}$$

where $\boldsymbol{\theta}_0$ is the estimand, $\boldsymbol{\psi}_0(\cdot)$ is an influence function (whose exact form is given below), and where $\rightsquigarrow$ denotes weak convergence. Moreover, the nonparametric bootstrap approximation to the distribution in (1.3) is consistent in the sense that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n^* - \widehat{\boldsymbol{\theta}}_n) \rightsquigarrow_{\mathbb{P}} \mathsf{N}(\mathbf{0}, \mathbb{E}[\boldsymbol{\psi}_0(\mathbf{z})\boldsymbol{\psi}_0(\mathbf{z})']), \tag{1.4}$$

where $\widehat{\boldsymbol{\theta}}_n^*$ is the bootstrap analogue of $\widehat{\boldsymbol{\theta}}_n$ and where $\rightsquigarrow_{\mathbb{P}}$ denotes weak convergence in probability.

The main results of this paper generalize (1.2)-(1.4) by combining three ideas:

1. *Small Bandwidth Asymptotics.* Utilizing the framework introduced by Cattaneo, Crump, and Jansson (2014a), we obtain a Gaussian distributional approximation for the pairwise difference estimator without imposing the condition $nh_n^d \to \infty$, thereby allowing for higher levels of localization. This generalized distributional approximation shows in particular that, while the localization restriction $nh_n^d \to \infty$ is necessary for establishing asymptotic linearity, a valid Gaussian approximation can be obtained under the substantially weaker condition $n^2 h_n^d \to \infty$, albeit with a convergence rate (and approximate variance) that depends explicitly on the level of localization used.

2. *Debiasing.* Following Honoré and Powell (2005), we debias the pairwise difference estimator using the method of *generalized jackknifing* introduced by Schucany and Sommers (1977). Doing so allows for (larger) bandwidths that violate the bias condition $nh_n^4 \to 0$. The same goal could be achieved by replacing the (second-order) kernel K with a higher-order kernel, but a higher-order kernel annihilates the convexity of the objective function because higher-order kernels take negative values. In contrast, generalized jackknifing retains the convexity of objective functions, which in turn is attractive for both theoretical (weaker regularity conditions) and practical (faster computation) reasons.

3. *Bootstrapping.* Building on insights from Cattaneo, Crump, and Jansson (2014b), we develop a valid bootstrap-based distributional approximation for the debiased pairwise difference estimator by rescaling the localization parameter. The nonparametric bootstrap distributional approximation exhibits a mismatch in its asymptotic variance under small bandwidth asymptotics. The mismatch is characterized by a known multiplicative factor involving the localization parameter h_n. As a result, bootstrapping the (debiased) pairwise difference estimator with a different localization parameter (namely, $3^{1/d} h_n$ rather than h_n) leads to a valid bootstrap-based inference procedure also under small bandwidth asymptotics.

In combination, these three ideas enable us to offer a novel resampling-based inference method for (convex) pairwise difference estimators that is demonstrably more robust to the choice of the localization parameter h_n than methods based on (1.2)-(1.4).

Our theoretical work is carefully developed to retain and leverage convexity of the objective function defining the pairwise difference estimator. This feature not only allows for fast implementation of the estimator and resampling-based methods, but also enables us to proceed under relatively weak conditions when obtaining theoretical results. When developing our theoretical results, we rely heavily on the foundational work of Hjort and Pollard (1993) and Pollard (1991), which we apply to the case of U-processes.

This paper is connected to several strands of the literature. Contributions to the pairwise difference estimation literature include Ahn, Ichimura, Powell, and Ruud (2018), Ahn and Powell (1993), Aradillas-Lopez (2012), Blundell and Powell (2004), Hong and Shum (2010), Honoré (1992), Honoré, Kyriazidou, and Udry (1997), Honoré and Powell (1994), Jochmans (2013), Kyriazidou (1997), and

Powell (2001). The theoretical and practical features of small bandwidth asymptotics, and their connection with resampling methods for inference, are discussed in Cattaneo, Crump, and Jansson (2010, 2014a,b), Cattaneo, Farrell, Jansson, and Masini (2025), Cattaneo and Jansson (2018, 2022), Cattaneo, Jansson, and Newey (2018), Matsushita and Otsu (2021), and references therein. The generalized jackknife has been successfully used for debiasing in density weighted average derivative estimation (Powell, Stock, and Stoker, 1989), asymptotically linear pairwise difference estimation (Honoré and Powell, 2005), nonlinear semiparametric estimation (Cattaneo, Crump, and Jansson, 2013), monotone estimation (Cattaneo, Jansson, and Nagasawa, 2024), and random forest estimation (Cattaneo, Klusowski, and Underwood, 2026), among other settings. Shao and Tu (2012) give a textbook introduction to jackknifing, bootstrapping, and other resampling methods.

The rest of the paper proceeds as follows. Section 2 introduces the three examples that are used throughout the paper to motivate our work and to illustrate the verification of the high-level assumptions imposed. Section 3 presents our main results. The proofs of these results are given in Section 4. Section 5 revisits the three motivating examples and gives primitive conditions under which these examples are covered by our general theory. Simulation evidence is reported in Section 6. Section 7 gives final remarks.

2 Motivating Examples

We use three examples to motivate and illustrate our work. The first example involves an estimator that can be written in closed form (because it has a quadratic-in-$\boldsymbol{\theta}$ function $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$), while the other two examples do not. The second example has a smooth-in-$\boldsymbol{\theta}$ function $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$, while the third example does not. All three examples have convex-in-$\boldsymbol{\theta}$ functions $m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ and employ the following notation: $\mathbf{z}_i = (y_i, \mathbf{x}_i', \mathbf{w}_i')'$, with y_i a scalar outcome variable, $\mathbf{x}_i$ a k-dimensional covariate, and $\mathbf{w}_i$ a d-dimensional covariate. For more details on the examples, see Powell (1994), Honoré and Powell (2005), and Aradillas-Lopez et al. (2007).

2.1 Partially Linear Regression Model

The partially linear regression model is of the form

$$y_i = \mathbf{x}_i'\boldsymbol{\theta}_0 + \gamma_0(\mathbf{w}_i) + \varepsilon_i,$$

where $\boldsymbol{\theta}_0$ is the parameter of interest, $\gamma_0(\cdot)$ is an unknown function, and where $\mathbb{E}[\varepsilon_i|\mathbf{x}_i, \mathbf{w}_i] = 0$. Defining $\dot{y}_{i,j} = y_i - y_j$ and $\dot{\mathbf{x}}_{i,j} = \mathbf{x}_i - \mathbf{x}_j$, a pairwise difference estimator of $\boldsymbol{\theta}_0$ can be based on

$$m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = m_{\texttt{PLR}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = \frac{1}{2}(\dot{y}_{i,j} - \dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta})^2.$$

Setting $\Theta = \mathbb{R}^k$, the minimization problem in (1.1) admits a closed form solution (provided that a non-negative kernel function is used), namely

$$\widehat{\boldsymbol{\theta}}_n = \left(\sum_{i<j} \dot{\mathbf{x}}_{i,j} \dot{\mathbf{x}}_{i,j}' K_{h_n}(\mathbf{w}_i - \mathbf{w}_j) \right)^{-1} \sum_{i<j} \dot{\mathbf{x}}_{i,j} \dot{y}_{i,j} K_{h_n}(\mathbf{w}_i - \mathbf{w}_j).$$

2.2 Partially Linear Logit Model

The partially linear logit model studied here is of the form

$$y_i = \mathbb{1}\{\mathbf{x}_i'\boldsymbol{\theta}_0 + \gamma_0(\mathbf{w}_i) + \varepsilon_i \geq 0\},$$

where $\boldsymbol{\theta}_0$ is the parameter of interest, $\gamma_0(\cdot)$ is an unknown function, and where

$$\mathbb{P}\left[\varepsilon_i \leq u | \mathbf{x}_i, \mathbf{w}_i\right] = \Lambda(u), \qquad \Lambda(u) = \frac{\exp(u)}{1+\exp(u)}.$$

The parameter $\boldsymbol{\theta}_0$ can be estimated using a pairwise difference estimator with $\Theta = \mathbb{R}^k$ and

$$m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = m_{\texttt{PLL}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = -\mathbb{1}\{\dot{y}_{i,j} \neq 0\} \left(y_i \ln \Lambda(\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta}) + y_j \ln \Lambda(-\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta}) \right).$$

The minimization problem in (1.1) does not admit a closed form solution, but it is convex (provided that a non-negative kernel function is used) because $u \mapsto -\ln \Lambda(u)$ is.

2.3 Partially Linear Tobit Model

The partially linear censored regression model studied here is of the form

$$y_i = \max\{\mathbf{x}_i'\boldsymbol{\theta}_0 + \gamma_0(\mathbf{w}_i) + \varepsilon_i, 0\},$$

where $\boldsymbol{\theta}_0$ is the parameter of interest, $\gamma_0(\cdot)$ is an unknown function, $\mathbf{x}_i \perp\!\!\!\perp \varepsilon_i | \mathbf{w}_i$, and the conditional distribution of ε_i given $\mathbf{w}_i$ admits a Lebesgue density. A pairwise difference estimator of $\boldsymbol{\theta}_0$ can be obtained by setting $\Theta = \mathbb{R}^k$ and employing

$$m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = m_{\texttt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = \tilde{m}_{\texttt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) - \tilde{m}_{\texttt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \mathbf{0}),$$

where

$$\tilde{m}_{\texttt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = \begin{cases} |y_i| - \left(\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} + y_j\right) \operatorname{sgn}(y_i) & \text{if } \dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} \leq -y_j \\ \left|\dot{y}_{i,j} - \dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta}\right| & \text{if } -y_j < \dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} < y_i \\ |y_j| + \left(\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} - y_i\right) \operatorname{sgn}(y_j) & \text{if } y_i \leq \dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} \end{cases}.$$

Because $\tilde{m}_{\texttt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \mathbf{0})$ does not depend on $\boldsymbol{\theta}$, its presence in $m_{\texttt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ does not affect the minimization problem defining the estimator. Nevertheless, it is theoretically attractive to work

with $m_{\texttt{PLT}}$ rather than $\tilde{m}_{\texttt{PLT}}$, as doing so allows for weaker regularity conditions for the existence of the expectation of the objective function.

For future reference, we note that $m_{\texttt{PLT}}$ admits the alternative representation

$$m_{\texttt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = \begin{cases} \left| \dot{y}_{i,j} - \dot{\mathbf{x}}_{i,j}' \boldsymbol{\theta} \right| - |\dot{y}_{i,j}| & \text{if } y_i > 0, y_j > 0 \\ \max\{y_i - \dot{\mathbf{x}}_{i,j}' \boldsymbol{\theta}, 0\} - y_i & \text{if } y_i > 0, y_j = 0 \\ \max\{y_j + \dot{\mathbf{x}}_{i,j}' \boldsymbol{\theta}, 0\} - y_j & \text{if } y_i = 0, y_j > 0 \\ 0 & \text{if } y_i = 0, y_j = 0 \end{cases}.$$

The function $\boldsymbol{\theta} \mapsto m_{\texttt{PLT}}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})$ is convex and therefore so is the minimization problem in (1.1) (provided that a non-negative kernel function is used).

3 Distributional Approximation and Bootstrap Inference

As is standard in the literature, we generalize (1.1) slightly and define our estimator $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(h_n)$ to be any approximate minimizer of $\widehat{M}_n(\boldsymbol{\theta}; h_n)$, where

$$\widehat{M}_n(\boldsymbol{\theta}; h) = \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) K_h(\mathbf{w}_i - \mathbf{w}_j).$$

To be specific, we require

$$\widehat{M}_n(\widehat{\boldsymbol{\theta}}_n(h); h) \leq \inf_{\boldsymbol{\theta} \in \Theta} \widehat{M}_n(\boldsymbol{\theta}; h) + o_{\mathbb{P}}(n^{-1}).$$

The objective function $\widehat{M}_n$ is a sample counterpart of the function M given by

$$M(\boldsymbol{\theta}; h) = \mathbb{E}\left[\widehat{M}_n(\boldsymbol{\theta}; h)\right] = \mathbb{E}\left[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) K_h(\mathbf{w}_1 - \mathbf{w}_2)\right].$$

Under regularity conditions, this function approximates, as $h \downarrow 0$, a function M_0, which (does not depend on K and) admits a unique minimizer, namely the parameter of interest $\boldsymbol{\theta}_0$.

For the purposes of analyzing $\widehat{\boldsymbol{\theta}}_n$ it is convenient to define $\boldsymbol{\theta}_n = \boldsymbol{\theta}(h_n)$, where

$$\boldsymbol{\theta}(h) \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} M(\boldsymbol{\theta}; h)$$

is interpretable as a (fixed-h) "pseudo" parameter. With the help of $\boldsymbol{\theta}_n$ we can decompose the estimation error $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0$ into a (non-stochastic) "bias" component $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0$ and a "noise" component $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n$. Each component can be analyzed separately and in both cases the analysis will leverage convexity.

3.1 Regularity Conditions

The following assumption guarantees, among other things, that $\boldsymbol{\theta}_n$ is well defined for large n and that the bias component $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0$ vanishes asymptotically; for details, see Lemma 1 of Section 4.1.

Assumption 1. (i) The kernel function K is a symmetric, bounded probability density.

(ii) $\Theta \subseteq \mathbb{R}^k$ is convex, $(\mathbf{z}, \bar{\mathbf{z}}) \mapsto m(\mathbf{z}, \bar{\mathbf{z}}; \boldsymbol{\theta})$ is permutation symmetric, and $\boldsymbol{\theta} \mapsto m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})$ is convex with probability one.

(iii) The distribution of $\mathbf{w}$ admits a Lebesgue density $f_{\mathbf{w}}$, which is bounded and continuous on its support $\mathcal{W}$.

(iv) For each $\boldsymbol{\theta} \in \Theta$,

$$\mathbb{E}\left[\sup_{\mathbf{w}_2 \in \mathcal{W}} |\mathbb{E}[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})|\mathbf{w}_1, \mathbf{w}_2]| f_{\mathbf{w}}(\mathbf{w}_2)\right] < \infty$$

and (with probability one)

$$\lim_{\mathbf{u} \to \mathbf{0}} \mathbb{E}[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})|\mathbf{w}_1 = \mathbf{w}, \mathbf{w}_2 = \mathbf{w} + \mathbf{u}] = \mathbb{E}[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})|\mathbf{w}_1 = \mathbf{w}, \mathbf{w}_2 = \mathbf{w}].$$

(v) On Θ, the function M_0 given by

$$M_0(\boldsymbol{\theta}) = \int_{\mathcal{W}} \mathbb{E}[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})|\mathbf{w}_1 = \mathbf{w}, \mathbf{w}_2 = \mathbf{w}] f_{\mathbf{w}}(\mathbf{w})^2 d\mathbf{w}$$

is uniquely minimized at an interior point $\boldsymbol{\theta}_0$.

The next assumption enables us to analyze the asymptotic properties of the noise component $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n$. To accommodate examples (such as the partially linear Tobit model) where $\boldsymbol{\theta} \mapsto m(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ is not fully differentiable, we assume the existence of derivative-like functions $\mathbf{s}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) \in \mathbb{R}^k$ and $\mathbf{H}(\mathbf{w}_i, \mathbf{w}_j; \boldsymbol{\theta}, \mathbf{t}) \in \mathbb{R}^{k \times k}$ such that, for any direction $\mathbf{t} \in \mathbb{R}^k$, the (remainder) terms

$$r_{\mathbf{t}}(\boldsymbol{\theta}, \tau) = \frac{m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta} + \mathbf{t}\tau) - m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})}{\tau} - \mathbf{s}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta})'\mathbf{t}$$

and

$$R_{\mathbf{t}}(\boldsymbol{\theta}, \tau) = \frac{\mathbb{E}[r_{\mathbf{t}}(\boldsymbol{\theta}, \tau)|\mathbf{w}_1, \mathbf{w}_2]}{\tau} - \frac{1}{2}\mathbf{t}'\mathbf{H}(\mathbf{w}_1, \mathbf{w}_2; \boldsymbol{\theta}, \mathbf{t})\mathbf{t}$$

are suitably small for $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$, $\tau > 0$ near zero, and $\mathbf{w}_1 \approx \mathbf{w}_2$. As further discussed below, functions $\mathbf{s}$ and $\mathbf{H}$ satisfying the following assumption exist (and are relatively easy to find) in each of our motivating examples.

