Bias Correction and Robust Inference in Semiparametric Models

Jungjun Choi and Xiye Yang

Department of Economics
Rutgers University

February, 2021

Abstract

This paper analyzes several different biases that emerge from the (possibly) low-precision nonparametric ingredient in a semiparametric model. We show that both the variance part and the bias part of the nonparametric ingredient can lead to some biases in the semiparametric estimator, under conditions weaker than typically required in the literature. We then propose two bias-robust inference procedures, based on multi-scale jackknife and analytical bias correction, respectively. We also extend our framework to the case where the semiparametric estimator is constructed by some discontinuous functionals of the nonparametric ingredient. The simulation study shows that both bias-correction methods have good finite-sample performance.

JEL classification: C13, C14.
Keywords: Semiparametric two-step estimation, nonparametric estimator, bias, robust inference, multi-scale jackknife, analytical bias correction.

1 Introduction

Recently, increasing attention has been drawn to the interplay between the asymptotic properties of semiparametric estimators and their nonparametric ingredients that could have relatively low precision (e.g., the nonparametric ingredient can have a slower-than-$n^{1/4}$ convergence rate), which may render the previously established asymptotic results invalid. Significant progress has been made by one branch of literature (Cattaneo et al., 2010, 2013, 2014; Calonico et al., 2014; Cattaneo and Jansson, 2018) about "small bandwidth asymptotics" for kernel-based semiparametric estimators and establishes bootstrap inference procedure robust to a bias that has non-negligible impacts when the bandwidth is "small." Another branch of literature (Ichimura and Newey, 2017; Chernozhukov et al., 2017, 2018a,c,b) has creatively introduced an influence function to the GMM semiparametric two-step estimator, to ensure local robustness to the first-step nonparametric ingredient, a property which, as pointed out by (Cattaneo and Jansson, 2018), can be interpreted as "large bandwidth asymptotics" in the case of kernel-based semiparametric estimators.

Motivated by these new results, this paper proposes a general framework to analyze the impacts of several different biases that emerge from the low-precision nonparametric ingredient, including kernel and sieve estimators, on the distributional approximations of the associated semiparametric estimator. We generalize the framework used by (Andrews, 1994), (Newey, 1994), and (Newey and McFadden, 1994), by allowing the nonparametric ingredient to have a convergence rate slower than what is required by the original papers (i.e., a faster-than-$n^{1/4}$ convergence rate). In short, we consider the case where the key Condition (2.8) in (Andrews, 1994) fails to hold. More specifically, we first replace the linear approximation (Assumption 5.1 in (Newey, 1994) and Condition (i) of Theorem 8.1 in (Newey and McFadden, 1994)) in the last two cited papers by a quadratic one. Although this requires a higher-order differentiability condition, it enables us to account for a nonlinear bias, which may appear when the nonparametric ingredient converges slower than $n^{1/4}$. Second, we also relax a restriction jointly implied by the stochastic equicontinuity condition and the mean-square continuity condition (Assumptions 5.2 and 5.3 in (Newey, 1994), and Conditions (ii) and (iii) of Theorem 8.1 in (Newey and McFadden, 1994)), to account for another "linear" bias (see Remarks 2.6 and 2.10 below). Both biases can have non-negligible (in the sense of not being $o_{\mathbb{P}}(n^{-1/2})$) impacts on the distributional approximation of the semiparametric estimator.

As for the sources of the above biases, recall the well-known bias-variance tradeoff in the nonparametric literature. Our analysis shows that the nonlinear bias is related to the variance part of the nonparametric ingredient, while the other bias comes from the nonparametric bias. Theoretically speaking, it is possible to impose certain restriction(s) on the tuning parameter of the nonparametric ingredient so that one bias becomes $o_{\mathbb{P}}(n^{-1/2})$ (e.g., under- or over-smoothing in the kernel case), just like the above-cited recent literature. However, it is often hard to verify such restriction(s) in practice. Besides, even though one bias could be $o_{\mathbb{P}}(n^{-1/2})$ in an asymptotic sense, its effects may not be sufficiently small to be negligible in finite or small samples. Therefore, we do not impose such restriction(s) and allow the possibility that either one or both of them could be larger than $o_{\mathbb{P}}(n^{-1/2})$. By doing so, our distributional approximation will be robust to a larger range of values of the tuning parameter. When specialized to the kernel-based case, this is equivalent to establishing asymptotic results without distinguishing small and large bandwidths. Consequently, the finite sample performance of the corresponding inference procedures will be less sensitive to the choice of the tuning parameter.

In addition to the above two biases that appear in general cases, our analysis also indicates that there can be another special bias for the kernel-based semiparametric estimators. We refer to it as the "singularity bias," which, in our view, is the same as the "leave-in bias" studied by (Cattaneo and Jansson, 2018). In the cited paper, the "leave-in bias" highlights the fundamental difference between the asymptotic separability condition and the stochastic equicontinuity condition therein (see Remark 2.6 for more discussions). Since the framework we adopted is somewhat different, we discuss the "singularity bias" mainly from the perspectives of U-statistics and V-statistics. If we use the same empirical measure to construct the nonparametric and the semiparametric estimators, then the first-order term in our quadratic approximation is a V-statistic. In contrast, if we either use the "leave-one-out" version of the empirical measure to construct the nonparametric estimator, or use a smoothed measure to construct the semiparametric estimator, then the

first-order term becomes a U-statistic. Typically, the difference between a V-statistic and its corresponding U-statistic is very small, often of order $O_{\mathbb{P}}(n^{-1})$. However, the special structure (we believe it is the convolution structure that matters here) of the kernel-based nonparametric estimator can lead to a potentially much larger difference, yielding this special bias. As a comparison, there is no such bias in the sieve-based case.

The second main result of this paper is that we propose two different inference procedures that are robust to the aforementioned biases. The first one is the multi-scale jackknife (MSJ) method, which utilizes the tuning parameter of the nonparametric ingredient in the role of sample size as in the original jackknife method introduced by (Quenouille, 1949). Similar ideas have been adopted by, for example, (Schucany and Sommers, 1977), (Bierens, 1987), (Powell et al., 1989), and (Li et al., 2019). Theoretically speaking, this method can remove *all* aforementioned biases, provided that an appropriate weighting scheme is chosen. In the kernel-based case, this method can automatically remove the "singularity bias," for that it has the same order as the nonlinear bias. If one knows the orders of other smaller biases, one can use more scales to remove these biases as well (refer to the simulation results). The second one is the analytical-based bias correction (ABC) method. It requires a twice Fréchet differentiable assumption (so that one can get the analytical form of the nonlinear bias) and some consistent estimators of both the variance part and the bias part of the nonparametric ingredient. Provided that some other regularity conditions are satisfied, this method can remove or reduce those biases (the remaining bias, if any, will be negligible at a root-n rate).

Last but not least, we show that our framework can be extended to the family of semiparametric estimators that are constructed from discontinuous functionals of the nonparametric ingredients. The requirement is that those discontinuous functionals must have smooth projections, which can be well approximated by quadratic functionals of the nonparametric ingredients. Under certain regularity conditions, the multi-scale jackknife method can yield valid and robust inference. However, the analytical bias correction in this case is more involved, for that one needs to take into account the estimation error and/or bias associated with the unknown smooth projection. Hence, we leave this to future exploration.

The rest of this paper is organized as follows. Section 2 discusses several key properties of a general class of semiparametric estimators and present our first main result, i.e., a distributional approximation that accounts for various biases. In Section 3, we present two inference procedures that are robust to those biases and provide some sufficient conditions to extend the results from the class of twice differentiable functionals to certain discontinuous functionals. Section 4 demonstrates the finite sample performance of the two inference procedures through some simulation results. Section 5 concludes.

2 Asymptotically Linear Semiparametric Estimators

Throughout this paper, any random sequence that is $o_{\mathbb{P}}(n^{-1/2})$ will be referred to as "root-n negligible." We will use C to denote some finite positive number, the value of which may change from line to line. Denote by $\|\|$ the Euclidean norm.

2.1 Asymptotic linearity

Let $\theta_0 \in \Theta$ be a finite-dimensional parameter of interest, where Θ is a subset of some Euclidean space. Suppose that the identification of θ_0 depends on an unknown function $\gamma_0 \in \Gamma$, where Γ represents certain infinite-dimensional functional space. Let $z_1, \cdots, z_n$ be an i.i.d. copies of a random vector $z \in \mathbb{R}^{d_z}$. We shall use x to denote a real vector in $\mathbb{R}^{d_z}$. Suppose that we can sequentially construct two consistent estimators $\hat{\gamma}_n$ and $\hat{\theta}_n$ from this sample.

Let $\mathbb{P}$ and $\mathbb{P}_n$ be the true probability measure and the empirical probability measure, respectively. For any signed measure $\mathbb{Q}$, let $\mathbb{Q}f \coloneqq \int f d\mathbb{Q}$ for any function f. Then for any functional g of (z, θ, γ), define

$$G(\theta, \gamma) \coloneqq \mathbb{P}g = \mathbb{E}[g(z, \theta, \gamma)] \quad \text{and} \quad \widehat{G}_n(\theta, \gamma) \coloneqq \mathbb{P}_n g = \frac{1}{n}\sum_{i=1}^{n} g(z_i, \theta, \gamma).$$

Here the notation $g(z_i, \theta, \gamma)$ is to stress that the moment function is evaluated at the sample point z_i under the empirical measure. The functional g can directly and/or indirectly (i.e., through γ) depend on z_i.

Assumption 2.1 (AL—Asymptotic Linearity in g)**.** *Assume that the estimator $\hat{\theta}_n$ is asymptotically linear. That is,*

$$\hat{\theta}_n - \theta_0 = \mathcal{J}_n \widehat{G}_n(\theta_0, \hat{\gamma}_n) + o_{\mathbb{P}}(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{J}_n g(z_i, \theta_0, \hat{\gamma}_n) + o_{\mathbb{P}}(n^{-1/2}), \tag{2.1}$$

where $\mathcal{J}_n \xrightarrow{\mathbb{P}} \mathcal{J}_0$ for some non-random, finite and non-degenerate $\mathcal{J}_0$ (when it is a matrix, all of its eigenvalues are finite and bounded below from zero), and the functional g satisfies that $G(\theta_0, \gamma_0) = \mathbb{E}[g(z, \theta_0, \gamma_0)] = 0$, which uniquely determines θ_0.

Remark 2.2. *Another way to formulate $\widehat{G}_n$ is to use an estimated probability measure, which is absolutely continuous with respect to the Lebesgue measure. Denote such a measure by $\mathbb{P}_n^{AC}$. For instance, it can be obtained by using a kernel-based method. Now consider the case of estimating the average density $\theta_0 = \mathbb{E}[\gamma_0(z)]$, which implies that $g(z, \theta, \gamma) = \gamma(z) - \theta$. We can then have two different formulations for $\hat{\theta}_n - \theta_0$: one for the average density estimator $\hat{\theta}_n^{AD}$:*

$$\hat{\theta}_n^{AD} - \theta_0 = \widehat{G}_n(\theta_0, \hat{\gamma}_n) = \mathbb{P}_n g = \frac{1}{n}\sum_{i=1}^{n} \big(\hat{\gamma}_n(z_i) - \theta_0\big),$$

and the other one for the integrated squared density estimator $\hat{\theta}_n^{ISD}$ (recall that x is a real vector):

$$\hat{\theta}_n^{ISD} - \theta_0 = \widehat{G}_n(\theta_0, \hat{\gamma}_n) = \mathbb{P}_n^{AC} g = \int \hat{\gamma}_n^2(x) dx - \theta_0.$$

In both cases, $\mathcal{J}_n = \mathcal{J}_0 = I$.

Remark 2.3. *The requirement on $\mathcal{J}_0$ excludes the possibility of weak identification of θ. This may seem to be restrictive. However, we are going to extend the classic theory in a different direction.*

As pointed out by (Andrews and Mikusheva, 2016), the empirical process theory typically implies that the root-n re-scaled sample moment function converges in distribution to the sum of three parts (refer to Equation (1) therein): a mean function, which may allow for various types of identification; a mean-zero Gaussian process, which establishes the central limit theorem; and a residual term, which is typically assumed to be negligible at the root-n rate. While we assume the mean function gives strong identification of θ, we are going to relax the assumption on the residual term and allow it to be non-negligible at the root-n rate.

We note that $\mathcal{J}_n g(z_i, \theta_0, \hat{\gamma}_n)$ gives the influence of a single observation in the leading term of the estimation error $\hat{\theta}_n - \theta_0$. In this sense, it can be viewed as the influence function, following (Hampel, 1974). (Ichimura and Newey, 2017) adopt a very similar definition of asymptotic linearity in their equation (2.1). The only difference is that we introduce the term $\mathcal{J}_n$, in order to focus on the more essential part g of the influence function. As pointed out by (Ichimura and Newey, 2017), under sufficient regularity conditions, almost all root-n consistent semiparametric estimators satisfy Assumption 2.1.

