



Contents lists available at ScienceDirect

# Journal of Econometrics

journal homepage: [www.elsevier.com/locate/jeconom](http://www.elsevier.com/locate/jeconom)



## Higher-order refinements of small bandwidth asymptotics for density-weighted average derivative estimators<sup>☆</sup>

Matias D. Cattaneo <sup>a,\*</sup>, Max H. Farrell <sup>b</sup>, Michael Jansson <sup>c</sup>, Ricardo P. Masini <sup>d</sup>

<sup>a</sup> Department of Operations Research and Financial Engineering, Princeton University, United States of America

<sup>b</sup> Department of Economics, UC Santa Barbara, United States of America

<sup>c</sup> Department of Economics, UC Berkeley, United States of America

<sup>d</sup> Department of Statistics, UC Davis, United States of America



### ARTICLE INFO

**Keywords:**

Density weighted average derivatives  
Edgeworth expansions  
Small bandwidth asymptotics

### ABSTRACT

The density weighted average derivative (DWAD) of a regression function is a canonical parameter of interest in economics. Classical first-order large sample distribution theory for kernel-based DWAD estimators relies on tuning parameter restrictions and model assumptions that imply an asymptotic linear representation of the point estimator. These conditions can be restrictive, and the resulting distributional approximation may not be representative of the actual sampling distribution of the statistic of interest. In particular, the approximation is not robust to bandwidth choice. Small bandwidth asymptotics offers an alternative, more general distributional approximation for kernel-based DWAD estimators that allows for, but does not require, asymptotic linearity. The resulting inference procedures based on small bandwidth asymptotics were found to exhibit superior finite sample performance in simulations, but no formal theory justifying that empirical success is available in the literature. Employing Edgeworth expansions, this paper shows that small bandwidth asymptotic approximations lead to inference procedures with higher-order distributional properties that are demonstrably superior to those of procedures based on asymptotic linear approximations.

### 1. Introduction

Identification, estimation, and inference in the context of semiparametric models has a long tradition in econometrics (Powell, 1994). Canonical two-step semiparametric estimands are finite dimensional functionals of some other unknown infinite dimensional parameters in the model (e.g., a density or regression function). A leading example of such a finite dimensional estimand is the density weighted average derivative (DWAD) of a regression function. This paper seeks to honor the many contributions of Jim Powell to semiparametric theory in econometrics by juxtaposing the higher-order distributional properties of Powell et al.'s (1989) two-step kernel-based DWAD estimator under two alternative large sample approximation regimes: one based on the classical

<sup>☆</sup> Prepared for the Conference in Honor of James L. Powell at UC Berkeley, March 25–26, 2022. We thank the conference participants for their comments, as well as two reviewers, an Associate Editor, and co-Editor Elie Tamer for their feedback. Cattaneo gratefully acknowledges financial support from the National Science Foundation, United States through grants SES-1947805 and DMS-2210561, Jansson gratefully acknowledges financial support from the National Science Foundation, United States through grant SES-1947662, and Masini gratefully acknowledges financial support from the National Science Foundation, United States through grant DMS-2210561.

\* Corresponding author.

E-mail address: [cattaneo@princeton.edu](mailto:cattaneo@princeton.edu) (M.D. Cattaneo).

asymptotic linear representation, and the other based on a more general quadratic distributional approximation known as *small bandwidth asymptotics*.<sup>1</sup>

In a landmark contribution, Powell et al. (1989) proposed a kernel-based DWAD estimator and obtained first-order, asymptotically linear distribution theory employing ideas from the U-statistics literature in statistics to develop valid inference procedures in large samples. This work sparked a wealth of subsequent developments in the econometrics literature: Robinson (1995) obtained Berry–Esseen bounds, Powell and Stoker (1996) considered mean square error expansions, Nishiyama and Robinson (2000, 2001, 2005) developed Edgeworth expansions, and Newey et al. (2004) investigated bias properties, just to mention a few contributions. The two-step semiparametric estimator in this literature employs a preliminary kernel-based estimator of a density function, which requires choosing two main tuning parameters: a bandwidth and a kernel function. The “optimal” choices for these tuning parameters depend on the goal of interest (e.g., point estimation vs. inference), as well as on the features of the underlying data generating process (e.g., smoothness of the unknown density and dimension of the covariates).

Classical first-order distribution theory for kernel-based DWAD estimators has focused on cases where tuning parameter restrictions and model assumptions imply an asymptotic linear representation of the two-step semiparametric point estimator (see Bickel et al., 1993; Newey and McFadden, 1994; Ichimura and Todd, 2007, for overviews). That is, the two-step estimator is approximated by a sample average based on an influence function. This approach can be used to construct semiparametrically efficient inference procedures, but requires potentially high smoothness levels of the underlying unknown functions, thereby forcing the use of higher-order kernels or other debiasing techniques. Further, the implied distributional approximation may not be “robust” to tuning parameter choices and/or model features. More specifically, the limiting distribution emerging from the asymptotic linear representation of the centered and scaled point estimator is invariant to the way that the preliminary nonparametric estimators are constructed. At its core, an asymptotic linear approximation assumes away the contribution of additional terms forming the statistic of interest, despite the fact that these terms do contribute to the sampling variability of the two-step semiparametric estimator and, more importantly, do reflect the impact of tuning parameter choices in finite samples.

Cattaneo et al. (2014b) proposed an alternative distributional approximation for kernel-based DWAD estimators that allows for, but does not require, asymptotic linearity. The idea is to capture the joint contribution to the sampling distribution of both linear and quadratic terms forming the kernel-based DWAD estimator, because the quadratic term explicitly captures the effect of the choice of bandwidth and kernel function. To operationalize this idea, Cattaneo et al. (2014b) introduced an asymptotic experiment where the bandwidth sequence is allowed (but not required) to vanish at a speed that would render the classical asymptotic linear representation invalid because the quadratic term becomes first order even in large samples, which they termed *small bandwidth asymptotics*. This framework was carefully developed to obtain a distributional approximation that explicitly depends on both linear and quadratic terms, thereby forcing a more careful analysis of how the nonparametric first stage contributes to the sampling distribution of the statistic.

Inference methods based on small bandwidth asymptotics for kernel-based DWAD estimators were found to perform well in simulations (Cattaneo et al., 2010, 2014b,a), but no formal justification for this finite sample success is available in the literature. Methodologically, this alternative distributional approximation leads to a new way of conducting inference (e.g., constructing confidence intervals) because the original standard error formula proposed by Powell et al. (1989) must be modified to make the asymptotic approximation valid across the full range of allowable bandwidths (including the region where asymptotic linearity fails). Theoretically, however, the empirical success of small bandwidth asymptotics could come from two distinct sources: (i) it could deliver a better distributional approximation to the sampling distribution of the point estimator; or (ii) it could deliver a better distributional approximation to the sampling distribution of the studentized t-statistic because the standard error formula is modified.

Employing Edgeworth expansions (Bhattacharya and Rao, 1976; Hall, 1992), this paper shows that the higher-order distributional properties of inference procedures motivated by the small bandwidth asymptotics approximation framework are demonstrably superior to those of procedures motivated by asymptotic linear approximations. We study both standardized and studentized estimators and show that those emerging from the small bandwidth regime offer higher-order corrections, as measured by the second cumulant underlying their Edgeworth expansions. An immediate implication of our results is that the small bandwidth asymptotic framework simultaneously enjoys two advantages: delivering a better distributional approximation (Theorem 1, standardized t-statistic) and leading to a better standard error construction (Theorem 2, studentized t-statistic). Therefore, our results have theoretical and practical implications for empirical work in economics, in addition to providing a theory-based explanation for prior simulation findings documenting better numerical performance of inference procedures motivated by small bandwidth asymptotics relative to those motivated by classical asymptotically linear distributional approximations.

The closest antecedent to our work is Nishiyama and Robinson (2000, 2001), who also studied Edgeworth expansions for kernel-based DWAD estimators. Their expansions, however, were motivated by the asymptotic linear approximation of the point estimator, and hence cannot be used to evaluate the distributional approximation emerging from the alternative small bandwidth asymptotic regime. Therefore, from a technical perspective, this paper also offers novel Edgeworth expansions that allow for different standardization and studentization schemes, thereby allowing us to plug-and-play when juxtaposing the two asymptotic approximation frameworks. More specifically, Theorem 1 below concerns a generic standardized t-statistic and is proven based on

<sup>1</sup> Jim Powell’s contributions to semiparametric theory are numerous. Honoré and Powell (1994), Powell and Stoker (1996), Blundell and Powell (2004), Aradillas-Lopez et al. (2007), Ahn et al. (2018), and Graham et al. (2024) are some of the most closely connected to the our work: these papers employ U-statistics methods for two-step kernel-based estimators similar to those considered herein. See Powell (2017) for more discussion and references.

[Appendix A](#), which may be of independent technical interest due to its generality. In contrast, [Theorem 2](#) below concerns a more specialized class of studentized t-statistics because establishing valid Edgeworth expansions is considerably harder when dealing with studentization.

The idea of employing more general asymptotic approximation frameworks that do not enforce asymptotic linearity for two-step semiparametric estimators has also featured in other contexts: (i) semi-linear series-based, many covariates, and many instruments estimation ([Cattaneo et al., 2018a,b](#)), (ii) non-linear two-step semiparametric estimation ([Cattaneo et al., 2013; Cattaneo and Jansson, 2018; Cattaneo et al., 2019; Cattaneo and Jansson, 2022](#)), and (iii) network estimation ([Matsushita and Otsu, 2021](#)). While our theoretical developments and results focus specifically on the case of kernel-based DWAD estimation, our main conceptual conclusions can be extrapolated to those settings as well. The main takeaway is that employing alternative asymptotic frameworks can deliver improved inference with smaller higher-order distributional approximation errors, thereby offering more robust inference procedures in finite samples. Furthermore, our theoretical and methodological results can also be leveraged to study the higher-order distributional properties of bootstrap-based methods for inference. Although a complete theoretical analysis is beyond the scope of this paper, we provide further discussion about the bootstrap in Section 4.

The paper continues as follows. Section 2 introduces the setup and main assumptions. Section 3 reviews the classical first-order distributional approximation based on asymptotic linearity and the more general small bandwidth distributional approximation, along with their corresponding choices of standard error formulas. Section 4 presents the main results of our paper. Section 5 concludes. The appendix is organized in four parts: [Appendix A](#) provides a self-contained generic Edgeworth expansion for second-order U-statistics, which may be of independent technical interest, and is proven in [Appendix B](#), [Appendix C](#) gives the proof of [Theorem 1](#), and [Appendix D](#) gives the proof of [Theorem 2](#).

## 2. Setup and assumptions

Suppose  $Z_i = (Y_i, X'_i)'$ ,  $i = 1, \dots, n$ , is a random sample from the distribution of the random vector  $Z = (Y, X')'$ , where  $Y$  is an outcome variable and  $X$  takes values on  $\mathbb{R}^d$  with Lebesgue density  $f$ . We consider

$$\theta := \mathbb{E}[f(X)\dot{g}(X)], \quad g(X) := \mathbb{E}[Y|X],$$

the DWAD of the regression function  $g$ , where, for any (differentiable) function  $a$ ,  $\dot{a}(x)$  denotes  $\partial a(x)/\partial x$ , and where existence of  $\theta$  is implied by parts (b) and (c) of the following assumption, which collects the regularity conditions under which our subsequent analysis will proceed.

**Assumption 1.** For some  $S \geq 2$ , the following are satisfied:

- (a)  $\mathbb{E}[|Y|^3] < \infty$ ;
- (b)  $f$  is  $(S+1)$  times differentiable, and  $f$  and its first  $(S+1)$  derivatives are bounded;
- (c)  $g$  is  $(S+1)$  times differentiable and its first three derivatives are bounded;
- (d)  $e$  and its first  $(S+1)$  derivatives are bounded, where  $e(X) := f(X)g(X)$ ;
- (e)  $\mathbb{E}[\mathbb{V}(Y|X)f(X)] > 0$  and  $\Sigma := \mathbb{V}[\psi(Z)]$  is positive definite, where  $\psi(Z) := 2[\dot{e}(X) - Y\dot{f}(X) - \theta]$  and where  $\mathbb{V}[\cdot]$  denotes the variance;
- (f)  $v$  is twice differentiable, its first two derivatives are bounded, and  $v\dot{f}$  and  $\mathbb{E}[|Y|^3|X]f(X)$  are bounded, where  $v(X) := \mathbb{E}[Y^2|X]$ ;
- (g)  $\limsup_{\|x\| \rightarrow \infty} [1 + v(x)]f(x) = 0$ , where  $\|\cdot\|$  is the Euclidean norm; and
- (h) Cramér Condition: For every  $v \in \mathbb{R}^d$ ,

$$\limsup_{|t| \rightarrow \infty} \left| \mathbb{E} [\exp(it\psi_v(Z))] \right| < 1,$$

where  $\psi_v(Z) := v'\psi(Z)$  and  $i^2 := -1$ .

Under [Assumption 1](#) and using integration by parts, the DWAD vector can be expressed as

$$\theta = -2\mathbb{E}[Y\dot{f}(X)],$$

which motivates the celebrated plug-in analog estimator of [Powell et al. \(1989\)](#) given by

$$\hat{\theta} = -2n^{-1} \sum_{1 \leq i \leq n} Y_i \frac{\partial}{\partial x} \hat{f}_i(X_i), \quad \hat{f}_i(x) = (n-1)^{-1} \sum_{\substack{1 \leq j \leq n \\ j \neq i}} \frac{1}{h^d} K\left(\frac{X_j - x}{h}\right),$$

where  $\hat{f}_i(\cdot)$  is a “leave-one-out” kernel density estimator employing a symmetric and differentiable kernel function  $K : \mathbb{R}^d \rightarrow \mathbb{R}$  and a positive vanishing (bandwidth) sequence  $h$ .