Assumption 2. (i) For each $\mathbf{t} \in \mathbb{R}^k$, there is some $\delta > 0$ such that

$$\mathbb{E}\left[\sup_{\tau\in(0,\delta),\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta,\mathbf{w}_2\in\mathcal{W}} |\mathbb{E}[r_{\mathbf{t}}(\boldsymbol{\theta},\tau)|\mathbf{z}_1,\mathbf{w}_2]|\, f_{\mathbf{w}}(\mathbf{w}_2)^2\right] < \infty,$$

$$\mathbb{E}\left[\sup_{\tau\in(0,\delta),\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta,\mathbf{w}_2\in\mathcal{W}} \mathbb{E}[r_{\mathbf{t}}(\boldsymbol{\theta},\tau)^2|\mathbf{w}_1,\mathbf{w}_2] f_{\mathbf{w}}(\mathbf{w}_2)\right] < \infty,$$

$$\mathbb{E}\left[\sup_{\tau\in(0,\delta),\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta,\mathbf{w}_2\in\mathcal{W}} |R_{\mathbf{t}}(\boldsymbol{\theta},\tau)|\, f_{\mathbf{w}}(\mathbf{w}_2)\right] < \infty,$$

and (with probability one)

$$\lim_{\tau\downarrow 0,(\boldsymbol{\theta},\mathbf{u})\to(\boldsymbol{\theta}_0,\mathbf{0})} \mathbb{E}[r_{\mathbf{t}}(\boldsymbol{\theta},\tau)|\mathbf{z}_1=\mathbf{z},\mathbf{w}_2=\mathbf{w}+\mathbf{u}] = 0,$$

$$\lim_{\tau\downarrow 0,(\boldsymbol{\theta},\mathbf{u})\to(\boldsymbol{\theta}_0,\mathbf{0})} \mathbb{E}[r_{\mathbf{t}}(\boldsymbol{\theta},\tau)^2|\mathbf{w}_1=\mathbf{w},\mathbf{w}_2=\mathbf{w}+\mathbf{u}] = 0,$$

$$\lim_{\tau\downarrow 0,(\boldsymbol{\theta},\mathbf{u})\to(\boldsymbol{\theta}_0,\mathbf{0})} \mathbb{E}[R_{\mathbf{t}}(\boldsymbol{\theta},\tau)|\mathbf{w}_1=\mathbf{w},\mathbf{w}_2=\mathbf{w}+\mathbf{u}] = 0.$$

(ii) There is some $\delta > 0$ and some function b with

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta} \|\mathbf{s}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta})\| \le b(\mathbf{z}_1)b(\mathbf{z}_2),$$

such that

$$\mathbb{E}[b(\mathbf{z})^4] + \sup_{\mathbf{w}\in\mathcal{W}} \mathbb{E}[b(\mathbf{z})^4|\mathbf{w}] f_{\mathbf{w}}(\mathbf{w}) < \infty$$

and

$$\mathbb{E}\left[\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|<\delta,\mathbf{w}_2\in\mathcal{W}} \|\mathbf{H}(\mathbf{w}_1,\mathbf{w}_2;\boldsymbol{\theta},\mathbf{t})\| f_{\mathbf{w}}(\mathbf{w}_2)\right] < \infty \qquad \text{for each } \mathbf{t}\in\mathbb{R}^k.$$

(iii) There exist functions $\mathbf{G}_0, \boldsymbol{\xi}_0$, and $\boldsymbol{\Xi}_0$ such that, for each $\mathbf{t} \in \mathbb{R}^k$ (and with probability one),

$$\lim_{(\boldsymbol{\theta},\mathbf{u})\to(\boldsymbol{\theta}_0,\mathbf{0})} \mathbf{H}(\mathbf{w},\mathbf{w}+\mathbf{u};\boldsymbol{\theta},\mathbf{t}) f_{\mathbf{w}}(\mathbf{w}) = \mathbf{G}_0(\mathbf{w}),$$

$$\lim_{(\boldsymbol{\theta},\mathbf{u})\to(\boldsymbol{\theta}_0,\mathbf{0})} -2\mathbb{E}[\mathbf{s}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta})|\mathbf{z}_1=\mathbf{z},\mathbf{w}_2=\mathbf{w}+\mathbf{u}] f_{\mathbf{w}}(\mathbf{w}) = \boldsymbol{\xi}_0(\mathbf{z}),$$

and

$$\lim_{(\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\mathbf{u})\to(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0,\mathbf{0})} \mathbb{E}[\mathbf{s}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta})\mathbf{s}(\mathbf{z}_1,\mathbf{z}_2;\bar{\boldsymbol{\theta}})'|\mathbf{w}_1=\mathbf{w},\mathbf{w}_2=\mathbf{w}+\mathbf{u}] f_{\mathbf{w}}(\mathbf{w}) = \boldsymbol{\Xi}_0(\mathbf{w}).$$

(iv) $\boldsymbol{\Gamma}_0 = \mathbb{E}[\mathbf{G}_0(\mathbf{w})], \boldsymbol{\Sigma}_0 = \mathbb{E}[\boldsymbol{\xi}_0(\mathbf{z})\boldsymbol{\xi}_0(\mathbf{z})']$, and $\mathbb{E}[\boldsymbol{\Xi}_0(\mathbf{w})]$ are positive definite.

3.2 Small Bandwidth Asymptotics

Defining

$$\mathbf{V}_n = \mathbf{V}_n(h_n) = \boldsymbol{\Gamma}_0^{-1}\left[n^{-1}\boldsymbol{\Sigma}_0 + \binom{n}{2}^{-1} h_n^{-d}\boldsymbol{\Delta}_0(K)\right]\boldsymbol{\Gamma}_0^{-1}, \qquad \boldsymbol{\Delta}_0(K) = \mathbb{E}[\boldsymbol{\Xi}_0(\mathbf{w})]\int_{\mathbb{R}^d} K^2(\mathbf{u})d\mathbf{u},$$

and letting Φ_k denote the distribution function of a k-dimensional standard Gaussian random vector, we have the following result.

Theorem 1. *Suppose Assumptions 1 and 2 hold. If $n^2h_n^d \to \infty$ and if $h_n \to 0$, then*

$$\sup_{\mathbf{t}\in\mathbb{R}^k}\left|\mathbb{P}\left[\mathbf{V}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n) \le \mathbf{t}\right] - \Phi_k(\mathbf{t})\right| \to 0.$$

Under the assumptions of Theorem 1, the convergence rate of $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n$ is given by the magnitude of $\mathbf{V}_n^{-1/2}$, namely

$$\rho_n = \sqrt{\min\left(n, \binom{n}{2}h_n^d\right)}.$$

Provided that the bias is "small" in the sense that $\rho_n\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| \to 0$, Theorem 1 therefore encompasses the following three distinct large-sample regimes:

- *Asymptotic Linearity*: If $nh_n^d \to \infty$, then $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ satisfies (1.3) with $\boldsymbol{\psi}_0 = \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\xi}_0$. In particular, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in law to a centered Gaussian distribution with variance

$$\lim_{n\to\infty} n\mathbf{V}_n(h_n) = \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Gamma}_0^{-1}.$$

- *Root-n Consistency without Asymptotic Linearity*: If $nh_n^d \to 2c \in (0,\infty)$, then $\widehat{\boldsymbol{\theta}}_n$ is not asymptotically linear, but it is $\sqrt{n}$-consistent, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converging in law to a centered Gaussian distribution with variance

$$\lim_{n\to\infty} n\mathbf{V}_n(h_n) = \boldsymbol{\Gamma}_0^{-1}\left[\boldsymbol{\Sigma}_0 + \frac{1}{c}\boldsymbol{\Delta}_0(K)\right]\boldsymbol{\Gamma}_0^{-1}.$$

- *Slower than Root-n Consistency*: If $nh_n^d \to 0$ (but $n^2h_n^d \to \infty$), then $\widehat{\boldsymbol{\theta}}_n$ is neither asymptotically linear nor $\sqrt{n}$-consistent, but $\sqrt{n^2h_n^d/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in law to a centered Gaussian distribution with variance

$$\lim_{n\to\infty}\binom{n}{2}h_n^d\mathbf{V}_n(h_n) = \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Delta}_0(K)\boldsymbol{\Gamma}_0^{-1}.$$

The small bandwidth component (i.e., the term involving $\boldsymbol{\Delta}_0(K)$) in $\mathbf{V}_n$ captures the additional estimation uncertainty generated from increasing the localization of the observation pairs. Incorporating this component in the approximate variance is key to enabling us to replace the condition

$nh_n^d \to \infty$ by the weaker condition $n^2 h_n^d \to \infty$ when obtaining a Gaussian approximation. As demonstrated by Cattaneo et al. (2025) in a related context, incorporating the small bandwidth component can furthermore lead to a more accurate distributional approximation (in a higher-order asymptotic sense) even under asymptotic linearity.

3.3 Debiasing

In Theorem 1, we centered the estimator $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(h_n)$ at $\boldsymbol{\theta}_n = \boldsymbol{\theta}(h_n)$ to circumvent bias issues. This section focuses on the bias term $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0$ and introduces an automatic debiasing approach under the assumption that $\boldsymbol{\theta}_n - \boldsymbol{\theta}_0$ can be expanded in even powers of h_n. To be specific, we follow Honoré and Powell (2005, Section 3.3) and discuss debiasing under the following high-level condition.

Assumption 3. For some even $\mathtt{S} \geq 0$, $\boldsymbol{\theta}(\cdot)$ admits $\mathbf{b}_{2l} \in \mathbb{R}^k$ (for $l = 1, \ldots, \mathtt{S}/2$) such that

$$\boldsymbol{\theta}(h) - \boldsymbol{\theta}_0 = \sum_{l=1}^{\mathtt{S}/2} \mathbf{b}_{2l} h^{2l} + o(h^{\mathtt{S}}) \qquad \text{as } h \downarrow 0.$$

The ease with which Assumption 3 can be verified depends on the magnitude of $\mathtt{S}$. For instance, Assumption 1 implies that Assumption 3 holds with $\mathtt{S} = 0$. Under additional smoothness conditions and using symmetry of K, the following result gives conditions under which Assumption 3 holds with $\mathtt{S} = 2$. When stating the result, we employ the following standard multi-index notation: for $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)' \in \mathbb{Z}_+^d$, $\bar{\mathbf{w}} = (\bar{w}_1, \ldots, \bar{w}_d)' \in \mathbb{R}^d$, and a sufficiently smooth-in-$\bar{\mathbf{w}}$ function $f(\mathbf{w}, \bar{\mathbf{w}})$,

$$\partial_{\bar{\mathbf{w}}}^{\boldsymbol{\nu}} f(\mathbf{w}, \bar{\mathbf{w}}) = \frac{\partial^{|\boldsymbol{\nu}|}}{\partial \bar{w}_1^{\nu_1} \cdots \partial \bar{w}_d^{\nu_d}} f(\mathbf{w}, \bar{\mathbf{w}}), \qquad |\boldsymbol{\nu}| = \sum_{j=1}^{d} \nu_j.$$

Proposition 1. *Suppose Assumptions 1-2 hold and that*

(i) $\int_{\mathbb{R}^d} \|\mathbf{u}\|^2 K(\mathbf{u}) d\mathbf{u} < \infty$, *and*

(ii) with probability one, $\bar{\mathbf{w}} \mapsto \boldsymbol{\varphi}(\mathbf{w}, \bar{\mathbf{w}}) = \mathbb{E}[\mathbf{s}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}_0) | \mathbf{w}_1 = \mathbf{w}, \mathbf{w}_2 = \bar{\mathbf{w}}] f_{\mathbf{w}}(\bar{\mathbf{w}})$ *is twice continuously differentiable with* $\mathbb{E}[\sup_{\bar{\mathbf{w}} \in \mathcal{W}} \|\partial_{\bar{\mathbf{w}}}^{\boldsymbol{\nu}} \boldsymbol{\varphi}(\mathbf{w}, \bar{\mathbf{w}})\|] < \infty$ *for all* $\boldsymbol{\nu} \in \mathbb{Z}_+^d$ *with* $|\boldsymbol{\nu}| \leq 2$.

Then $\boldsymbol{\theta}(\cdot)$ *admits a* $\mathbf{b}_2 \in \mathbb{R}^k$ *such that*

$$\boldsymbol{\theta}(h) - \boldsymbol{\theta}_0 = \mathbf{b}_2 h^2 + o(h^2) \qquad \text{as } h \downarrow 0.$$

The proof of Proposition 1 leverages convexity and may therefore be of independent interest. The convexity argument in question can furthermore be adapted to form the basis of a verification by induction of Assumption 3 with $\mathtt{S} > 2$. Details are provided in Section 4.4, which describes the induction step for general $\mathtt{S}$ and states explicit (smoothness) conditions under which Assumption 3 holds with $\mathtt{S} = 4$.

To describe the debiasing procedure based on generalized jackknifing, we maintain Assumption 3, define $c_0 = 1$, choose a non-negative integer L, and let $\mathbf{c} = (c_0, \ldots, c_L)'$ be a vector of (distinct) positive constants such that the following vector is well defined:

$$\begin{pmatrix} \lambda_0(\mathbf{c}) \\ \lambda_1(\mathbf{c}) \\ \vdots \\ \lambda_L(\mathbf{c}) \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & c_1^2 & \cdots & c_L^2 \\ \vdots & & \ddots & \\ 1 & c_1^{2L} & \cdots & c_L^{2L} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

The debiased estimator is

$$\widetilde{\boldsymbol{\theta}}_n = \widetilde{\boldsymbol{\theta}}_n(\mathbf{c}, h_n) = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) \widehat{\boldsymbol{\theta}}_n(h_{n,l}), \qquad h_{n,l} = c_l h_n,$$

the construction of which involves solving $L + 1$ convex optimization problems. As defined, the debiased estimator is a generalization of the original pairwise difference estimator because if $L = 0$, then $\mathbf{c} = 1 = \lambda_0$ and therefore $\widetilde{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n$.

The next theorem generalizes Theorem 1 by establishing the small bandwidth Gaussian approximation for $\widetilde{\boldsymbol{\theta}}_n$. To state the theorem, let

$$\bar{\boldsymbol{\theta}}_n = \bar{\boldsymbol{\theta}}_n(\mathbf{c}, h_n) = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) \boldsymbol{\theta}(h_{n,l})$$

and

$$\bar{\mathbf{V}}_n = \bar{\mathbf{V}}_n(\mathbf{c}, h_n) = \boldsymbol{\Gamma}_0^{-1} \left[n^{-1} \boldsymbol{\Sigma}_0 + \binom{n}{2}^{-1} h_n^{-d} \boldsymbol{\Delta}_0(\bar{K}) \right] \boldsymbol{\Gamma}_0^{-1}, \qquad \bar{K}(\mathbf{u}) = \bar{K}(\mathbf{u}; \mathbf{c}) = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) K_{c_l}(\mathbf{u}).$$

As they should, the expressions have the feature that if $L = 0$, then $\bar{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_n$ and $\bar{\mathbf{V}}_n = \mathbf{V}_n$. Another noteworthy feature of the expressions is that debiasing via generalized jackknifing affects the variance $\bar{\mathbf{V}}_n$ only through the kernel shape entering its small bandwidth component.

Theorem 2. *Suppose Assumptions 1 and 2 hold. If $n^2 h_n^d \to \infty$ and if $h_n \to 0$, then*

$$\sup_{\mathbf{t} \in \mathbb{R}^k} \left| \mathbb{P}\left[\bar{\mathbf{V}}_n^{-1/2} (\widetilde{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}_n) \leq \mathbf{t} \right] - \Phi_k(\mathbf{t}) \right| \to 0.$$

As a consequence, if also Assumption 3 holds with $\mathtt{S} \geq 2L + 2$, and if $n h_n^{4(L+1)} \to 0$, then

$$\sup_{\mathbf{t} \in \mathbb{R}^k} \left| \mathbb{P}\left[\bar{\mathbf{V}}_n^{-1/2} (\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \leq \mathbf{t} \right] - \Phi_k(\mathbf{t}) \right| \to 0.$$

The magnitude of $\bar{\mathbf{V}}_n^{-1/2}$ is the same as that of $\mathbf{V}_n^{-1/2}$. With obvious modifications, the discussion of $\widehat{\boldsymbol{\theta}}_n$ following Theorem 1 therefore applies to $\widetilde{\boldsymbol{\theta}}_n$, the only noteworthy difference being that (by design) the relevant "small bias" condition is different (and typically milder) in the case of $\widetilde{\boldsymbol{\theta}}_n$.

When $L = 0$ (i.e., one uses the original pairwise difference estimator), the second part of Theorem 2 imposes $nh_n^4 \to 0$, which coincides with the standard small bias condition in (1.2), while the first part of the theorem still accommodates the small bandwidth distributional approximation.

It is worth noting that for $L \geq 1$, the equivalent kernel $\bar{K}$ is of higher order, even though the debiased estimator $\widetilde{\boldsymbol{\theta}}_n$ only employs estimators constructed using second-order kernels, thereby retaining the desired convexity for implementation. To be specific, if $\int_{\mathbb{R}^d} \|\mathbf{u}\|^{2L+2} K(\mathbf{u}) d\mathbf{u} < \infty$, then $\bar{K}$ is of order $2L+2$ because

$$\int_{\mathbb{R}^d} \bar{K}(\mathbf{u}) d\mathbf{u} = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) \int_{\mathbb{R}^d} K_{c_l}(\mathbf{u}) d\mathbf{u} = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) = 1$$

and, for $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)' \in \mathbb{Z}_+^d$ with $0 < |\boldsymbol{\nu}| \leq 2L+1$,

$$\int_{\mathbb{R}^d} \mathbf{u}^{\boldsymbol{\nu}} \bar{K}(\mathbf{u}) d\mathbf{u} = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) \int_{\mathbb{R}^d} \mathbf{u}^{\boldsymbol{\nu}} K_{c_l}(\mathbf{u}) d\mathbf{u} = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) c_l^{|\boldsymbol{\nu}|} \int_{\mathbb{R}^d} \mathbf{u}^{\boldsymbol{\nu}} K(\mathbf{u}) d\mathbf{u} = 0,$$

where the last equality uses the defining property of $\{\lambda_l(\mathbf{c})\}$ and symmetry of K, and where $\mathbf{u}^{\boldsymbol{\nu}}$ denotes $\prod_{j=1}^d u_j^{\nu_j}$ for $\mathbf{u} = (u_1, \ldots, u_d)' \in \mathbb{R}^d$.