Example (GMM Semiparametric Estimator). Consider a GMM-type estimator $\hat{\theta}_n$:

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} -\frac{1}{2} \widehat{G}_n(\theta, \hat{\gamma}_n)^\intercal W_n \widehat{G}_n(\theta, \hat{\gamma}_n),$$

where $W_n \xrightarrow{\mathbb{P}} W_0$, representing the weighting matrix and its limit. Suppose that g is first-order differentiable at θ_0, then one can readily get

$$\begin{aligned}
\mathcal{J}_n &= [\partial_\theta \widehat{G}_n(\theta_0, \hat{\gamma}_n)^\intercal W_n \partial_\theta \widehat{G}_n(\theta_0, \hat{\gamma}_n)]^{-1} \partial_\theta \widehat{G}_n(\theta_0, \hat{\gamma}_n)^\intercal W_n, \\
\mathcal{J}_0 &= [\partial_\theta G(\theta_0, \gamma_0)^\intercal W_0 \partial_\theta G(\theta_0, \gamma_0)]^{-1} \partial_\theta G(\theta_0, \gamma_0)^\intercal W_0.
\end{aligned}$$

We have $\mathcal{J}_n \xrightarrow{\mathbb{P}} \mathcal{J}_0$, if $\partial_\theta g(\theta_0, \gamma)$ is continuous with respect to γ in a neighborhood of γ_0.

The above example shows a subtle difference in the definition of asymptotic linearity between this paper and those in (Ichimura and Newey, 2017) and (Cattaneo and Jansson, 2018). In this paper, the term $\mathcal{J}_n$ can be random, hence can be different from $\mathcal{J}_0$ in a non-trivial way. However, in the GMM examples of the two cited papers, the authors set $\mathcal{J}_n \equiv \mathcal{J}_0$ (cf. (2.2) in (Ichimura and Newey, 2017) and the discussion following Condition AL in (Cattaneo and Jansson, 2018)). It is easy to see that if the following condition holds

$$(\mathcal{J}_n - \mathcal{J}_0) \widehat{G}_n(\theta_0, \hat{\gamma}_n) = o_{\mathbb{P}}(n^{-1/2}), \tag{2.2}$$

then the above definition can be modified to be exactly the same as the two cited papers. A sufficient condition for (2.2) is $\widehat{G}_n(\theta_0, \hat{\gamma}_n) = O_{\mathbb{P}}(n^{-1/2})$, which indeed holds in a lots of applications. This sufficient condition may not hold in the current paper, since we are going to consider the general case where $\widehat{G}_n(\theta_0, \hat{\gamma}_n)$ could have some bias(es) that can be larger than $O_{\mathbb{P}}(n^{-1/2})$ in order. However, eventually, we will make sure that Condition (2.2) is satisfied (see Lemma 2.13 for details).

2.2 Quadratic approximation of $\widehat{G}_n(\theta_0, \hat{\gamma}_n)$

To begin with, we have the following decomposition (recall that $G(\theta_0, \gamma_0) = 0$)

$$\widehat{G}_n(\theta_0, \hat{\gamma}_n) = \widehat{G}_n(\theta_0, \hat{\gamma}_n) - \widehat{G}_n(\theta_0, \gamma_0) + \widehat{G}_n(\theta_0, \gamma_0) - G(\theta_0, \gamma_0).$$

The first difference is the impact of replacing γ_0 by its estimator in the empirical moment condition, while the second one is the difference between a sample average and its expectation, to which we can apply the central limit theorem (CLT) for i.i.d. random variables.

We introduce the following assumption on g, in order to get a more detailed evaluation of the first term.

Assumption 2.4 (Quadraticity)**.** *Suppose that the following (stochastic) quadratic approximation of the functional g holds around (θ_0, γ_0) for sufficiently large n:*

$$\begin{aligned} g(z_i, \theta_0, \hat{\gamma}_n) = {} & g(z_i, \theta_0, \gamma_0) + g'_\gamma(z_i, \theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0) + \frac{1}{2} g''_{\gamma\gamma}(z_i, \theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0, \hat{\gamma}_n - \gamma_0) \\ & + g_R(z_i, \theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0), \end{aligned}$$

where $g'_\gamma(z_i, \theta_0, \gamma_0, \cdot)$ is a linear functional, $g''_{\gamma\gamma}(z_i, \theta_0, \gamma_0, \cdot, \cdot)$ is a bi-linear functional and symmetric in its two inputs (the subscript γ indicates that these functionals are from the expansion with respect to γ, not z or θ), and the functional g_R captures the remainder of this expansion. We assume that $\mathbb{E}[\|g'_\gamma(z_i, \theta_0, \gamma_0, \gamma - \gamma_0)\|] \leqslant C\,\mathbb{E}[\|\gamma(z_i) - \gamma(z_i)\|]$, $\mathbb{E}[\|g''_{\gamma\gamma}(z_i, \theta_0, \gamma_0, \gamma - \gamma_0, \gamma - \gamma_0)\|] \leqslant C\,\mathbb{E}[\|\gamma(z_i) - \gamma(z_i)\|^2]$, and $\mathbb{E}[\|g_R(z_i, \theta_0, \gamma_0, \gamma - \gamma_0)\|] \leqslant C\,\mathbb{E}[\|\gamma(z_i) - \gamma_0(z_i)\|^3]$ for γ sufficiently close to γ_0 and some finite number C.

Compared to Assumption 5.1 (Linearization) in (Newey, 1994) and Condition (i) of Theorem 8.1 in (Newey and McFadden, 1994), the above assumption requires a second-order, instead of first-order, differentiability of g with respect to γ, which could be a random function, such as $\hat{\gamma}_n$. However, the two cited papers both require that $\|\hat{\gamma}_n(z_i) - \gamma_0(z_i)\|^2 = o_{\mathbb{P}}(n^{-1/2})$. In other words, the nonparametric estimator $\hat{\gamma}_n$ must have a faster-than-$n^{1/4}$ convergence rate (i.e., $r > 1/4$ and $s > 1/4$ in Assumption 2.12 below). Yet, as to be shown later, we just need $\|\hat{\gamma}_n(z_i) - \gamma_0(z_i)\|^3 = o_{\mathbb{P}}(n^{-1/2})$, which only requires a faster-than-$n^{1/6}$ convergence rate for $\hat{\gamma}_n$. With this slower convergence rate, we may have some non-root-n-negligible biases.

Define the following terms using the empirical measure $\mathbb{P}_n$:

$$\begin{aligned} \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \eta) &:= \frac{1}{n} \sum_{i=1}^n g'_\gamma(z_i, \theta_0, \gamma_0, \eta), \\ \widehat{G}''_{n,\gamma\gamma}(\theta_0, \gamma_0, \eta, \phi) &:= \frac{1}{n} \sum_{i=1}^n g''_{\gamma\gamma}(z_i, \theta_0, \gamma_0, \eta, \phi). \end{aligned}$$

The quadraticity assumption implies that, for sufficiently large n, we have

$$\widehat{G}_n(\theta_0, \hat{\gamma}_n) = \widehat{G}_n(\theta_0, \gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0) + \frac{1}{2} \widehat{G}''_{n,\gamma\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0, \hat{\gamma}_n - \gamma_0)$$

$$+ \widehat{G}_{n,R}(\theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0),$$

where $\widehat{G}_{n,R}(\theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0) = \frac{1}{n}\sum_{i=1}^n g_R(z_i, \theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0)$.

Remark 2.5. *In the case where we use the measure $\mathbb{P}_n^{AC}$, instead of $\mathbb{P}_n$, to construct $\widehat{G}$, we apply Assumption 2.4 to an equivalent functional $\tilde{g}$, which will be evaluated at a real vector x, defined as follows. Let $\mathbb{L}$ be the Lebesgue measure, ν_0 be the true density function of z, which may or may not be part of γ_0, and $\hat{\nu}_n = d\mathbb{P}_n^{AC}/d\mathbb{L}$. Then we have $\mathbb{P}g = \mathbb{E}[g] = \mathbb{L}[g(\cdot, \theta_0, \gamma_0)\nu_0(\cdot)]$ and $\mathbb{P}_n^{AC}g = \mathbb{L}(g(\cdot, \theta_0, \hat{\gamma}_n)\hat{\nu}_n(\cdot))$. Hence, we set $\tilde{g}(\theta, \gamma, \nu) := \mathbb{L}[g(\cdot, \theta, \gamma)\nu(\cdot)]$. In the special case where ν_0 is part of γ_0, we can write $\tilde{g}(\theta, \gamma, \nu)$ as $\tilde{g}(\theta, \gamma)$. In the end, we suppose that Assumption 2.4 holds true for the functional $\tilde{g}$ with respect to (γ, ν) around (γ_0, ν_0).*

Throughout this paper, we assume that $\hat{\gamma}_n$ is a consistent estimator of the unknown function γ_0. Yet, such a nonparametric estimator is often biased, leading to the well-known bias-variance tradeoff in the nonparametric literature. In the semiparametric literature, it is often assumed that the nonparametric bias is sufficiently small so that this bias is root-n negligible, causing no problems for the associated semiparametric estimator (that is, $G'_\gamma(\theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0) := \mathbb{E}[\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0)] = o_{\mathbb{P}}(n^{-1/2})$). Since we aim at relaxing such an assumption, we are going to separate the bias part from the variance part. The idea is to introduce a function $\bar{\gamma}_n$ such that $G'_\gamma(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n) := \mathbb{E}[\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n)]$ is identically zero or at least $o_{\mathbb{P}}(n^{-1/2})$, no matter how one chooses the tuning parameter. Then we obtain a more detailed decomposition:

$$\begin{aligned}\widehat{G}_n(\theta_0, \hat{\gamma}_n) = &\widehat{G}_n(\theta_0, \gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n) + \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0) \\ &+ \frac{1}{2}\widehat{G}''_{n,\gamma\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n, \hat{\gamma}_n - \bar{\gamma}_n) + \widehat{G}''_{n,\gamma\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n, \bar{\gamma}_n - \gamma_0) \\ &+ \frac{1}{2}\widehat{G}''_{n,\gamma\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0, \bar{\gamma}_n - \gamma_0) + \widehat{G}_{n,R}(\theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0).\end{aligned} \tag{2.3}$$

Here, we would expect to establish a central limit theorem for the sum of the first two terms. The third and fourth terms are the two main biases that we are going to analyze. Intuitively, we may defined $\bar{\gamma}_n$ as $\bar{\gamma}_n := \mathbb{E}[\hat{\gamma}_n]$. However, this may not necessarily lead to the desired result. Instead, we are going to use the definition $\bar{\gamma}_n(z_i) := \mathbb{E}[\hat{\gamma}_n(z_i)|z_i]$, especially when there is a "singularity bias."

2.3 V-statistic and U-statistic

To begin with, consider the case where we also use the empirical measure $\mathbb{P}_n$ to construct $\hat{\gamma}_n$. Without much loss of generality, suppose that there exists some function ψ such that $\hat{\gamma}_n(\cdot) = \mathbb{P}_n\psi(\cdot) = \frac{1}{n}\sum_{j=1}^n \psi(\cdot, z_j)$ ((Newey and McFadden, 1994) adopt a similar representation in Section 8 therein). Moreover, it is reasonable to assume that $g'_\gamma(z_i, \theta_0, \gamma_0, \hat{\gamma}_n)$ can be reduced to $g'_\gamma\big(z_i, \theta_0, \gamma_0, \hat{\gamma}_n(z_i)\big)$. Consequently, the linearity of $g'_\gamma(z, \theta_0, \gamma_0, \cdot)$ implies that

$$\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n) = \frac{1}{n}\sum_{i=1}^n g'_\gamma\big(z_i, \theta_0, \gamma_0, \hat{\gamma}_n(z_i)\big)$$

$$= \frac{1}{n}\sum_{i=1}^{n} g'_\gamma\big(z_i, \theta_0, \gamma_0, \frac{1}{n}\sum_{i=1}^{n}\psi(z_i, z_j)\big) = \frac{1}{n^2}\sum_{i,j=1}^{n} g'_\gamma\big(z_i, \theta_0, \gamma_0, \psi(z_i, z_j)\big)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} g'_\gamma\big(z_i, \theta_0, \gamma_0, \psi(z_i, z_i)\big) + \frac{1}{n^2}\sum_{i\neq j} g'_\gamma\big(z_i, \theta_0, \gamma_0, \psi(z_i, z_j)\big),$$

where the sum $\sum_{i\neq j}$ is taken over $1 \leqslant i, j \leqslant n$ with $i \neq j$.

It is then clear that $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n)$ is a V-statistic in this case. Typically, the difference between a V-statistic and its corresponding U-statistic is rather small, often of order $O_{\mathbb{P}}(1/n)$. However, as to be shown in the following example of the kernel density estimator, it sometimes can be larger than $O_{\mathbb{P}}(1/n)$, or even $O_{\mathbb{P}}(n^{-1/2})$. The following example highlights the potentially "large" difference between V- and U-statistics, when the nonparametric ingredient has sufficiently low precision.