The estimator  $\hat{\theta}$  can be expressed as a second-order U-statistic with an  $n$ -varying kernel:

$$\hat{\theta} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} U_{ij}, \quad U_{ij} := -h^{-d-1} \dot{K}\left(\frac{X_i - X_j}{h}\right)(Y_i - Y_j). \quad (2.1)$$

Our analysis of  $\hat{\theta}$  is based on this representation and proceeds under the following assumption about the kernel function.<sup>2</sup>

**Assumption 2.** For some  $P \geq 2$  and some  $\{\mu_a : a \in \mathbb{Z}_+^d, |a| = P\} \subseteq \mathbb{R}$ , the following are satisfied:

- (a)  $K$  is even, differentiable, and  $\dot{K}$  is bounded;
- (b)  $\int_{\mathbb{R}^d} \dot{K}(u)\dot{K}(u)'du$  is positive definite; and
- (c)  $\int_{\mathbb{R}^d} |K(u)|(1 + \|u\|^P)du + \int_{\mathbb{R}^d} \|\dot{K}(u)\|(1 + \|u\|^2)du < \infty$  and

$$\int_{\mathbb{R}^d} u^a K(u)du = \begin{cases} 1, & \text{if } |a| = 0, \\ 0, & \text{if } 0 < |a| < P \\ \mu_a, & \text{if } |a| = P, \end{cases}$$

where  $a \in \mathbb{Z}_+^d$  is a multi-index.

### 3. First-order distribution theory

Before presenting our main results concerning the higher-order distributional properties of different statistics based on  $\hat{\theta}$ , we review conventional and alternative first-order asymptotic distributional approximations, as well as the distinct variance estimation methods emerging from each of those approximation frameworks. Limits are taken as  $h \rightarrow 0$  and  $n \rightarrow \infty$  unless otherwise noted,  $\rightarrow_{\mathbb{P}}$  denotes convergence in probability, and  $\rightsquigarrow$  denotes convergence in law.

#### 3.1. Distributional approximation

Under appropriate restrictions on  $h$  and  $K$ , the estimator  $\hat{\theta}$  is asymptotically linear with influence function  $\psi$  and asymptotic variance  $\Sigma$ . More precisely, Powell et al. (1989) showed that if Assumptions 1 and 2 hold and if  $nh^{2(P \wedge S)} \rightarrow 0$  and  $nh^{d+2} \rightarrow \infty$  (where  $a \wedge b$  denotes  $\min(a, b)$ ), then

$$\sqrt{n}(\hat{\theta} - \theta) = n^{-1/2} \sum_{1 \leq i \leq n} \psi(Z_i) + o_{\mathbb{P}}(1) \rightsquigarrow \mathcal{N}(0, \Sigma). \quad (3.1)$$

A proof of (3.1) can be based on the  $U$ -statistic representation in (2.1) and its Hoeffding decomposition  $\hat{\theta} = \mathbb{E}[U_{ij}] + \bar{L} + \bar{Q}$ , where  $\bar{L}$  and  $\bar{Q}$  are mean zero random vectors given by

$$\bar{L} := n^{-1} \sum_{1 \leq i \leq n} L_i, \quad L_i := 2(\mathbb{E}[U_{ij}|Z_i] - \mathbb{E}[U_{ij}]),$$

and

$$\bar{Q} := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} Q_{ij}, \quad Q_{ij} := U_{ij} - \mathbb{E}[U_{ij}|Z_i] - \mathbb{E}[U_{ij}|Z_j] + \mathbb{E}[U_{ij}],$$

respectively: because  $\mathbb{E}[U_{ij}] = \theta + O(h^{P \wedge S})$  and  $\mathbb{V}[\bar{Q}] = O(n^{-2}h^{-d-2})$ , we have

$$\sqrt{n}(\hat{\theta} - \theta) = n^{-1/2} \sum_{i=1}^n L_i + O_{\mathbb{P}}\left(\sqrt{nh^{P \wedge S}} + \frac{1}{\sqrt{nh^{d+2}}}\right),$$

from which the result (3.1) follows upon noting that  $\mathbb{V}[L_i - \psi(Z_i)] = O(h^{P \wedge S})$ . Using Edgeworth expansions, Nishiyama and Robinson (2000, 2001) studied the quality of the distributional approximation implied by (3.1); their result is contained as a special case of our Theorem 1.

The Hoeffding decomposition and subsequent analysis of each of its terms shows that the estimator admits a bilinear form representation in general, which then is reduced to a sample average approximation by assuming a bandwidth sequence and kernel shape that makes both the misspecification error (smoothing bias) and the variability introduced by  $\bar{Q}$  (a “quadratic” term) negligible in large samples. As a result, provided that such tuning parameter choices are feasible, the estimator will be asymptotically linear.

Asymptotic linearity of a semiparametric estimator has several distinct features that may be considered attractive from a theoretical point of view. In particular, it is a necessary condition for semiparametric efficiency, and it leads to a limiting distribution that is invariant to the choice of the first-step nonparametric estimator entering the two-step semiparametric procedure (Newey, 1994). However, insisting on asymptotic linearity may also have its drawbacks because it requires several potentially strong assumptions, and because it leads to a large sample theory that may not accurately represent the finite sample behavior of the statistic. In the case of  $\hat{\theta}$ , asymptotic linearity requires  $P > 2$  unless  $d = 1$ ; that is, the use of higher-order kernels or similar debiasing techniques (see, e.g., Chernozhukov et al., 2022, and references therein) is necessary in order to achieve asymptotic linearity. In addition, asymptotic linearity leads to a limiting experiment which is invariant to the particular choices of smoothing ( $K$ ) and bandwidth ( $h$ ) tuning parameters involved in the construction of the estimator. As a result, large sample distribution theory

<sup>2</sup> In Assumption 2(c) and elsewhere, we employ standard multi-index notation: For  $a := (a_1, \dots, a_d)' \in \mathbb{Z}_+^d$ , we have (i)  $|a| := a_1 + \dots + a_d$ , (ii)  $a! := a_1! \dots a_d!$ , (iii)  $x^a := x_1^{a_1} \dots x_d^{a_d}$  for  $x := (x_1, \dots, x_d)' \in \mathbb{R}^d$ , and (iv)  $\partial^a q(x)/\partial x^a := \partial^{|a|} q(x)/(x_1^{a_1} \dots x_d^{a_d})$  for (sufficiently smooth)  $q : \mathbb{R}^d \rightarrow \mathbb{R}$ .

based on (or implying) asymptotic linearity is silent with respect to the impact that tuning parameter choices may have on the finite sample behavior of the two-step semiparametric statistic.

To address the aforementioned limitations of distribution theory based on asymptotic linearity, Cattaneo et al. (2014b) proposed a more general distributional approximation for kernel-based DWAD estimators that accommodates, but does not enforce, asymptotic linearity. The idea is to characterize the joint asymptotic distributional features of both the linear ( $\bar{L}$ ) and quadratic ( $\bar{Q}$ ) terms, and in the process develop a more general first-order asymptotic theory that allows for weaker assumptions than those imposed in the classical asymptotically linear distribution theory. Formally, if Assumptions 1 and 2 hold, and if  $(nh^{d+2} \wedge 1)nh^{2(P \wedge S)} \rightarrow 0$  and  $n^2h^d \rightarrow \infty$ , then

$$\mathbb{V}[\hat{\theta}]^{-1/2}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, I), \quad (3.2)$$

where  $\mathbb{V}[\hat{\theta}] = \mathbb{V}[\bar{L}] + \mathbb{V}[\bar{Q}]$  with

$$\mathbb{V}[\bar{L}] = n^{-1} [\Sigma + O(h^{P \wedge S})]$$

and

$$\mathbb{V}[\bar{Q}] = \binom{n}{2}^{-1} h^{-d-2} [\Delta + O(h^2)], \quad \Delta := 2\mathbb{E}[\mathbb{V}(Y|X)f(X)] \int_{\mathbb{R}^d} \bar{K}(u)\bar{K}(u)'du.$$

This more general distributional approximation was developed explicitly in an attempt to better characterize the finite sample behavior of  $\hat{\theta}$ . The result in (3.2) shows that the conditions on the bandwidth sequence may be considerably weakened without invalidating the limiting Gaussian distribution, although the asymptotic variance formula changes. Importantly, if  $nh^{d+2}$  is bounded then  $\hat{\theta}$  is no longer asymptotically linear and its limiting distribution will cease to be invariant with respect to the underlying preliminary nonparametric estimator. In particular, if  $nh^{d+2} \rightarrow c > 0$  then  $\hat{\theta}$  is root- $n$  consistent, but not asymptotically linear. The bias of the estimator is also controlled in a different way because the bandwidth is allowed to be “smaller” than usual, which may remove the need for higher-order kernels. Interestingly, (3.2) allows for the point estimator to not even be consistent for  $\theta$ , which occurs for sufficiently small bandwidth sequences.

Beyond the aforementioned technical considerations, the result in (3.2) can conceptually be interpreted as a more refined first-order distributional approximation for  $\hat{\theta}$ , which by relying on a quadratic approximation (i.e., accounting for the contributions of both  $\bar{L}$  and  $\bar{Q}$ ) is expected to offer a “better” distributional approximation than approximations relying on asymptotic linearity (i.e., accounting only for the contribution of  $\bar{L}$ ). The idea of standardizing a U-statistic by the joint variance of the linear and quadratic terms underlying its Hoeffding decomposition can be traced back to the original paper of Hoeffding (1948, p. 307). Simulation evidence reported in Cattaneo et al. (2010, 2014b,a) corroborated those conceptual interpretations numerically, but no formal justification is available in the literature. Theorem 1 below will offer the first theoretical result in the literature highlighting specific robustness features of the distributional approximation in (3.2) by showing that such approximation has a demonstrably smaller higher-order distributional approximation error.

### 3.2. Variance estimation

Motivated by the asymptotic linearity result (3.1), Powell et al. (1989) also proposed the “plug-in” variance estimator

$$\hat{\Sigma} := n^{-1} \sum_{1 \leq i \leq n} \hat{L}_i \hat{L}'_i, \quad \hat{L}_i := 2 \left[ (n-1)^{-1} \sum_{\substack{1 \leq j \leq n \\ j \neq i}} U_{ij} - \hat{\theta} \right],$$

and proved its consistency (i.e.,  $\hat{\Sigma} \rightarrow_{\mathbb{P}} \Sigma$ ) under the same bandwidth sequences required for asymptotic linearity (i.e., assuming  $nh^{2(P \wedge S)} \rightarrow 0$  and  $nh^{d+2} \rightarrow \infty$ ). Combining this consistency result with (3.1), we obtain the following result about a studentized version of  $\hat{\theta}$ :

$$\hat{V}_{AL}^{-1/2}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, I), \quad \hat{V}_{AL} := n^{-1} \hat{\Sigma}. \quad (3.3)$$

Using Edgeworth expansions, Nishiyama and Robinson (2000, 2001) studied the quality of the distributional approximation implied by (3.3); their result is contained as a special case of our Theorem 2.

Complementing the small bandwidth asymptotic representation (3.2), Cattaneo et al. (2014b) showed that

$$\hat{V}_{AL} = n^{-1} [\Sigma + o_{\mathbb{P}}(1)] + 2 \binom{n}{2}^{-1} h^{-d-2} [\Delta + o_{\mathbb{P}}(1)],$$

which implies among other things that the consistency result  $\hat{\Sigma} \rightarrow_{\mathbb{P}} \Sigma$  is valid only if  $nh^{d+2} \rightarrow \infty$ ; otherwise,  $\hat{\Sigma}$  is in general asymptotically upwards biased relative to  $\mathbb{V}[\hat{\theta}]$  in (3.2). Because  $\hat{\Sigma}$  is asymptotically equivalent to the jackknife variance estimator of  $\hat{\theta}$ , Cattaneo et al. (2014a) also noted that the asymptotic bias of  $\hat{\Sigma}$  is a consequence of a more generic phenomena underlying jackknife variance estimators studied in Efron and Stein (1981). See also Matsushita and Otsu (2021) for related discussion.

To conduct asymptotically valid inference under the more general small bandwidth asymptotic regime, Cattaneo et al. (2014b) proposed several “debiased” variance estimators, including

$$\hat{V}_{SB} := n^{-1} \hat{\Sigma} - \binom{n}{2}^{-1} h^{-d-2} \hat{\Delta}, \quad \hat{\Delta} := h^{d+2} \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} U_{ij} U'_{ij},$$

and showed that  $\hat{\Delta} \rightarrow_{\mathbb{P}} \Delta$  under the same bandwidth sequences required for (3.2) to hold (i.e., assuming  $nh^{2(P \wedge S)} \rightarrow 0$  and  $n^2 h^d \rightarrow \infty$ ). The estimator  $\hat{\Delta}$  is asymptotically equivalent to the debiasing procedure proposed in Efron and Stein (1981). By design, the result

$$\hat{V}_{\text{SB}}^{-1/2}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}(0, I) \quad (3.4)$$

holds under more general conditions than those required for (3.3), suggesting that inference procedures based on  $\hat{V}_{\text{SB}}$  are more “robust” than procedures based on  $\hat{V}_{\text{AL}}$ .

Conceptually, robustness manifests itself in two distinct ways. First, the underlying Gaussian distributional approximation holds under weaker bandwidth restrictions, a property achieved in part by employing a standardization factor depending explicitly on tuning parameter choices. Second, the new variance estimator  $\hat{V}_{\text{SB}}$  is obtained from the more general small bandwidth approximation and explicitly accounts for the contribution of terms regarded as higher-order under asymptotic linearity.

While not reproduced here to conserve space, the in-depth Monte Carlo evidence reported in Cattaneo et al. (2010, 2014b,a) also showed that employing inference procedures based on (3.4) lead to remarkable improvements in terms of “robustness” to bandwidth choice and other tuning inputs, when compared to classical asymptotically linear inference procedures based on (3.3). Theorem 2 below will show formally that the distributional approximation (3.4) has demonstrably smaller higher-order errors than the distributional approximation (3.3), thereby providing a theory-based explanation for the empirical success of feasible inference procedures developed under the small bandwidth approximation framework.