3.4 Bootstrapping

To develop feasible inference procedures that do not require (explicit) estimation of $\bar{\mathbf{V}}_n$, we consider nonparametric bootstrap-based approximations to the distribution of $\widetilde{\boldsymbol{\theta}}_n$. (Since $\widetilde{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n$ when $L = 0$, results for $\widehat{\boldsymbol{\theta}}_n$ can be extracted by setting $L = 0$ in what follows.)

Letting $\mathbf{z}_{1,n}^*, \ldots, \mathbf{z}_{n,n}^*$ denote a random sample from the empirical distribution of $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the defining property of $\widehat{\boldsymbol{\theta}}_n^*(h)$, the nonparametric bootstrap analogue of $\widehat{\boldsymbol{\theta}}_n(h)$, is the following:

$$\widehat{M}_n^*(\widehat{\boldsymbol{\theta}}_n^*(h); h) \leq \inf_{\boldsymbol{\theta} \in \Theta} \widehat{M}_n^*(\boldsymbol{\theta}; h) + o_{\mathbb{P}}(n^{-1}),$$

where

$$\widehat{M}_n^*(\boldsymbol{\theta}; h) = \binom{n}{2}^{-1} \sum_{i<j} m(\mathbf{z}_{i,n}^*, \mathbf{z}_{j,n}^*; \boldsymbol{\theta}) K_h(\mathbf{w}_{i,n}^* - \mathbf{w}_{j,n}^*).$$

Similarly, the nonparametric bootstrap analogue of $\widetilde{\boldsymbol{\theta}}_n$ is

$$\widetilde{\boldsymbol{\theta}}_n^* = \widetilde{\boldsymbol{\theta}}_n^*(\mathbf{c}) = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) \widehat{\boldsymbol{\theta}}_n^*(h_{n,l}).$$

The following theorem characterizes the large sample properties of $\widetilde{\boldsymbol{\theta}}_n^* - \widetilde{\boldsymbol{\theta}}_n$, the bootstrap counterpart of $\widetilde{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}_n$. In perfect analogy with the results in Cattaneo et al. (2014b), we find that the bootstrap distribution estimator is consistent only when $nh_n^d \to \infty$, but otherwise exhibits a variance inflation making the distributional approximation inconsistent. To state the result, let

$\mathbb{P}_n^*[\cdot]$ denote $\mathbb{P}[\cdot|\mathbf{z}_1,\ldots,\mathbf{z}_n]$, let $\to_{\mathbb{P}}$ denote convergence in probability, and define

$$\bar{\mathbf{V}}_n^* = \bar{\mathbf{V}}_n^*(\mathbf{c},h_n) = \boldsymbol{\Gamma}_0^{-1}\left[n^{-1}\boldsymbol{\Sigma}_0 + 3\binom{n}{2}^{-1} h_n^{-d}\boldsymbol{\Delta}_0(\bar{K})\right]\boldsymbol{\Gamma}_0^{-1}.$$

Theorem 3. *Suppose Assumptions 1-2 hold and that, for $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$ (and with probability one), $m(\mathbf{z},\mathbf{z};\boldsymbol{\theta}) = 0$ and $\mathbf{s}(\mathbf{z},\mathbf{z};\boldsymbol{\theta}) = \mathbf{0}$. If $n^2h_n^d \to \infty$ and if $h_n \to 0$, then*

$$\sup_{\mathbf{t}\in\mathbb{R}^k}\left|\mathbb{P}_n^*\left[\bar{\mathbf{V}}_n^{*-1/2}(\widetilde{\boldsymbol{\theta}}_n^* - \widetilde{\boldsymbol{\theta}}_n) \le \mathbf{t}\right] - \Phi_k(\mathbf{t})\right| \to_{\mathbb{P}} 0.$$

Because $\bar{\mathbf{V}}_n^{-1}\bar{\mathbf{V}}_n^* \to \mathbf{I}_k$ if and only if $nh_n^d \to \infty$ (where $\mathbf{I}_k$ denotes the k-dimensional identity matrix), under the assumptions of Theorem 3,

$$\sup_{\mathbf{t}\in\mathbb{R}^k}\left|\mathbb{P}_n^*\left[\widetilde{\boldsymbol{\theta}}_n^* - \widetilde{\boldsymbol{\theta}}_n \le \mathbf{t}\right] - \mathbb{P}\left[\widetilde{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}_n \le \mathbf{t}\right]\right| \to_{\mathbb{P}} 0$$

if and only if $nh_n^d \to \infty$. In particular, if $\liminf_{n\to\infty} nh_n^d < \infty$, then the nonparametric bootstrap is inconsistent, albeit conservative in the sense that the (approximate) variance under the bootstrap distribution is larger than the (approximate) variance of the asymptotic distribution: $\bar{\mathbf{V}}_n^* > \bar{\mathbf{V}}_n$ in a positive definite sense.

The variance inflation problem associated with the nonparametric bootstrap under the small bandwidth regime can be easily fixed by appropriately rescaling the bandwidth used for the bootstrap implementation of the pairwise difference estimator. Employing

$$\breve{\boldsymbol{\theta}}_n^* = \breve{\boldsymbol{\theta}}_n^*(\mathbf{c},h_n) = \sum_{l=0}^{L}\lambda_l(\mathbf{c})\widehat{\boldsymbol{\theta}}_n^*(3^{1/d}h_{n,l})$$

and centering its distribution at

$$\breve{\boldsymbol{\theta}}_n = \breve{\boldsymbol{\theta}}_n(\mathbf{c},h_n) = \sum_{l=0}^{L}\lambda_l(\mathbf{c})\widehat{\boldsymbol{\theta}}_n(3^{1/d}h_{n,l})$$

automatically adjusts the bootstrap variance, leading to a consistent distributional approximation. Indeed, the following result is an immediate consequence of Theorems 2 and 3.

Corollary 1. *If the assumptions of Theorem 3 hold, then*

$$\sup_{\mathbf{t}\in\mathbb{R}^k}\left|\mathbb{P}_n^*\left[\breve{\boldsymbol{\theta}}_n^* - \breve{\boldsymbol{\theta}}_n \le \mathbf{t}\right] - \mathbb{P}\left[\widetilde{\boldsymbol{\theta}}_n - \bar{\boldsymbol{\theta}}_n \le \mathbf{t}\right]\right| \to_{\mathbb{P}} 0.$$

As a consequence, if also Assumption 3 holds with $\mathtt{S} \ge 2L+2$, and if $nh_n^{4(L+1)} \to 0$, then

$$\sup_{\mathbf{t}\in\mathbb{R}^k}\left|\mathbb{P}_n^*\left[\breve{\boldsymbol{\theta}}_n^* - \breve{\boldsymbol{\theta}}_n \le \mathbf{t}\right] - \mathbb{P}\left[\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \le \mathbf{t}\right]\right| \to_{\mathbb{P}} 0.$$

The statement of Corollary 1 emphasizes the rate-adaptive nature of the consistency property enjoyed by the bootstrap distributional approximation.

3.5 Discussion

Our results provide a simple practical message: bootstrap-based inference for convex pairwise difference estimators is highly sensitive to bandwidth choice under conventional implementations, but this sensitivity can be substantially reduced by combining debiasing with a bandwidth rescaling. In particular, the classical bootstrap is valid only over a relatively narrow range of bandwidth sequences, whereas generalized jackknifing enlarges the set of admissible large bandwidths, small bandwidth asymptotics justifies an implementation enlarging the set of admissible small bandwidths, and the combination of the two yields the broadest range of valid inference.

To summarize these conclusions, suppose $\mathsf{S} \geq 2L+2$, let $\alpha \in (0,1)$ and $\mathbf{a} \in \mathbb{R}^k$ be fixed, and consider the following family of percentile bootstrap confidence intervals (in the terminology of van der Vaart, 1998) for $\mathbf{a}'\boldsymbol{\theta}_0$:

$$\mathsf{CI}^*_{n,1-\alpha}(B,L) = \left[\mathbf{a}'\widetilde{\boldsymbol{\theta}}_n(\mathbf{c},h_n) - \widetilde{q}^*_{1-\alpha/2,n}(Bh_n,L)\ ,\ \mathbf{a}'\widetilde{\boldsymbol{\theta}}_n(\mathbf{c},h_n) - \widetilde{q}^*_{\alpha/2,n}(Bh_n,L)\right],$$

where

$$\widetilde{q}^*_{t,n}(h,L) = \inf\left\{q \in \mathbb{R} : \mathbb{P}^*_n[\mathbf{a}'\widetilde{\boldsymbol{\theta}}^*_n(\mathbf{c},h) - \mathbf{a}'\widetilde{\boldsymbol{\theta}}_n(\mathbf{c},h) \leq q] \geq t\right\}.$$

The notation highlights the two tuning parameters that vary across the different inference procedures: given a choice of bandwidth h used in the point estimator, the scaling constant B used in the bootstrap approximation and the debiasing order $L = \dim(\mathbf{c}) - 1$, which together with $\mathbf{c}$ determines the generalized jackknife weights.

To sharpen the practical interpretation of these conclusions, it is useful to view each procedure through its coverage function

$$\mathsf{CF}_{n,1-\alpha}(B,L) = \mathbb{P}\left[\mathbf{a}'\boldsymbol{\theta}_0 \in \mathsf{CI}^*_{n,1-\alpha}(B,L)\right].$$

Our theory implies that the main differences across procedures can be understood in terms of how $\mathsf{CF}_{n,1-\alpha}(B,L)$ behaves as the bandwidth sequence moves from the small-bandwidth region to the large-bandwidth region. In particular, the classical bootstrap tends to become conservative when the bandwidth is too small, because the bootstrap variance is no longer correctly matched under small bandwidth asymptotics, and it tends to undercover when the bandwidth is too large, because smoothing bias becomes non-negligible. Debiasing mitigates the latter problem by shifting the onset of bias-driven undercoverage toward larger bandwidths, whereas bootstrap rescaling mitigates the former problem by restoring correct variance matching in the small-bandwidth region. Combining the two therefore produces the flattest coverage profile, in the sense that the coverage function remains close to its nominal level over the widest range of bandwidth sequences.

We formally compare the four implementations, ordered from the conventional procedure to our recommended method.

- *Classical Method.* This corresponds to $(B, L) = (1, 0)$. If Assumptions 1-3 hold and

$$nh_n^d \to \infty \qquad \text{and} \qquad nh_n^4 \to 0,$$

then

$$\lim_{n\to\infty} \mathbb{P}\left[\mathbf{a}'\boldsymbol{\theta}_0 \in \mathsf{CI}^*_{n,1-\alpha}(1,0)\right] = 1-\alpha.$$

The admissible range of bandwidth sequences is narrow, and requires $d < 4$. In terms of the coverage function, the classical percentile bootstrap has two distinct failure modes. When the bandwidth is too small, so that nh_n^d is no longer large while $n^2h_n^d \to \infty$ still holds, the procedure becomes conservative: $\mathsf{CF}_{n,1-\alpha}(1,0)$ rises above its nominal level because the bootstrap variance is too large relative to the sampling variance. When the bandwidth is too large, so that the bias condition fails, $\mathsf{CF}_{n,1-\alpha}(1,0)$ falls below its nominal level because smoothing bias is no longer negligible. Thus, the classical method delivers accurate coverage only over a narrow range of bandwidth choices.

- *Classical Debiased Method.* This corresponds to $(B, L) = (1, L)$ for some $L \geq 1$. If Assumptions 1-3 hold and

$$nh_n^d \to \infty \qquad \text{and} \qquad nh_n^{4(L+1)} \to 0,$$

then

$$\lim_{n\to\infty} \mathbb{P}\left[\mathbf{a}'\boldsymbol{\theta}_0 \in \mathsf{CI}^*_{n,1-\alpha}(1,L)\right] = 1-\alpha.$$

The range of allowable bandwidth sequences is wider on the large-bandwidth side than for the classical method, and requires $d < 4(L+1)$. Relative to the classical method, debiasing leaves the small-bandwidth behavior essentially unchanged: if the bandwidth is too small, the procedure may still be conservative because the bootstrap variance mismatch remains. Its gain is instead on the large-bandwidth side, where the onset of bias-driven undercoverage is pushed outward from the nh_n^4 scale to the weaker $nh_n^{4(L+1)}$ scale. Consequently, $\mathsf{CF}_{n,1-\alpha}(1,L)$ stays close to nominal coverage over a wider range of large bandwidth sequences than $\mathsf{CF}_{n,1-\alpha}(1,0)$.

- *Small Bandwidth Method.* This corresponds to $(B, L) = (3^{1/d}, 0)$. If Assumptions 1-3 hold and

$$n^2h_n^d \to \infty \qquad \text{and} \qquad nh_n^4 \to 0,$$

then

$$\lim_{n\to\infty} \mathbb{P}\left[\mathbf{a}'\boldsymbol{\theta}_0 \in \mathsf{CI}^*_{n,1-\alpha}(3^{1/d}, 0)\right] = 1-\alpha.$$

The range of allowable bandwidth sequences is wider on the small-bandwidth side than for the classical approach, and requires $d < 8$. Relative to the classical method, bootstrap rescaling corrects the variance mismatch responsible for small-bandwidth conservativeness. As a result, $\mathsf{CF}_{n,1-\alpha}(3^{1/d}, 0)$ remains close to its nominal level throughout the admissible small-bandwidth region. Its limitation is on the large-bandwidth side: because the procedure is not debiased, coverage still deteriorates once the bias condition $nh_n^4 \to 0$ fails.

- *Small Bandwidth Debiased Method.* This corresponds to $(B, L) = (3^{1/d}, L)$ for some $L \geq 1$. If Assumptions 1-3 hold and

$$n^2 h_n^d \to \infty \qquad \text{and} \qquad nh_n^{4(L+1)} \to 0,$$

then

$$\lim_{n\to\infty} \mathbb{P}\left[\mathbf{a}'\boldsymbol{\theta}_0 \in \mathsf{CI}^*_{n,1-\alpha}(3^{1/d}, L)\right] = 1-\alpha.$$

The range of allowable bandwidth sequences is the widest among the four procedures, and requires $d < 8(L+1)$. This method combines the two improvements described above. On the small-bandwidth side, bootstrap rescaling removes the variance mismatch that would otherwise induce conservativeness. On the large-bandwidth side, debiasing delays the onset of bias-driven undercoverage. Accordingly, $\mathsf{CF}_{n,1-\alpha}(3^{1/d}, L)$ remains close to its nominal level over the broadest range of bandwidth sequences considered in this paper.

Figure 1 presents a schematic plot summarizing the main theoretical conclusions. Overall, the comparison suggests a clear practical recommendation: bootstrap inference based on small bandwidth and debiasing corrections provides the most robust coverage accuracy as a function of bandwidth choice. Thus, for empirical settings where bandwidth choice is uncertain or data-driven, this method provides the most reliable default.

4 Proofs and Other Technical Results

4.1 A Useful Lemma

The following lemma is used in the proofs of Theorems 1-3.

Lemma 1. *Suppose that Assumption 1 holds. Then* $\arg\min_{\boldsymbol{\theta}\in\Theta} M(\boldsymbol{\theta}; h)$ *is non-empty for* $h > 0$ *near zero and*

$$\boldsymbol{\theta}(h) - \boldsymbol{\theta}_0 = o(1) \qquad \text{as } h \downarrow 0.$$

If also Assumption 2 holds, then $\mathbb{E}[\mathbf{s}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}(h))K_h(\mathbf{w}_1 - \mathbf{w}_2)] = \mathbf{0}$ *for* $h > 0$ *near zero.*

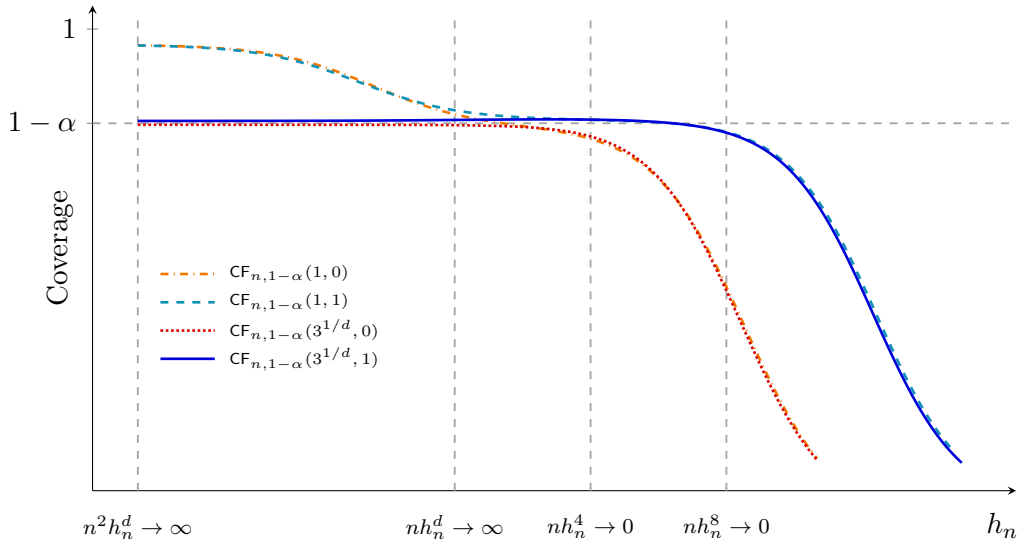


Figure 1: Schematic coverage functions for four bootstrap confidence interval procedures.