Example (Kernel Density Estimator). Suppose that γ_0 is the density function of each z_i. Le K be a kernel function with order m and $K_h(\cdot) \coloneqq K(\cdot/h)/h^{d_z}$. The kernel density estimator $\hat{\gamma}_n$ at a real vector $x \in \mathbb{R}^{d_z}$ and at a sampling point z_i are given by

$$\hat{\gamma}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - z_i) \quad \text{and} \quad \hat{\gamma}_n(z_i) = \frac{K(0)}{nh^{d_z}} + \frac{1}{n}\sum_{\substack{j=1\\ i\neq j}}^{n} K_h(z_i - z_j),$$

respectively. In this case, we have $\psi(x, y) = K_h(x - y)$ (note that the kernel method is closely related to convolution). In the expression of $\hat{\gamma}_n(z_i)$, the term $\psi(z_i, z_i) = K_h(z_i - z_i) = K(0)/(nh^{d_z})$ is non-random. This shows a difference between $\hat{\gamma}_n(x)$ and $\hat{\gamma}_n(z_i)$, which is quite important when $1/(nh^{d_z})$ is not $o(n^{-1/2})$. It is easy to see that $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n)$ becomes

$$\frac{1}{nh^{d_z}}\frac{1}{n}\sum_{i=1}^{n} g'_\gamma\big(z_i, \theta_0, \gamma_0, K(0)\big) + \frac{1}{n^2}\sum_{i\neq j} g'_\gamma\big(z_i, \theta_0, \gamma_0, K_h(z_i - z_j)\big).$$

In general, the first term is of order $O_{\mathbb{P}}(1/(nh^{d_z}))$, which may not be root-$n$ negligible. Since it is from $K_h(z_i - z_i)$, which behaves differently from $K_h(z_i - z_j)$ with $j \neq i$, we refer to it as the "singularity bias" (or maybe "non-smoothing bias").

On the other hand, we have $\bar{\gamma}_n(x) = \mathbb{E}[\hat{\gamma}_n(x)] = \int K(u)\gamma_0(x - hu)du$. The plug-in definition then leads to $\bar{\gamma}_n(z_i) = \int K(u)\gamma_0(z_i - hu)du$. According to the Law of Iterated Expectation, we readily get

$$G'_\gamma(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n) = \frac{1}{nh^{d_z}}\mathbb{E}\big[\widehat{G}'_{n,\gamma}\big(\theta_0, \gamma_0, K(0)\big)\big] + O(\frac{1}{n}) = O\big(\frac{1}{nh^{d_z}}\big).$$

The sufficient and necessary condition for this term to be root-n negligible is $n^{1/4} = o(\sqrt{nh^{d_z}})$, which is equivalent to a faster-than-$n^{1/4}$ convergence rate for the kernel density estimator $\hat{\gamma}_n$. Since we aim at relaxing this requirement, the above plug-in definition of $\bar{\gamma}_n$ does not suit our purpose.

To address this problem, we can modify the definition of $\bar{\gamma}_n$ at sample points $\{z_i\}_{i=1}^n$, which are more important when we use the empirical measure $\mathbb{P}_n$ to construct $\widehat{G}_n$. More specifically, we define ($\bar{\gamma}_n(x)$ remains the same as above for any real vector x)

$$\bar{\gamma}_n(z_i) \coloneqq \mathbb{E}[\hat{\gamma}_n(z_i)|z_i] = \frac{1}{nh^{d_z}}K(0) + \frac{n-1}{n}\int K(u)\gamma_0(z_i - hu)du,$$

With this modified $\bar{\gamma}_n$, we move the "singularity bias" to $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0)$. One can check that $G'_\gamma(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n) = \mathbb{E}[\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n)] = 0$.

With the modified definition of $\bar{\gamma}_n$, we readily get

$$\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n) = \frac{1}{n(n-1)} \sum_{i \neq j} g'_\gamma\big(z_i, \theta_0, \gamma_0, \phi(z_i, z_j)\big) \times \big(1 - \frac{1}{n}\big),$$

where $\phi(z_i, z_j) := \psi(z_i, z_j) - \mathbb{E}[\psi(z_i, z_j)|z_i]$. Its difference with the associated U-statistic is at most $O_{\mathbb{P}}(n^{-1})$, which is always root-n negligible. However, in this case, we may still have the "singularity bias" in $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0)$, if $\hat{\gamma}_n$ is a kernel-based estimator.

Example (Sieve Estimator). Let $z = (Y, X^\intercal)^\intercal$. Consider a conditional mean model for Y and X: $\gamma_0(z, \theta) = \mathbb{E}[\rho(Y, \theta)|X]$. Following the notation used by (Chen, 2007), we denote by $\{p_{0j}(X), j = 1, 2, \cdots, k_{m,n}\}$ a sequence of known basis functions in the space of square integrable functions. Let $p^{k_{m,n}}(X) = (p_{01}(X), \cdots, p_{0k_{m,n}}(X))^\intercal$ and $P = (p^{k_{m,n}}(X_1), \cdots, p^{k_{m,n}}(X_n))^\intercal$. Then the sieve estimator of γ_0 is given by

$$\hat{\gamma}_n(z_i, \theta) = \frac{1}{n} \sum_{j=1}^n \rho(Y_j, \theta) p^{k_{m,n}}(X_j)^\intercal \big(P^\intercal P\big)^+ p^{k_{m,n}}(X_i) = \frac{1}{n} \sum_{j=1}^n \psi(z_i, z_j),$$

where $(P^\intercal P)^+$ is the Moore-Penrose inverse of $P^\intercal P$. In this case, $\psi(z_i, z_i)$ does not lead to a "singularity bias."

The above two examples show that only the kernel-based estimator may suffer from the "singularity bias" problem. In certain cases, such as the average density estimator to be discussed in the next subsection, it might be desirable to remove this bias in advance. As implied by the example of the sieve estimator, one way to get rid of this bias is to use a global nonparametric estimator. Besides, there are two alternative solutions. However, we stress that it is not always necessary to remove the "singularity bias" in advance (see the discussions in Section 3.1).

One (possible) solution is to use the measure $\mathbb{P}_n^{\texttt{AC}}$, instead of $\mathbb{P}_n$, to construct $\widehat{G}_n$. For simplicity, recall the integrated density estimator $\hat{\theta}_n^{\texttt{ISD}}$. In this case, the linear functional

$$\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n) = 2 \int \gamma_0(x) \hat{\gamma}_n(x) dx = \frac{2}{n} \sum_{i=1}^n \int \gamma_0(x) \psi(x, z_i) dx$$

is a U-statistic of degree 1. In general, even when ν_0 is not part of γ_0 (recall Remark 2.5), the above functional is still a U-statistic, hence is not subject to the "singularity bias." Hence, we don't have to make any adjustment to $\bar{\gamma}_n$, as we do not evaluate $\hat{\gamma}_n$ at the sample points. However, as to be shown in the next subsection, this solution increases the level of nonlinearity, hence may bring additional nonlinear bias.

Another solution is to replace the above V-statistic by its corresponding U-statistic. In other words, we can use the "leave-one-out" empirical measure $\mathbb{P}_n^{\texttt{LOO}}$ to construct the nonparametric estimator $\hat{\gamma}_n$. That is, let $\hat{\gamma}_n(z_i) := \mathbb{P}_n^{\texttt{LOO}} \psi(z_i, \cdot) = \frac{1}{n-1} \sum_{j=1, j \neq i}^n \psi(z_i, z_j)$. It is then obvious that

$$\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n) = \frac{1}{n(n-1)} \sum_{i \neq j} g'_\gamma\big(z_i, \theta_0, \gamma_0, \psi(z_i, z_j)\big)$$

is a U-statistic of degree 2, following the terminology of (Hoeffding, 1948). It then follows that $\hat{\gamma}_n(z_i) - \bar{\gamma}_n(z_j) = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \phi(z_i, z_j)$ and

$$\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n) = \frac{1}{n(n-1)} \sum_{i \neq j} g'_\gamma\big(z_i, \theta_0, \gamma_0, \phi(z_i, z_j)\big).$$

That is, the term $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n)$ is also a U-statistic of degree 2. In addition, there is no "singularity bias" in $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0)$. Moreover, this will not bring any additional nonlinear biases. Hence, we recommend this method whenever it is feasible.

Remark 2.6 (Stochastic Equicontinuity Condition)**.** *(Cattaneo and Jansson, 2018) have insightfully observed that, in the kernel-based case, the "singularity bias" is a key in understanding the difference between the stochastic equicontinuity (SE) condition and the asymptotic separability (AS) condition. We note that the AS condition in the cited paper may involve quadratic terms. Below, we offer a different perspective that is only based on the first-order term in the approximation of g.*

The stochastic equicontinuity condition given in Assumption 5.2 in (Newey, 1994) or Condition (ii) in (Newey and McFadden, 1994) (the formulation given by (Andrews, 1994) is a bit different. So we defer the discussion to Remark 2.11) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\Big(g'_\gamma(z_i, \theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0) - \int g'_\gamma(z, \theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0) dF_0\Big) = o_{\mathbb{P}}(n^{-1/2}), \tag{2.4}$$

where F_0 is the true distribution function of z. The integral does not involve the "singularity bias" because one evaluates the functional $g'_{n,\gamma}$ at a real vector x, not a sample point z_i, when calculating the integral. Therefore, when $\hat{\gamma}_n$ is the original kernel density estimator, the "singularity bias" only appears in the first term. The sample average of the "singularity bias" is of order $O_{\mathbb{P}}(\frac{1}{nh^{d_z}})$ (if g only depends on z_i through γ, this becomes $O(\frac{1}{nh^{d_z}})$, which is not $o_{\mathbb{P}}(n^{-1/2})$ when $\hat{\gamma}_n$ does not have a faster-than-$n^{1/4}$ converges rate.

If one uses the "leave-one-out" kernel estimator or a sieve estimator, then there is no "singularity bias" (this might also be achieved by replacing the input z in the integrand by z_i). Hence, it might be possible that the above SE condition also holds true with a low precision $\hat{\gamma}_n$. However, as to be shown in Remark 2.10, the mean-square continuity condition will fail in such case, when the convergence rate of $\hat{\gamma}_n$ is relatively slow.

As a summary of the above discussion, no matter how we construct $\widehat{G}_n$ and $\hat{\gamma}_n$, we can always find $\bar{\gamma}_n$ such that $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_n, \hat{\gamma}_n - \bar{\gamma}_n)$ is a U-statistic, or its difference with a U-statistic is always root-n negligible. Given such a suitable $\bar{\gamma}_n$, we are ready to introduce the following assumption on the asymptotic behavior of the sum of the first two terms in (2.3).

Assumption 2.7 (AN—Asymptotic Normality)**.** *For some non-random and positive definite Σ_g, we have*

$$\sqrt{n}\left(\widehat{G}_n(\theta_0, \gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n)\right) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \Sigma_g).$$

Remark 2.8. *The first two terms in (2.3) have been intensively studied in the literature, mostly under the assumption that all biases are root-n negligible. Recall that*

$$\widehat{G}_n(\theta_0,\gamma_0)+\widehat{G}'_{n,\gamma}(\theta_0,\gamma_0,\hat{\gamma}_n-\bar{\gamma}_n)=\frac{1}{n}\sum_{i=1}^{n}\big(g(z_i,\theta_0,\gamma_0)+g'_\gamma(z_i,\theta_0,\gamma_0,\hat{\gamma}_n-\bar{\gamma}_n)\big).$$

The functionals $g(z,\theta_0,\gamma_0)$ and $g'_\gamma(z,\theta_0,\gamma_0,\hat{\gamma}_n-\bar{\gamma}_n)$ are respectively very similar to, for instance, $m(z,h_0)$ and $D(z,h-h_0)$ studied by (Newey, 1994), or $g(z,\gamma_0)$ and $G(z,\gamma-\gamma_0)$ analyzed by (Newey and McFadden, 1994). Note that when all biases are root-n negligible, the terms $h-h_0$ and $\gamma-\gamma_0$ in the cited papers behave essentially the same as $\hat{\gamma}_n-\bar{\gamma}_n$ in the current paper.

The previous discussion suggests that both $\widehat{G}_n(\theta_0,\gamma_0)$ and $\widehat{G}'_{n,\gamma}(\theta_0,\gamma_0,\hat{\gamma}_n-\bar{\gamma}_n)$ can be essentially viewed as U-statistics. Hence, although Assumption 2.7 is a high-level assumption, it is a direct result from the well-established theory on U-statistic (see, e.g., (Hoeffding, 1948), (Korolyuk and Borovskich, 1994), and (Borovskikh, 1996)) in most if not all cases. Therefore, we would expect it to be true under quite general conditions. In particular, it may also hold true for weakly dependent observations. Refer to (Dehling, 2006) and the references therein for more details.

Remark 2.9. *When $\hat{\gamma}_n(\cdot)=\mathbb{P}_n\psi(\cdot)=\frac{1}{n}\sum_{j=1}^{n}\psi(\cdot,z_j)$, let $\psi_g(z_i,z_j)\coloneqq g'_\gamma(z_i,\theta_0,\gamma_0,\psi(z_i,z_j))$ and $\phi_g(z_i,z_j)\coloneqq\psi_g(z_i,z_j)-\mathbb{E}[\psi_g(z_i,z_j)|z_i]$.*

According to the previous discussions, the term $\widehat{G}'_{n,\gamma}(\theta_0,\gamma_0,\hat{\gamma}_n-\bar{\gamma}_n)$ is (approximately) a U-statistic:

$$U_n=\frac{1}{n(n-1)}\sum_{\substack{i,j=1\\ j\neq i}}^{n}g'_\gamma(z_i,\theta_0,\gamma_0,\phi(z_i,z_j))=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}\frac{1}{2}[\phi_g(z_i,z_j)+\phi_g(z_j,z_i)].$$

Its projection $\widehat{U}_n$ is given by

$$\widehat{U}_n=\frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}[\psi_g(z_j,z_i)|z_i]-\mathbb{E}[\psi_g(z_j,z_i)]\Big),\ \text{where } j\neq i.$$

The U-statistic projection theory implies that $\sqrt{n}(U_n-\widehat{U}_n)\xrightarrow{\mathbb{P}}0$. On the other hand, the statistic $\widehat{U}_n$ is a sum of i.i.d. random variables with zero mean. Hence, the asymptotic normality of $\widehat{G}'_{n,\gamma}(\theta_0,\gamma_0,\hat{\gamma}_n-\bar{\gamma}_n)$ can be established. If we also know its correlation with $\widehat{G}_n(\theta_0,\gamma_0)$, then Assumption 2.7 readily follows.