#### 4. Higher-order distribution theory

Letting  $v \in \mathbb{R}^d$  be a fixed vector and defining  $\hat{\theta}_v := v' \hat{\theta}$ , this section presents Edgeworth expansions for standardized and studentized statistics based on  $\hat{\theta}_v$ . Section 4.1 studies standardized statistics of the form  $(\hat{\theta}_v - \theta_v)/\vartheta_v$ , where  $\theta_v := v' \theta$  and  $\vartheta_v$  is an approximate standard deviation of  $\hat{\theta}_v$ ; that is,  $\vartheta_v$  is positive, non-random, and such that  $(\hat{\theta}_v - \theta_v)/\vartheta_v$  is asymptotically standard normal. The main purpose of studying standardized statistics is to allow us to compare the quality of the distributional approximations (3.1) and (3.2) based on asymptotic linearity and small bandwidth asymptotics, respectively. Section 4.2 then studies studentized statistics of the form  $(\hat{\theta}_v - \theta_v)/\hat{\delta}_v^2$ , where  $\hat{\delta}_v^2$  is (random and) equal to either  $v' \hat{V}_{\text{AL}} v$  or  $v' \hat{V}_{\text{SB}} v$ . The main purpose of studying studentized statistics is to allow us to investigate the impact of variance estimation on the quality of the distributional approximations (3.3) and (3.4) based on asymptotic linearity and small bandwidth asymptotics, respectively.

Nishiyama and Robinson (2000, 2001) obtained valid Edgeworth expansions for the distribution of the standardized and studentized statistics employing  $\vartheta_v^2 = v' \Sigma v/n$  and  $\hat{\delta}_v^2 = v' \hat{\Sigma} v/n$ , respectively. Those results were obtained under assumptions implying asymptotic linearity. Although our main interest is in standardization and studentization schemes whose (first-order) validity does not require asymptotic linearity, we retain the assumption of asymptotic linearity to ensure a fair comparison; that is, our results are derived under the same assumptions as those imposed in prior work, in which case all inference procedures are asymptotically valid, and therefore amenable to juxtaposition. While beyond the scope of this paper, allowing for departures from asymptotic linearity is an interesting topic for future research.

##### 4.1. Standardized statistics

Suppressing the dependence on  $v$  and  $\vartheta_v$ , let

$$F(x) := \mathbb{P}\left[\frac{\hat{\theta}_v - \theta_v}{\vartheta_v} \leq x\right], \quad x \in \mathbb{R},$$

be the cumulative distribution function (cdf) of  $(\hat{\theta}_v - \theta_v)/\vartheta_v$ , where  $\vartheta_v$  is positive and non-random. Letting  $\Phi$  denote the standard normal cdf, it follows from (3.2) that

$$\sup_{x \in \mathbb{R}} |F(x) - \Phi(x)| = o(1) \quad (4.1)$$

under assumptions implying in particular that  $\omega_v^2/\vartheta_v^2 \rightarrow 1$ , where

$$\omega_v^2 := \mathbb{V}[\hat{\theta}_v] = n^{-1} [\sigma_v^2 + O(h^{P \wedge S})] + \binom{n}{2}^{-1} h^{-d-2} [\delta_v^2 + O(h^2)],$$

with  $\sigma_v^2 := v' \Sigma v$  and  $\delta_v^2 := v' \Delta v$ .

Our first theorem provides a refinement of (4.1). To state the theorem, let

$$\dot{f}_v(X) := v' \dot{f}(X), \quad \varphi_v(Z) := \psi_v(Z) + 2\theta_v, \quad \eta_v(Z_2) := \lim_{n \rightarrow \infty} \mathbb{E}[\varphi_v(Z_1)v' U_{12}|Z_2],$$

and define the following quantities (all of which are finite under the assumptions of Theorem 1):

$$\beta_v := 2(-1)^P \sum_{a \in \mathbb{Z}_+^d, |a|=P} \frac{\mu_a}{a!} \mathbb{E}\left[g(X) \frac{\partial^a}{\partial X^a} \dot{f}_v(X)\right], \quad \kappa_{1,v} := \mathbb{E}[\psi_v(Z)^3],$$

$$\kappa_{2,v} := \mathbb{E}[\varphi_v(Z)\eta_v(Z)] - \mathbb{E}[\varphi_v(Z)^2]\theta_v - \mathbb{V}[\varphi_v(Z)]\theta_v.$$

Also, let  $\phi$  denote the standard normal probability density function.

**Theorem 1 (Standardization).** Suppose Assumptions 1 and 2 hold with  $S \geq P$ , and that  $nh^{2P} \rightarrow 0$  and  $nh^{d+2}/\log^9 n \rightarrow \infty$ . If  $\theta_v$  is positive and non-random with  $\omega_v^2/\theta_v^2 \rightarrow 1$ , then

$$\sup_{x \in \mathbb{R}} |F(x) - G(x)| = o(r_n), \quad r_n := \sqrt{nh^P} + \frac{1}{nh^{d+2}} + \frac{1}{\sqrt{n}},$$

with

$$G(x) := \Phi(x) - \phi(x) \left[ \frac{\sqrt{nh^P} \beta_v}{\sigma_v} + \frac{\omega_v^2/\theta_v^2 - 1}{2} x + \frac{\kappa_{1,v} + \kappa_{2,v}}{6\sqrt{n}\sigma_v^3} (x^2 - 1) \right].$$

The proof of the theorem proceeds by verifying the high-level conditions of more general results presented in Appendix A. The general results establish a valid Edgeworth expansion for a generic class of U-statistics with  $n$ -varying kernels and may be of independent theoretical interest. Theorem 1 generalizes Nishiyama and Robinson (2000, Theorem 1) by allowing for a generic standardization factors  $\theta_v$  instead of their specific choice  $\sqrt{v' \Sigma v/n} = \sigma_v/\sqrt{n}$ . The latter generalization is important for our purposes, as it enables us to compare the different distributional approximations implied by (3.1) and (3.2).

As is customary with Edgeworth expansions, the square-bracketed term in the function  $G$  is a “correction” term capturing the extent to which the first three cumulants of the statistic differ from those of the standard normal distribution. To be specific, the first and third terms correct for bias and skewness, respectively. None of these correction terms depend on the particular  $\theta_v$  used for standardization purposes. In contrast, and as was to be expected, the variance correction term does depend on  $\theta_v$ , being proportional to  $\omega_v^2/\theta_v^2 - 1$ .

The asymptotic linearity result (3.1) suggests setting  $\theta_v^2 = \sigma_v^2/n$ . Doing so, and in agreement with Nishiyama and Robinson (2000, Theorem 1), we have

$$\omega_v^2/\theta_v^2 - 1 \approx \frac{2\delta_v^2}{nh^{d+2}\sigma_v^2},$$

the approximation error being  $o(r_n)$ . In the display, the term on the right hand side involves  $\delta_v^2 = \lim_{n \rightarrow \infty} h^{d+2} \mathbb{V}[v' Q_{ij} v]$  and is therefore interpretable as a variance correction term intrinsically associated with approximations based on asymptotic linearity, as such approximations ignore the contribution of the “quadratic” terms  $Q_{ij}$  to the variability of  $\hat{\theta}$ . Unlike (3.1), the small bandwidth formulation (3.2) explicitly accounts for the presence of “quadratic” terms and the standardization factor  $\theta_v^2 = \mathbb{V}[\hat{\theta}_v] = \omega_v^2$  suggested by the small bandwidth formulation is one for which the variance correction term in  $G$  vanishes altogether.

Although the details of the results reported here are specific to DWAD estimation, one important qualitative conclusion appears to generalize: the feature that it can be advantageous (in a higher-order sense) to capture the full variability of a statistic when approximating its distribution is known to be shared by certain statistics arising in the context of nonparametric kernel-based density and local polynomial regression inference; for details, see Calonico et al. (2018, 2022).

In isolation, Theorem 1 is mostly of theoretical interest, the reason being that it is concerned with standardized (as opposed to studentized) estimators. The consequences of employing studentization (i.e., replacing  $\theta_v$  with an estimator) will be explored in the next subsection. One important qualitative conclusion of that subsection concerns inference. That conclusion can be anticipated with the help of Theorem 1. We conclude this subsection by doing so.

For any  $\alpha \in (0, 1)$ , a natural (albeit infeasible)  $100(1 - \alpha)\%$  two-sided confidence interval for  $\theta_v$  has endpoints given by  $\hat{\theta}_v \pm c_\alpha \theta_v$ , where  $c_\alpha := \Phi^{-1}(1 - \alpha/2)$  and where  $\theta_v^2$  is an approximate variance of  $\hat{\theta}_v$ . Under the assumptions of Theorem 1, the coverage probability of this interval satisfies

$$\mathbb{P} \left[ \hat{\theta}_v - c_\alpha \theta_v \leq \theta_v \leq \hat{\theta}_v + c_\alpha \theta_v \right] = 1 - \alpha - (\omega_v^2/\theta_v^2 - 1) \phi(c_\alpha) c_\alpha + o(r_n),$$

so to the order considered the coverage error is proportional to the term  $\omega_v^2/\theta_v^2 - 1$  discussed previously and our conclusions about this term therefore apply directly. In particular, the coverage error of an infeasible interval using  $\theta_v^2 = \mathbb{V}[\hat{\theta}_v]$  (as suggested by the small bandwidth asymptotic result (3.2)) is  $o(r_n)$ , while the coverage errors of an infeasible intervals using  $\theta_v^2 = \sigma_v^2/n$  (as suggested by the asymptotic linearity result (3.1)) or its pre-asymptotic counterpart  $\theta_v^2 = \mathbb{V}[v' L_i]/n$  are of larger magnitude.

#### 4.2. Studentized statistics

Next, we investigate the role of variance estimation by obtaining Edgeworth expansions for studentized versions of  $\hat{\theta}$ . For specificity, and inspired by (3.3) and (3.4), we compare

$$\hat{F}_{AL}(x) := \mathbb{P} \left[ \frac{\hat{\theta}_v - \theta_v}{\hat{\theta}_{AL,v}} \leq x \right], \quad \hat{\theta}_{AL,v}^2 := v' \hat{V}_{AL} v$$

and

$$\hat{F}_{SB}(x) := \mathbb{P} \left[ \frac{\hat{\theta}_v - \theta_v}{\hat{\theta}_{SB,v}} \leq x \right], \quad \hat{\theta}_{SB,v}^2 := v' \hat{V}_{SB} v.$$

Studying  $\hat{F}_{AL}$ , Nishiyama and Robinson (2000, Theorem 3) found that if the assumptions of Theorem 2 below are satisfied, then

$$\sup_{x \in \mathbb{R}} |\hat{F}_{AL}(x) - \hat{G}_{AL}(x)| = o(r_n),$$

with

$$\hat{G}_{\text{AL}}(x) := \Phi(x) - \phi(x) \left[ \left( \frac{\sqrt{n}h^P \beta_v}{\sigma_v} - \frac{3\kappa_{1,v} + 2\kappa_{2,v}}{6\sqrt{n}\sigma_v^3} \right) - \frac{\delta_v^2}{nh^{d+2}\sigma_v^2} x - \frac{2\kappa_{1,v} + \kappa_{2,v}}{6\sqrt{n}\sigma_v^3} (x^2 - 1) \right],$$

where, once again, the three terms in square brackets correct for bias, variance, and skewness, respectively. In light of the results of the previous subsection, one would expect the Edgeworth approximation to  $\hat{F}_{\text{SB}}$  to be similar to  $\hat{G}_{\text{AL}}$ , the only (possible) difference being the variance correction term. The following result shows that this is indeed the case.

**Theorem 2 (Studentization).** Suppose Assumptions 1 and 2 hold with  $S \geq P$  and  $\mathbb{E}[Y^6] < \infty$ , and that  $nh^{2P} \rightarrow 0$  and  $nh^{d+2}/\log^9 n \rightarrow \infty$ . Then

$$\sup_{x \in \mathbb{R}} |\hat{F}_{\text{SB}}(x) - \hat{G}_{\text{SB}}(x)| = o(r_n),$$

with

$$\hat{G}_{\text{SB}}(x) := \Phi(x) - \phi(x) \left[ \left( \frac{\sqrt{n}h^P \beta_v}{\sigma_v} - \frac{3\kappa_{1,v} + 2\kappa_{2,v}}{6\sqrt{n}\sigma_v^3} \right) - \frac{2\kappa_{1,v} + \kappa_{2,v}}{6\sqrt{n}\sigma_v^3} (x^2 - 1) \right]$$

This theorem shows that employing studentization based on small bandwidth asymptotics offers demonstrable improvements in terms of distributional approximations for the resulting feasible  $t$ -test: the variance correction term present in  $\hat{G}_{\text{AL}}$  is absent from  $\hat{G}_{\text{SB}}$ . As in Nishiyama and Robinson (2000, 2001), the result is obtained under the somewhat stronger moment condition  $\mathbb{E}[Y^6] < \infty$  than the condition  $\mathbb{E}[|Y|^3] < \infty$  of Theorem 1, the purpose of the strengthened condition being to help control the contribution of the random denominator of the studentized version of  $\hat{\theta}$ .