Proof. For every $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned} M(\boldsymbol{\theta}; h) &= \int_{\mathcal{W}} \int_{\mathbb{R}^d} \mathbb{E}[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) | \mathbf{w}_1 = \mathbf{w}, \mathbf{w}_2 = \mathbf{w} - \mathbf{u}h] f_{\mathbf{w}}(\mathbf{w}) f_{\mathbf{w}}(\mathbf{w} - \mathbf{u}h) K(\mathbf{u}) d\mathbf{u} d\mathbf{w} \\ &\to \int_{\mathcal{W}} \mathbb{E}[m(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}) | \mathbf{w}_1 = \mathbf{w}, \mathbf{w}_2 = \mathbf{w}] f_{\mathbf{w}}(\mathbf{w})^2 d\mathbf{w} = M_0(\boldsymbol{\theta}) \qquad \text{as } h \downarrow 0, \end{aligned}$$

the convergence being uniform on compact subsets of Θ because $\boldsymbol{\theta} \mapsto M(\boldsymbol{\theta}; h)$ is convex (e.g., Hjort and Pollard, 1993, Lemma 1).

Take any $\epsilon > 0$ with $\Theta_0^\epsilon = \{\boldsymbol{\theta} \in \mathbb{R}^k : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \epsilon\} \subseteq \Theta$. By the preceding paragraph,

$$\sup_{\boldsymbol{\theta} \in \Theta_0^\epsilon} |M(\boldsymbol{\theta}; h) - M_0(\boldsymbol{\theta})| \le \frac{1}{2} \left(\inf_{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = \epsilon} M_0(\boldsymbol{\theta}) - M_0(\boldsymbol{\theta}_0) \right)$$

for $h > 0$ near zero. For any such h and any $\boldsymbol{\theta} \in \Theta \setminus \Theta_0^\epsilon$, we have

$$\epsilon_{\boldsymbol{\theta}} = \frac{\epsilon}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|} \in (0, 1),$$

and therefore, by convexity of $\boldsymbol{\theta} \mapsto M(\boldsymbol{\theta}; h)$,

$$M(\epsilon_{\boldsymbol{\theta}} \boldsymbol{\theta} + (1 - \epsilon_{\boldsymbol{\theta}}) \boldsymbol{\theta}_0; h) \le \epsilon_{\boldsymbol{\theta}} M(\boldsymbol{\theta}; h) + (1 - \epsilon_{\boldsymbol{\theta}}) M(\boldsymbol{\theta}_0; h),$$

which rearranges as

$$M(\boldsymbol{\theta}; h) - M(\boldsymbol{\theta}_0; h) \ge \frac{1}{\epsilon_{\boldsymbol{\theta}}} [M(\epsilon_{\boldsymbol{\theta}} \boldsymbol{\theta} + (1 - \epsilon_{\boldsymbol{\theta}}) \boldsymbol{\theta}_0; h) - M(\boldsymbol{\theta}_0; h)] \ge 0.$$

As a consequence,

$$\inf_{\boldsymbol{\theta}\in\Theta} M(\boldsymbol{\theta};h) = \inf_{\boldsymbol{\theta}\in\Theta_0^\epsilon} M(\boldsymbol{\theta};h) = \min_{\boldsymbol{\theta}\in\Theta_0^\epsilon} M(\boldsymbol{\theta};h),$$

where the last equality uses continuity of $\boldsymbol{\theta}\mapsto M(\boldsymbol{\theta};h)$ and compactness of Θ_0^ϵ.

The above argument shows in particular that $\boldsymbol{\theta}(h)\in\Theta_0^\epsilon$ for $h>0$ near zero.

If also Assumption 2 holds, then, for $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$, $h>0$ near zero, and for any $\mathbf{t}\in\mathbb{R}^k$,

$$\begin{aligned}\left|\frac{M(\boldsymbol{\theta}+\mathbf{t}\tau;h)-M(\boldsymbol{\theta};h)}{\tau} - \mathbb{E}[\mathbf{s}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta})K_h(\mathbf{w}_1-\mathbf{w}_2)]'\mathbf{t}\right| &\leq |\mathbb{E}[r_{\mathbf{t}}(\boldsymbol{\theta},\tau)K_h(\mathbf{w}_1-\mathbf{w}_2)]| \\ &\to 0 \qquad \text{as } \tau\downarrow 0,\end{aligned}$$

implying that for $h>0$ near zero, $\boldsymbol{\theta}\mapsto M(\boldsymbol{\theta};h)$ is (directionally) differentiable near $\boldsymbol{\theta}_0$, the directional derivative $\mathbb{E}[\mathbf{s}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta})K_h(\mathbf{w}_1-\mathbf{w}_2)]'\mathbf{t}$ being zero when $\boldsymbol{\theta}=\boldsymbol{\theta}(h)$ because $\boldsymbol{\theta}(h)$ minimizes $M(\boldsymbol{\theta};h)$. □

4.2 Proof of Theorems 1 and 2

Theorem 1 can be obtained from Theorem 2 by setting $L=0$ in the latter, so it suffices to prove Theorem 2. To do so, for $l\in\{0,\dots,L\}$, let $\widehat{\boldsymbol{\theta}}_{n,l}=\widehat{\boldsymbol{\theta}}_n(h_{n,l}), \boldsymbol{\theta}_{n,l}=\boldsymbol{\theta}(h_{n,l})$, and

$$\widehat{\mathbf{U}}_{n,l} = \binom{n}{2}^{-1}\sum_{i<j}\mathbf{s}^{\mu}_{n,l}(\mathbf{z}_i,\mathbf{z}_j), \qquad \mathbf{s}^{\mu}_{n,l}(\mathbf{z}_i,\mathbf{z}_j)=\mathbf{s}_{n,l}(\mathbf{z}_i,\mathbf{z}_j)-\mathbb{E}[\mathbf{s}_{n,l}(\mathbf{z}_1,\mathbf{z}_2)],$$

where

$$\mathbf{s}_{n,l}(\mathbf{z}_i,\mathbf{z}_j)=\mathbf{s}(\mathbf{z}_i,\mathbf{z}_j;\boldsymbol{\theta}_{n,l})K_{h_{n,l}}(\mathbf{w}_i-\mathbf{w}_j).$$

By Lemma 1, $\lim_{n\to\infty}\boldsymbol{\theta}_{n,l}=\boldsymbol{\theta}_0$ and $\mathbb{E}[\mathbf{s}_{n,l}(\mathbf{z}_i,\mathbf{z}_j)]=0$ for large n.

Suppose that

$$\widehat{\boldsymbol{\theta}}_{n,l}-\boldsymbol{\theta}_{n,l} = -\boldsymbol{\Gamma}_0^{-1}\widehat{\mathbf{U}}_{n,l}+o_{\mathbb{P}}(\rho_n^{-1}) \quad \text{for} \quad l\in\{0,\dots,L\}. \tag{4.1}$$

Then

$$\widetilde{\boldsymbol{\theta}}_n-\bar{\boldsymbol{\theta}}_n = -\boldsymbol{\Gamma}_0^{-1}\widetilde{\mathbf{U}}_n+o_{\mathbb{P}}(\rho_n^{-1}),$$

where

$$\widetilde{\mathbf{U}}_n=\sum_{l=0}^{L}\lambda_l(\mathbf{c})\widehat{\mathbf{U}}_{n,l}$$

satisfies

$$\left[n^{-1}\boldsymbol{\Sigma}_0+\binom{n}{2}^{-1}h_n^{-d}\boldsymbol{\Delta}_0(\bar{K})\right]^{-1/2}\widetilde{\mathbf{U}}_n \rightsquigarrow \mathsf{N}(\mathbf{0}_{k\times 1},\mathbf{I}_k) \tag{4.2}$$

because, letting $\sum_i$ denote $\sum_{i=1}^n$,

$$\widetilde{\mathbf{U}}_n=\widetilde{\mathbf{L}}_n+\widetilde{\mathbf{W}}_n,$$

where

$$\widetilde{\mathbf{L}}_n = \frac{2}{n}\sum_i \widetilde{\boldsymbol{\ell}}_n(\mathbf{z}_i), \qquad \widetilde{\boldsymbol{\ell}}_n(\mathbf{z}_i) = \sum_{l=0}^{L} \lambda_l(\mathbf{c})\mathbb{E}[\mathbf{s}_{n,l}^{\mu}(\mathbf{z}_i,\mathbf{z}_j)|\mathbf{z}_i] \qquad (j \neq i),$$

and

$$\widetilde{\mathbf{W}}_n = \binom{n}{2}^{-1}\sum_{i<j}\widetilde{\boldsymbol{\omega}}_n(\mathbf{z}_i,\mathbf{z}_j), \qquad \widetilde{\boldsymbol{\omega}}_n(\mathbf{z}_i,\mathbf{z}_j) = \sum_{l=0}^{L}\lambda_l(\mathbf{c})\left[\mathbf{s}_{n,l}^{\mu}(\mathbf{z}_i,\mathbf{z}_j) - \widetilde{\boldsymbol{\ell}}_n(\mathbf{z}_i) - \widetilde{\boldsymbol{\ell}}_n(\mathbf{z}_j)\right],$$

satisfy

$$\begin{pmatrix} \sqrt{n}\widetilde{\mathbf{L}}_n \\ \sqrt{\binom{n}{2}h_n^d}\widetilde{\mathbf{W}}_n \end{pmatrix} \rightsquigarrow \mathsf{N}\left(\begin{bmatrix} \mathbf{0}_{k\times 1} \\ \mathbf{0}_{k\times 1}\end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_0 & \mathbf{0}_{k\times k} \\ \mathbf{0}_{k\times k} & \boldsymbol{\Delta}_0(\bar{K})\end{bmatrix}\right),$$

as can be shown by means of the Cramér-Wold device and the central limit theorem of Heyde and Brown (1970), the latter being applicable because it follows from routine calculations that for every $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^k$, we have

$$\sigma_n^2 = \sum_i \mathbb{V}[g_{i,n}] = \boldsymbol{\mu}_1'\boldsymbol{\Sigma}_0\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2'\boldsymbol{\Delta}_0(\bar{K})\boldsymbol{\mu}_2 + o(1),$$

$$\sum_i \mathbb{E}[g_{i,n}^4] = o(1),$$

and

$$\mathbb{V}\left[\sum_i \sigma_{i,n}^2 - \sigma_n^2\right] = o(1), \qquad \sigma_{i,n}^2 = \mathbb{V}[g_{i,n}|\mathbf{z}_1,\ldots,\mathbf{z}_{i-1}],$$

where

$$g_{i,n} = g_{i,n}(\boldsymbol{\mu}) = \frac{2}{\sqrt{n}}\boldsymbol{\mu}_1'\widetilde{\boldsymbol{\ell}}_n(\mathbf{z}_i) + \sqrt{\binom{n}{2}^{-1}h_n^d}\sum_{j=1}^{i-1}\boldsymbol{\mu}_2'\widetilde{\boldsymbol{\omega}}_n(\mathbf{z}_i,\mathbf{z}_j).$$

The proof of Theorem 2 can therefore be completed by verifying (4.1).

To do so, we leverage convexity. For any $l \in \{0,\ldots,L\}$ and any $\mathbf{t} \in \mathbb{R}^k$, it can be shown that

$$\lim_{\tau\downarrow 0, h\downarrow 0, \boldsymbol{\theta}\to\boldsymbol{\theta}_0} \mathbb{E}\left[\mathbb{E}[r_{\mathbf{t}}(\boldsymbol{\theta},\tau)K_h(\mathbf{w}_1-\mathbf{w}_2)|\mathbf{z}_1]^2\right] = 0$$

and

$$\lim_{\tau\downarrow 0, h\downarrow 0, \boldsymbol{\theta}\to\boldsymbol{\theta}_0} h^d\mathbb{E}[r_{\mathbf{t}}(\boldsymbol{\theta},\tau)^2K_h(\mathbf{w}_1-\mathbf{w}_2)^2] = 0,$$

and it therefore follows from a Hoeffding decomposition that

$$\begin{aligned}&\rho_n^2\left[\widehat{M}_n(\boldsymbol{\theta}_{n,l}+\mathbf{t}\rho_n^{-1};h_{n,l}) - \widehat{M}_n(\boldsymbol{\theta}_{n,l};h_{n,l})\right] \\ &= \rho_n^2[M(\boldsymbol{\theta}_{n,l}+\mathbf{t}\rho_n^{-1};h_{n,l}) - M(\boldsymbol{\theta}_{n,l};h_{n,l})] + \mathbf{t}'\rho_n\widehat{\mathbf{U}}_{n,l} + o_{\mathbb{P}}(1).\end{aligned}$$

Moreover,

$$\rho_n^2[M(\boldsymbol{\theta}_{n,l}+\mathbf{t}\rho_n^{-1};h_{n,l})-M(\boldsymbol{\theta}_{n,l};h_{n,l})]\to\frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}_0\mathbf{t},$$

and proceeding as in the proof of (4.2) it can be shown that $\rho_n\widehat{\mathbf{U}}_{n,l}=O_{\mathbb{P}}(1)$. Because $\boldsymbol{\Gamma}_0$ is positive definite and because $\mathbf{t}\mapsto\widehat{M}_n(\boldsymbol{\theta}_{n,l}+\mathbf{t}\rho_n^{-1};h_{n,l})$ is convex (almost surely), the corollary following Hjort and Pollard (1993, Lemma 2) implies that (4.1) holds.

4.3 Proof of Theorem 3

The proof of Theorem 3 is a natural bootstrap analog of the proof of Theorem 2.

For $l\in\{0,\dots,L\}$, let $\widehat{\boldsymbol{\theta}}^*_{n,l}=\widehat{\boldsymbol{\theta}}^*_n(h_{n,l})$ and

$$\widehat{\mathbf{U}}^*_{n,l}=\binom{n}{2}^{-1}\sum_{i<j}\mathbf{s}^{\mu,*}_{n,l}(\mathbf{z}^*_{i,n},\mathbf{z}^*_{j,n}),\qquad \mathbf{s}^{\mu,*}_{n,l}(\mathbf{z}^*_{i,n},\mathbf{z}^*_{j,n})=\mathbf{s}_{n,l}(\mathbf{z}^*_{i,n},\mathbf{z}^*_{j,n})-\mathbb{E}^*_n[\mathbf{s}_{n,l}(\mathbf{z}^*_{1,n},\mathbf{z}^*_{2,n})],$$

where $\mathbb{E}^*_n[\cdot]$ denotes $\mathbb{E}[\cdot|\mathbf{z}_1,\dots,\mathbf{z}_n]$.