Consider the average density example, in which $g(z,\theta,\gamma)=\gamma(z)-\theta$. It can be shown that

$$\sqrt{n}\,\widehat{G}_n(\theta_0,\gamma_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\gamma_0(z_i)-\theta_0],$$
$$\sqrt{n}\,\widehat{G}'_{n,\gamma}(\theta_0,\gamma_0,\hat{\gamma}_n-\bar{\gamma}_n)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\gamma_0(z_i)-\theta_0]+o_{\mathbb{P}}(1).$$

Hence, Assumption 2.7 holds with $\Sigma_g = 4\,Var[\gamma_0(z)]$. *As a comparison, if* γ_0 *were known, then we would be able to estimate* θ_0 *by* $\widehat{G}_n(\theta_0, \gamma_0)$, *the asymptotic variance of which is* $Var[\gamma_0(z)]$. *This shows the efficiency loss due to not knowing* γ_0.

It is worth mentioning that the main advantage of this U-statistic perspective is that the asymptotic normality result with a root-n rate can be established (provided that the U-statistic is not degenerate), regardless of the convergence rate of $\hat{\gamma}_n - \bar{\gamma}_n$, which has no (asymptotic) biases by construction. Hence, if we can correct for those biases, then we can have asymptotic normality result for $\hat{\theta}_n$ even in the case of having a low precision nonparametric ingredient.

2.4 Possibly non-root-n-negligible biases

Most previous asymptotic results for semiparametric two-step estimators, e.g., (Andrews, 1994), (Newey, 1994), (Newey and McFadden, 1994), (Chen, 2007), and (Ichimura and Todd, 2007), impose certain conditions so that all the biases are root-n negligible. Recent literature (recall the cited papers in the beginning of introduction) has started to relax such an assumption, so that some biases may have non-trivial impacts on the asymptotic distribution of $\hat{\theta}_n$.

Intuitively, one would expect the following two terms dominate the last three terms in the decomposition (2.3):

$$\mathcal{B}_n^{\texttt{ANB}} := \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0) \quad \text{and} \quad \mathcal{B}_n^{\texttt{NL}} := \frac{1}{2}\widehat{G}''_{n,\gamma\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n, \hat{\gamma}_n - \bar{\gamma}_n).$$

The term $\mathcal{B}_n^{\texttt{ANB}}$ represents the sample average of the nonparametric bias(es), while $\mathcal{B}_n^{\texttt{NL}}$ is a nonlinear bias.

Remark 2.10 (Mean-square Continuity Condition)**.** *Together with the stochastic equicontinuity condition (refer to Remark 2.6 for the equivalent formulation in the current context), Assumption 5.3 in (Newey, 1994) and Condition (iii) of Theorem 8.1 in (Newey and McFadden, 1994) imply that there exists* $\alpha(z)$ *(or* $\delta(z)$ *in the latter paper) such that* $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \gamma_0) = \frac{1}{n}\sum_{i=1}^n \alpha(z_i) + o_{\mathbb{P}}(n^{-1/2})$ *(we modified the original expression to adapt to the current context) and* $\mathbb{E}[\alpha(z)] = 0$.

It is easy to see that $\alpha(z) \equiv g'_\gamma(z, \theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n)$ *satisfies the second requirement (this can also be verified from a comparison of the asymptotic variances in the cited papers and in Assumption 2.7). Then the first condition essentially requires* $\mathcal{B}_n^{ANB} = \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0) = o_{\mathbb{P}}(n^{-1/2})$. *However, we are going to relax this restriction and allow* $\mathcal{B}_n^{ANB}$, *which may or may not include the "singularity bias," to be non-root-n-negligible. Following the discussion in Remark 2.6, even though it might be possible to reformulate the original stochastic equicontinuity condition in the two above-cited papers to make it hold true, the mean-square continuity condition will not hold in the current setting.*

Remark 2.11 (Condition (2.8) in (Andrews, 1994))**.** *A main result that (Andrews, 1994) intended to derive from the SE condition is (2.8) therein. Using the notation of the current*

paper, it can be written as:

$$\widehat{G}_n(\theta_0, \hat{\gamma}_n) - \widehat{G}_n(\theta_0, \gamma_0) = o_{\mathbb{P}}(n^{-1/2}).$$

However, both $\mathcal{B}_n^{ANB}$ *and* $\mathcal{B}_n^{NL}$*, two components of the left hand side difference, can be non-root-*n*-negligible, when the precision of* $\hat{\gamma}_n$ *is low.*

Different from the previous discussion about asymptotic normality, the analysis of the above possibly non-root-n-negligible biases critically hinges on the order of $\hat{\gamma}_n - \bar{\gamma}_n$ and/or $\bar{\gamma}_n - \gamma_0$. Therefore, given a suitably defined $\bar{\gamma}_n$, we introduce the following high-level assumption on the asymptotic behavior of the nonparametric estimator $\hat{\gamma}_n$.

Assumption 2.12 (Bias Order)**.** *Suppose that* $\mathcal{B}^{NL} = \mathbb{E}[\mathcal{B}_n^{NL}] = O(n^{-2r})$ *and* $\mathcal{B}^{ANB} = \mathbb{E}[\mathcal{B}_n^{ANB}] = O(n^{-s})$*, where* $r, s > 0$ *such that*

$$\|\mathcal{B}_n^{ANB} - \mathcal{B}^{ANB}\| = o_{\mathbb{P}}(n^{-1/2}) \quad \textit{and} \quad \|\mathcal{B}_n^{NL} - \mathcal{B}^{NL}\| = o_{\mathbb{P}}(n^{-1/2}).$$

Here, we allow $2r$ *and/or* s *to be smaller than or equal to* $1/2$*.*

Typically, the above rates should depend on the tuning parameter of the nonparametric estimator $\hat{\gamma}_n$. Since it is a common practice to set the tuning parameter as a function of n eventually, we express all the rates in the above assumption in terms of a power of n, for convenience.

Compared with the previous requirement that both $\mathcal{B}_n^{\mathtt{NL}}$ and $\mathcal{B}_n^{\mathtt{ANB}}$ are $o_{\mathbb{P}}(n^{-1/2})$, Assumption 2.12 is much weaker. It requires no more than splitting each (asymptotically negligible) bias into two components: one is $o_{\mathbb{P}}(n^{-1/2})$, while the other is not. In this sense, it should be satisfied under very general conditions. For example, when $\hat{\gamma}_n(z_i) - \bar{\gamma}_n(z_i) = \frac{1}{n-1}\sum_{j\neq i}\phi(z_i, z_j)$ as above, we can obtain

$$\widehat{G}''_{n,\gamma\gamma}\big(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n, \hat{\gamma}_n - \bar{\gamma}_n\big) = \frac{1}{n-1}U_{n,1} + \frac{n-2}{n-1}U_{n,2},$$

where $U_{n,1}$ and $U_{n,2}$ are two U-statistics:

$$U_{n,1} = \frac{1}{n(n-1)}\sum_{i\neq j} g''_{\gamma\gamma}(z_i, \theta_0, \gamma_0, \phi(z_i, z_j), \phi(z_i, z_j)),$$
$$U_{n,2} = \frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq l} g''_{\gamma\gamma}(z_i, \theta_0, \gamma_0, \phi(z_i, z_j), \phi(z_i, z_l)).$$

By using the U-statistic theory (refer to the appendix for details), we can show that

$$\mathcal{B}^{\mathtt{NL}} = \frac{1}{n-1}\mathbb{E}[U_{n,1}], \quad \mathrm{Var}(U_{n,1}) \leqslant \frac{C}{n}\Big(\mathbb{E}\big(\mathbb{E}[\psi(z_i, z_j)^2|z_i]^2\big) + \mathbb{E}\big(\mathbb{E}[\psi(z_i, z_j)^2|z_j]^2\big)\Big),$$
$$\mathbb{E}[U_{n,2}] = 0, \quad \mathrm{Var}(U_{n,2}) \leqslant \frac{C}{n^2}\mathbb{E}\big(\psi(z_i, z_j)^2\psi(z_i, z_l)^2\big).$$

Hence, the first condition of Assumption 2.12 holds if the assumptions of Lemma A.1 are true. As for the second one, note that $\mathcal{B}_n^{\texttt{ANB}}$ is the average of a sequence of i.i.d. random variables with negligible variance:

$$\mathcal{B}_n^{\texttt{ANB}} := \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0) = \frac{1}{n}\sum_{i=1}^{n} g'_\gamma(z_i, \theta_0, \gamma_0, \bar{\gamma}_n(z_i) - \gamma_0(z_i)).$$

Let $\mathcal{B}^{\texttt{ANB}} = \mathbb{E}[g'_\gamma(z_i, \theta_0, \gamma_0, \bar{\gamma}_n(z_i) - \gamma_0(z_i))]$. We then have

$$\|\mathcal{B}^{\texttt{ANB}}\| \leqslant \mathbb{E}[\|g'_\gamma(z_i, \theta_0, \gamma_0, \bar{\gamma}_n(z_i) - \gamma_0(z_i))\|] \leqslant C\,\mathbb{E}[\|\bar{\gamma}_n(z_i) - \gamma_0(z_i)\|],$$
$$\mathrm{Var}(\mathcal{B}_n^{\texttt{ANB}}) = \frac{1}{n}\mathrm{Var}\big(g'_\gamma(z_i, \theta_0, \gamma_0, \bar{\gamma}_n(z_i) - \gamma_0(z_i))\big) \leqslant \frac{C}{n}\mathrm{Var}\big(\bar{\gamma}_n(z_i) - \gamma_0(z_i)\big).$$

Hence, the desired result readily follows.

Example (Kernel Density Estimator Continued). Suppose $\hat{\gamma}_n$ is the "leave-one-out" estimator. We have $\psi(z_i, z_j) = K_h(z_i - z_j)$. It then follows that (refer to the appendix for detailed calculation)

$$\frac{1}{n-1}\mathbb{E}[\psi(z_i, z_j)^2] = \frac{1}{(n-1)h^{d_z}}\int K^2(u)\gamma_0(x)\gamma_0(x - hu)dudx = O\Big(\frac{1}{nh^{d_z}}\Big),$$
$$\mathbb{E}\big(\mathbb{E}[\psi(z_i, z_j)^2|z_i]^2\big) = \mathbb{E}\big(\mathbb{E}[\psi(z_i, z_j)^2|z_j]^2\big) = O\Big(\frac{1}{h^{2d_z}}\Big),$$
$$\mathbb{E}\big[\mathbb{E}[\psi(z_i, z_j)\psi(z_i, z_l)|z_j, z_l]^2\big] = O\Big(\frac{1}{h^{d_z}}\Big).$$

Hence, the corresponding conditions given in (A.1) only require that

$$nh^{d_z} = n^{2r} \to \infty.$$

Note that the convergence rate of $\hat{\gamma}_n - \bar{\gamma}_n$ is given by $\sqrt{nh^{dz}}$. Hence, the above condition merely requires that $\hat{\gamma}_n - \bar{\gamma}_n$ converges to zero.

On the other hand, we have

$$\bar{\gamma}_n(z_i) - \gamma_0(z_i) = \int K(u)[\gamma_0(z_i - hu) - \gamma_0(z_i)]du.$$

It the easy to see that the second last condition given in (A.1) is satisfied with $h^m = n^{-s}$, where m is the order of the kernel K. The last condition in (A.1) only requires that the nonparametric bias $\bar{\gamma}_n - \gamma_0$ is asymptotically negligible.

To briefly sum up, in the kernel density case, the conditions in (A.1) essentially requires $\hat{\gamma}_n$ to be a consistent nonparametric estimator of γ_0.

For the (leave-one-out) average density estimator $\hat{\theta}_n^{\texttt{AD}} = \frac{1}{n}\sum_{i=1}^n \hat{\gamma}_n(z_i)$. We have $\mathcal{B}_n^{\texttt{NL}} = 0$ and

$$\mathcal{B}_n^{\texttt{ANB}} = \frac{1}{n}\sum_{i=1}^{n}[\bar{\gamma}_n(z_i) - \gamma_0(z_i)] = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} K(u)[\gamma_0(z_i - hu) - \gamma_0(z_i)]du = O_{\mathbb{P}}(h^m),$$
$$\mathcal{B}^{\texttt{ANB}} = \int_{\mathbb{R}}\int_{\mathbb{R}} K(u)[\gamma_0(x - hu) - \gamma_0(x)]\gamma_0(x)dudx = O(h^m).$$

For the integrated squared density estimator $\hat{\theta}_n^{\mathtt{ISD}} = \int \hat{\gamma}_n^2(x)dx$, we have

$$\mathcal{B}_n^{\mathtt{NL}} = \int_{\mathbb{R}} [\hat{\gamma}_n(x) - \bar{\gamma}_n(x)]^2 dx = O\big(\frac{1}{nh^{d_z}}\big),$$
$$\mathcal{B}_n^{\mathtt{ANB}} = 2\int_{\mathbb{R}} \gamma_0(x)[\bar{\gamma}_n(x) - \gamma_0(x)]dx = O(h^m).$$

and

$$\mathcal{B}^{\mathtt{NL}} = \frac{1}{nh^{d_z}} \int_{\mathbb{R}} \Big(\gamma_0(x) \int_{\mathbb{R}} K(u)^2 du\Big)dx = O\Big(\frac{1}{nh^{d_z}}\Big),$$
$$\mathcal{B}^{\mathtt{ANB}} = \int_{\mathbb{R}} \int_{\mathbb{R}} K(u)[\gamma_0(x - hu) - \gamma_0(x)]\gamma_0(x)dudx = O(h^m).$$

Refer to the appendix for an example with the Nadaraya–Watson estimator.