The main practical implication of Theorem 2 can be illustrated by analyzing the coverage error of  $100(1-\alpha)\%$  confidence intervals with endpoints  $\hat{\theta}_v \pm c_\alpha \hat{\delta}_v$ . Setting  $\hat{\theta}_v = \hat{\theta}_{\text{AL},v}$  and applying Nishiyama and Robinson (2000, Theorem 3), we have

$$\mathbb{P} \left[ \hat{\theta}_v - c_\alpha \hat{\delta}_{\text{AL},v} \leq \theta_v \leq \hat{\theta}_v + c_\alpha \hat{\delta}_{\text{AL},v} \right] = 1 - \alpha + \frac{2\delta_v^2}{nh^{d+2}\sigma_v^2} \phi(c_\alpha)c_\alpha + o(r_n),$$

whereas setting  $\hat{\delta}_v = \hat{\delta}_{\text{SB},v}$  and applying Theorem 2 gives

$$\mathbb{P} \left[ \hat{\theta}_v - c_\alpha \hat{\delta}_{\text{SB},v} \leq \theta_v \leq \hat{\theta}_v + c_\alpha \hat{\delta}_{\text{SB},v} \right] = 1 - \alpha + o(r_n).$$

In other words, confidence intervals based on (3.4) are demonstrably superior to those based on (3.3) from a higher-order asymptotic point of view. This finding provides a theoretical explanation of the simulation evidence reported in Cattaneo et al. (2014b,a, 2010), where feasible confidence intervals based on small bandwidth asymptotics were shown to offer better finite sample performance in terms of coverage error than their counterparts based on classical asymptotic linear approximations.

#### 4.3. Discussion and bootstrap-based inference

In the previous subsections, we investigated the higher-order performance of large sample distributional approximations under two alternative asymptotic frameworks: asymptotic linearity and small bandwidth asymptotics. We found that the choice of studentization matters in terms of distributional approximation errors, even under conditions guaranteeing that asymptotic linearity holds ( $nh^{d+2} \rightarrow \infty$ ), in which case both asymptotic frameworks are first-order valid. As a consequence, our Edgeworth expansions (reported in Theorems 1 and 2) provide alternative validation of the main conclusions obtained by Cattaneo et al. (2010, 2014b) using first-order distributional approximations: in terms of distributional approximation accuracy, the small bandwidth framework justifying (3.2) and (3.4) dominates the asymptotic linear framework justifying (3.1) and (3.3).

It is natural to ask whether a similar ranking emerges when employing bootstrap-based inference procedures. Cattaneo et al. (2014a) studied the first-order properties of the nonparametric bootstrap under small bandwidth asymptotics for the kernel-based DWAD estimator, and showed that a similar first-order pattern emerges in that case: Under slightly stronger assumptions than Assumptions 1 and 2, they showed that if  $(nh^{d+2} \wedge 1)nh^{2(P+S)} \rightarrow 0$  and if  $n^2h^d \rightarrow \infty$ , then

$$\mathbb{V}^*[\hat{\theta}^*]^{-1/2}(\hat{\theta}^* - \hat{\theta}) \rightsquigarrow_{\mathbb{P}} \mathcal{N}(0, I),$$

where  $\rightsquigarrow_{\mathbb{P}}$  denotes weak convergence in probability,

$$\mathbb{V}^*[\hat{\theta}^*] = \mathbb{V}^*[\bar{L}^*] + \mathbb{V}^*[\bar{Q}^*], \quad \mathbb{V}^*[\bar{L}^*] = n^{-1} [\Sigma + o_{\mathbb{P}}(1)], \quad \mathbb{V}^*[\bar{Q}^*] = 3 \binom{n}{2}^{-1} h^{-d-2} [\Delta + o_{\mathbb{P}}(1)],$$

and

$$n^{-1}\hat{\Sigma}^* = n^{-1}[\Sigma + o_{\mathbb{P}}(1)] + 4 \binom{n}{2}^{-1} h^{-d-2} [\Delta + o_{\mathbb{P}}(1)], \quad \hat{\Delta}^* = \Delta + o_{\mathbb{P}}(1),$$

with  $\hat{\theta}^*$ ,  $\bar{L}^*$ ,  $\bar{Q}^*$ ,  $\hat{\Sigma}^*$  and  $\hat{\Delta}^*$  denoting nonparametric bootstrap analogs of  $\hat{\theta}$ ,  $\bar{L}$ ,  $\bar{Q}$ ,  $\hat{\Sigma}$  and  $\hat{\Delta}$ , respectively, and  $\mathbb{V}^*[\cdot]$  denoting the variance computed conditional on the original data. It follows from those results that under small bandwidth asymptotics the bootstrap consistently estimates the distribution of a studentized version of  $\hat{\theta}$  when  $\hat{V}_{\text{SB}}$  is used for studentization purposes, but not

when  $\hat{V}_{AL}$  is. These conclusions are in perfect agreement with those discussed in Section 3, the only notable difference in the details being that the variability induced by the bootstrap is larger outside the asymptotic linear regime because  $\mathbb{V}^*[\bar{Q}^*]/\mathbb{V}[\bar{Q}] = 3 + o_{\mathbb{P}}(1)$ . Furthermore, the second term of the bootstrap-based jackknife variance estimator  $\hat{\Sigma}^*$  is asymptotically doubled relative to the second term of the jackknife variance estimator  $\hat{\Sigma}$ .

Nishiyama and Robinson (2005) used Edgeworth expansions to study the properties of inference procedures based on the nonparametric bootstrap under assumptions implying asymptotic linearity. Based on the findings in this paper and those in Cattaneo et al. (2014a), we conjecture that analogous conclusions to those obtained herein will be valid for the case of bootstrap-based inference. While the conceptual parallelism between bootstrap-based inference and the results reported in this paper are clear, formalizing our conjecture requires substantial additional technical work due to the added complications associated with the data resampling, and hence we leave the theoretical analysis for future work.

## 5. Conclusion

Employing Edgeworth expansions, we compared the higher-order properties of two first-order distributional approximations and their associated confidence intervals for the kernel-based DWAD estimator of Powell et al. (1989). We showed that small bandwidth asymptotics not only give demonstrably better distributional approximations than those implied by asymptotic linearity, but also justifies employing a variance estimator for studentization purposes that improves the distributional approximation. The main takeaway from our results is that in two-step semiparametric settings, and related problems, alternative asymptotic approximations that capture higher-order terms, which are ignored by more traditional asymptotic linearity-based approximations, can deliver better distributional approximations and, by implication, more accurate inference procedures in finite samples. See Cattaneo et al. (2018a) for related discussion.

While beyond the scope of this paper, it would be of interest to develop analogous Edgeworth expansions for more general linear and non-linear two-step semiparametric procedures employing either Gaussian or resampling approximations under both conventional and alternative asymptotic frameworks (Cattaneo et al., 2013; Cattaneo and Jansson, 2018; Cattaneo et al., 2019; Cattaneo and Jansson, 2022). In particular, for the special case of kernel-based DWAD estimators, which is a linear two-step kernel-based semiparametric estimator, Nishiyama and Robinson (2005) already obtained Edgeworth expansions for bootstrap-based inference procedures under asymptotic linearity that could be contrasted with those obtained under small bandwidth asymptotics (Cattaneo et al., 2014a), after establishing more general Edgeworth expansions accounting for the bootstrap.

## Appendix A. Second-order U-statistics

Let  $U_n$  be a second-order U-statistic with  $n$ -varying kernel:

$$U_n := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} u_n(Z_i, Z_j),$$

where  $Z_1, Z_2, \dots$  are *i.i.d.* copies of a random vector  $Z$  and where, for each  $n \geq 2$ ,  $u_n$  is a permutation symmetric  $\mathbb{R}$ -valued function. The main result of this section concerns the distribution of the (approximately) standardized statistic  $(U_n - \theta_n)/\vartheta_n$ , where  $\theta_n \in \mathbb{R}$  and  $\vartheta_n > 0$  are non-random. To be specific, dropping the subscript  $n$  to simplify notation and defining

$$F(x) := \mathbb{P}\left[\frac{U - \theta}{\vartheta} \leq x\right], \quad x \in \mathbb{R},$$

our objective is to obtain a valid Edgeworth expansion for  $F$ .

When stating and proving the result, it is useful to employ the Hoeffding decomposition

$$U - \theta = B + L + Q,$$

where

$$B := \mathbb{E}[U] - \theta, \quad L := n^{-1} \sum_{i=1}^n \ell_i, \quad Q := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} q_{ij},$$

and where, for  $1 \leq i < j \leq n$ ,

$$\ell_i := 2(\mathbb{E}[u(Z_i, Z_j)|Z_i] - \mathbb{E}[u(Z_i, Z_j)])$$

and

$$q_{ij} := u(Z_i, Z_j) - \frac{1}{2}(\ell(Z_i) + \ell(Z_j)) - \mathbb{E}[u(Z_i, Z_j)].$$

Also, it is useful to define  $\sigma_\ell^2 := \mathbb{E}[\ell^2]$ ,  $\sigma_q^2 := \mathbb{E}[q_{12}^2]$ ,  $x_1 := \mathbb{E}[\ell_1^3]$ , and  $x_2 := \mathbb{E}[\ell_1 \ell_2 q_{12}]$ , and to note that

$$\omega^2 := \mathbb{V}[U] = \mathbb{V}[L] + \mathbb{V}[Q] = n^{-1} \sigma_\ell^2 + \binom{n}{2}^{-1} \sigma_q^2 = n^{-1} \sigma_\ell^2 \left[ 1 + O(n^{-1} \sigma_q^2 / \sigma_\ell^2) \right].$$

**Theorem A.1.** For some  $p \in (2, 3]$ , let the following conditions hold:

- (a)  $\mathbb{E} \left[ |\ell_1/\sigma_\ell|^3 \right] = O(1)$  and  $\mathbb{E}[|q_{12}/\sigma_\ell|^p] < \infty$ ;  
(b)  $n^{-1}\sigma_\ell^2/\omega^2 \rightarrow 1$  and  $\omega^2/\theta^2 \rightarrow 1$ ;  
(c) For every  $\underline{c}, \bar{c} > 0$ ,

$$\limsup_{n \rightarrow \infty} \sup_{\underline{c} < |t| \leq \bar{c} \log n} |\mathbb{E}[\exp(it\ell_1/\sigma_\ell)]| < 1.$$

Then

$$\sup_{x \in \mathbb{R}} |F(x) - G(x)| = O \left( \mathcal{E} + \frac{1}{\sqrt{n} \log n} \right),$$

where  $G$  is the distribution function with characteristic function

$$\chi_G(t) := \exp \left( it\gamma_1 - \frac{t^2}{2} \right) \left[ 1 + \sum_{2 \leq j \leq 9} (it)^j \gamma_j \right],$$

with

$$\begin{aligned} \gamma_1 &:= \frac{B}{\theta}, \quad \gamma_2 := \frac{\omega^2 - \theta^2}{2\theta^2}, \quad \gamma_3 := \frac{x_1 + 6x_2}{6n^2\theta^3}, \\ \gamma_4 &:= \frac{\omega^2 - \theta^2}{4\theta^4} \binom{n}{2}^{-1} \sigma_q^2, \quad \gamma_5 := \frac{1}{12n^2\theta^5} \left[ \binom{n}{2}^{-1} x_1 \sigma_q^2 + 6(n^{-1}\sigma_\ell^2 - \theta^2)x_2 \right], \\ \gamma_6 &:= \frac{1}{6n^4\theta^6} \left[ x_1 x_2 + 12 \binom{n}{2}^{-2} \binom{n}{4} x_2^2 \right], \quad \gamma_7 := 0, \\ \gamma_8 &:= \frac{n^{-1}\sigma_\ell^2 - \theta^2}{4n^4\theta^8} \binom{n}{2}^{-2} \binom{n}{4} x_2^2, \quad \gamma_9 := \frac{1}{12n^6\theta^9} \binom{n}{2}^{-2} \binom{n}{4} x_1 x_2^2, \end{aligned}$$

and where

$$\begin{aligned} \mathcal{E} &:= \frac{1}{n\sigma_\ell^4} \mathbb{E}[|\ell_1^2 \ell_2 q_{12}|] + \frac{1}{n^{3/2}\sigma_\ell^5} \mathbb{E}[|\ell_1^2 \ell_2^2 q_{12}|] + \frac{1}{n^{3/2}\sigma_\ell^3} \mathbb{E}[|\ell_1 q_{12}^2|] + \frac{1}{n^{3/2}\sigma_\ell^5} \mathbb{E}[|\ell_1 \ell_2 \ell_3 q_{13} q_{23}|] \\ &\quad + \frac{1}{n^{3/2}\sigma_\ell^7} |x_2| \mathbb{E}[|\ell_1^2 \ell_2 q_{12}|] + \frac{1}{n^2\sigma_\ell^8} |x_2| \mathbb{E}[|\ell_1^2 \ell_2^2 q_{12}|] + \frac{\log^p n}{n^p \sigma_\ell^p} \mathcal{M}_p(\log n), \end{aligned}$$

with

$$\mathcal{M}_p(m) := (mn\mathbb{E}[q_{12}^2])^{p/2} + m\mathbb{E}[(n\mathbb{E}[q_{12}^2|Z_1])^{p/2}] + mn\mathbb{E}[|q_{12}|^p]$$

**Corollary A.1.** If the assumptions of [Theorem A.1](#) hold and if  $\gamma_1 \rightarrow 0$ , then

$$\sup_{x \in \mathbb{R}} |F(x) - \bar{G}(x)| = O \left( \gamma_1^2 + \mathcal{E} + \frac{1}{\sqrt{n} \log n} \right),$$

where  $\bar{G}$  is the distribution function with characteristic function

$$\chi_{\bar{G}}(t) := \exp \left( -\frac{t^2}{2} \right) \left[ 1 + \sum_{1 \leq j \leq 9} (it)^j \gamma_j \right], \quad \gamma_1 := B.$$

**Remark A.1.** For every  $k \in \mathbb{Z}_+$ ,

$$\frac{1}{2\pi} \int_{\mathbb{R}} \exp \left( -itx - \frac{t^2}{2} \right) (it)^k dt = \phi(x) H_k(x), \quad k \in \mathbb{Z}_+,$$

where  $H_k(x)$  is the  $k$ th order Hermite polynomial (i.e.,  $H_0(k) = 1$ ,  $H_1(x) = x$ ,  $H_2(x) = x^2 - 1$ , and so on). Therefore, the characteristic function  $\chi_{\bar{G}}$  from [Corollary A.1](#) can be inverted to obtain the following closed form expression for  $\bar{G}$ :

$$\bar{G}(x) = \Phi(x) - \phi(x) \left[ \sum_{1 \leq j \leq 9} \gamma_j H_{j-1}(x) \right].$$