It suffices to show that

$$\widehat{\boldsymbol{\theta}}^*_{n,l}-\boldsymbol{\theta}_{n,l}=-\boldsymbol{\Gamma}_0^{-1}\left(\widehat{\mathbf{U}}^*_{n,l}+\widehat{\mathbf{U}}_{n,l}\right)+o_{\mathbb{P}}(\rho_n^{-1})\quad\text{for}\quad l\in\{0,\dots,L\}\tag{4.3}$$

and that

$$\left[n^{-1}\boldsymbol{\Sigma}_0+3\binom{n}{2}^{-1}h_n^{-d}\boldsymbol{\Delta}_0(\bar{K})\right]^{-1/2}\widetilde{\mathbf{U}}^*_n\rightsquigarrow_{\mathbb{P}}\mathsf{N}(\mathbf{0}_{k\times1},\mathbf{I}_k),\tag{4.4}$$

where

$$\widetilde{\mathbf{U}}^*_n=\sum_{l=0}^{L}\lambda_l(\mathbf{c})\widehat{\mathbf{U}}^*_{n,l}.$$

For every $\mathbf{t}\in\mathbb{R}^k$, using a Hoeffding decomposition and the fact that $m(\mathbf{z},\mathbf{z};\boldsymbol{\theta}_{n,l})=0$ and $\mathbf{s}(\mathbf{z},\mathbf{z};\boldsymbol{\theta}_{n,l})=\mathbf{0}$ for large n, we have

$$\begin{aligned}&\rho_n^2\left[\widehat{M}^*_n(\boldsymbol{\theta}_{n,l}+\mathbf{t}\rho_n^{-1};h_{n,l})-\widehat{M}^*_n(\boldsymbol{\theta}_{n,l};h_{n,l})\right]\\&=\rho_n^2(1+o(1))[\widehat{M}_n(\boldsymbol{\theta}_{n,l}+\mathbf{t}\rho_n^{-1};h_{n,l})-\widehat{M}_n(\boldsymbol{\theta}_{n,l};h_{n,l})]+\mathbf{t}'\rho_n\widehat{\mathbf{U}}^*_{n,l}+o_{\mathbb{P}}(1),\end{aligned}$$

where it can be shown that $\rho_n\widehat{\mathbf{U}}^*_{n,l}=O_{\mathbb{P}}(1)$ and where it follows from the proof of (4.1) that

$$\rho_n^2\left[\widehat{M}_n(\boldsymbol{\theta}_{n,l}+\mathbf{t}\rho_n^{-1};h_{n,l})-\widehat{M}_n(\boldsymbol{\theta}_{n,l};h_{n,l})\right]=\frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}_0\mathbf{t}+\mathbf{t}'\rho_n\widehat{\mathbf{U}}_{n,l}+o_{\mathbb{P}}(1),$$

where $\rho_n\widehat{\mathbf{U}}_{n,l}=O_{\mathbb{P}}(1)$. In other words,

$$\begin{aligned}&\rho_n^2\left[\widehat{M}^*_n(\boldsymbol{\theta}_{n,l}+\mathbf{t}\rho_n^{-1};h_{n,l})-\widehat{M}^*_n(\boldsymbol{\theta}_{n,l};h_{n,l})\right]\\&=\frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}_0\mathbf{t}+\mathbf{t}'\rho_n\left(\widehat{\mathbf{U}}^*_{n,l}+\widehat{\mathbf{U}}_{n,l}\right)+o_{\mathbb{P}}(1)\qquad\text{for every }\mathbf{t}\in\mathbb{R}^k.\end{aligned}$$

Because $\mathbf{\Gamma}_0$ is positive definite and because $\mathbf{t} \mapsto \widehat{M}_n^*(\boldsymbol{\theta}_{n,l} + \mathbf{t}\rho_n^{-1}; h_{n,l})$ is convex (almost surely), the corollary following Hjort and Pollard (1993, Lemma 2) implies that (4.3) holds.

To prove (4.4), we begin by decomposing $\widetilde{\mathbf{U}}_n^*$ as

$$\widetilde{\mathbf{U}}_n^* = \widetilde{\mathbf{L}}_n^* + \widetilde{\mathbf{W}}_n^*,$$

where

$$\widetilde{\mathbf{L}}_n^* = \frac{2}{n}\sum_i \widetilde{\boldsymbol{\ell}}_n^*(\mathbf{z}_{i,n}^*), \qquad \widetilde{\boldsymbol{\ell}}_n^*(\mathbf{z}_{i,n}^*) = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) \mathbb{E}_n^*[\mathbf{s}_{n,l}^{\mu,*}(\mathbf{z}_{i,n}^*, \mathbf{z}_{j,n}^*) | \mathbf{z}_{i,n}^*] \qquad (j \neq i),$$

and

$$\widetilde{\mathbf{W}}_n^* = \binom{n}{2}^{-1} \sum_{i<j} \widetilde{\boldsymbol{\omega}}_n^*(\mathbf{z}_{i,n}^*, \mathbf{z}_{j,n}^*), \qquad \widetilde{\boldsymbol{\omega}}_n^*(\mathbf{z}_{i,n}^*, \mathbf{z}_{j,n}^*) = \sum_{l=0}^{L} \lambda_l(\mathbf{c}) \left[\mathbf{s}_{n,l}^{\mu,*}(\mathbf{z}_{i,n}^*, \mathbf{z}_{j,n}^*) - \widetilde{\boldsymbol{\ell}}_n^*(\mathbf{z}_{i,n}^*) - \widetilde{\boldsymbol{\ell}}_n^*(\mathbf{z}_{j,n}^*) \right].$$

Defining

$$\pi_n = \frac{\sqrt{nh_n^d}}{1 + \sqrt{nh_n^d}},$$

routine calculations can be used to show that for every $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^k$, we have

$$\widehat{\sigma}_n^2 = \sum_i \mathbb{V}_n^*[g_{i,n}^*] = \boldsymbol{\mu}_1' \left[\pi_n^2 \mathbf{\Sigma}_0 + 4(1-\pi_n)^2 \mathbf{\Delta}_0(\bar{K}) \right] \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2' \mathbf{\Delta}_0(\bar{K}) \boldsymbol{\mu}_2 + o_{\mathbb{P}}(1),$$

$$\sum_i \mathbb{E}_n^*[g_{i,n}^{*4}] = o_{\mathbb{P}}(1),$$

and

$$\mathbb{V}_n^* \left[\sum_i \widehat{\sigma}_{i,n}^2 - \widehat{\sigma}_n^2 \right] = o_{\mathbb{P}}(1), \qquad \widehat{\sigma}_{i,n}^2 = \mathbb{V}_n^*[g_{i,n}^* | \mathbf{z}_{1,n}^*, \ldots, \mathbf{z}_{i-1,n}^*],$$

where $\mathbb{V}_n^*[\cdot]$ denotes $\mathbb{V}[\cdot | \mathbf{z}_1, \ldots, \mathbf{z}_n]$ and where

$$g_{i,n}^* = g_{i,n}^*(\boldsymbol{\mu}) = \frac{\pi_n}{\sqrt{n}} 2\boldsymbol{\mu}_1' \widetilde{\boldsymbol{\ell}}_n^*(\mathbf{z}_{i,n}^*) + \sqrt{\binom{n}{2}^{-1} h_n^d} \sum_{j=1}^{i-1} \boldsymbol{\mu}_2' \widetilde{\boldsymbol{\omega}}_n^*(\mathbf{z}_{i,n}^*, \mathbf{z}_{j,n}^*).$$

The Cramér-Wold device and the central limit theorem of Heyde and Brown (1970) therefore imply that if $\pi_n \to \pi_0 \in [0,1]$, then

$$\begin{pmatrix} \sqrt{n}\pi_n \widetilde{\mathbf{L}}_n^* \\ \sqrt{\binom{n}{2} h_n^d} \widetilde{\mathbf{W}}_n^* \end{pmatrix} \rightsquigarrow_{\mathbb{P}} \mathsf{N}\left(\begin{bmatrix} \mathbf{0}_{k\times 1} \\ \mathbf{0}_{k\times 1} \end{bmatrix}, \begin{bmatrix} \pi_0^2 \mathbf{\Sigma}_0 + 4(1-\pi_0)^2 \mathbf{\Delta}_0(\bar{K}) & \mathbf{0}_{k\times k} \\ \mathbf{0}_{k\times k} & \mathbf{\Delta}_0(\bar{K}) \end{bmatrix} \right).$$

Whether or not π_n is convergent, the result (4.4) can be obtained from the preceding display by arguing along subsequences (if necessary).

4.4 Verifying Assumption 3

It follows from Lemma 1 that if Assumption 1 holds, then so does Assumption 3 with $\mathtt{S} = 0$. This observation provides the base case for an induction argument. To describe the induction step, suppose that for some even $S \geq 0$, we have

$$\boldsymbol{\theta}(h) - \boldsymbol{\theta}_0 = \sum_{l=1}^{S/2} \mathbf{b}_{2l} h^{2l} + o(h^S) \qquad \text{as } h \downarrow 0.$$

Suppose also that, for every $\mathbf{t} \in \mathbb{R}^k$ and some $\boldsymbol{\beta}_{S+2} \in \mathbb{R}^k$, we have

$$\begin{aligned} &h^{-2(S+2)} \left[M\left(\boldsymbol{\theta}_0 + \sum_{l=1}^{S/2} \mathbf{b}_{2l} h^{2l} + \mathbf{t} h^{S+2}; h \right) - M\left(\boldsymbol{\theta}_0 + \sum_{l=1}^{S/2} \mathbf{b}_{2l} h^{2l}; h \right) \right] \\ &= \mathbf{t}' \boldsymbol{\beta}_{S+2} + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_0 \mathbf{t} + o(1) \qquad \text{as } h \downarrow 0. \end{aligned}$$

Then, the corollary following Hjort and Pollard (1993, Lemma 2) implies that

$$\begin{aligned} h^{-(S+2)} \left(\boldsymbol{\theta}(h) - \boldsymbol{\theta}_0 - \sum_{l=1}^{S/2} \mathbf{b}_{2l} h^{2l} \right) &= \operatorname*{arg\,min}_{\mathbf{t} \in \mathbb{R}^k} M\left(\boldsymbol{\theta}_0 + \sum_{l=1}^{S/2} \mathbf{b}_{2l} h^{2l} + \mathbf{t} h^{S+2}; h \right) \\ &= -\boldsymbol{\Gamma}_0^{-1} \boldsymbol{\beta}_{S+2} + o(1) \qquad \text{as } h \downarrow 0; \end{aligned}$$

that is, defining $\mathbf{b}_{S+2} = -\boldsymbol{\Gamma}_0^{-1} \boldsymbol{\beta}_{S+2}$, we have

$$\boldsymbol{\theta}(h) - \boldsymbol{\theta}_0 = \sum_{l=1}^{(S+2)/2} \mathbf{b}_{2l} h^{2l} + o(h^{S+2}) \qquad \text{as } h \downarrow 0.$$

To complete the proof of Proposition 1, it therefore suffices to note that (for every $\mathbf{t} \in \mathbb{R}^k$ and) under the assumptions of the proposition, we have

$$h^{-4} \left[M(\boldsymbol{\theta}_0 + \mathbf{t} h^2; h) - M(\boldsymbol{\theta}_0; h) \right] = \mathbf{t}' \boldsymbol{\beta}_2 + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_0 \mathbf{t} + o(1) \qquad \text{as } h \downarrow 0,$$

where

$$\boldsymbol{\beta}_2 = \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \left(\int_{\mathcal{W}} \frac{\partial^2 \boldsymbol{\varphi}(\mathbf{w}, \bar{\mathbf{w}})}{\partial \bar{w}_i \partial \bar{w}_j} \bigg|_{\bar{\mathbf{w}} = \mathbf{w}} f_{\mathbf{w}}(\mathbf{w}) d\mathbf{w} \right) \left(\int_{\mathbb{R}^d} u_i u_j K(\mathbf{u}) d\mathbf{u} \right).$$

Similarly, if in addition to the assumptions of Proposition 1 it is assumed that for every $\mathbf{t} \in \mathbb{R}^k$ and for some $\boldsymbol{\beta}_4 \in \mathbb{R}^k$, we have

$$h^{-8} \left[M(\boldsymbol{\theta}_0 + \mathbf{b}_2 h^2 + \mathbf{t} h^4; h) - M(\boldsymbol{\theta}_0 + \mathbf{b}_2 h^2; h) \right] = \mathbf{t}' \boldsymbol{\beta}_4 + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_0 \mathbf{t} + o(1) \qquad \text{as } h \downarrow 0,$$

then Assumption 3 holds with $\mathtt{S} = 4$. One set of sufficient conditions for this to occur is that

Assumptions 1-2 hold and that, for every $\mathbf{t} \in \mathbb{R}^k$, the following are satisfied (with probability one):

(i) $\int_{\mathbb{R}^d} \|\mathbf{u}\|^4 K(\mathbf{u}) d\mathbf{u} < \infty$.

(ii) $\bar{\mathbf{w}} \mapsto \boldsymbol{\varphi}(\mathbf{w}, \bar{\mathbf{w}}) = \mathbb{E}[\mathbf{s}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}_0) | \mathbf{w}_1 = \mathbf{w}, \mathbf{w}_2 = \bar{\mathbf{w}}] f_{\mathbf{w}}(\bar{\mathbf{w}})$ is four times continuously differentiable with $\mathbb{E}[\sup_{\bar{\mathbf{w}} \in \mathcal{W}} \|\partial_{\bar{\mathbf{w}}}^{\boldsymbol{\nu}} \boldsymbol{\varphi}(\mathbf{w}, \bar{\mathbf{w}})\|] < \infty$ for all $\boldsymbol{\nu} \in \mathbb{Z}_+^d$ with $|\boldsymbol{\nu}| \leq 4$.

(iii) $f_{\mathbf{w}}$ is twice continuously differentiable and $\bar{\mathbf{w}} \mapsto \mathbf{H}(\mathbf{w}, \bar{\mathbf{w}}; \boldsymbol{\theta}_0, \mathbf{t})$ is twice continuously differentiable with $\mathbb{E}[\sup_{\bar{\mathbf{w}} \in \mathcal{W}} \|\partial_{\bar{\mathbf{w}}}^{\boldsymbol{\nu}} \mathbf{H}(\mathbf{w}, \bar{\mathbf{w}}; \boldsymbol{\theta}_0, \mathbf{t}) f_{\mathbf{w}}(\bar{\mathbf{w}})\|] < \infty$ for all $\boldsymbol{\nu} \in \mathbb{Z}_+^d$ with $|\boldsymbol{\nu}| \leq 2$.

(iv) For some function $\dot{\mathbf{H}}(\mathbf{w}, \bar{\mathbf{w}}; \boldsymbol{\theta}, \mathbf{t}) \in \mathbb{R}^{k \times k}$, $\bar{\mathbf{w}} \mapsto \dot{\mathbf{H}}(\mathbf{w}, \bar{\mathbf{w}}; \boldsymbol{\theta}_0, \mathbf{t})$ is continuous,

$$\mathbb{E}\left[\sup_{\bar{\mathbf{w}} \in \mathcal{W}} \|\dot{\mathbf{H}}(\mathbf{w}, \bar{\mathbf{w}}; \boldsymbol{\theta}_0, \mathbf{t}) f_{\mathbf{w}}(\bar{\mathbf{w}})\|\right] < \infty,$$

$$\lim_{\tau \downarrow 0, (\boldsymbol{\theta}, \mathbf{u}) \to (\boldsymbol{\theta}_0, \mathbf{0})} \left\| \frac{\mathbf{H}(\mathbf{w}, \mathbf{w} + \mathbf{u}; \boldsymbol{\theta} + \tau \mathbf{t}, \mathbf{t}) - \mathbf{H}(\mathbf{w}, \mathbf{w} + \mathbf{u}; \boldsymbol{\theta}, \mathbf{t})}{\tau} - \dot{\mathbf{H}}(\mathbf{w}, \mathbf{w} + \mathbf{u}; \boldsymbol{\theta}, \mathbf{t}) \right\| = 0,$$

and, for some $\delta > 0$,

$$\mathbb{E}\left[\sup_{\tau \in (0, \delta), \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta, \mathbf{w}_2 \in \mathcal{W}} \left\| \frac{\mathbf{H}(\mathbf{w}_1, \mathbf{w}_2; \boldsymbol{\theta} + \tau \mathbf{t}, \mathbf{t}) - \mathbf{H}(\mathbf{w}_1, \mathbf{w}_2; \boldsymbol{\theta}, \mathbf{t})}{\tau} - \dot{\mathbf{H}}(\mathbf{w}_1, \mathbf{w}_2; \boldsymbol{\theta}, \mathbf{t}) \right\|\right] < \infty.$$

5 Sufficient Conditions for Motivating Examples

To demonstrate the plausibility of Assumptions 1 and 2, we revisit the examples of Section 2. In each example, Assumption 1(ii) holds and Assumption 1(iii) is fairly primitive, so we focus on giving primitive sufficient conditions for Assumptions 1(iv)-(v) and 2.

5.1 Partially Linear Regression Model

We take $\mathbf{s} = \mathbf{s}_{\mathtt{PLR}}$ and $\mathbf{H} = \mathbf{H}_{\mathtt{PLR}}$, where

$$\mathbf{s}_{\mathtt{PLR}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} m_{\mathtt{PLR}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = -\dot{\mathbf{x}}_{i,j}(\dot{y}_{i,j} - \dot{\mathbf{x}}_{i,j}' \boldsymbol{\theta})$$

and

$$\begin{aligned} \mathbf{H}_{\mathtt{PLR}}(\mathbf{w}_i, \mathbf{w}_j) &= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathbb{E}[m_{\mathtt{PLR}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) | \mathbf{w}_i, \mathbf{w}_j] = \frac{\partial}{\partial \boldsymbol{\theta}'} \mathbb{E}[\mathbf{s}_{\mathtt{PLR}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) | \mathbf{w}_i, \mathbf{w}_j] \\ &= \mathbb{E}[\dot{\mathbf{x}}_{i,j} \dot{\mathbf{x}}_{i,j}' | \mathbf{w}_i, \mathbf{w}_j], \end{aligned}$$

the latter depending on neither $\boldsymbol{\theta}$ nor $\mathbf{t}$ (because $m_{\mathtt{PLR}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$).