As mentioned in the previous subsection, there is no "singularity bias" (even with the kernel-based method) when we use the smooth measure $\mathbb{P}_n^{\mathtt{AC}}$ (recall Remark 2.2) in the construction of $\widehat{G}_n$ (this gives the integrated square density estimator in the above example). However, it may bring an additional nonlinear bias, when the alternative estimator is linear in $\hat{\gamma}_n$. Besides, we note that the nonlinear bias (when it exists) and the "singularity bias" are of the same order. Hence, they can be corrected simultaneously by using the multi-scale jackknife method (see Section 3.1).

To make both biases shrink faster than the root-n rate, we need both $r > 1/4$ and $s > 1/2$, which are consistent with the prevalent requirement of a faster-than-$n^{1/4}$ convergence rate for the nonparametric estimator. Some complications may arise if we have more than one source of bias in $\bar{\gamma}_n - \gamma_0$, like in the average density example. Once these conditions are satisfied, one can use some well-established empirical process results, such as the stochastic equicontinuity condition (Andrews, 1994; Newey, 1994). However, if $r \leqslant 1/4$ or $s \leqslant 1/2$, then either $\mathcal{B}_n^{\mathtt{NL}}$ or $\mathcal{B}_n^{\mathtt{ANB}}$ will not be $o_{\mathbb{P}}(n^{-1/2})$. In such cases, such bias(es) will have some non-trivial impact(s) on the asymptotic behavior of $\hat{\theta}_n$.

Example (Kernel Density Estimator Continued). In view of the above discussion, no matter we use the original kernel density estimator or its "leave-one-out" version, the necessary and sufficient condition for both $\mathcal{B}_n^{\mathtt{NL}}$ and $\mathcal{B}_n^{\mathtt{ANB}}$ to be root-n negligible is $1/(2m) < \kappa < 1/(2d_z)$, which requires $d_z < m$, i.e., the dimension of the random vector should be smaller than the order of the kernel. If this condition fails, then at least one of the two biases will not be asymptotically negligible at the root-n rate. To some extent, this observation also reflects the curse of dimensionality: if $d_z \geqslant m$, then there is no way to make both biases root-n negligible. In fact, when $d_z > m$, if the bandwidth satisfies $1/(2d_z) < \kappa < 1/(2m)$, then neither $\mathcal{B}_n^{\mathtt{NL}}$ nor $\mathcal{B}_n^{\mathtt{ANB}}$ is root-n negligible. Motivated by this possibility, we are going to keep both biases in our analysis. This observation also indicates that our bias correction methods may help ameliorate the curse of dimensionality.

The following lemma gives the sufficient conditions for the remaining terms in (2.3), as well as the impact of $\mathcal{J}_n - \mathcal{J}_0$ on $\hat{\theta}_n$, to be root-n negligible,

Lemma 2.13. *Suppose that Assumptions 2.4 (about g) and 2.12 both hold true. Additionally, assume that $\mathcal{J}_n - \mathcal{J}_0 = O_{\mathbb{P}}\big(\widehat{G}_n(\theta_0, \hat{\gamma}_n)\big)$.*

We have the following conclusions: (i) if $s + 2r > 1/2$ and $r > 1/8$, then $(\mathcal{J}_n - \mathcal{J}_0)\mathcal{B}_n^{NL} = o_{\mathbb{P}}(n^{-1/2})$; (ii) if $s + 2r > 1/2$ and $s > 1/4$, then $(\mathcal{J}_n - \mathcal{J}_0)\mathcal{B}_n^{ANB} = o_{\mathbb{P}}(n^{-1/2})$; (iii) if $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\hat{\gamma}_n(z_i) - \gamma_0(z_i)\|^3] \leqslant Cn^{-3(r\wedge s)}$ for some finite number C, $s > 1/4$ and $r > 1/6$, then

$$\widehat{G}_n(\theta_0, \hat{\gamma}_n) - \widehat{G}_n(\theta_0, \gamma_0) - \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n - \bar{\gamma}_n) - \mathcal{B}^{NL} - \mathcal{B}^{ANB} = o_{\mathbb{P}}(n^{-1/2}).$$

The assumption $\mathcal{J}_n - \mathcal{J}_0 = O_{\mathbb{P}}\big(\widehat{G}_n(\theta_0, \hat{\gamma}_n)\big)$ is to accommodate the possibility that $\mathcal{J}_n - \mathcal{J}_0$ may depend on or be related to $\widehat{G}_n(\theta_0, \hat{\gamma}_n)$, which complicates the proof a bit. In general, the above lemma will also hold if one assumes $\mathcal{J}_n - \mathcal{J}_0 = O_{\mathbb{P}}(n^{-\iota})$, and then let $\iota + 2r > 1/2$ in part (i), and $\iota + s > 1/2$ in part (ii). The same conclusions can be verified rather straightforwardly. In such case, the parameter ι is essentially equivalent to $1/\rho$ in Lemma 1 of (Cattaneo and Jansson, 2018).

As discussed above, most previous papers on semiparametric estimators require both $\mathcal{B}_n^{\mathtt{NL}}$ and $\mathcal{B}_n^{\mathtt{ANB}}$ to be root-n negligible. Although recent works relax this requirement, they often require one of $\mathcal{B}_n^{\mathtt{NL}}$ and $\mathcal{B}_n^{\mathtt{ANB}}$ is root-n negligible. For instance, Theorem 2 of (Cattaneo and Jansson, 2018) effectively require the bias $\mathcal{B}_n^{\mathtt{ANB}}$ to be root-n negligible (small bandwidth asymptotics), while (Chernozhukov et al., 2018b) implicitly assume the nonlinear bias $\mathcal{B}_n^{\mathtt{NL}}$ is root-n negligible (large bandwidth asymptotics).

However, it is often not easy to check whether such restrictions hold or not in practice. Moreover, recall the previous example of the kernel density estimator. It is possible that both biases are non-root-n-negligible. In view of these results, we keep both $\mathcal{B}_n^{\mathtt{NL}}$ and $\mathcal{B}_n^{\mathtt{ANB}}$ in our analysis. In a different setup with the non-stationary underlying process and in-fill asymptotics, (Yang, 2020) adopts a similar approach. The following theorem gives the first main result of this paper.

Theorem 2.14 (Asymptotic Normality for $\hat{\theta}_n$)**.** *Suppose that Assumptions 2.1 to 2.7 hold true. Assume that $\mathcal{J}_n - \mathcal{J}_0 = O_{\mathbb{P}}\big(\widehat{G}_n(\theta_0, \hat{\gamma}_n)\big)$. If $s > 1/4$ and $r > 1/6$, then we have*

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0 - \mathcal{J}_n\mathcal{B}^{NL} - \mathcal{J}_n\mathcal{B}^{ANB}\big) \xrightarrow{\mathscr{L}} \mathcal{N}\big(0, \Sigma_\theta\big),$$

where $\Sigma_\theta = \mathcal{J}_0\,\Sigma_g\,\mathcal{J}_0^\intercal$ with Σ_g given in Assumption 2.7.

The conditions $s > 1/4$ and $r > 1/6$ only require a faster-than-$n^{1/6}$ convergence rate for the nonparametric estimator $\hat{\gamma}_n$, consistent with the conclusion of (Cattaneo and Jansson, 2018) in the kernel-based case. This is a weaker condition than the typical requirement of a faster-than-$n^{1/4}$ convergence rate (see those cited papers at the beginning of this subsection).

Besides, we also note that the above central limit theorem (CLT) is infeasible, for that the two biases are evaluated at (θ_0, γ_0), both of which are unknown. In the next section, we are going to discuss how to correct for these biases and conduct robust inference.

Remark 2.15. *It might happen that the bias $\mathcal{B}_n^{ANB}$ is identically zero. For example, in the continuous-time setting (with in-fill asymptotics), (Yang, 2020) has shown that, when*

estimating integrated volatility functionals, the counterpart of $\mathcal{B}_n^{ANB}$, which is the first-order effect of the nonparametric bias, is canceled by the discretization error. In the cited paper, what left is the counterpart of the following second-order effect of the nonparametric bias:

$$\frac{1}{2}\widehat{G}''_{n,\gamma\gamma}(\theta_0, \gamma_0, \bar{\gamma}_n - \gamma_0, \bar{\gamma}_n - \gamma_0) = \sum_{l=1}^{L} O_{\mathbb{P}}(n^{-2s_l}).$$

In such case, then one can replace the first-order effect by the above second-order one and replace s by $2s$ in Lemma 2.13 and Theorem 2.14.

3 Bias-Robust Inference

(Cattaneo and Jansson, 2018) propose a bootstrap-based inference procedure that is robust to the nonlinear bias. We believe that if the bootstrap version of all the above assumptions hold, then the corresponding inference should also be robust to the average nonparametric bias. Since it has been proposed in the literature, we will not discuss it here.

In this section, we are going to discuss two alternative methods to conduct inference that is robust to the possibly non-root-n-negligible bias(es). At the end of this section, we will also discuss an extension of our framework to the case where $\hat{\theta}_n$ is constructed as the sample average of some discontinuous functionals of $\hat{\gamma}_n$.

For simplicity, we illustrate the ideas using kernel-based estimators. The linear sieve case would be characterized in a similar manner. Yet, the nonlinear sieve case may require extra non-trivial efforts.

3.1 Multi-scale jackknife

The original jackknife estimator, first introduced by (Quenouille, 1949), is essentially a linear combination of estimators computed from samples with different sizes, for that the biases in many estimators depend on the sample size. While in the current context, the biases depend on the tuning parameter. Thus, it is natural to utilize the tuning parameter in the role of the sample size (see, e.g., (Schucany and Sommers, 1977), (Bierens, 1987), and (Powell et al., 1989) among others). However, there is only one bias in these papers. In the context of in-fill asymptotics, (Li et al., 2019) has developed a multi-scale jackknife (MSJ) estimator to correct for various biases for integrated volatility functionals.

In this subsection, we are going to show that MSJ can remove various biases in the current context, provided that we have some knowledge about the structure of the nonparametric estimator, i.e., knowing how the rates in Assumption 2.12 depend on the tuning parameter.

In the kernel-based case, the semiparametric estimator $\hat{\theta}_n$ depends on the bandwidth h. Let Q be a finite positive integer. Then consider a sequence of estimators $\{\hat{\theta}_n(h_q)\}_{q=1}^{Q}$ and a sequence of real numbers $\{w_q\}_{q=1}^{Q}$. For example, define the following three-scale jackknife (3SJ) estimator:

$$\hat{\theta}_n^w = \sum_{q=1}^{3} w_q \hat{\theta}_n(h_q),$$

where

$$\sum_{q=1}^{3} w_q = 1, \quad \sum_{q=1}^{3} w_q h_q^m = o(n^{-1/2}), \quad \sum_{q=1}^{3} \frac{w_q}{n h_q^{d_z}} = o(n^{-1/2}). \tag{3.1}$$

In practice, for example, we can choose $h_q = \eta_q h$, where $\{\eta_q\}_{q=1}^{Q}$ is a sequence of positive numbers. In the above three-scale case, the weights $\{w_q\}_{q=1}^{3}$ are solved as

$$\begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ \eta_1^m & \eta_2^m & \eta_3^m \\ \eta_1^{-d_z} & \eta_2^{-d_z} & \eta_3^{-d_z} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Moreover, one can choose a larger Q to remove/reduce more biases. For instance, in the kernel case, the smoothing bias may also have components that are $O_{\mathbb{P}}(h^{m+1})$, $O_{\mathbb{P}}(h^{m+2})$, or of even higher orders (for symmetric kernels, the odd-order terms will be zero).

We consider the general case where we have the smoothing bias $\mathcal{B}_{n,1}^{\mathtt{ANB}}$, the "singularity bias" $\mathcal{B}_{n,2}^{\mathtt{ANB}}$ and the nonlinear bias $\mathcal{B}_n^{\mathtt{NL}}$. The reason is that the "singularity bias" may be unavoidable when estimating the asymptotic variance using the bootstrap method. Recall that $\mathcal{B}_{n,2}^{\mathtt{ANB}}$ and $\mathcal{B}_n^{\mathtt{NL}}$ are of the same order when both exist. The key is to show that, under condition (3.1), the following three terms

$$\widetilde{\mathcal{B}}_{n,1}^{\mathtt{ANB}} = \sum_{q=1}^{Q} w_q \, \mathcal{B}_{n,1}^{\mathtt{ANB}}(h_q), \quad \widetilde{\mathcal{B}}_{n,2}^{\mathtt{ANB}} = \sum_{q=1}^{Q} w_q \, \mathcal{B}_{n,2}^{\mathtt{ANB}}(h_q), \quad \widetilde{\mathcal{B}}_{n}^{\mathtt{NL}} = \sum_{q=1}^{Q} w_q \, \mathcal{B}_{n}^{\mathtt{NL}}(h_q).$$

are all root-n negligible. Then the following CLT readily follows.