**Remark A.2.** Condition (c) of [Theorem A.1](#) is implied by the following condition:

- (c') For every  $\underline{c} > 0$ ,

$$\limsup_{n \rightarrow \infty} \sup_{|t| > \underline{c}} |\mathbb{E}[\exp(it\ell_1/\sigma_\ell)]| < 1.$$

Moreover, by the proposition following [Petrov \(1995, Lemma 1.4\)](#), condition (c') is in turn implied by the following condition:

$$(c'') \quad \limsup_{n,|t|\rightarrow\infty} \mathbb{E}[\exp(it\ell_1/\sigma_\ell)] < 1.$$

When  $\ell_1$  does not depend on  $n$ , conditions (c) and (c') are equivalent and it is customary to replace these conditions by (c''), which itself reduces to the familiar Cramér condition on  $\ell_1$ , namely

$$\limsup_{|t|\rightarrow\infty} \mathbb{E}[\exp(it\ell_1)] < 1. \quad (\text{A.1})$$

In contrast, when  $\ell_1$  does depend on  $n$ , condition (c) is potentially easier to verify than (c''). An example where this potential is realized is given by the DWAD estimator studied in [Theorems 1](#) and [2](#). In that example, it appears difficult to formulate simple conditions under which (c'') holds, but we are able to verify (c) with the help of the following observation: By [\(B.1\)](#), (c) holds whenever there exists a fixed function  $\lambda$  satisfying

$$\mathbb{E}[(\ell_1 - \lambda(Z_1))^2] = o(1/\log^2 n) \quad \text{and} \quad \limsup_{|t|\rightarrow\infty} \mathbb{E}[\exp(it\lambda(Z_1))] < 1, \quad (\text{A.2})$$

a condition which is itself equivalent to [\(A.1\)](#) when  $\ell_1$  does not depend on  $n$ .

To summarize, the condition [\(A.2\)](#) is equivalent to (c'') when  $\ell_1$  does not depend on  $n$  and more generally it provides a simple sufficient condition for (c) when  $\ell_1$  is mean square convergent.

**Remark A.3.** Suppose  $u_n$  does not depend on  $n$  and that

$$\mathbb{E}[|\ell_1|^3] < \infty, \quad \mathbb{E}[|q_{12}|^3] < \infty, \quad \text{and} \quad \limsup_{|t|\rightarrow\infty} |\mathbb{E}[\exp(it\ell_1)]| < 1.$$

If  $\theta = \mathbb{E}[u(Z_1, Z_2)]$  and if  $\theta^2 = n^{-1}\sigma_\ell^2$ , then the assumptions of [Corollary A.1](#) are satisfied with  $\gamma_1 = 0$  and  $\mathcal{E} = O(n^{-1})$ . Also,  $\gamma_2 = O(n^{-1})$  and  $\gamma_j = O(n^{-1})$  for  $4 \leq j \leq 9$ , so

$$\bar{G}(x) = \Phi(x) - \phi(x) \frac{\gamma_3}{6\sqrt{n}}(x^2 - 1) + O(n^{-1}), \quad \gamma_3 = \frac{\kappa_1 + 6\kappa_2}{\sigma_\ell^3},$$

uniformly in  $x \in \mathbb{R}$ . In other words, we recover a variant of [Bickel et al. \(1986, Theorem 1.2\)](#). See also [Jing and Wang \(2003\)](#).

**Remark A.4.** If  $\mathbb{E}[|\ell_1/\sigma_\ell|^3] = O(1)$  and if  $n^{-1}\sigma_q^2/\sigma_\ell^2 \rightarrow 0$ , then the Hölder inequality implies

$$\begin{aligned} \mathcal{E} &\lesssim \frac{1}{n} \sqrt[3]{\mathbb{E}[|q_{12}/\sigma_\ell|^3]} + \left( \frac{1}{n^{3/4}} \sqrt[3]{\mathbb{E}[|q_{12}/\sigma_\ell|^3]} \right)^2 \\ &\quad + \left( \frac{\log^3 n}{n} \mathbb{E}[|q_{12}/\sigma_\ell|^2] \right)^{p/2} + \frac{\log^{1+p} n}{n^{p/2}} \mathbb{E}[(\mathbb{E}[(q_{12}/\sigma_\ell)^2 | Z_1])^{p/2}] + \frac{\log^{1+p} n}{n^{p-1}} \mathbb{E}[|q_{12}/\sigma_\ell|^p], \end{aligned}$$

where  $a \lesssim b$  denotes  $a \leq Cb$  for some positive constant  $C$ . Thus, for  $p = 3$  the majorant side of  $\mathcal{E}$  is  $o(1)$  iff

$$\mathbb{E}[|q_{12}/\sigma_\ell|^2] = o(n/\log^3 n), \quad \mathbb{E}[|q_{12}/\sigma_\ell|^3] = o(n^2/\log^4 n),$$

and if

$$\mathbb{E}[(\mathbb{E}[(q_{12}/\sigma_\ell)^2 | Z_1])^{3/2}] = o(n^{3/2}/\log^4 n).$$

## Appendix B. Proof of [Theorem A.1](#)

Before presenting the formal proof, we outline the main steps involved, and compare our proof strategy to the approach taken in [Jing and Wang \(2003\)](#) for second-order U-statistics with fixed kernels (i.e.,  $u_n$  not depending on  $n$ ) and, more broadly, to the classical Edgeworth expansion theory for sums of independent random variables (e.g., [Bhattacharya and Rao, 1976](#); [Hall, 1992](#)).

We start from the following bound on the Kolmogorov distance between  $F$  and  $G$ :

$$\rho(F, G) := \sup_{t \in \mathbb{R}} |F(t) - G(t)| \lesssim \int_{|t| \leq \sqrt{n} \log n} \left| \frac{\chi_F(t) - \chi_G(t)}{t} \right| dt + \frac{1}{\sqrt{n} \log n},$$

where  $\chi_F$  is the characteristic function of  $F$ . The integral is then upper bounded over three different frequency domains: Low frequency (LF),  $|t| \leq \log n$ ; medium frequency (MF),  $\log n < |t| \leq c\sqrt{n}$  (for a judiciously chosen  $c$ ); and high frequency (HF),  $c\sqrt{n} < |t| \leq \sqrt{n} \log n$ . In the case of MF and HF, we further use the bound  $|\chi_F(t) - \chi_G(t)| \leq |\chi_F(t)| + |\chi_G(t)|$  (i.e., the triangle inequality) and deal with each term separately. At this level of generality the proof strategy is similar to the canonical case of first-order U-statistics (e.g., [Bhattacharya and Rao, 1976](#); [Hall, 1992](#)). However, we need to proceed differently to control the influence of the quadratic term, for which a bound follows almost exclusively from the linear part. For instance, for HF, we use Assumption (c) to find a  $b > 0$  such that for large  $n$  and for  $|t| \in (c\sqrt{n}, \sqrt{n} \log n]$ ,

$$\left| \chi_\ell \left( \frac{t}{n\delta} \right) \right| \leq 1 - b < \exp(-b), \quad \chi_\ell(t) := \mathbb{E}[\exp(it\ell_1)].$$

For the most interesting part, the LF domain, we begin by approximating  $\chi_F(t)$  as follows:

$$\begin{aligned}\chi_F(t) &= \exp\left(\mathrm{i} \frac{t}{\theta} B\right) \mathbb{E}\left[\exp\left(\mathrm{i} \frac{t}{\theta} L\right) \exp\left(\mathrm{i} \frac{t}{\theta} Q\right)\right] \\ &\approx \exp\left(\mathrm{i} \frac{t}{\theta} B\right) \mathbb{E}\left[\left(\mathrm{i} \frac{t}{\theta} L\right)\left(1 + \mathrm{i} \frac{t}{\theta} Q - \frac{t^2}{2\theta^2} Q^2\right)\right],\end{aligned}$$

and then examine each term on the right hand side separately while controlling the approximation error using [Lemma E.1](#) and the bound

$$\left|\exp(\mathrm{i}x) - \sum_{0 \leq j \leq 2} \frac{(\mathrm{i}x)^j}{j!}\right| \leq |x|^p. \quad (\text{B.1})$$

When doing this, our proof departs from [Jing and Wang \(2003\)](#) because we do not know *ex ante* which term(s) are higher-order due to the possibly  $n$ -varying structure of the U-statistic kernel. Therefore, we keep track of all terms, with special attention to the contribution of the terms involving  $Q$ . Finally, we collect all (possibly) leading terms in the approximation to  $\chi_F$  in  $\chi_G$  and arrive at a result of the form

$$|\chi_F(t) - \chi_G(t)| \lesssim \exp\left(-\frac{t^2}{4}\right) \mathcal{R}(t) + \frac{|t|^p}{n^{3p/2} \theta^p \log^{p/2} n} \mathcal{M}_p(\log n), \quad (\text{B.2})$$

where

$$\begin{aligned}\mathcal{R}(t) &:= \frac{t^4}{n^3 \theta^4} \mathbb{E}[|\ell_1^2 \ell_2 q_{12}|] + \frac{|t|^5}{n^4 \theta^5} \mathbb{E}[|\ell_1^2 \ell_2^2 q_{12}|] + \frac{|t|^3}{n^3 \theta^3} \mathbb{E}[|\ell_1 q_{12}^2|] + \frac{|t|^5}{n^4 \theta^5} \mathbb{E}[|\ell_1 \ell_2 \ell_3 q_{13} q_{23}|] \\ &\quad + \frac{|t|^7}{n^5 \theta^7} |\chi_2| \mathbb{E}[|\ell_1^2 \ell_2 q_{12}|] + \frac{t^8}{n^6 \theta^8} |\chi_2| \mathbb{E}[|\ell_1^2 \ell_2^2 q_{12}|],\end{aligned}$$

and where  $n^{-1} \sigma_\ell^2 / \theta^2 \rightarrow 1$ .

**Technical Details.** Letting  $g$  denote the Lebesgue density of  $G$ , application of a “smoothing inequality” (e.g., [Theorem 5.1 of Petrov, 1995](#)) gives

$$\rho(F, G) \lesssim \int_{|t| \leq v} \left| \frac{\chi_F(t) - \chi_G(t)}{t} \right| dt + \frac{\sup_{x \in \mathbb{R}} |g(x)|}{v}, \quad v > 0.$$

Setting  $v = \sqrt{n} \log n$  and using the fact that  $g$  is bounded (because  $\sum_{2 \leq j \leq 9} |\gamma_j| \rightarrow 0$ ), it follows from the triangle inequality that

$$\rho(F, G) \lesssim \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3 + \mathcal{I}_4 + \frac{1}{\sqrt{n} \log n}, \quad (\text{B.3})$$

where, with  $c > 0$  a constant to be chosen later,

$$\begin{aligned}\mathcal{I}_1 &:= \int_{|t| \leq \log n} \left| \frac{\chi_F(t) - \chi_G(t)}{t} \right| dt, \quad \mathcal{I}_2 := \int_{\log n < |t| \leq c\sqrt{n}} \left| \frac{\chi_F(t)}{t} \right| dt, \\ \mathcal{I}_3 &:= \int_{c\sqrt{n} < |t| \leq \sqrt{n} \log n} \left| \frac{\chi_F(t)}{t} \right| dt, \quad \mathcal{I}_4 := \int_{|t| > \log n} \left| \frac{\chi_G(t)}{t} \right| dt.\end{aligned}$$

In what follows, we bound each of these integrals in turn.

*Bound for  $\mathcal{I}_1$ .* We start by approximating

$$\chi_{L+Q}\left(\frac{t}{\theta}\right) := \mathbb{E}\left[\exp\left(\mathrm{i} \frac{t}{\theta} L\right) \exp\left(\mathrm{i} \frac{t}{\theta} Q\right)\right].$$

First, using [\(B.1\)](#), we have

$$\left| \chi_{L+Q}\left(\frac{t}{\theta}\right) - \mathbb{E}\left[\exp\left(\mathrm{i} \frac{t}{\theta} L\right) \left(1 + \mathrm{i} \frac{t}{\theta} Q - \frac{t^2}{2\theta^2} Q^2\right)\right] \right| \leq |\tau|^p n^p \mathbb{E}[|Q|^p], \quad \tau := \frac{t}{n\theta}.$$

Also, since  $\ell_1, \dots, \ell_n$  are *i.i.d.*,

$$\mathbb{E}\left[\exp\left(\mathrm{i} \frac{t}{\theta} L\right)\right] = \mathbb{E}\left[\exp\left(\mathrm{i} \tau \sum_{1 \leq i \leq n} \ell_i\right)\right] = \mathbb{E}\left[\prod_{1 \leq i \leq n} \exp(\mathrm{i} \tau \ell_i)\right] = \chi_\ell(\tau)^n.$$