Under mild conditions, Assumptions 1(iv)-(v) and 2 hold with

$$\boldsymbol{\xi}_0(\mathbf{z}) = -2\mathbb{E}[\mathbf{s}_{\mathtt{PLR}}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}_0) | \mathbf{z}_1 = \mathbf{z}, \mathbf{w}_2 = \mathbf{w}] f_{\mathbf{w}}(\mathbf{w}) = 2(\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathbf{w}]) f_{\mathbf{w}}(\mathbf{w}) \varepsilon,$$

$$\begin{aligned}\boldsymbol{\Xi}_0(\mathbf{w}) &= \mathbb{E}[\mathbf{s}_{\text{PLR}}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta}_0)\mathbf{s}_{\text{PLR}}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta}_0)'|\mathbf{w}_1=\mathbf{w},\mathbf{w}_2=\mathbf{w}]f_{\mathbf{w}}(\mathbf{w})\\ &= \mathbb{E}[\dot{\mathbf{x}}_{1,2}\dot{\mathbf{x}}_{1,2}'(\varepsilon_1^2+\varepsilon_2^2)|\mathbf{w}_1=\mathbf{w},\mathbf{w}_2=\mathbf{w}]f_{\mathbf{w}}(\mathbf{w}),\end{aligned}$$

and

$$\mathbf{G}_0(\mathbf{w}) = \mathbf{H}_{\text{PLR}}(\mathbf{w},\mathbf{w})f_{\mathbf{w}}(\mathbf{w}) = 2\mathbb{V}[\mathbf{x}|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w}).$$

For instance, it suffices to set $b(\mathbf{z}) = (1+\|\mathbf{x}\|)(1+\|\mathbf{x}\|+|\gamma_0(\mathbf{w})|+|\varepsilon|)$ and to assume that

(i) The functions $\mathbf{w}\mapsto\gamma_0(\mathbf{w})$, $\mathbf{w}\mapsto\mathbb{E}[\mathbf{x}|\mathbf{w}]$, $\mathbf{w}\mapsto\mathbb{E}[\mathbf{x}\mathbf{x}'|\mathbf{w}]$, $\mathbf{w}\mapsto\mathbb{E}[\varepsilon^2|\mathbf{w}]$, $\mathbf{w}\mapsto\mathbb{E}[\mathbf{x}\varepsilon^2|\mathbf{w}]$, and $\mathbf{w}\mapsto\mathbb{E}[\mathbf{x}\mathbf{x}'\varepsilon^2|\mathbf{w}]$ are continuous on $\mathcal{W}$.

(ii) $\mathbb{E}[(1+\|\mathbf{x}\|^4)\varepsilon^4] + \sup_{\mathbf{w}\in\mathcal{W}}\mathbb{E}[(1+\|\mathbf{x}\|^4)\varepsilon^4|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w}) < \infty$ and

$$\mathbb{E}[(1+\|\mathbf{x}\|^4)\gamma_0(\mathbf{w})^4+\|\mathbf{x}\|^8] + \sup_{\mathbf{w}\in\mathcal{W}}\mathbb{E}[(1+\|\mathbf{x}\|^4)\gamma_0(\mathbf{w})^4+\|\mathbf{x}\|^8|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w}) < \infty.$$

(iii) With probability one, $\mathbb{V}[\mathbf{x}|\mathbf{w}]$ is positive definite and $\mathbb{V}[\varepsilon|\mathbf{x},\mathbf{w}] > 0$.

Under the additional assumptions of Theorem 1 and if $nh_n^d \to \infty$, the pairwise difference estimator is asymptotically linear with influence function

$$\mathbf{z}\mapsto\mathbb{E}[\mathbb{V}[\mathbf{x}|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w})]^{-1}(\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{w}])f_{\mathbf{w}}(\mathbf{w})\varepsilon.$$

Unless the distribution of $\mathbf{w}$ is uniform on $\mathcal{W}$, the pairwise difference estimator is therefore asymptotically distinct from the estimators studied by Robinson (1988) and Donald and Newey (1994), the influence function of these estimators being given by

$$\mathbf{z}\mapsto\mathbb{E}[\mathbb{V}[\mathbf{x}|\mathbf{w}]]^{-1}(\mathbf{x}-\mathbb{E}[\mathbf{x}|\mathbf{w}])\varepsilon.$$

When $nh_n^d \to 2c < \infty$, the pairwise difference estimator is still $\sqrt{n}$-normal, but it ceases to be asymptotically linear. Similarly, the estimator of Donald and Newey (1994) is $\sqrt{n}$-normal, but not asymptotically linear, when the associated tuning parameter is chosen appropriately (Cattaneo et al., 2018), and in light of Linton (1995) it stands to reason that the same is true for the estimator of Robinson (1988). Another $\sqrt{n}$-normal estimator that is not asymptotically linear was proposed by Yatchew (1997). However, even under simplifying assumptions (e.g., uniformity of the distribution of $\mathbf{w}$ and/or conditional homoskedasticity of ε), it appears difficult to make insightful comparisons between the various estimators just mentioned.

5.2 Partially Linear Logit Model

We take $\mathbf{s} = \mathbf{s}_{\text{PLL}}$ and $\mathbf{H} = \mathbf{H}_{\text{PLL}}$, where

$$\mathbf{s}_{\text{PLL}}(\mathbf{z}_i,\mathbf{z}_j;\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}m_{\text{PLL}}(\mathbf{z}_i,\mathbf{z}_j;\boldsymbol{\theta}) = -\dot{\mathbf{x}}_{i,j}(y_i-\Lambda(\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta}))\mathbb{1}\{\dot{y}_{i,j}\neq 0\}$$

and, defining $\lambda(u) = \partial\Lambda(u)/\partial u = \exp(u)/[1+\exp(u)]^2$,

$$\begin{aligned}\mathbf{H}_{\mathtt{PLL}}(\mathbf{w}_i, \mathbf{w}_j; \boldsymbol{\theta}) &= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathbb{E}[m_{\mathtt{PLL}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})|\mathbf{w}_i, \mathbf{w}_j] = \frac{\partial}{\partial\boldsymbol{\theta}'}\mathbb{E}[\mathbf{s}_{\mathtt{PLL}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})|\mathbf{w}_i, \mathbf{w}_j] \\ &= \mathbb{E}[\dot{\mathbf{x}}_{i,j}\dot{\mathbf{x}}_{i,j}'\lambda(\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta})\mathbb{1}\{\dot{y}_{i,j} \neq 0\}|\mathbf{w}_i, \mathbf{w}_j],\end{aligned}$$

where the latter does not depend on $\mathbf{t}$ (because $m_{\mathtt{PLL}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})$ is twice differentiable in $\boldsymbol{\theta}$).

Under mild conditions, Assumptions 1(iv)-(v) and 2 hold with

$$\boldsymbol{\xi}_0(\mathbf{z}) = -2\mathbb{E}[\mathbf{s}_{\mathtt{PLL}}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}_0)|\mathbf{z}_1 = \mathbf{z}, \mathbf{w}_2 = \mathbf{w}]f_{\mathbf{w}}(\mathbf{w}),$$

$$\boldsymbol{\Xi}_0(\mathbf{w}) = \mathbb{E}[\mathbf{s}_{\mathtt{PLL}}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}_0)\mathbf{s}_{\mathtt{PLL}}(\mathbf{z}_1, \mathbf{z}_2; \boldsymbol{\theta}_0)'|\mathbf{w}_1 = \mathbf{w}, \mathbf{w}_2 = \mathbf{w}]f_{\mathbf{w}}(\mathbf{w}),$$

and

$$\mathbf{G}_0(\mathbf{w}) = \mathbf{H}_{\mathtt{PLL}}(\mathbf{w}, \mathbf{w}; \boldsymbol{\theta}_0)f_{\mathbf{w}}(\mathbf{w}).$$

For instance, it suffices to set $b(\mathbf{z}) = 1 + \|\mathbf{x}\|$ and to assume that, for some $\delta > 0$,

(i) The function $\mathbf{w} \mapsto \gamma_0(\mathbf{w})$ is continuous on $\mathcal{W}$. Also, the conditional distribution of $\mathbf{x}$ given $\mathbf{w}$ admits a density $f_{\mathbf{x}|\mathbf{w}}$ with respect to some measure ρ such that $\mathbf{w} \mapsto f_{\mathbf{x}|\mathbf{w}}(\mathbf{x}|\mathbf{w})$ is continuous on $\mathcal{W}$ (with probability one) and

$$\int_{\mathbb{R}^k}(1+\|\mathbf{x}\|^2)\sup_{\|\mathbf{u}\|\leq\delta} f_{\mathbf{x}|\mathbf{w}}(\mathbf{x}|\mathbf{w}+\mathbf{u})d\rho(\mathbf{x}) < \infty \qquad \text{for every } \mathbf{w} \in \mathcal{W}.$$

(ii) $\mathbb{E}[\|\mathbf{x}\|^4] + \sup_{\mathbf{w}\in\mathcal{W}}\mathbb{E}[\|\mathbf{x}\|^4|\mathbf{w}]f_{\mathbf{w}}(\mathbf{w}) < \infty$.

(iii) With probability one, $\mathbb{V}[\mathbf{x}|\mathbf{w}]$ is positive definite.

5.3 Partially Linear Tobit Model

We take $\mathbf{s} = \mathbf{s}_{\mathtt{PLT}}$ and $\mathbf{H} = \mathbf{H}_{\mathtt{PLT}}$, where

$$\mathbf{s}_{\mathtt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta}) = -\dot{\mathbf{x}}_{i,j}\left(\mathbb{1}\{y_i > \max(y_j + \dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta}, 0)\} - \mathbb{1}\{y_j > \max(y_i - \dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta}, 0)\}\right)$$

and

$$\begin{aligned}\mathbf{H}_{\mathtt{PLT}}(\mathbf{w}_i, \mathbf{w}_j; \boldsymbol{\theta}, \mathbf{t}) = \mathbb{E}\left[\dot{\mathbf{x}}_{i,j}\dot{\mathbf{x}}_{i,j}'\left(\mathbb{1}\{\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} > 0\} + \mathbb{1}\{\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} = 0, \dot{\mathbf{x}}_{i,j}'\mathbf{t} \geq 0\}\right)\eta_{\mathtt{PLT}}(\mathbf{z}_i, \mathbf{z}_j; \boldsymbol{\theta})\middle|\, \mathbf{w}_i, \mathbf{w}_j\right] \\ + \mathbb{E}\left[\dot{\mathbf{x}}_{i,j}\dot{\mathbf{x}}_{i,j}'\left(\mathbb{1}\{\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} < 0\} + \mathbb{1}\{\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta} = 0, \dot{\mathbf{x}}_{i,j}'\mathbf{t} < 0\}\right)\eta_{\mathtt{PLT}}(\mathbf{z}_j, \mathbf{z}_i; \boldsymbol{\theta})\middle|\, \mathbf{w}_i, \mathbf{w}_j\right],\end{aligned}$$

with

$$\eta_{\texttt{PLT}}(\mathbf{z}_i,\mathbf{z}_j;\boldsymbol{\theta}) = 2\int_0^\infty f_{\varepsilon|\mathbf{w}}(\varepsilon-\mathbf{x}_i'\boldsymbol{\theta}_0-\gamma_0(\mathbf{w}_i)+\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta}|\mathbf{w}_i)f_{\varepsilon|\mathbf{w}}(\varepsilon-\mathbf{x}_j'\boldsymbol{\theta}_0-\gamma_0(\mathbf{w}_j)|\mathbf{w}_j)d\varepsilon$$
$$+ f_{\varepsilon|\mathbf{w}}(-\mathbf{x}_i'\boldsymbol{\theta}_0-\gamma_0(\mathbf{w}_i)+\dot{\mathbf{x}}_{i,j}'\boldsymbol{\theta}|\mathbf{w}_i)\int_{-\infty}^0 f_{\varepsilon|\mathbf{w}}(\varepsilon-\mathbf{x}_j'\boldsymbol{\theta}_0-\gamma_0(\mathbf{w}_j)|\mathbf{w}_j)d\varepsilon.$$

Under mild conditions, Assumptions 1(iv)-(v) and 2 hold with

$$\boldsymbol{\xi}_0(\mathbf{z}) = -2\mathbb{E}[\mathbf{s}_{\texttt{PLT}}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta}_0)|\mathbf{z}_1=\mathbf{z},\mathbf{w}_2=\mathbf{w}]f_{\mathbf{w}}(\mathbf{w}),$$

$$\boldsymbol{\Xi}_0(\mathbf{w}) = \mathbb{E}[\mathbf{s}_{\texttt{PLT}}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta}_0)\mathbf{s}_{\texttt{PLT}}(\mathbf{z}_1,\mathbf{z}_2;\boldsymbol{\theta}_0)'|\mathbf{w}_1=\mathbf{w},\mathbf{w}_2=\mathbf{w}]f_{\mathbf{w}}(\mathbf{w}),$$

and

$$\mathbf{G}_0(\mathbf{w}) = \mathbf{H}_{\texttt{PLT}}(\mathbf{w},\mathbf{w};\boldsymbol{\theta}_0,\boldsymbol{\theta}_0)f_{\mathbf{w}}(\mathbf{w}).$$

For instance, it suffices to set $b(\mathbf{z}) = 1+\|\mathbf{x}\|$ and to assume that, for some $\delta>0$,

(i) The function $\mathbf{w}\mapsto\gamma_0(\mathbf{w})$ is continuous on $\mathcal{W}$. Also, the conditional distribution of $\mathbf{x}$ given $\mathbf{w}$ admits a density $f_{\mathbf{x}|\mathbf{w}}$ with respect to some measure ρ such that $\mathbf{w}\mapsto f_{\mathbf{x}|\mathbf{w}}(\mathbf{x}|\mathbf{w})$ is continuous on $\mathcal{W}$ (with probability one) and

$$\int_{\mathbb{R}^k}(1+\|\mathbf{x}\|^2)\sup_{\|\mathbf{u}\|\le\delta} f_{\mathbf{x}|\mathbf{w}}(\mathbf{x}|\mathbf{w}+\mathbf{u})d\rho(\mathbf{x}) < \infty \qquad \text{for every } \mathbf{w}\in\mathcal{W}.$$

In addition, the function $(\varepsilon,\mathbf{w})\mapsto f_{\varepsilon|\mathbf{w}}(\varepsilon|\mathbf{w})$ is continuous and bounded and the function

$$(\mathbf{x},\mathbf{w})\mapsto\int_{\mathbb{R}}\sup_{|u|+\|\mathbf{u}\|\le\delta} f_{\varepsilon|\mathbf{w}}(\varepsilon-\mathbf{x}'\boldsymbol{\theta}_0-\gamma_0(\mathbf{w})+u|\mathbf{w}+\mathbf{u})d\varepsilon$$

is bounded.

(ii) $\mathbb{E}[\|\mathbf{x}\|^4]+\sup_{\mathbf{w}\in\mathcal{W}}(1+\mathbb{E}[\|\mathbf{x}\|^4|\mathbf{w}])f_{\mathbf{w}}(\mathbf{w})<\infty$.

(iii) With probability one, $\mathbb{V}[\mathbf{x}|\mathbf{w}]$ is positive definite.

6 Simulation Evidence

We present simulation evidence for the partially linear regression and partially linear logit models. We compare the four bootstrap-based confidence intervals discussed in Section 3.5: $\mathsf{CI}^*_{n,1-\alpha}(1,0)$, $\mathsf{CI}^*_{n,1-\alpha}(1,L)$, $\mathsf{CI}^*_{n,1-\alpha}(3^{1/d},0)$, and $\mathsf{CI}^*_{n,1-\alpha}(3^{1/d},L)$. We set $\alpha=0.05$, and for the debiasing procedure, we use $L=1$ and $\mathbf{c}=(1,2)'$.

6.1 Simulation Design: Partially Linear Regression Model

We base the partially linear regression designs on equations (6.1)–(6.2) of Robinson (1988). We consider

$$y_i = x_i + \gamma_0(\mathbf{w}_i) + \varepsilon_i.$$

The three designs are as follows. In Model 1, $x_i = 2v_i$, $v_i \sim \mathsf{N}(0,1)$, $w_i = v_i + u_i$, $u_i \sim \mathrm{N}(0,2)$, $\gamma_0(w) = w^2 + 1$, $\varepsilon_i \sim \mathsf{N}(0,1)$, and $\{v_i, u_i, \varepsilon_i\}$ are jointly independent. In Model 2, $d = \dim(\mathbf{w}) = 2$, $\gamma_0(\mathbf{w}) = \mathbf{w}'\mathbf{w} + 1$, $(x_i, \mathbf{w}_i')'$ is normal with each component having mean 1 and variance 3, and any pair in $(x_i, \mathbf{w}_i')'$ has correlation 2/3. Model 3 uses the same joint normal specification for $(x_i, \mathbf{w}_i')'$ as Model 2, but with $d = \dim(\mathbf{w}) = 3$. In all models, $\theta_0 = 1$. The regressor of interest x_i is correlated with $\mathbf{w}_i$, and thus, ignoring the $\gamma_0(\mathbf{w}_i)$ term will induce bias in the estimator of θ_0.