Theorem 3.1 (Multi-scale jackknife)**.** *Suppose that all assumptions of Theorem 2.14 hold true and that $\hat{\gamma}_n(h_q)$ is a kernel-based nonparametric estimator depending on the bandwidth h_q, where $q = 1, \cdots, Q$ for some finite Q. In addition, assume $h_q \to 0$, $n^2 h_q^{3d_z} \to \infty$, $n h_q^{4m} \to 0$, and that the general version of condition (3.1) is satisfied. Then we have*

$$\sqrt{n}\big(\hat{\theta}_n^w - \theta_0\big) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \Sigma_\theta^w).$$

The asymptotic variance is given by $\Sigma_\theta^w := \mathcal{J}_0 \, \Sigma_g^w \, \mathcal{J}_0^\intercal$ and Σ_g^w is the asymptotic variance of the following (exact or approximate) U-statistic

$$\widehat{G}(\theta_0, \gamma_0) + \widehat{G}'_{n,\gamma}\big(\theta_0, \gamma_0, \hat{\gamma}_n^w - \bar{\gamma}_n^w\big),$$

where $\hat{\gamma}_n^w = \sum_{q=1}^{Q} w_q \hat{\gamma}_n(h_q)$ and $\bar{\gamma}_n^w = \sum_{q=1}^{Q} w_q \bar{\gamma}(h_q)$.

Suppose that the following column vector

$$\sqrt{n}\Big(\widehat{G}(\theta_0, \gamma_0) + \widehat{G}_n\big(\theta_0, \gamma_0, \hat{\gamma}_n(h_q) - \bar{\gamma}_n(h_q)\big)\Big)_{q=1,\cdots,Q}^\intercal$$

converges in distribution to $\mathcal{N}(0, \Sigma_g^Q)$, then we have $\Sigma_\theta^w = \mathcal{J}_0 \, w \, \Sigma_g^Q w^\intercal \mathcal{J}_0^\intercal$.

For illustration purpose, consider the case where $h_q \propto n^{-\kappa}$ for all $q = 1, \cdots, Q$. Then we have $r = (1 - \kappa d_z)/2$, $s_1 = \kappa m$ and $s_2 = 2r$ (if there is "singularity bias") for the kernel-based estimators. The requirements $r > 1/6$ and $s > 1/4$ in Theorem 2.14 are equivalent to $n^2 h_q^{3d_z} \to \infty$ and $n h_q^{4m} \to 0$ (the conditions in the above theorem). To put it differently, we need $\kappa \in (1/(4m), 2/(3d_z))$. This set is non-empty if and only if $3d_z < 8m$, which is weaker than $d_z < m$ (recall the previous discussion on the curse of dimensionality). As a comparison, we note that $r > 1/4 \Leftrightarrow \kappa < 1/(2d_z) \Leftrightarrow n h_q^{2d_z} \to \infty$ and $s_1 > 1/2 \Leftrightarrow \kappa > 1/(2m) \Leftrightarrow n h_q^{2m}$.

Intuitively, the statistics $\{\widehat{G}(\theta_0, \gamma_0) + \widehat{G}_n\big(\theta_0, \gamma_0, \hat{\gamma}_n(h_q) - \bar{\gamma}_n(h_q)\big)\}_{q=1}^Q$ are constructed from the same sample, hence are "highly" correlated. It would be reasonable to expect that, in some cases, their correlations are approximately one. If so, then the matrix Σ_g^Q becomes $\Sigma_g \mathbf{1}_Q$ (assuming Σ_g is a scalar for illustration purpose), where $\mathbf{1}_Q$ is a Q-by-Q matrix with all the elements being one. Then the asymptotic variance $\Sigma_\theta^w = \mathcal{J}_0 \Sigma_g w \mathbf{1}_Q w^\intercal \mathcal{J}_0^\intercal = \Sigma_\theta$ (note that $w \mathbf{1}_Q w^\intercal = (\sum_{q=1}^Q w_q)^2 = 1$). That is to say, when these estimators are approximately perfectly correlated, there is no efficiency loss by using the MSJ estimator.

In some cases, it may not be very easy to find the analytical form of the functional $g'_\gamma(\theta_0, \gamma_0, \cdot)$ or its variance. Hence, it may not always be possible to estimate Σ_g^w directly. In such cases, one can use the following algorithm to estimate the asymptotic variance Σ_θ^w.

Algorithm 1 (Bootstrap variance estimator)**.** *The procedure consists of the following steps: (1) Draw a bootstrap sample $\{z_i^*\}_{i=1}^n$ and calculate $\hat{\theta}_n^{w*}$. (2) Repeat Step (1) a large number of times, say P, and get $\{\hat{\theta}_{n,p}^{w*}\}_{p=1}^P$. (3) Compute Σ_θ^{w*} as the sample variance-covariance of $\{\hat{\theta}_{n,p}^{w*}\}_{p=1}^P$.*

Theorem 3.2 (Bootstrap variance)**.** *Suppose that the assumptions of Theorem 3.1 hold true. In addition, assume that $g^* \equiv g$, $g_\gamma^{*\prime} \equiv g'_\gamma$, and both $g(\theta, \gamma)$ and $g'_\gamma(\theta, \gamma, \cdot)$ are Lipschitz continuous with respect to θ and γ in a neighborhood of (θ_0, γ_0). Then $\Sigma_\theta^{w*} \xrightarrow{\mathbb{P}} \Sigma_\theta^w$.*

Since the "singularity bias" can always be removed together with the nonlinear bias, the bootstrap estimator $\hat{\theta}_n^{w*}$ will have no such bias, even if the re-sampled data may include several replicates of the same observation.

If certain bias(es) is/are root-n negligible, then some of the requirements in Condition (3.1) will not be binding, which can then be simplified. For instance, if the smoothing bias is root-n negligible, i.e., $h_q^m = o(n^{-1/2})$ for $q = 1, 2$, then we only need

$$\sum_{q=1}^2 w_q = 1 \quad \text{and} \quad \sum_{q=1}^2 \frac{w_q}{n h_q^{d_z}} = o(n^{-1/2}).$$

On the other hand, if the nonlinear bias and the "singularity bias" are root-n negligible, i.e., $h_q^{-d_z} = o(n^{1/2})$ for $q = 1, 2$, then we only need

$$\sum_{q=1}^2 w_q = 1 \quad \text{and} \quad \sum_{q=1}^2 w_q h_q^m = o(n^{-1/2}).$$

In these two cases, the two-scale jackknife (2SJ) estimators are asymptotically normal with a root-n rate.

3.2 Analytical bias correction

The analytical bias correction method requires more assumptions on the semiparametric model. The idea is to introduce some sufficient conditions so that we can construct consistent estimators of the average nonparametric bias $\mathcal{B}^{\texttt{ANB}}$ and the nonlinear bias $\mathcal{B}^{\texttt{NL}}$.

Suppose that the functional g is twice Fréchet differentiable with respect to γ around γ_0. Consider the general case where γ is a matrix-valued function, with the row and column numbers being r_γ and c_γ, respectively. Define the following matrix representation of the partial derivative (Kollo and von Rosen, 2006):

$$\Big(\frac{\partial}{\partial \mathrm{vec}(\gamma)}\Big)^\intercal = \frac{\partial}{\partial [\mathrm{vec}(\gamma)]^\intercal} = \Big(\frac{\partial}{\partial \gamma_{11}}, \cdots, \frac{\partial}{\partial \gamma_{r_\gamma 1}}, \cdots, \frac{\partial}{\partial \gamma_{1c_\gamma}}, \cdots, \frac{\partial}{\partial \gamma_{r_\gamma c_\gamma}}\Big).$$

Let $\mathbb{D}_\gamma g = \frac{\partial g}{\partial [\mathrm{vec}(\gamma)]^\intercal}$ and $\mathbb{D}^2_{\gamma\gamma} g = \frac{\partial}{\partial \mathrm{vec}(\gamma)} \otimes \frac{\partial g}{\partial [\mathrm{vec}(\gamma)]^\intercal}$. Assume that

$$g'_\gamma(z, \theta_0, \gamma_0, \gamma - \gamma_0) = \mathbb{D}_\gamma g(z, \theta_0, \gamma_0)\, \mathrm{vec}\big(\gamma(z) - \gamma_0(z)\big),$$
$$g''_{\gamma\gamma}(z, \theta_0, \gamma_0, \gamma - \gamma_0) = \big[\mathrm{vec}\big(\gamma(z) - \gamma_0(z)\big)^{\otimes 2} \otimes I_{d_g}\big]^\intercal \mathrm{vec}\big(\mathbb{D}^2_{\gamma\gamma} g(z, \theta_0, \gamma_0)\big).$$

Under these assumptions, the two biases can be written as

$$\mathcal{B}^{\texttt{ANB}}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{D}_\gamma g(z_i, \theta_0, \gamma_0)\, \mathrm{vec}\big(\bar\gamma_n(z_i) - \gamma_0(z_i)\big),$$
$$\mathcal{B}^{\texttt{NL}}_n = \frac{1}{n}\sum_{i=1}^n \big[\mathrm{vec}\big(\hat\gamma_n(z_i) - \bar\gamma_n(z_i)\big)^{\otimes 2} \otimes I_{d_g}\big]^\intercal \mathrm{vec}\big(\mathbb{D}^2_{\gamma\gamma} g(z_i, \theta_0, \gamma_0)\big).$$

Suppose that $n^r \mathrm{vec}\big(\hat\gamma_n(x) - \bar\gamma_n(x)\big) \xrightarrow{\mathscr{L}} \mathcal{N}(0, V(x))$ for any $x \in \mathbb{R}^{d_z}$. Then, when Assumption 2.12 holds true, we would expect that $\mathcal{B}^{\texttt{NL}}_n - \mathcal{B}^{\texttt{NL}} = o_{\mathbb{P}}(n^{-1/2})$ with the following $\mathcal{B}^{\texttt{NL}}$:

$$\mathcal{B}^{\texttt{NL}} := \mathbb{E}\Big(\big[\mathrm{vec}\big(V(z)\big) \otimes I_{d_g}\big]^\intercal \mathrm{vec}\big(\mathbb{D}^2_{\gamma\gamma} g(z, \theta_0, \gamma_0)\big)\Big).$$

Suppose that we have a consistent estimator $\widehat{V}_n(\cdot)$ of the asymptotic variance $V(\cdot)$. It then follows that we can estimate $\mathcal{B}^{\texttt{NL}}$ by

$$\widehat{\mathcal{B}}^{\texttt{NL}}_n = \frac{1}{n^{1+2r}}\sum_{i=1}^n \big[\mathrm{vec}\big(\widehat{V}_n(z_i)\big) \otimes I_{d_g}\big]^\intercal \mathrm{vec}\big(\mathbb{D}^2_{\gamma\gamma} g(z_i, \hat\theta_n, \hat\gamma_n)\big). \tag{3.2}$$

On the other hand, suppose that there exists a (point-wise) consistent estimator $\hat{\bar\gamma}_n$ of $\bar\gamma_n$. Then we can estimate $\mathcal{B}^{\texttt{ANB}}$ by

$$\widehat{\mathcal{B}}^{\texttt{ANB}}_n = \widehat{G}'_{n,\gamma}(\hat\theta_n, \hat\gamma_n, \hat{\bar\gamma}_n - \hat\gamma_n) = \frac{1}{n}\sum_{i=1}^n \mathbb{D}_\gamma g(z_i, \hat\theta_n, \hat\gamma_n)\, \mathrm{vec}\big(\hat{\bar\gamma}_n(z_i) - \hat\gamma_n(z_i)\big). \tag{3.3}$$

For simplicity, we assume that there is no "singularity bias" in $\mathcal{B}^{\texttt{ANB}}_n$, since it can be easily removed using the methods discussed in Section 2.3.

Assumption 3.3. *Suppose that Assumption 2.12 holds with real numbers r and s. Assume that the functional g is twice Fréchet differentiable with respect to γ around γ_0, with $\mathbb{E}\big(\|\mathbb{D}^2_{\gamma\gamma} g(z,\theta_0,\gamma_0)\|^2\big) < \infty$ and*

$$\mathbb{E}\big(\|\mathbb{D}_\gamma g(z,\theta_0,\gamma_0) - \mathbb{D}_\gamma g(z,\hat{\theta}_n,\hat{\gamma}_n)\|^2\big) = O(n^{-2(r\wedge s)}),$$

for sufficiently large n.