Finally, because  $\ell_k$  is independent of  $q_{ij}$  when  $k \notin \{i, j\}$ ,

$$\begin{aligned}\binom{n}{2} \mathbb{E}\left[\exp\left(\mathrm{i} \frac{t}{\theta} L\right) Q\right] &= \sum_{1 \leq i < j \leq n} \mathbb{E}\left[q_{ij} \prod_{1 \leq k \leq n} \exp(\mathrm{i} \tau \ell_k)\right] \\ &= \sum_{1 \leq i < j \leq n} \mathbb{E}\left[\exp(\mathrm{i} \tau (\ell_i + \ell_j)) q_{ij} \prod_{\substack{1 \leq k \leq n \\ k \notin \{i, j\}}} \exp(\mathrm{i} \tau \ell_k)\right] \\ &= \binom{n}{2} \chi_\ell(\tau)^{n-2} \mathbb{E}\left[\exp(\mathrm{i} \tau (\ell_1 + \ell_2)) q_{12}\right]\end{aligned}$$

and

$$\begin{aligned}
& \binom{n}{2}^2 \mathbb{E} \left[ \exp \left( i \frac{t}{\theta} L \right) Q^2 \right] \\
&= \sum_{1 \leq i < j \leq n} \mathbb{E} \left[ q_{ij}^2 \prod_{1 \leq m \leq n} \exp(i\tau\ell_m) \right] + \sum_{1 \leq i < j < l \leq n} \mathbb{E} \left[ q_{ij} q_{jl} \prod_{1 \leq m \leq n} \exp(i\tau\ell_m) \right] \\
&\quad + \sum_{1 \leq i < j < k < l \leq n} \mathbb{E} \left[ q_{ij} q_{kl} \prod_{1 \leq m \leq n} \exp(i\tau\ell_m) \right] \\
&= \binom{n}{2} \chi_\ell(\tau)^{n-2} \mathbb{E} [\exp(i\tau(\ell_1 + \ell_2)) q_{12}^2] + \binom{n}{3} \chi_\ell(\tau)^{n-3} \mathbb{E} [\exp(i\tau(\ell_1 + \ell_2 + \ell_3)) q_{12} q_{23}] \\
&\quad + \binom{n}{4} \chi_\ell(\tau)^{n-4} (\mathbb{E} [\exp(i\tau(\ell_1 + \ell_2)) q_{12}])^2.
\end{aligned}$$

Using the four preceding displays, we therefore have, uniformly in  $|t| \leq \log n$ ,

$$\begin{aligned}
\chi_{L+Q} \left( \frac{t}{\theta} \right) &= \chi_\ell(\tau)^n \\
&\quad + \chi_\ell(\tau)^{n-2} \left[ i\tau n \mathbb{E} [\exp(i\tau(\ell_1 + \ell_2)) q_{12}] - \frac{\tau^2}{2} n^2 \binom{n}{2}^{-1} \mathbb{E} [\exp(i\tau(\ell_1 + \ell_2)) q_{12}^2] \right] \\
&\quad - \chi_\ell(\tau)^{n-3} \frac{\tau^2}{2} n^2 \binom{n}{2}^{-2} \binom{n}{3} \mathbb{E} [\exp(i\tau(\ell_1 + \ell_2 + \ell_3)) q_{13} q_{23}] \\
&\quad - \chi_\ell(\tau)^{n-4} \frac{\tau^2}{2} n^2 \binom{n}{2}^{-2} \binom{n}{4} (\mathbb{E} [\exp(i\tau(\ell_1 + \ell_2)) q_{12}])^2 \\
&\quad + |\tau|^p O(n^p \mathbb{E} [|Q|^p]). \tag{B.4}
\end{aligned}$$

Next, using degeneracy of  $q_{ij}$  and (B.1), we have

$$\begin{aligned}
\mathbb{E} [\exp(i\tau(\ell_1 + \ell_2)) q_{12}] &= -\tau^2 \mathbb{E} [\ell_1 \ell_2 q_{12}] \\
&\quad + i\tau \mathbb{E} [\ell_1 (\exp(i\tau\ell_2) - 1 - i\tau\ell_2) q_{12} + \ell_2 (\exp(i\tau\ell_1) - 1 - i\tau\ell_1) q_{12}] \\
&\quad + \mathbb{E} [(\exp(i\tau\ell_1) - 1 - i\tau\ell_1) (\exp(i\tau\ell_2) - 1 - i\tau\ell_2) q_{12}] \\
&= -\tau^2 \kappa_2 + |\tau|^3 O(\mathbb{E} [|\ell_1^2 \ell_2 q_{12}|]) + \tau^4 O(\mathbb{E} [|\ell_1^2 \ell_2^2 q_{12}|]).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E} [\exp(i\tau(\ell_1 + \ell_2)) q_{12}^2] &= \mathbb{E} [q_{12}^2] + \mathbb{E} [(\exp(i\tau(\ell_1 + \ell_2)) - 1) q_{12}^2] \\
&= \sigma_q^2 + |\tau| O(\mathbb{E} [|\ell_1 q_{12}^2|])
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E} [\exp(i\tau(\ell_1 + \ell_2 + \ell_3)) q_{13} q_{23}] &= \mathbb{E} \left[ \left( \prod_{1 \leq i \leq 3} (\exp(i\tau\ell_i) - 1) \right) q_{13} q_{23} \right] \\
&= |\tau|^3 O(\mathbb{E} [|\ell_1 \ell_2 \ell_3 q_{13} q_{23}|]).
\end{aligned}$$

Also, using arguments familiar from the Edgeworth expansion theory for sum of *i.i.d.* random variables (e.g., Bhattacharya and Rao, 1976; Hall, 1992), we have, for  $k \in \{0, 2, 3, 4\}$ ,

$$\begin{aligned}
\chi_\ell(\tau)^{n-k} &= \exp \left( -\frac{t^2}{2} \right) \left[ 1 - \left( \frac{n^{-1} \sigma_\ell^2 - \theta^2}{\theta^2} \right) \frac{t^2}{2} - i \frac{\kappa_1}{n^2 \theta^3} \frac{t^3}{6} + \left( \frac{n^{-1} \sigma_\ell^2 - \theta^2}{\theta^2} \right)^2 O(t^4) \right] \\
&\quad + \exp \left( -\frac{t^2}{4} \right) o \left( \frac{|t|^3}{\sqrt{n}} + \frac{t^6}{\sqrt{n}} \right).
\end{aligned}$$

Finally, using Lemma E.1, we have

$$\mathbb{E} [|Q|^p] \lesssim \frac{\mathcal{M}_p(n)}{n^{2p}} \lesssim \frac{\mathcal{M}_p(\log n)}{n^{3p/2} \log^{p/2} n}.$$

Plugging the displays from the previous paragraph into (B.4), we obtain (B.2) and therefore

$$\begin{aligned}
I_1 &\lesssim \int_{|t| \leq \log n} \exp \left( -\frac{t^2}{4} \right) |t|^{-1} \mathcal{R}(t) dt + \frac{\mathcal{M}_p(\log n)}{n^{3p/2} \theta^p \log^{p/2} n} \int_{|t| \leq \log n} |t|^{p-1} dt \\
&\lesssim \mathcal{R}(1) + \frac{\log^{p/2} n}{n^{3p/2} \theta^p} \mathcal{M}_p(\log n)
\end{aligned}$$

$$\begin{aligned} &\lesssim \frac{1}{n\sigma_\ell^4} \mathbb{E}[|\ell_1^2 \ell_2 q_{12}|] + \frac{1}{n^{3/2}\sigma_\ell^5} \mathbb{E}[|\ell_1^2 \ell_2^2 q_{12}|] + \frac{1}{n^{3/2}\sigma_\ell^3} \mathbb{E}[|\ell_1 q_{12}^2|] + \frac{1}{n^{3/2}\sigma_\ell^5} \mathbb{E}[|\ell_1 \ell_2 \ell_3 q_{13} q_{23}|] \\ &\quad + \frac{1}{n^{3/2}\sigma_\ell^7} |\chi_2| \mathbb{E}[|\ell_1^2 \ell_2 q_{12}|] + \frac{1}{n^2\sigma_\ell^8} |\chi_2| \mathbb{E}[|\ell_1^2 \ell_2^2 q_{12}|] + \frac{\log^{p/2} n}{n^p \sigma_\ell^p} \mathcal{M}_p(\log n), \end{aligned}$$

where the last  $\lesssim$  uses  $n^{-1}\sigma_\ell^2/\theta^2 \rightarrow 1$ .

*Bound for  $I_2$ .* For  $1 \leq m < n$ , defining

$$Q(m) := \binom{n}{2}^{-1} \sum_{\substack{1 \leq i < j \leq n \\ i \leq m}} q_{ij}$$

and using (B.1), we have

$$|\chi_F(t)| = \left| \chi_{L+Q} \left( \frac{t}{\theta} \right) \right| \leq \left| \mathbb{E} \left[ \exp \left( i \frac{t}{\theta} (L + Q - Q(m)) \right) \sum_{0 \leq k \leq 2} \frac{(it)^k}{k! \theta^k} Q(m)^k \right] \right| + \frac{|t|^p}{\theta^p} \mathbb{E}[|Q(m)|^p],$$

where, using the fact that  $Q - Q(m)$  is a function of  $X_{m+1}, \dots, X_n$ , it can be shown that

$$\left| \mathbb{E} \left[ \exp \left( i \frac{t}{\theta} (L + Q - Q(m)) \right) Q(m)^k \right] \right| \lesssim \left( \frac{m}{n} \right)^k |\chi_\ell(\tau)|^{m-2k} \mathbb{E}[|q_{12}|^k], \quad k \in \{0, 1, 2\},$$

and therefore, using  $n^{-1}\sigma_\ell^2/\theta^2 \rightarrow 1$ ,

$$|\chi_F(t)| \lesssim \sum_{0 \leq k \leq 2} \left( \frac{|t|m}{\sqrt{n}} \right)^k |\chi_\ell(\tau)|^{m-2k} \mathbb{E} \left[ \left| \frac{q_{12}}{\sigma_\ell} \right|^k \right] + \frac{|t|^p}{\theta^p} \mathbb{E}[|Q(m)|^p]. \quad (\text{B.5})$$

Next, because  $\mathbb{E}[|\ell_1/\sigma_\ell|^3] = O(1)$  and  $n^{-1}\sigma_\ell^2/\theta^2 \rightarrow 1$ , there exists a  $c > 0$  such that, for  $n$  sufficiently large,

$$|\chi_\ell(\tau)| \leq 1 - \frac{t^2}{3n} \leq \exp \left( -\frac{t^2}{3n} \right), \quad t \leq c\sqrt{n}.$$

Setting  $m = \lfloor 15n \log n / t^2 \rfloor$  in (B.5), where  $\lfloor \cdot \rfloor$  denotes the floor operator, and using the preceding display, we obtain (for  $n$  sufficiently large)

$$|\chi_F(t)| \lesssim \sum_{0 \leq k \leq 2} \frac{|t|^k}{n^{5-k}} \mathbb{E} \left[ \left| \frac{q_{12}}{\sigma_\ell} \right|^k \right] + \frac{|t|^p}{\theta^p} \mathbb{E}[|Q(m)|^p], \quad \log n < |t| \leq c\sqrt{n},$$

where, using Lemma E.1 and  $n^{-1}\sigma_\ell^2/\theta^2 \rightarrow 1$ ,

$$\begin{aligned} \frac{|t|^p}{\theta^p} \mathbb{E}[|Q(m)|^p] &\lesssim \frac{|t|^p}{n^{2p}\theta^p} \mathcal{M}_p(m) \\ &\lesssim \frac{\log^{p/2} n}{n^{p/2}\sigma_\ell^p} \sigma_q^p + |t|^{p-2} \frac{\log n}{n^{p-1}\sigma_\ell^p} \left[ \mathbb{E}[(\mathbb{E}[q_{12}^2 | Z_1])^{p/2}] + \frac{\mathbb{E}[|q_{12}|^p]}{n^{p/2-1}} \right], \quad \log n < |t| \leq c\sqrt{n}. \end{aligned}$$

As a consequence, using  $n^{-1}\sigma_q^2/\sigma_\ell^2 \rightarrow 0$ ,

$$\begin{aligned} I_2 &\lesssim \sum_{0 \leq k \leq 2} \frac{1}{n^{5-3k/2}} \mathbb{E} \left[ \left| \frac{q_{12}}{\sigma_\ell} \right|^k \right] + \frac{\log^{1+p/2} n}{n^{p/2}\sigma_\ell^p} \sigma_q^p + \frac{\log n}{n^{p/2}\sigma_\ell^p} \mathbb{E}[(\mathbb{E}[q_{12}^2 | Z_1])^{p/2}] + \frac{\log n}{n^{p-1}\sigma_\ell^p} \mathbb{E}[|q_{12}|^p] \\ &= o(n^{-1}) + o \left( \frac{\log^p n}{n^p \sigma_\ell^p} \mathcal{M}_p(\log n) \right). \end{aligned}$$

*Bound for  $I_3$ .* By condition (c), there exists  $b > 0$  such that, for  $n$  sufficiently large,

$$|\chi_\ell(\tau)| \leq 1 - b \leq \exp(-b), \quad c\sqrt{n} < |t| < \sqrt{n} \log n.$$

Setting  $m = \lfloor 4 \log n / b \rfloor$  in (B.5) and using the preceding display, we obtain (for  $n$  sufficiently large)

$$|\chi_F(t)| \lesssim \sum_{0 \leq k \leq 2} \frac{|t|^k \log^k n}{n^{4+k/2}} \mathbb{E} \left[ \left| \frac{q_{12}}{\sigma_\ell} \right|^k \right] + \frac{|t|^p}{\theta^p} \mathbb{E}[|Q(m)|^p], \quad c\sqrt{n} < |t| < \sqrt{n} \log n,$$

where, using Lemma E.1 and  $n^{-1}\sigma_\ell^2/\theta^2 \rightarrow 1$ ,

$$\frac{|t|^p}{\theta^p} \mathbb{E}[|Q(m)|^p] \lesssim \frac{|t|^p}{n^{2p}\theta^p} \mathcal{M}_p(m) \lesssim \frac{|t|^p}{n^{3p/2}\sigma_\ell^p} \mathcal{M}_p(\log n), \quad c\sqrt{n} < |t| < \sqrt{n} \log n.$$

As a consequence, using  $n^{-1}\sigma_q^2/\sigma_\ell^2 \rightarrow 0$ ,

$$I_3 = o(n^{-2}) + O \left( \frac{\mathcal{M}_p(\log n)}{n^{3p/2}\sigma_\ell^p} \int_{c\sqrt{n} \leq |t| \leq \sqrt{n} \log n} |t|^{p-1} dt \right)$$

$$= o(n^{-2}) + O\left(\frac{\log^p n}{n^p \sigma_\ell^p} \mathcal{M}_p(\log n)\right).$$