6.2 Simulation Design: Partially Linear Logit Model

We base the partially linear logit designs on those of Honoré and Powell (2005). As in the partially linear regression simulations, the three designs differ mainly in the dimension of $\mathbf{w}$. We consider

$$y_i = \mathbb{1}\left\{x_{1i} + x_{2i} + \gamma_0(\mathbf{w}_i) + \varepsilon_i \geq 0\right\}, \qquad \gamma_0(\mathbf{w}) = \mathbf{w}'\mathbf{w} - (1 + \dim(\mathbf{w})),$$

where the CDF of ε_i is Λ, x_{2i} has a discrete distribution with $\mathbb{P}[x_{2i} = 1] = 1/2 = \mathbb{P}[x_{2i} = -1]$, $x_{1i} = v_i + \mathbf{w}_i'\mathbf{w}_i$ with $v_i \sim \mathsf{N}(0,1)$, and $\{\varepsilon_i, \mathbf{w}_i, x_{2i}, v_i\}$ are jointly independent. For Model d ($d \in \{1,2,3\}$), $\dim(\mathbf{w}) = d$, and $\mathbf{w}_i$ has a normal distribution with each element having mean zero and variance one, equicorrelated with correlation 0.2. In all models, $\boldsymbol{\theta}_0 = (1,1)'$.

As noted by Honoré and Powell (2005), ignoring the presence of $\gamma_0(\mathbf{w}_i)$ induces bias in the estimator of $\boldsymbol{\theta}_0$, although the bias for the second element tends to be negligible relative to the standard error. We focus on constructing 95% confidence intervals for the first element of $\boldsymbol{\theta}_0$.

6.3 Simulation Results

Tables 1–3 display the coverage probabilities and average lengths of the four confidence intervals for the partially linear regression model, and Tables 4–6 for the partially linear logit model. Figure 2 complements the tables by plotting coverage probability as a function of bandwidth h for all six DGPs. For each data generating process, the sample size was $n = 2{,}000$, we conducted 2,000 simulation replications, and for each replication we drew 2,000 bootstrap samples to compute bootstrap quantiles.

The panel headed "$B = 1$" refers to $\mathsf{CI}^*_{n,0.95}(1, L)$, in which the bandwidth used to compute bootstrapped quantiles equals the bandwidth used for estimation. The panel headed "$B = 3^{1/d}$" refers to our proposed methods $\mathsf{CI}^*_{n,0.95}(3^{1/d}, L)$, in which the bandwidth for bootstrap estimation is rescaled by the factor $3^{1/d}$. Within each panel, $L = 0$ is based on the estimator $\widehat{\boldsymbol{\theta}}_n$ and $L = 1$ is based on the debiased estimator $\widetilde{\boldsymbol{\theta}}_n$ with $\mathbf{c} = (1,2)'$.

To form the bandwidth grid, we first estimate $h_{L=0}$ and $h_{L=1}$ by auxiliary simulations. These

are the MSE-minimizing bandwidths for $\widehat{\boldsymbol{\theta}}_n$ and $\widetilde{\boldsymbol{\theta}}_n$, respectively. The grid is then taken as the union of $h_{L=0} \times \{0.5, \ldots, 1.5\}$ and $h_{L=1} \times \{0.5, \ldots, 1.5\}$ (increments of 0.1). Rows highlighted in yellow and orange in the tables mark $h_{L=0}$ and $h_{L=1}$, respectively.

The results are broadly consistent with our theoretical predictions. We focus on two forces of particular interest: small bandwidth regions, where the standard bootstrap variance may be inflated, and large bandwidth regions, where smoothing bias may become non-negligible. The simulations also include one instructive qualification: in the partially linear regression model with $d = 1$, the conventional procedures are already very well calibrated, while our proposed procedure $\mathsf{CI}^*_{n,0.95}(3^{1/d}, 1)$ is conservative. This is a specialized design in which the smoothing bias targeted by jackknife debiasing is negligible, so bandwidth rescaling and debiasing mostly widen the intervals without offsetting a meaningful bias. This qualification is specific to the partially linear regression $d = 1$ design and does not describe the partially linear logit model with $d = 1$, where uncorrected intervals exhibit undercoverage as the bandwidth grows and bias correction is beneficial.

Small-bandwidth overcoverage. For small bandwidths, the confidence intervals $\mathsf{CI}^*_{n,0.95}(1, L)$ tend to overcover the nominal 95% level, whereas our proposed methods $\mathsf{CI}^*_{n,0.95}(3^{1/d}, L)$ achieve coverage close to the nominal probability. This observation is predicted by our small-bandwidth asymptotics: using the same h_n for both estimation and the bootstrap quantile computation causes the bootstrap to overestimate the sampling dispersion of $\widehat{\boldsymbol{\theta}}_n$. The average interval lengths in the tables are in line with this prediction. At small bandwidths, $\mathsf{CI}^*_{n,0.95}(1, L)$ are systematically longer than their $\mathsf{CI}^*_{n,0.95}(3^{1/d}, L)$ counterparts.

The severity of this distortion grows sharply with $d = \dim(\mathbf{w})$, consistent with our theoretical prediction that the larger d, the wider the small-bandwidth region (based on the condition $\liminf_{n\to\infty} nh_n^d < \infty$). For the partially linear regression model with $d = 2$ and $d = 3$, the coverage probabilities for $\mathsf{CI}^*_{n,0.95}(1, L)$ are well above the nominal level for smaller bandwidths, whereas our methods $\mathsf{CI}^*_{n,0.95}(3^{1/d}, L)$ achieve the nominal level (Tables 2 and 3). The same small-bandwidth variance distortion is visible in the logit model. Overcoverage under $\mathsf{CI}^*_{n,0.95}(1, 0)$ is modest for $d = 1$, rises to 0.982 for $d = 2$, and reaches 0.996 for $d = 3$ at the smallest bandwidth used in the simulation (Tables 4-6), with analogous patterns for $\mathsf{CI}^*_{n,0.95}(1, 1)$. Figure 2 makes this dimension-dependence visually apparent: the coverage curves are well above the nominal level (the horizontal dotted line) for $d > 1$.

Large-bandwidth undercoverage and bias correction. At larger bandwidths, except in the partially linear regression $d = 1$ design discussed above, the bias may become non-negligible, and $\mathsf{CI}^*_{n,0.95}(B, 0)$ tends to exhibit severe undercoverage. For the partially linear regression model with $d = 2$, the coverage of $\mathsf{CI}^*_{n,0.95}(1, 0)$ falls to 0.468 at $h = 0.75$ and to 0.007 at $h = 1.12$. For $d = 3$, it falls to 0.051 at $h = 1.00$ and to essentially zero for $h \geq 1.20$. The undercoverage is similarly pronounced in the logit model: for $d = 2$, coverage collapses to 0.858 at $h = 0.68$, and for $d = 3$, it falls to 0.720 at $h = 0.75$.

The debiased intervals correct this distortion effectively. At larger bandwidths, the $(B = 1, L = 1)$ procedure is closer to the nominal 95% level than the $(B = 1, L = 0)$ procedure: for the partially linear regression model with $d = 2$, coverage is 0.965 at $h_{L=1} = 0.75$, and for $d = 3$ it is 0.970 at $h_{L=1} = 1.00$. Similar patterns hold for the logit model, where coverage is 0.963 at $h_{L=1} = 0.45$ for $d = 2$ and 0.976 at $h_{L=1} = 0.50$ for $d = 3$.

Combining bandwidth rescaling and bias correction. Under the procedure that combines bandwidth rescaling and bias correction, $\mathsf{CI}^*_{n,0.95}(3^{1/d}, 1)$, coverage remains close to the nominal level across a wide bandwidth range. For the partially linear regression model with $d = 2$, the coverage probabilities lie between 0.949 and 0.959 across all bandwidths on the grid. For $d = 3$, coverage ranges from 0.950 to 0.970 over most of the grid and remains at 0.955 even at $h = 1.50$. The cost of bias correction is a modest increase in average interval length relative to the uncorrected $L = 0$ intervals: comparing the $(B = 3^{1/d}, L = 0)$ and $(B = 3^{1/d}, L = 1)$ columns at the MSE-optimal bandwidth $h_{L=0}$, the debiased intervals are roughly 4% wider for $d = 2$ and roughly 11% wider for $d = 3$ in the partially linear regression model, a small price given the improvement in coverage reliability for larger bandwidths. For the logit model, the qualitative picture is the same for $d = 2$ and $d = 3$. For $d = 1$, the proposed method is not conservative in the same way as in the partially linear regression case; instead, the main issue is the emergence of bias as the bandwidth grows, which bias correction partly offsets. Under the proposed procedure, coverage remains close to the nominal 95% across a wide range of bandwidths considered.

Overall, the simulation evidence supports combining bandwidth rescaling $(B = 3^{1/d})$ with jackknife-based bias correction $(L = 1)$, especially when smoothing bias is present or when the dimension of the nonparametric component is moderate. The partially linear regression $d = 1$ design provides a useful reminder that when conventional methods are already well calibrated and the relevant smoothing bias is essentially absent, the robust procedure can be conservative.

7 Conclusion

This paper develops bandwidth-robust distribution theory and bootstrap-based inference procedures for a broad class of convex pairwise difference estimators. Our theoretical work is based on small bandwidth asymptotics and carefully leverages convexity. We illustrate the theory with three prominent examples. In addition to expanding the scope of small bandwidth asymptotics, our results lay the groundwork for several promising avenues of future research. First, our methods could be generalized to develop bandwidth selection based on higher-order stochastic expansions. Second, consistent variance estimators could be developed as an alternative to bootstrap-based inference. Third, it would be of interest to expand our theory to allow for pairwise difference estimators based on generated regressors, a class of estimators that sometimes arises in the context of control function and related econometric methods. Finally, it seems plausible that there exist settings where the objective function is sufficiently non-smooth to result in non-Gaussian distributional

approximations. We leave these extensions for future work.

References

Ahn, H., H. Ichimura, J. L. Powell, and P. A. Ruud (2018): "Simple Estimators for Invertible Index Models," *Journal of Business & Economic Statistics*, 36, 1–10.

Ahn, H. and J. L. Powell (1993): "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3–29.

Aradillas-Lopez, A. (2012): "Pairwise-Difference Estimation of Incomplete Information Games," *Journal of Econometrics*, 168, 120–140.

Aradillas-Lopez, A., B. E. Honoré, and J. L. Powell (2007): "Pairwise Difference Estimation with Nonparametric Control Variables," *International Economic Review*, 48, 1119–1158.

Blundell, R. W. and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655–679.

Cattaneo, M. D., R. K. Crump, and M. Jansson (2010): "Robust Data-Driven Inference for Density-Weighted Average Derivatives," *Journal of the American Statistical Association*, 105, 1070–1083.

——— (2013): "Generalized Jackknife Estimators of Weighted Average Derivatives (with Discussions and Rejoinder)," *Journal of the American Statistical Association*, 108, 1243–1268.

——— (2014a): "Small Bandwidth Asymptotics for Density-Weighted Average Derivatives," *Econometric Theory*, 30, 176–200.

——— (2014b): "Bootstrapping Density-Weighted Average Derivatives," *Econometric Theory*, 30, 1135–1164.

Cattaneo, M. D., M. H. Farrell, M. Jansson, and R. P. Masini (2025): "Higher-Order Refinements of Small Bandwidth Asymptotics for Density-Weighted Average Derivative Estimators," *Journal of Econometrics*, 252, 105855.

Cattaneo, M. D. and M. Jansson (2018): "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency," *Econometrica*, 86, 955–995.

——— (2022): "Average Density Estimators: Efficiency and Bootstrap Consistency," *Econometric Theory*, 38, 1140–1174.

Cattaneo, M. D., M. Jansson, and K. Nagasawa (2024): "Bootstrap-Assisted Inference for Generalized Grenander-type Estimators," *Annals of Statistics*, 52, 1509–1533.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 34, 277–301.

Cattaneo, M. D., J. M. Klusowski, and W. G. Underwood (2026): "Inference with Mondrian Random Forests," *Journal of the Royal Statistical Society Series B, forthcoming*.

Donald, S. G. and W. K. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis*, 50, 30–40.

Heyde, C. C. and B. M. Brown (1970): "On the Departure from Normality of a Certain Class of Martingales," *Annals of Mathematical Statistics*, 41, 2161–2165.

Hjort, N. L. and D. Pollard (1993): "Asymptotics for Minimisers of Convex Processes," *arXiv preprint arXiv:1107.3806*.

Hong, H. and M. Shum (2010): "Pairwise-Difference Estimation of a Dynamic Optimization Model," *Review of Economic Studies*, 77, 273–304.

Honoré, B. E. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60, 533–565.

Honoré, B. E., E. Kyriazidou, and C. Udry (1997): "Estimation of Type 3 Tobit Models using Symmetric Trimming and Pairwise Comparisons," *Journal of Econometrics*, 76, 107–128.

Honoré, B. E. and J. L. Powell (1994): "Pairwise Difference Estimators of Censored and Truncated Regression Models," *Journal of Econometrics*, 64, 241–278.

——— (2005): "Pairwise Difference Estimators for Nonlinear Models," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews and J. H. Stock, Cambridge University Press, 520–553.

Jochmans, K. (2013): "Pairwise-Comparison Estimation with Non-parametric Controls," *Econometrics Journal*, 16, 340–372.

Kyriazidou, E. (1997): "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335–1364.

Linton, O. (1995): "Second Order Approximation in the Partially Linear Regression Model," *Econometrica*, 63, 1079–1112.

Matsushita, Y. and T. Otsu (2021): "Jackknife Empirical Likelihood: Small Bandwidth, Sparse Network and High-Dimensional Asymptotics," *Biometrika*, 108, 661–674.

Pollard, D. (1991): "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199.

Powell, J. L. (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics, Volume IV*, ed. by R. F. Engle and D. L. McFadden, Elsevier, 2443–2521.

——— (2001): "Semiparametric estimation of censored selection models," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. L. Powell, Cambridge University Press, 165—-196.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

Schucany, W. R. and J. P. Sommers (1977): "Improvement of Kernel Type Density Estimators," *Journal of the American Statistical Association*, 72, 420–423.

Shao, J. and D. Tu (2012): *The Jackknife and Bootstrap*, Springer.

van der Vaart, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press.

Yatchew, A. (1997): "An Elementary Estimator of the Partial Linear Model," *Economics Letters*, 57, 135–143.

Table 1: Bootstrap 95% Confidence Intervals for Partially Linear Regression Model: DGP 1 ($d = 1$).

	$B = 1$				$B = 3^{1/d}$			
	$L = 0$		$L = 1$		$L = 0$		$L = 1$	
h	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
0.35	0.958	0.058	0.958	0.059	0.956	0.058	0.956	0.058
0.42	0.958	0.058	0.958	0.058	0.958	0.059	0.955	0.058
0.49	0.956	0.058	0.958	0.058	0.963	0.060	0.955	0.058
0.50	0.956	0.058	0.958	0.058	0.963	0.060	0.956	0.058
0.56	0.955	0.058	0.957	0.058	0.966	0.061	0.958	0.058
0.60	0.954	0.058	0.956	0.058	0.969	0.063	0.959	0.059
0.63	0.953	0.058	0.956	0.058	0.970	0.064	0.961	0.059
0.70	0.954	0.058	0.956	0.058	0.977	0.067	0.965	0.061
0.77	0.955	0.058	0.956	0.058	0.986	0.070	0.970	0.063
0.80	0.956	0.058	0.955	0.058	0.987	0.072	0.972	0.064
0.84	0.956	0.058	0.956	0.058	0.990	0.074	0.976	0.066
0.90	0.954	0.058	0.955	0.058	0.993	0.078	0.982	0.069
0.91	0.954	0.058	0.955	0.058	0.993	0.078	0.984	0.070
0.98	0.953	0.058	0.953	0.058	0.997	0.083	0.987	0.074
1.00	0.954	0.058	0.953	0.058	0.998	0.084	0.989	0.075
1.05	0.955	0.058	0.952	0.058	0.999	0.088	0.992	0.078
1.10	0.955	0.058	0.952	0.058	1.000	0.091	0.994	0.081
1.20	0.955	0.058	0.953	0.058	1.000	0.098	0.999	0.088
1.30	0.953	0.059	0.954	0.058	1.000	0.105	1.000	0.095
1.40	0.955	0.059	0.954	0.058	1.000	0.112	1.000	0.102
1.50	0.957	0.060	0.954	0.058	1.000	0.119	1.000	0.108

Note: The table shows coverage probability and average length of confidence intervals over 2,000 simulation replications. The sample size is $n = 2{,}000$ and the number of bootstrap draws is 2,000. The rows highlighted in yellow and orange correspond to the estimated MSE-optimal bandwidths for the point estimators with $L = 0$ and $L = 1$, respectively.

Table 2: Bootstrap 95% Confidence Intervals for Partially Linear Regression Model: DGP 2 ($d = 2$).