Moreover, there exist $\widehat{V}_n$ and $\hat{\bar{\gamma}}_n$ such that $\hat{\bar{\gamma}}_n - \bar{\gamma}_n \xrightarrow{\mathbb{P}} 0$, $\mathbb{E}\big(\|\hat{\bar{\gamma}}_n(z) - \hat{\gamma}_n(z)\|^2\big) = o(n^{-2t})$, and

$$\mathbb{E}\big(\|n^{2r}\, vec\big(\hat{\gamma}_n(z) - \bar{\gamma}_n(z)\big)^{\otimes 2} - vec\big(\widehat{V}_n(z)\big)\|^2\big) = o(n^{-2v}),$$

where t and v are some positive real numbers.

Assumption 3.3 is a strengthened version of the combination of Assumptions 2.4 and 2.12. The twice Fréchet differentiable condition implies the quadratic approximation in Assumption 2.4, with a more detailed structure on the first- and second-order derivatives. In addition, Assumption 3.3 also imposes certain conditions on the estimators of V and $\bar{\gamma}_n$ in Assumption 2.12.

Theorem 3.4 (Analytical bias correction)**.** *Suppose that Assumptions 2.1 and 3.3 hold true. Define $\bar{\bar{\gamma}}_n(z_i) := \mathbb{E}[\hat{\bar{\gamma}}_n(z_i)|z_i]$. Assume that $s > 1/4$, $r > 1/6$, $t + r \wedge s > 1/2$, $v + 2r > 1/2$, and*

$$\sqrt{n}\Big(\widehat{G}_n(\theta_0,\gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0,\gamma_0, 2\hat{\gamma}_n - \hat{\bar{\gamma}}_n - 2\bar{\gamma}_n + \bar{\bar{\gamma}}_n)\Big) \xrightarrow{\mathscr{L}} \mathcal{N}(0,\tilde{\Sigma}_g), \tag{3.4}$$

$$\widehat{G}'_{n,\gamma}(\theta_0,\gamma_0, 2\bar{\gamma}_n - \bar{\bar{\gamma}}_n - \gamma_0) = o_{\mathbb{P}}(n^{-1/2}). \tag{3.5}$$

Then we have

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0 - \mathcal{J}_n \widehat{\mathcal{B}}^{NL}_n - \mathcal{J}_n \widehat{\mathcal{B}}^{ANB}_n\big) \xrightarrow{\mathscr{L}} \mathcal{N}\big(0, \mathcal{J}_0\, \tilde{\Sigma}_g\, \mathcal{J}_0^\intercal\big).$$

where $\widehat{\mathcal{B}}^{NL}_n$ and $\widehat{\mathcal{B}}^{ANB}_n$ are given by and (3.3), respectively.

A possible choice for $\hat{\bar{\gamma}}_n$ is $\hat{\gamma}_n$, which then yields $\bar{\bar{\gamma}}_n \equiv \bar{\gamma}_n$. In this case, condition 3.4 reduces to Assumption 2.7. Condition 3.5 is then equivalent to $\mathcal{B}^{\texttt{ANB}}_n = o_{\mathbb{P}}(n^{-1/2})$. That is to say, when we couldn't estimate $\mathcal{B}^{\texttt{ANB}}_n$, we can obtain an analytical-based inference only if $\mathcal{B}^{\texttt{ANB}}_n$ is root-n negligible.

In some cases, it is possible to have an estimator $\hat{\bar{\gamma}}_n$ different from $\hat{\gamma}_n$. Then Condition 3.5 requires that this estimator can reduce the average nonparametric bias to the extent that the remaining bias becomes root-n negligible. Conditions 3.5 and 3.4 together imply that

$$\widehat{G}_n(\theta_0,\gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0,\gamma_0, 2\hat{\gamma}_n - \hat{\bar{\gamma}}_n - \gamma_0) \xrightarrow{\mathscr{L}} \mathcal{N}(0,\tilde{\Sigma}_g).$$

That is, the asymptotic variance is determined by the updated estimator $2\hat{\gamma}_n - \hat{\bar{\gamma}}_n$. We expect that, in most cases, the left hand side can be written as a U-statistic. Then the above asymptotic normality result shall be satisfied under very general conditions.

Example (Kernel density estimator continued). Let $\hat{\gamma}_n$ be the "leave-one-out" kernel density estimator. In this case, $V(x) = \gamma_0(x)\int K^2(u)du$, which can be easily estimated. Recall that $\bar{\gamma}_n(\cdot) = \int K(u)\gamma_0(\cdot - hu)du$. It then follows that

$$\hat{\bar{\gamma}}_n(\cdot) = \int K(u)\hat{\gamma}_n(\cdot - hu)du,\ \bar{\bar{\gamma}}_n(\cdot) = \int\int K(u)K(v)\gamma_0(\cdot - hu - hv)dudv.$$

The updated estimator becomes

$$2\hat{\gamma}_n(z_i) - \hat{\bar{\gamma}}_n(z_i) = \frac{1}{n-1}\sum_{j\neq i}\left(2K_h(z_i - z_j) - \int K_h(z_i - x)K_h(x - z_j)dx\right)$$
$$= \frac{1}{n-1}\sum_{j\neq i}\left(2K_h(z_i - z_j) - \int K_h(z_i - z_j - y)K_h(y)dy\right) = \frac{1}{n-1}\sum_{j\neq i}\tilde{K}_h(z_i - z_j),$$

where $\tilde{K}_h(u) = \frac{1}{h^{d_z}}\tilde{K}(u/h)$ and $\tilde{K}(u) = 2K(u) - \int K(u-v)K(v)dv$ is the twicing kernel studied by (Stuetzle and Mittal, 1979) and (Newey et al., 2004).

According to (Newey et al., 2004), the twicing kernel enjoys a small bias property, which makes Condition (3.5) less stringent than requiring that $\mathcal{B}_n^{\texttt{ANB}}$ is root-n negligible. For instance, if γ_0 is at least $2m$ times differentiable and the order of K is m, then $\widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, 2\bar{\gamma}_n - \bar{\bar{\gamma}}_n - \gamma_0) = O_{\mathbb{P}}(h^{2m}) = O_{\mathbb{P}}(n^{-2\kappa m})$. Hence, Condition (3.5) only requires $\kappa > 1/(4m)$ (cf. $\kappa > 1/(2m)$ for $\mathcal{B}_n^{\texttt{ANB}}$ to be root-n negligible). If Condition (2.4) in (Newey et al., 2004) is satisfied with some function ν, then the requirement that γ_0 is at least $2m$ times differentiable can be replaced by both ν and γ_0 are at least m times differentiable.

The limitation of the analytical bias correction method is that it requires explicit expressions of $\mathbb{D}_\gamma g$, which is the influence function (refer to (Ichimura and Newey, 2017) for more discussions on the calculation of the influence function), and $\mathbb{D}^2_{\gamma\gamma}g$. In some cases, it can be very challenging to compute these derivatives. However, when they are available in analytical forms, the computation cost is lower than the multi-scale jackknife method, for that one only needs to conduct the estimation with one bandwidth.

3.3 Extension to discontinuous functionals

In many applications, the semiparametric estimator is a sample average of some discontinuous functional of the first-step nonparametric estimator. In this subsection, we are going to demonstrate that our framework can be extended to such case if there exists a sufficiently smooth projection of the discontinuous functional.

Assumption 3.5 (ALQP—Asymptotic Linearity in $\breve{g}$ with a Quadratic Projection)**.** *Assume that the semiparametric estimator $\breve{\theta}_n$ is asymptotically linear in a discontinuous functional $\breve{g}$:*

$$\breve{\theta}_n - \theta_0 = \mathcal{J}_n\breve{G}_n(\theta_0, \hat{\gamma}_n) + o_{\mathbb{P}}(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^n \mathcal{J}_n\,\breve{g}(z_i, \theta_0, \hat{\gamma}_n) + o_{\mathbb{P}}(n^{-1/2}),$$

where $\mathcal{J}_n \xrightarrow{\mathbb{P}} \mathcal{J}_0$ for some non-random and non-zero $\mathcal{J}_0$, and the functional $\breve{g}$ satisfies that $\breve{G}(\theta_0, \gamma_0) = \mathbb{E}[\breve{g}(z, \theta_0, \gamma_0)] = 0$.

Moreover, there exists a continuous functional g satisfying Assumption 2.4 and $\mathbb{E}[\check{g}(z_i,\theta,\gamma)] = \mathbb{E}[g(z_i,\theta,\gamma)]$, $\forall i = 1,\cdots,n$, in an open set containing (θ_0,γ_0).

Intuitively, the functional g is a smooth projection of $\check{g}$ on some sub-σ-algebra of the σ-algebra generated by the sample. Let $\hat{\theta}_n$ be the corresponding estimator defined by g. Under Assumption 3.5 and those conditions of Lemma 2.13, we obtain

$$\begin{aligned}&\check{\theta}_n - \theta_0 = (\check{\theta}_n - \hat{\theta}_n) + (\hat{\theta}_n - \theta_0)\\ &= \mathcal{J}_n\Big(\check{G}_n(\theta_0,\hat{\gamma}_n) - \widehat{G}_n(\theta_0,\hat{\gamma}_n) + \widehat{G}_n(\theta_0,\gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0,\gamma_0,\hat{\gamma}_n - \bar{\gamma}_n) + \mathcal{B}_n^{\texttt{ANB}} + \mathcal{B}_n^{\texttt{NL}}\Big) + o_{\mathbb{P}}(n^{-1/2}).\end{aligned}$$

The property of g implies that $\mathbb{E}[\check{G}_n(\theta_0,\hat{\gamma}_n) - \widehat{G}_n(\theta_0,\hat{\gamma}_n)] = 0$. That is, the difference $\check{G}_n(\theta_0,\hat{\gamma}_n) - \widehat{G}_n(\theta_0,\hat{\gamma}_n)$ does not contain any biases. Intuitively, it is the sample average of the difference between $\check{g}$ and its smooth projection g. Hence, it is reasonable to expect that this difference is asymptotically normal, under certain regularity conditions.

Assumption 3.6 (AN′—Asymptotic Normality)**.** *Suppose that there exists a non-random and positive definite $\check{\Sigma}_g$ such that*

$$\check{G}_n(\theta_0,\hat{\gamma}_n) - \widehat{G}_n(\theta_0,\hat{\gamma}_n) + \widehat{G}_n(\theta_0,\gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0,\gamma_0,\hat{\gamma}_n - \bar{\gamma}_n) \xrightarrow{\mathscr{L}} \mathcal{N}(0,\check{\Sigma}_g).$$

Example (Hit Rates). Consider the hit rates example discussed by (Chen et al., 2003). Let $z = (y, x^\intercal)^\intercal$, where y is a scalar dependent variable and $x \in \mathbb{R}^{d_x}$ is a continuous covariate with density γ_0. The parameter of interest is $\theta_0 = \mathbb{E}[\mathbb{1}(y \geqslant \gamma_0(x))] = \mathbb{E}\big[1 - F_{y|x}\big(\gamma_0(x)|x\big)\big]$, where $F_{y|x}$ is the conditional distribution of y given x. Consider a kernel-based semiparametric estimator

$$\check{\theta}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\big(y_i \geqslant \hat{\gamma}_n(x_i)\big), \quad \hat{\gamma}(x_i) = \frac{1}{n}\sum_{j\neq i} K_h(z_i - z_j).$$

Let $\check{g}(z,\theta,\gamma) = \mathbb{1}\big(y \geqslant \gamma(x)\big) - \theta$ and $g(z,\theta,\gamma) = \mathbb{E}[\check{g}(z,\theta,\gamma)|x] = 1 - F_{y|x}\big(\gamma(x)|x\big) - \theta$. Let $\mathcal{X}_n$ be the σ-algebra generated by $\{x_i\}_{i=1}^n$. Then we have

$$\check{G}_n(\theta_0,\hat{\gamma}_n) - \widehat{G}_n(\theta_0,\hat{\gamma}_n) = \frac{1}{n}\sum_{i=1}^n \Big(\mathbb{1}\big(y_i \geqslant \hat{\gamma}_n(x)\big) - 1 + F_{y|x}\big(\hat{\gamma}_n(x_i)|\mathcal{X}_n\big)\Big).$$

The asymptotic normality of the above difference is a direct result of the central limit theory in the i.i.d. case. If we further know the correlation between this difference and $\widehat{G}_n(\theta_0,\gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0,\gamma_0,\hat{\gamma}_n - \bar{\gamma}_n)$, as well as the variance of the latter, we will be able to find $\check{\Sigma}_g$.

Theorem 3.7 (A Summary Theorem for $\check{\theta}_n$)**.** *(i) Suppose that Assumptions 2.12, 3.5, and 3.6 hold true. Assume that $\mathcal{J}_n - \mathcal{J}_0 = O_{\mathbb{P}}\big(\widehat{G}_n(\theta_0,\hat{\gamma}_n)\big)$. If $s > 1/4$ and $r > 1/6$, then we have*

$$\sqrt{n}\big(\check{\theta}_n - \theta_0 - \mathcal{J}_n\mathcal{B}^{\texttt{NL}} - \mathcal{J}_n\mathcal{B}^{\texttt{ANB}}\big) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \mathcal{J}_0\,\check{\Sigma}_g\,\mathcal{J}_0^\intercal).$$

(ii) The assumptions of part (i) and Theorem 3.1 are all true. Then $\sqrt{n}(\check{\theta}_n^w - \theta_0) \xrightarrow{\mathscr{L}} \mathcal{N}(0, \check{\Sigma}_\theta^w)$ *with* $\check{\Sigma}_\theta^w := \mathcal{J}_0 \check{\Sigma}_g^w \mathcal{J}_0^\intercal$*, where* $\check{\Sigma}_g^w$ *is the asymptotic variance of*

$$\sqrt{n}\Big(\sum_{q=1}^{Q} w_q \big(\check{G}_n(\theta_0, \hat{\gamma}_n(h_q)) - \widehat{G}_n(\theta_0, \hat{\gamma}_n(h_q)) \big) + \widehat{G}_n(\theta_0, \gamma_0) + \widehat{G}'_{n,\gamma}(\theta_0, \gamma_0, \hat{\gamma}_n^w - \bar{\gamma}_n^w) \Big).$$

The counterpart of Theorem 3.4 seems to be more complicated, for that the smooth projection g may be unknown, as shown in the hit rates example. In such case, we also need to account for the errors and biases that arise from the estimation of g, g'_γ and $g''_{\gamma\gamma}$. Hence, we leave this to future exploration.