Bound for  $\mathcal{I}_4$ . For every  $j$ , we have

$$\int_{t > \log n} t^{j-1} \exp\left(-\frac{t^2}{2}\right) dt \lesssim \int_{t > \log n} \exp\left(-\frac{t^2}{4}\right) dt = o(n^{-1}),$$

and therefore, using  $\sum_{2 \leq j \leq 9} |\gamma_j| \rightarrow 0$ ,

$$\begin{aligned} \mathcal{I}_4 &\lesssim \int_{|t| > \log n} |t|^{-1} \exp\left(-\frac{t^2}{2}\right) \left| 1 + \sum_{j=2}^9 (it)^j \gamma_j \right| dt \\ &\lesssim \left(1 + \sum_{j=2}^9 |\gamma_j|\right) \int_{t > \log n} t^{-1} \exp\left(-\frac{t^2}{4}\right) dt = o(n^{-1}). \quad \square \end{aligned}$$

### Appendix C. Proof of Theorem 1

We employ Corollary A.1 with  $u(Z_i, Z_j) = v' U_{ij}$  and  $p = 3$ . Proceeding as in Cattaneo et al. (2010, 2014b,a), condition (a) of Theorem A.1 can be verified by direct calculations. Also, condition (b) of Theorem A.1 holds because

$$\sigma_\ell^2 = \sigma_v^2 + o(1), \quad \sigma_q^2 = \frac{\delta_v^2 + o(1)}{h^{d+2}}, \quad \text{and} \quad nh^{d+2} \rightarrow \infty,$$

while condition (c) of Theorem A.1 holds because (A.2) is satisfied with  $\lambda = \psi_v$ . The additional condition  $\gamma_1 \rightarrow 0$  of Corollary A.1 holds if  $nh^{2P} \rightarrow 0$  because it follows from routine (bias) calculations that  $\mathbb{E}[\hat{\theta}_v] - \theta_v = h^P \beta_v + o(h^P)$ .

Next, the law of iterated expectations, integration by parts, and Taylor series expansions can be used to show that

$$\kappa_1 = \kappa_{1,v} + O(h^P), \quad \text{and} \quad \kappa_2 = \kappa_{2,v} + O(h^P),$$

and also that  $\gamma_4 \lesssim n^{-3} h^{-d-2}$ ,  $\gamma_5 \lesssim n^{-2}$ ,  $\gamma_6 \lesssim n^{-1}$ ,  $\gamma_8 \lesssim n^{-3}$ , and  $\gamma_9 \lesssim n^{-7/2}$ .

Finally, by Cattaneo et al. (2014a, Supplemental Appendix), we have

$$\mathbb{E}[q_{12}^2]^{3/2} \leq \mathbb{E}[(\mathbb{E}[q_{12}^2 | Z_1])^{3/2}] \lesssim h^{-3d/2-3} \quad \text{and} \quad \mathbb{E}[|q_{12}|^3] \lesssim h^{-2d-3}.$$

Using these bounds and the Hölder inequality, we find that  $\mathcal{E} = o(n^{-1} h^{-d-2})$ .  $\square$

### Appendix D. Proof of Theorem 2

Letting  $\ell_i := v' L_i$  and  $q_{ij} := v' Q_{ij}$ , occasionally suppressing the dependence on  $v$ , and using the Hoeffding decomposition of  $\hat{\theta}_v - \theta_v$  along with the identity

$$\frac{\vartheta}{\hat{\vartheta}} = 1 - \frac{\hat{\vartheta}^2 - \vartheta^2}{2\vartheta^2} + \frac{(\hat{\vartheta} + 2\vartheta)(\hat{\vartheta}^2 - \vartheta^2)^2}{2\vartheta^2 \hat{\vartheta}(\hat{\vartheta} + \vartheta)^2}, \quad (\text{D.1})$$

we have

$$\frac{\hat{\theta}_v - \theta_v}{\hat{\vartheta}_{\text{SB},v}} = \tilde{T}_{\text{SB}} + R_{\text{SB}},$$

where

$$\tilde{T}_{\text{SB}} := \frac{B + L + Q}{\vartheta_{\text{SB},v}} - \frac{\hat{\vartheta}_{\text{SB},v}^2 - \vartheta_{\text{SB},v}^2}{2\vartheta_{\text{SB},v}^2} \frac{L + Q}{\vartheta_{\text{SB},v}},$$

with  $\vartheta_{\text{SB},v}$  is a judiciously chosen positive scalar,

$$B := \mathbb{E}[\hat{\theta}_v] - \theta_v, \quad L := n^{-1} \sum_{1 \leq i \leq n} \ell_i, \quad \text{and} \quad Q := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} q_{ij},$$

and where

$$R_{\text{SB}} := -\frac{\hat{\vartheta}_{\text{SB},v}^2 - \vartheta_{\text{SB},v}^2}{2\vartheta_{\text{SB},v}^2} \frac{B}{\vartheta_{\text{SB},v}} + \frac{(\hat{\vartheta}_{\text{SB},v} + 2\vartheta_{\text{SB},v})(\hat{\vartheta}_{\text{SB},v}^2 - \vartheta_{\text{SB},v}^2)^2}{2\vartheta_{\text{SB},v}^2 \hat{\vartheta}_{\text{SB},v}(\hat{\vartheta}_{\text{SB},v} + \vartheta_{\text{SB},v})^2} \frac{B + L + Q}{\vartheta_{\text{SB},v}}$$

is a remainder term.

Defining

$$\tilde{F}_{\text{SB}}(x) := \mathbb{P}[\tilde{T}_{\text{SB}} \leq x]$$

and adapting the proof of [Theorem A.1](#), we obtain

$$\rho(\tilde{F}_{\text{SB}}, \hat{G}_{\text{SB}}) = o(r_n) \quad (\text{D.2})$$

by applying a smoothing inequality followed by a split of the frequency domain of the resulting integral, where bounding the various integrals requires some additional care due to the presence of the variance estimator.

Also, using the strengthened moment condition  $\mathbb{E}[Y^6] < \infty$ , we obtain

$$\mathbb{P}(|R_{\text{SB}}| > r_n / \log n) = o(r_n), \quad (\text{D.3})$$

implying in turn that  $\rho(\hat{F}_{\text{SB}}, \tilde{F}_{\text{SB}}) = o(r_n)$  and therefore also that

$$\rho(\hat{F}_{\text{SB}}, \hat{G}_{\text{SB}}) \leq \rho(\hat{F}_{\text{SB}}, \tilde{F}_{\text{SB}}) + \rho(\tilde{F}_{\text{SB}}, \hat{G}_{\text{SB}}) = o(r_n).$$

**Technical Details.** The identity [\(D.1\)](#) can be obtained as follows:

$$\begin{aligned} \frac{\vartheta}{\hat{\vartheta}} &= 1 - \frac{\hat{\vartheta} - \vartheta}{\vartheta} \frac{\hat{\vartheta} + \vartheta}{\hat{\vartheta} + \vartheta} + \frac{(\hat{\vartheta} - \vartheta)^2}{\vartheta \hat{\vartheta}} \frac{(\hat{\vartheta} + \vartheta)^2}{(\hat{\vartheta} + \vartheta)^2} \\ &= 1 - \frac{\hat{\vartheta}^2 - \vartheta^2}{2\vartheta^2} + \frac{\hat{\vartheta} - \vartheta}{2\vartheta^2(\hat{\vartheta} + \vartheta)} (\hat{\vartheta}^2 - \vartheta^2) + \frac{(\hat{\vartheta}^2 - \vartheta^2)^2}{\vartheta \hat{\vartheta}(\hat{\vartheta} + \vartheta)^2} \\ &= 1 - \frac{\hat{\vartheta}^2 - \vartheta^2}{2\vartheta^2} + \frac{(\hat{\vartheta} + 2\vartheta)(\hat{\vartheta}^2 - \vartheta^2)^2}{2\vartheta^2 \hat{\vartheta}(\hat{\vartheta} + \vartheta)^2}. \end{aligned}$$

Letting  $u_{ij} := v' U_{ij}$  and defining

$$U := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} u_{ij} = \hat{\theta}_v, \quad W_1 := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} u_{ij}^2,$$

and

$$W_2 := \binom{n}{3}^{-1} \sum_{1 \leq i < j < k \leq n} \frac{u_{ij}u_{ik} + u_{ij}u_{jk} + u_{ik}u_{jk}}{3},$$

it follows from [Cattaneo et al. \(2014a, Supplemental Appendix\)](#) that

$$\hat{\vartheta}_{\text{SB},v}^2 = \binom{n}{2}^{-1} W_1 + \frac{4}{n} \frac{n-2}{n-1} W_2 - \frac{4}{n} U^2.$$

Also, for  $k \in \{2, 3\}$ , using [Cattaneo et al. \(2014a, Supplemental Appendix\)](#) and [Lemma E.1](#),

$$\begin{aligned} \mathbb{E}[|U - \mathbb{E}[u_{12}]|^{2k}] &\lesssim n^{-k} + n^{-2k} h^{-(2p-1)d-2k}, \\ \mathbb{E}[|W_1 - \mathbb{E}[u_{12}^2]|^k] &\lesssim n^{-k/2} h^{-k(d+2)} + n^{-k} h^{-(2k-1)d-2p}, \\ \mathbb{E}[|W_2 - \mathbb{E}[(\mathbb{E}[u_{12}|Z_1])^2]|^k] &\lesssim n^{-k/2} + n^{-k} h^{-(k-1)d-2p} + n^{-3k/2} h^{-2(k-1)d-2k}, \end{aligned}$$

and therefore

$$\begin{aligned} \mathbb{E}\left[|\hat{\vartheta}_{\text{SB},v}^2 - \vartheta_{\text{SB},v}^2|^k\right] &\lesssim n^{-2k} \mathbb{E}[|W_1 - \mathbb{E}[u_{12}^2]|^k] + n^{-k} \mathbb{E}[|W_2 - \mathbb{E}[(\mathbb{E}[u_{12}|Z_1])^2]|^k] \\ &\quad + n^{-k} \sqrt{\mathbb{E}[|U - \mathbb{E}[u_{12}]|^{2k}]} \\ &\lesssim n^{-3k/2} + n^{-2k} h^{-(k-1/2)d-2k} + n^{-5k/2} h^{-2(k-1)d-2k}, \end{aligned} \quad (\text{D.4})$$

where

$$\vartheta_{\text{SB},v}^2 := \binom{n}{2}^{-1} \mathbb{E}[u_{12}^2] + \frac{4}{n} \mathbb{E}[(\mathbb{E}[u_{12}|Z_1])^2] - \frac{4}{n} \mathbb{E}[u_{12}]^2.$$

To prove [\(D.3\)](#), it suffices to show that  $\mathcal{V}_1 + \mathcal{V}_2 + \mathcal{V}_3 = o(r_n)$ , where

$$\begin{aligned} \mathcal{V}_1 &:= \mathbb{P}\left[\frac{(\hat{\vartheta}_{\text{SB},v} + 2\vartheta_{\text{SB},v})(\hat{\vartheta}_{\text{SB},v}^2 - \vartheta_{\text{SB},v}^2)^2}{\vartheta_{\text{SB},v}^2 \hat{\vartheta}_{\text{SB},v}(\hat{\vartheta}_{\text{SB},v} + \vartheta_{\text{SB},v})^2} > \frac{r_n}{\log^2 n}\right], \\ \mathcal{V}_2 &:= \mathbb{P}\left[\frac{|\hat{\theta}_v - \theta_v|}{\vartheta_{\text{SB},v}} > \log n\right], \\ \mathcal{V}_3 &:= \mathbb{P}\left[\frac{|\hat{\vartheta}_{\text{SB},v}^2 - \vartheta_{\text{SB},v}^2|}{\vartheta_{\text{SB},v}^2} \frac{|B|}{\vartheta_{\text{SB},v}} > \frac{\sqrt{nh}^p}{\log n}\right]. \end{aligned}$$

First, using (D.4) and the Markov inequality,

$$\begin{aligned}\mathcal{V}_1 &\leq \mathbb{P} \left[ (\widehat{\theta}_{\text{SB},v}^2 - \theta_{\text{SB},v}^2)^2 > \frac{r_n \sigma_v^4}{n^2 \log^2 n} \right] + o(r_n) \lesssim \frac{n^3 \log^3 n}{r_n^{3/2}} \mathbb{E} \left[ |\widehat{\theta}_{\text{SB},v}^2 - \theta_{\text{SB},v}^2|^3 \right] + o(r_n) \\ &\lesssim \frac{\log^3 n}{r_n^{3/2}} (n^{-3/2} + n^{-3} h^{-5d/2-6} + n^{-9/2} h^{-4d-6}) + o(r_n) = o(r_n).\end{aligned}$$

Next, using Theorem 1 and the properties of the standard normal distribution,

$$\begin{aligned}\mathcal{V}_2 &= 1 - \mathbb{P} \left[ \frac{\widehat{\theta}_v - \theta_v}{\theta_{\text{SB},v}} \leq \log n \right] + \mathbb{P} \left[ \frac{\widehat{\theta}_v - \theta_v}{\theta_{\text{SB},v}} < -\log n \right] \\ &= 1 - \Phi(\log n) + \Phi(-\log n) + o(r_n) = o(r_n).\end{aligned}$$

Finally, using (D.4) and the Markov inequality,

$$\begin{aligned}\mathcal{V}_3 &\lesssim n \mathbb{E} \left[ |\widehat{\theta}_{\text{SB},v}^2 - \theta_{\text{SB},v}^2|^2 \right] \log^2 n \\ &\lesssim (n^{-2} + n^{-3} h^{-3d/2-4} + n^{-4} h^{-2d-4}) \log^2 n = o(r_n).\end{aligned}$$

Next, to prove (D.2) we begin by using a “smoothing inequality” to obtain the bound

$$\rho(\widetilde{F}_{\text{SB}}, \widehat{G}_{\text{SB}}) \lesssim \widehat{I}_1 + \widehat{I}_2 + \widehat{I}_3 + \widehat{I}_4 + \frac{1}{\sqrt{n} \log n},$$

where

$$\begin{aligned}\widehat{I}_1 &:= \int_{|t| \leq \log n} \left| \frac{\chi_{\widetilde{F}_{\text{SB}}}(t) - \chi_{\widehat{G}_{\text{SB}}}(t)}{t} \right| dt, \quad \widehat{I}_2 := \int_{\log n < |t| \leq c\sqrt{n}} \left| \frac{\chi_{\widetilde{F}_{\text{SB}}}(t)}{t} \right| dt, \\ \widehat{I}_3 &:= \int_{c\sqrt{n} < |t| \leq \sqrt{n} \log n} \left| \frac{\chi_{\widetilde{F}_{\text{SB}}}(t)}{t} \right| dt, \quad \text{and} \quad \widehat{I}_4 := \int_{|t| > \log n} \left| \frac{\chi_{\widehat{G}_{\text{SB}}}(t)}{t} \right| dt.\end{aligned}$$

Proceeding as in the proof of Theorem A.1, it can be shown that  $\widehat{I}_2 + \widehat{I}_3 + \widehat{I}_4 = o(r_n)$ . The proof of (D.2) can therefore be completed by showing that  $\widehat{I}_1 = o(r_n)$ . We shall do so by adapting the proof of Theorem A.1 to also account for the contribution of  $\widehat{\theta}_{\text{SB},v}^2$  to  $\widetilde{F}_{\text{SB}}$ .