	$B = 1$				$B = 3^{1/d}$			
	$L = 0$		$L = 1$		$L = 0$		$L = 1$	
h	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
0.12	0.991	0.149	0.994	0.171	0.949	0.111	0.949	0.122
0.15	0.987	0.133	0.992	0.151	0.950	0.104	0.950	0.112
0.17	0.984	0.123	0.990	0.137	0.951	0.100	0.955	0.106
0.20	0.979	0.115	0.986	0.127	0.946	0.097	0.957	0.102
0.22	0.977	0.110	0.984	0.120	0.944	0.095	0.955	0.099
0.25	0.970	0.106	0.983	0.114	0.944	0.093	0.956	0.097
0.28	0.965	0.102	0.979	0.110	0.939	0.092	0.956	0.095
0.30	0.961	0.100	0.978	0.107	0.931	0.091	0.956	0.094
0.32	0.952	0.098	0.978	0.104	0.923	0.090	0.956	0.093
0.35	0.941	0.096	0.972	0.102	0.912	0.090	0.956	0.092
0.38	0.925	0.095	0.973	0.100	0.905	0.089	0.956	0.091
0.45	0.887	0.092	0.973	0.096	0.872	0.089	0.957	0.090
0.52	0.829	0.091	0.968	0.094	0.821	0.089	0.958	0.089
0.60	0.743	0.090	0.966	0.092	0.749	0.090	0.959	0.089
0.68	0.625	0.089	0.964	0.091	0.646	0.091	0.958	0.088
0.75	0.468	0.089	0.965	0.090	0.514	0.093	0.957	0.089
0.83	0.298	0.089	0.961	0.090	0.375	0.097	0.958	0.089
0.90	0.151	0.089	0.952	0.089	0.239	0.100	0.956	0.090
0.98	0.071	0.089	0.942	0.089	0.124	0.105	0.954	0.092
1.05	0.022	0.090	0.935	0.089	0.067	0.110	0.957	0.094
1.12	0.007	0.090	0.924	0.088	0.026	0.116	0.957	0.098

Note: The table shows coverage probability and average length of confidence intervals over 2,000 simulation replications. The sample size is $n = 2{,}000$ and the number of bootstrap draws is 2,000. The rows highlighted in yellow and orange correspond to the estimated MSE-optimal bandwidths for the point estimators with $L = 0$ and $L = 1$, respectively.

Table 3: Bootstrap 95% Confidence Intervals for Partially Linear Regression Model: DGP 3 ($d = 3$).

	$B = 1$				$B = 3^{1/d}$			
	$L = 0$		$L = 1$		$L = 0$		$L = 1$	
h	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
0.20	0.998	0.325	0.999	0.402	0.950	0.210	0.950	0.254
0.24	0.997	0.259	0.998	0.318	0.956	0.174	0.952	0.207
0.28	0.993	0.217	0.996	0.263	0.957	0.152	0.961	0.177
0.32	0.991	0.188	0.995	0.225	0.951	0.138	0.961	0.158
0.36	0.986	0.168	0.994	0.198	0.949	0.129	0.964	0.145
0.40	0.979	0.154	0.994	0.179	0.938	0.122	0.963	0.135
0.44	0.969	0.143	0.993	0.164	0.917	0.117	0.961	0.128
0.48	0.950	0.135	0.993	0.153	0.887	0.114	0.962	0.123
0.50	0.934	0.131	0.992	0.149	0.873	0.113	0.966	0.121
0.52	0.916	0.129	0.991	0.145	0.858	0.111	0.966	0.119
0.56	0.879	0.124	0.990	0.138	0.827	0.110	0.963	0.116
0.60	0.830	0.120	0.988	0.132	0.778	0.109	0.961	0.114
0.60	0.830	0.120	0.988	0.132	0.778	0.109	0.961	0.114
0.70	0.650	0.114	0.979	0.123	0.613	0.107	0.959	0.110
0.80	0.408	0.110	0.975	0.117	0.399	0.108	0.958	0.107
0.90	0.178	0.108	0.971	0.113	0.202	0.111	0.962	0.106
1.00	0.051	0.107	0.970	0.110	0.075	0.116	0.964	0.106
1.10	0.009	0.108	0.967	0.108	0.019	0.121	0.967	0.108
1.20	0.000	0.109	0.960	0.107	0.002	0.128	0.969	0.110
1.30	0.000	0.111	0.950	0.106	0.000	0.137	0.970	0.114
1.40	0.000	0.114	0.925	0.106	0.000	0.146	0.967	0.120
1.50	0.000	0.118	0.887	0.107	0.000	0.156	0.955	0.126

Note: The table shows coverage probability and average length of confidence intervals over 2,000 simulation replications. The sample size is $n = 2{,}000$ and the number of bootstrap draws is 2,000. The rows highlighted in yellow and orange correspond to the estimated MSE-optimal bandwidths for the point estimators with $L = 0$ and $L = 1$, respectively.

Table 4: Bootstrap 95% Confidence Intervals for Partially Linear Logit Model: DGP 1 ($d = 1$).

	$B = 1$				$B = 3^{1/d}$			
	$L = 0$		$L = 1$		$L = 0$		$L = 1$	
h	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
0.10	0.952	0.319	0.953	0.324	0.944	0.308	0.947	0.311
0.12	0.950	0.317	0.951	0.321	0.942	0.306	0.946	0.310
0.14	0.950	0.315	0.952	0.318	0.942	0.305	0.946	0.309
0.16	0.951	0.314	0.953	0.317	0.944	0.304	0.945	0.308
0.18	0.950	0.312	0.952	0.315	0.943	0.302	0.948	0.307
0.18	0.950	0.312	0.952	0.315	0.943	0.302	0.948	0.307
0.20	0.950	0.311	0.951	0.314	0.941	0.301	0.947	0.306
0.22	0.950	0.311	0.951	0.314	0.941	0.300	0.946	0.306
0.22	0.949	0.311	0.951	0.313	0.941	0.300	0.946	0.306
0.24	0.948	0.310	0.950	0.313	0.940	0.299	0.945	0.305
0.25	0.947	0.310	0.950	0.312	0.939	0.298	0.945	0.304
0.26	0.948	0.309	0.950	0.312	0.938	0.297	0.945	0.304
0.28	0.948	0.309	0.949	0.312	0.935	0.296	0.943	0.303
0.29	0.947	0.308	0.950	0.311	0.934	0.295	0.942	0.302
0.30	0.947	0.308	0.950	0.311	0.933	0.295	0.942	0.302
0.32	0.946	0.307	0.949	0.311	0.933	0.293	0.941	0.300
0.36	0.942	0.306	0.948	0.310	0.931	0.291	0.940	0.298
0.40	0.941	0.306	0.948	0.309	0.930	0.289	0.936	0.296
0.43	0.938	0.305	0.947	0.309	0.926	0.287	0.935	0.295
0.47	0.939	0.304	0.948	0.308	0.922	0.285	0.935	0.293
0.50	0.936	0.303	0.946	0.308	0.919	0.283	0.932	0.291
0.54	0.932	0.302	0.947	0.307	0.916	0.281	0.930	0.289

Note: The table shows coverage probability and average length of confidence intervals over 2,000 simulation replications. The sample size is $n = 2{,}000$ and the number of bootstrap draws is 2,000. The rows highlighted in yellow and orange correspond to the estimated MSE-optimal bandwidths for the point estimators with $L = 0$ and $L = 1$, respectively.

Table 5: Bootstrap 95% Confidence Intervals for Partially Linear Logit Model: DGP 2 ($d = 2$).

	$B = 1$				$B = 3^{1/d}$			
	$L = 0$		$L = 1$		$L = 0$		$L = 1$	
h	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
0.14	0.982	0.507	0.985	0.568	0.952	0.404	0.952	0.431
0.17	0.980	0.462	0.982	0.510	0.953	0.385	0.953	0.407
0.20	0.977	0.433	0.980	0.471	0.953	0.372	0.954	0.390
0.22	0.974	0.414	0.978	0.445	0.951	0.364	0.951	0.380
0.22	0.973	0.413	0.978	0.444	0.951	0.364	0.951	0.379
0.25	0.971	0.399	0.977	0.426	0.954	0.357	0.953	0.372
0.27	0.969	0.392	0.975	0.416	0.950	0.354	0.953	0.367
0.28	0.969	0.388	0.973	0.411	0.950	0.352	0.953	0.365
0.31	0.964	0.380	0.972	0.400	0.948	0.348	0.950	0.360
0.32	0.963	0.378	0.970	0.398	0.948	0.347	0.951	0.359
0.34	0.961	0.373	0.969	0.391	0.945	0.344	0.951	0.356
0.36	0.958	0.368	0.968	0.385	0.941	0.341	0.951	0.353
0.36	0.958	0.368	0.967	0.385	0.940	0.341	0.951	0.353
0.39	0.954	0.363	0.964	0.379	0.936	0.338	0.949	0.350
0.40	0.950	0.361	0.964	0.376	0.933	0.336	0.950	0.348
0.42	0.946	0.359	0.964	0.374	0.928	0.335	0.951	0.347
0.45	0.938	0.356	0.963	0.370	0.919	0.332	0.948	0.344
0.50	0.928	0.351	0.960	0.364	0.909	0.329	0.947	0.340
0.54	0.915	0.347	0.958	0.360	0.893	0.326	0.940	0.337
0.59	0.898	0.344	0.953	0.356	0.876	0.323	0.938	0.334
0.63	0.881	0.341	0.947	0.353	0.855	0.320	0.930	0.331
0.68	0.858	0.338	0.937	0.350	0.832	0.318	0.921	0.329

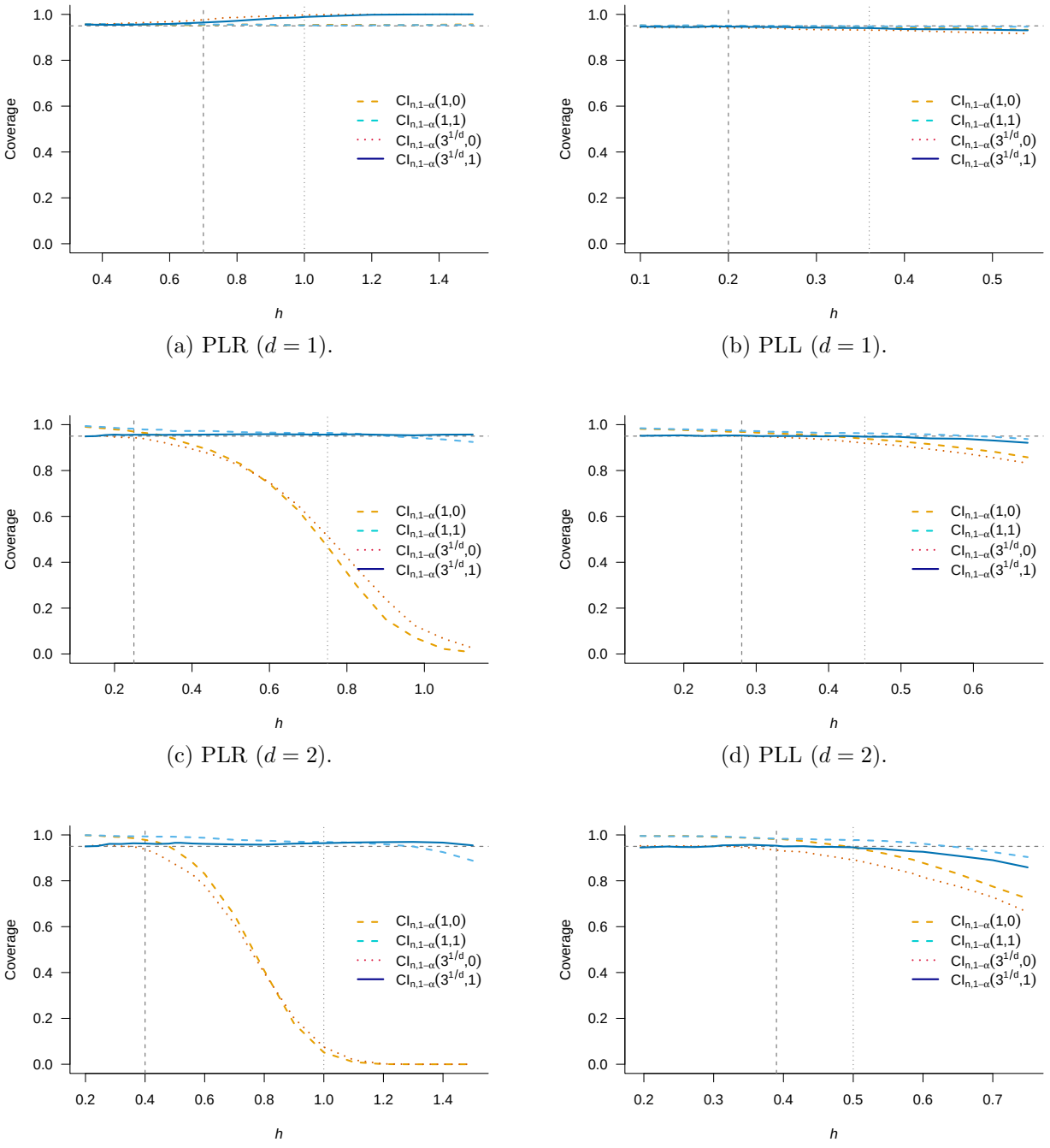
Note: The table shows coverage probability and average length of confidence intervals over 2,000 simulation replications. The sample size is $n = 2{,}000$ and the number of bootstrap draws is 2,000. The rows highlighted in yellow and orange correspond to the estimated MSE-optimal bandwidths for the point estimators with $L = 0$ and $L = 1$, respectively.

Table 6: Bootstrap 95% Confidence Intervals for Partially Linear Logit Model: DGP 3 ($d = 3$).

	$B = 1$				$B = 3^{1/d}$			
	$L = 0$		$L = 1$		$L = 0$		$L = 1$	
h	Coverage	Length	Coverage	Length	Coverage	Length	Coverage	Length
0.20	0.996	1.110	0.995	1.370	0.953	0.693	0.944	0.826
0.23	0.994	0.863	0.994	1.049	0.951	0.578	0.949	0.673
0.25	0.995	0.794	0.994	0.960	0.952	0.547	0.948	0.630
0.27	0.993	0.717	0.994	0.857	0.952	0.511	0.947	0.581
0.30	0.991	0.647	0.994	0.765	0.952	0.479	0.951	0.538
0.31	0.991	0.622	0.993	0.732	0.952	0.467	0.955	0.522
0.35	0.987	0.560	0.989	0.648	0.943	0.439	0.956	0.483
0.35	0.987	0.559	0.989	0.646	0.943	0.438	0.956	0.482
0.39	0.980	0.514	0.984	0.586	0.933	0.418	0.952	0.454
0.40	0.979	0.505	0.983	0.573	0.929	0.414	0.949	0.448
0.43	0.971	0.482	0.983	0.541	0.926	0.403	0.951	0.434
0.45	0.963	0.467	0.980	0.522	0.914	0.397	0.947	0.425
0.47	0.957	0.457	0.979	0.507	0.907	0.392	0.948	0.418
0.50	0.944	0.441	0.976	0.486	0.891	0.384	0.945	0.408
0.51	0.939	0.438	0.976	0.482	0.887	0.383	0.942	0.406
0.55	0.917	0.423	0.972	0.461	0.859	0.375	0.939	0.397
0.55	0.914	0.422	0.971	0.460	0.856	0.375	0.937	0.396
0.58	0.891	0.411	0.964	0.445	0.832	0.369	0.928	0.389
0.60	0.875	0.407	0.960	0.440	0.814	0.367	0.927	0.387
0.65	0.829	0.396	0.945	0.425	0.774	0.362	0.908	0.379
0.70	0.774	0.387	0.926	0.413	0.727	0.357	0.888	0.373
0.75	0.720	0.380	0.902	0.403	0.663	0.353	0.855	0.368

Note: The table shows coverage probability and average length of confidence intervals over 2,000 simulation replications. The sample size is $n = 2{,}000$ and the number of bootstrap draws is 2,000. The rows highlighted in yellow and orange correspond to the estimated MSE-optimal bandwidths for the point estimators with $L = 0$ and $L = 1$, respectively.

Figure 2: Coverage probabilities as a function of bandwidth h.



(a) PLR ($d = 1$). (b) PLL ($d = 1$).

(c) PLR ($d = 2$). (d) PLL ($d = 2$).

(e) PLR ($d = 3$). (f) PLL ($d = 3$).

Note: Each panel plots the empirical coverage probability of $\mathsf{CI}^*_{n,0.95}(B, L)$ against bandwidth h for the four procedures $(B, L) \in \{(1, 0), (1, 1), (3^{1/d}, 0), (3^{1/d}, 1)\}$. "PLR" stands for the partially linear regression model, and "PLL" stands for the partially linear logit model. d denotes the dimension of $\mathbf{w}$. The horizontal dotted line marks the 95% nominal level. The two vertical lines mark the estimated MSE-optimal bandwidths $h_{L=0}$ and $h_{L=1}$ for the point estimators $\widehat{\boldsymbol{\theta}}_n$ ($L = 0$) and $\widetilde{\boldsymbol{\theta}}_n$ ($L = 1$), respectively. The sample size is $n = 2{,}000$ and results are based on 2,000 simulation replications with 2,000 bootstrap samples.