Defining

$$V_{\text{SB}} := \frac{1}{2n\theta_{\text{SB},v}^2} n^{-1} \sum_{1 \leq i \leq n} (\ell_i^2 - \sigma_\ell^2 + 4\mathbb{E}[\ell_j q_{ij} | Z_i]) + \frac{2}{n\theta_{\text{SB},v}^2} \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \mathbb{E}[q_{ik} q_{jk} | Z_i, Z_j]$$

and using Callaert and Veraverbeke (1981) and the Hölder inequality, it can be shown that

$$\chi_{\widetilde{F}_{\text{SB}}}(t) = \mathbb{E} \left[ \exp(it\widetilde{F}_{\text{SB}}) \right] = \chi_F(t) - it\mathbb{E} \left[ \exp \left( it \frac{L}{\theta_{\text{SB},v}} \right) V_{\text{SB}} \frac{L}{\theta_{\text{SB},v}} \right] + O(\mathcal{T}_1(t)), \quad (\text{D.5})$$

where  $\chi_F$  was (defined and) analyzed in the proof of Theorem A.1 and where

$$\mathcal{T}_1(t) := \frac{|t|}{nh^{d/2+1} + n^{3/2}h^{3d/2+3}} + \frac{t^2}{nh^{d/2+1} + n^{3/2}h^{3d/2+3}}.$$

As in Nishiyama and Robinson (2001), the second term on the right-hand side of (D.5) admits an expansion of the form

$$\begin{aligned}\mathbb{E} \left[ \exp \left( it \frac{L}{\theta_{\text{SB},v}} \right) V_{\text{SB}} \frac{L}{\theta_{\text{SB},v}} \right] &= \chi_\ell(\tau)^{n-1} \frac{x_1 + 4x_2}{2n^2\theta_{\text{SB},v}^3} - \chi_\ell(\tau)^{n-2} \frac{t^2}{2} \frac{x_1 + 4x_2}{n^2\theta_{\text{SB},v}^3} + \chi_\ell(\tau)^{n-3} O(\mathcal{T}_2(t)), \quad \tau := \frac{t}{n\theta_{\text{SB},v}}, \\ &= \chi_\ell(\tau)^{n-1} \frac{x_1 + 4x_2}{2n^2\theta_{\text{SB},v}^3} - \chi_\ell(\tau)^{n-2} \frac{t^2}{2} \frac{x_1 + 4x_2}{n^2\theta_{\text{SB},v}^3} + \chi_\ell(\tau)^{n-3} O(\mathcal{T}_2(t)),\end{aligned} \quad (\text{D.6})$$

where  $\chi_\ell(\tau)^{n-k}$  was (defined and) analyzed in the proof of Theorem A.1 and where

$$\mathcal{T}_2(t) := \frac{|t|}{n + h^{d+2-P}} + \frac{t^2}{n + n^{3/2}h^{d+2}} + \frac{|t|^3}{n} + \frac{t^4}{n^{3/2} + n^2h^{d+2}} + \frac{|t|^5}{n^{5/2}h^{d+2}} + \frac{t^6}{n^3h^{d+2}}.$$

Combining (D.5) and (D.6) with the previously obtained expansions for  $\chi_F$  and  $\chi_\ell(\tau)^{n-k}$ , we obtain an expansion of the form

$$\chi_{\widetilde{F}_{\text{SB}}}(t) = \chi_{\widehat{G}_{\text{SB}}}(t) + \mathcal{F}(t),$$

where

$$\widehat{I}_1 := \int_{|t| \leq \log n} \left| \frac{\mathcal{F}(t)}{t} \right| dt = o(r_n). \quad \square$$

## Appendix E. Auxiliary lemma

**Lemma E.1.** If  $S \subseteq \{(i, j) : 1 \leq i, j \leq n\}$  and if  $p \geq 2$ , then

$$\mathbb{E} \left[ \left| \sum_{(i,j) \in S} q_{ij} \right|^p \right] \lesssim (\mathfrak{s}_1 \mathbb{E}[q_{12}^2])^{p/2} + \mathfrak{s}_{p/2} \mathbb{E}[(\mathbb{E}[q_{12}^2 | Z_1])^{p/2}] + \mathfrak{s}_1 \mathbb{E}[|q_{12}|^p],$$

where

$$\mathfrak{s}_s := \max \left[ \sum_{1 \leq i \leq n} \left( \sum_{1 \leq j \leq n} \mathbb{1}(\{i, j\} \in S) \right)^s, \sum_{1 \leq i \leq n} \left( \sum_{1 \leq j \leq n} \mathbb{1}(\{j, i\} \in S) \right)^s \right].$$

**Proof.** By [Giné et al. \(2000\)](#), Proposition 2.4), the inequality holds for the *decoupled* version of  $q_{ij}$ , defined as  $\tilde{q}_{ij} := q(Z_i^{(1)}, Z_j^{(2)})$  where  $\{Z_i^{(k)} : 1 \leq i \leq n, 1 \leq k \leq 2\}$  are *i.i.d.* copies of  $Z$ . Finally, we can apply the decoupling inequalities in [de la Peña and Montgomery-Smith \(1995\)](#) to obtain the result at the expense of increasing the constant without altering the order of the upper bound; for further details, see [Giné et al. \(2000\)](#), Section 2.5).  $\square$

## References

- Ahn, H., Ichimura, H., Powell, J.L., Ruud, P.A., 2018. Simple estimators for invertible index models. *J. Bus. Econom. Statist.* 36 (1), 1–10.
- Aradillas-Lopez, A., Honoré, B.E., Powell, J.L., 2007. Pairwise difference estimation with nonparametric control variables. *Internat. Econom. Rev.* 48 (4), 1119–1158.
- Battacharya, R.N., Rao, R.R., 1976. Normal Approximation and Asymptotic Expansions. Wiley, New York.
- Bickel, P.J., Götzte, F., van Zwet, W.R., 1986. The edgeworth expansion for  $U$ -statistics of degree two. *Ann. Statist.* 14 (4), 1463–1484.
- Bickel, P.J., Klaassen, C.A.J., Ritov, Y., Wellner, J.A., 1993. Efficient and Adaptive Estimation for Semiparametric Models. Springer, New York.
- Blundell, R.W., Powell, J.L., 2004. Endogeneity in semiparametric binary response models. *Rev. Econ. Stud.* 71 (3), 655–679.
- Callaert, H., Veraverbeke, N., 1981. The order of the normal approximation for a studentized  $U$ -statistic. *Ann. Statist.* 9 (1), 194–200.
- Calonico, S., Cattaneo, M.D., Farrell, M.H., 2018. On the effect of bias estimation on coverage accuracy in nonparametric inference. *J. Amer. Statist. Assoc.* 113 (522), 767–779.
- Calonico, S., Cattaneo, M.D., Farrell, M.H., 2022. Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli* 28 (4), 2998–3022.
- Cattaneo, M.D., Crump, R.K., Jansson, M., 2010. Robust data-driven inference for density-weighted average derivatives. *J. Amer. Statist. Assoc.* 105 (491), 1070–1083.
- Cattaneo, M.D., Crump, R.K., Jansson, M., 2013. Generalized jackknife estimators of weighted average derivatives (with discussions and rejoinder). *J. Amer. Statist. Assoc.* 108 (504), 1243–1268.
- Cattaneo, M.D., Crump, R.K., Jansson, M., 2014a. Bootstrapping density-weighted average derivatives. *Econometric Theory* 30 (6), 1135–1164.
- Cattaneo, M.D., Crump, R.K., Jansson, M., 2014b. Small bandwidth asymptotics for density-weighted average derivatives. *Econometric Theory* 30 (1), 176–200.
- Cattaneo, M.D., Jansson, M., 2018. Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency. *Econometrica* 86 (3), 955–995.
- Cattaneo, M.D., Jansson, M., 2022. Average density estimators: Efficiency and bootstrap consistency. *Econometric Theory* 38 (6), 1140–1174.
- Cattaneo, M.D., Jansson, M., Ma, X., 2019. Two-step estimation and inference with possibly many included covariates. *Rev. Econ. Stud.* 86 (3), 210–245.
- Cattaneo, M.D., Jansson, M., Newey, W.K., 2018a. Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory* 34 (2), 277–301.
- Cattaneo, M.D., Jansson, M., Newey, W.K., 2018b. Inference in linear regression models with many covariates and heteroscedasticity. *J. Amer. Statist. Assoc.* 113 (523), 1350–1361.
- Chernozhukov, V., Escanciano, J.C., Ichimura, H., Newey, W.K., Robins, J.M., 2022. Locally robust semiparametric estimation. *Econometrica* 90 (4), 1501–1535.
- Efron, B., Stein, C., 1981. The jackknife estimate of variance. *Ann. Statist.* 9 (3), 586–596.
- Giné, E., Latala, R., Zinn, J., 2000. Exponential and moment inequalities for  $U$ -statistics. In: Giné, E., Mason, D.M., Wellner, J.A. (Eds.), *High Dimensional Probability II*. Birkhäuser, Boston, pp. 13–38.
- Graham, B.S., Niu, F., Powell, J.L., 2024. Kernel density estimation for undirected dyadic data. *J. Econometrics* forthcoming.
- Hall, P., 1992. *The Bootstrap and Edgeworth Expansion*. Springer, New York.
- Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* 19 (3), 293–325.
- Honoré, B.E., Powell, J.L., 1994. Pairwise difference estimators of censored and truncated regression models. *J. Econometrics* 64 (1–2), 241–278.
- Ichimura, H., Todd, P.E., 2007. Implementing nonparametric and semiparametric estimators. In: Heckman, J., Leamer, E. (Eds.), *Handbook of Econometrics*, Volume VI. Elsevier, New York, pp. 5370–5468.
- Jing, B.-Y., Wang, Q., 2003. Edgeworth expansion for  $U$ -statistics under minimal conditions. *Ann. Statist.* 31 (4), 1376–1391.
- Matsushita, Y., Otsu, T., 2021. Jackknife empirical likelihood: Small bandwidth, sparse network and high-dimensional asymptotics. *Biometrika* 108 (3), 661–674.
- Newey, W.K., 1994. The asymptotic variance of semiparametric estimators. *Econometrica* 62 (6), 1349–1382.
- Newey, W.K., Hsieh, F., Robins, J.M., 2004. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* 72 (3), 947–962.
- Newey, W.K., McFadden, D.L., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D.L. (Eds.), *Handbook of Econometrics*, Volume IV. Elsevier, New York, pp. 2111–2245.
- Nishiyama, Y., Robinson, P.M., 2000. Edgeworth expansions for semiparametric averaged derivatives. *Econometrica* 68 (4), 931–979.
- Nishiyama, Y., Robinson, P.M., 2001. Studentization in edgeworth expansions for estimates of semiparametric index models. In: Hsiao, C., Morimune, K., Powell, J.L. (Eds.), *Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya*. Cambridge University Press, New York, pp. 197–240.
- Nishiyama, Y., Robinson, P.M., 2005. The bootstrap and the edgeworth correction for semiparametric averaged derivatives. *Econometrica* 73 (3), 197–240.
- de la Peña, V.H., Montgomery-Smith, S.J., 1995. Decoupling inequalities for the tail probabilities of multivariate  $U$ -statistics. *Ann. Probab.* 23 (2), 806–816.
- Petrov, V.V., 1995. *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford University Press, New York.
- Powell, J.L., 1994. Estimation of semiparametric models. In: Engle, R.F., McFadden, D.L. (Eds.), *Handbook of Econometrics*, Volume IV. Elsevier, New York, pp. 2443–2521.
- Powell, J.L., 2017. Identification and asymptotic approximations: Three examples of progress in econometric theory. *J. Econ. Perspect.* 31 (2), 107–124.
- Powell, J.L., Stock, J.H., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. *Econometrica* 57 (6), 1403–1430.
- Powell, J.L., Stoker, T.M., 1996. Optimal bandwidth choice for density-weighted averages. *J. Econometrics* 75 (2), 291–316.
- Robinson, P.M., 1995. The normal approximation for semiparametric averaged derivatives. *Econometrica* 63 (3), 667–680.