4 Simulation Study

We have conducted a Monte Carlo experiment to investigate the finite-sample performance of the multi-scale jackknife (MSJ) method and the analytical bias correction (ABC) method. We considered three different estimators: (1) the average density (AD) estimator, (2) the integrated squared density (ISD) estimator, and (3) the density-weighted average derivative (DWAD) estimator.

In the first two cases, we considered a one-dimensional mixed normal density given by

$$\gamma_0(x) = \alpha\phi(x; \mu_1, \sigma_1^2) + (1-\alpha)\phi(x; \mu_2, \sigma_2^2),$$

where $\mu_1 = -2$, $\sigma_1^2 = 0.5$, $\mu_2 = 1$, $\sigma_2^2 = 1$, and $\alpha = 0.4$. The true parameter of interest $\theta_0 = \mathbb{E}(\gamma_0(X))$ is given by

$$\theta_0 = \frac{\alpha^2}{\sqrt{4\sigma_1^2\pi}} + \frac{(1-\alpha)^2}{\sqrt{4\sigma_2^2\pi}} + 2\frac{\alpha(1-\alpha)}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\Big(-\frac{1}{2}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} \Big) = 0.0796.$$

In the last case, we are interested in estimating

$$\theta_0 = \mathbb{E}\big(\gamma_0(X)\, \partial_X \mathbb{E}(Y|X)\big) = -2\,\mathbb{E}(\partial_X \gamma_0(X) Y),$$

where $\gamma_0(\cdot)$ is the density of X. We considered a linear model

$$y_i = x_i^\intercal \beta + \epsilon_i,\ x_i \sim \mathcal{N}(0, I_d), \epsilon_i \sim \mathcal{N}(0,1).$$

For simplicity, we let $\beta = 1_d$, a d-dimensional vector with all the elements being one, and focus on estimating θ_{01}.

We employed a Gaussian kernel in all cases. So the order of the kernel is $m = 2$ across all cases. We considered three different sample sizes: $n = 50, 100$, and 200. In each case, we conducted 1,000 simulations. To save space, we only report the results with $n = 100$. Refer to the online supplement for more results.

Figure 1 shows the decomposition of mean squared error (MSE) for various AD estimators, at different bandwidth values. From left to right, it presents the result for the raw estimator without any bias correction, the analytical bias-corrected (ABC) estimator, and the two-scale jackknife (2SJ) estimator (with $\eta = (1, 5/4)$), respectively.

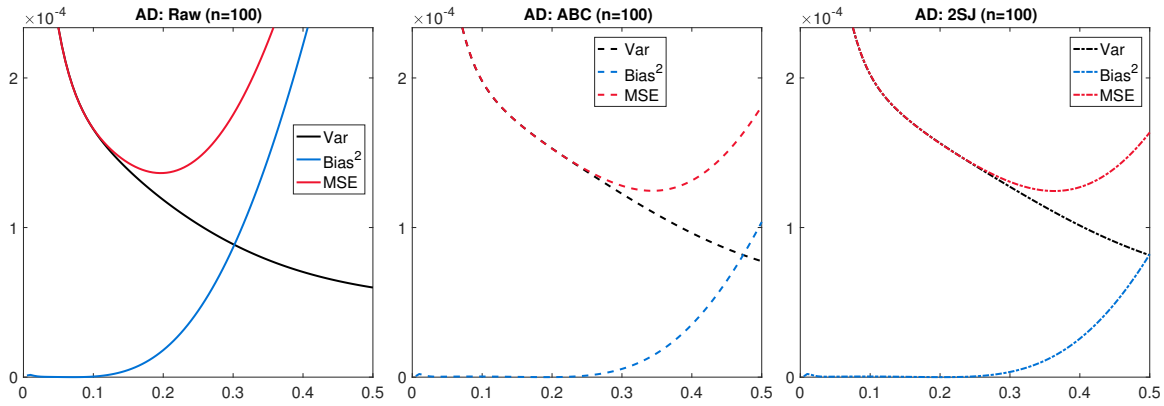

Figure 1: AD: Decomposition of Mean Squared Error

Since the raw estimator is linear in the kernel function, there is no nonlinear bias $\mathcal{B}_n^{\mathtt{NL}}$. As shown in the figure, the bias starts to increase with the bandwidth h when $h > 0.1$ for the raw estimator. While for the other two estimators, this only occurs approximately when $h > 0.25$. In other words, both ABC and 2SJ successfully removed the bias for a substantially large range of bandwidths. For larger values of h, although there is still bias left in the ABC and 2SJ estimators, it has been largely reduced. Consequently, the inference based on either ABC or 2SJ will be much less sensitive to the choice of bandwidth.

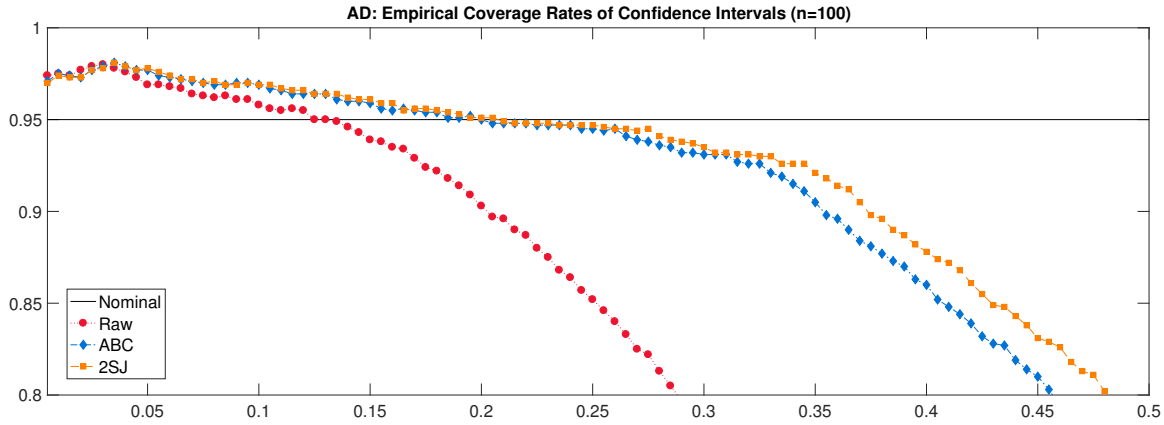

Figure 2: AD: Empirical Coverage Rates of Confidence Intervals

For any given bandwidth value, the variance parts of the ABC and 2SJ estimators are larger than that of the raw one. We think these are due to some finite sample effects. As shown by (Newey et al., 2004), the variance of the twicing-kernel-based semiparametric estimator only depends on the true function(s), not the kernel (cf. the notation following (2.2) therein). This implies that the asymptotic variances of the ABC and the raw estimators should be the same. However, the kernel may have some impacts on the finite-sample variance. While for the 2SJ estimator, it is probably because its two components are not perfectly correlated in such a finite sample. However, the increases are not that large. Hence, the ABC and 2SJ estimators can achieve slightly smaller minimum values for the MSE.

Figure 2 shows the empirical coverage rates for the 95% confidence intervals (CIs) associated with the raw, ABC, and 2SJ estimators. The x-axis is the bandwidth. The coverage rates are about two percentage points higher than the nominal level when h is small. This might be a result of slightly overestimating the asymptotic variance when h is very small. Not surprisingly, the coverage rates decrease, as bias increases (in absolute value). Since the ABC and 2SJ estimators can remove/reduce bias, their corresponding coverage-rate curves have much slower decreasing rates. More importantly, the curves are nearly flat and very close to the nominal level around the region $[0.2, 0.25]$. According to Figure 1, this is a region where the bias remains very close to zero. Besides, since h is not very small in this region, the variance estimators become more precise, compared to the cases with very small bandwidth values.

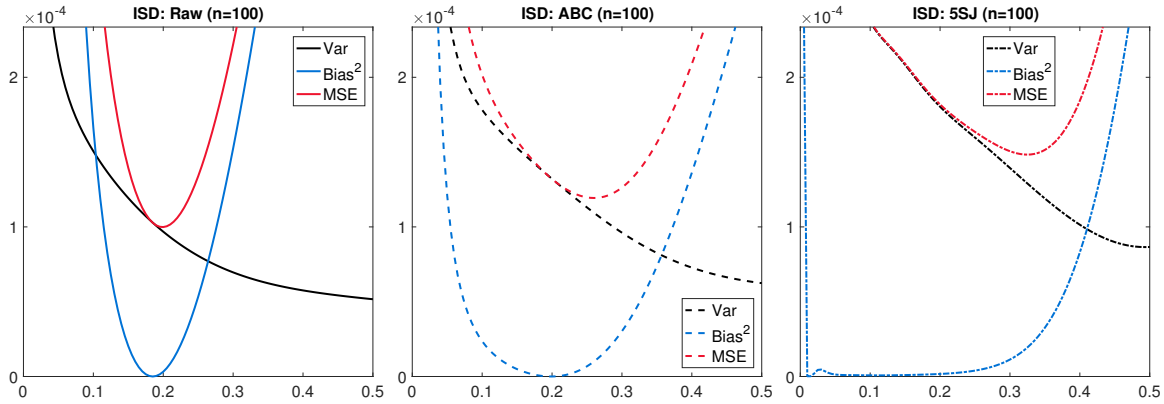

Figure 3: ISD: Decomposition of Mean Squared Error

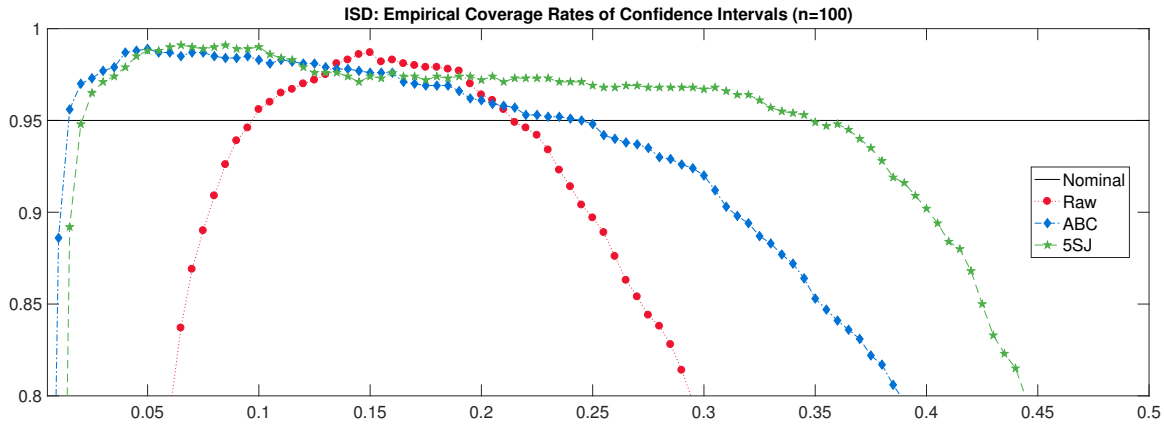

Figure 4: ISD: Empirical Coverage Rates of Confidence Intervals

Figure 3 presents the MSE decomposition results for various ISD estimators. In this case, both the two biases are non-zero. The nonlinear bias $\mathcal{B}_n^{\mathtt{NL}}$ is positive, while the average nonparametric bias $\mathcal{B}_n^{\mathtt{ANB}}$ is negative. This explains why there is a point where the overall bias is zero. Once deviating from this point, the overall bias increases rapidly in magnitude. The ABC method can substantially reduce both biases. One can construct 2SJ to remove/reduce either the nonlinear bias or the average nonparametric bias. However,

we found that 3SJ, which is the counterpart to ABC in this scenario, can only effectively remove the nonlinear bias. Hence, we tried higher-scale jackknife and found that 5SJ has a much better performance (we set $\eta = (3/5, 4/5, 1, 6/5, 7/5)$).

According to Figure 4, the coverage rates of the raw estimator are quite sensitive to the bandwidth, which is consistent with the MSE decomposition result. For the ABC and 5SJ estimators, the coverage rates are more robust to the bandwidth, especially in the latter case. This is not surprising, for that 5SJ can remove/reduce more biases by design. Generally speaking, the coverage rates are higher than the nominal level when the overall bias level is relatively small. One possible explanation is that although the true asymptotic variance of the ISD estimator is the same as that of the AD estimator, we employed a more nonlinear estimator, which may be subject to more sources of finite-sample biases, to estimate it in the ISD case.

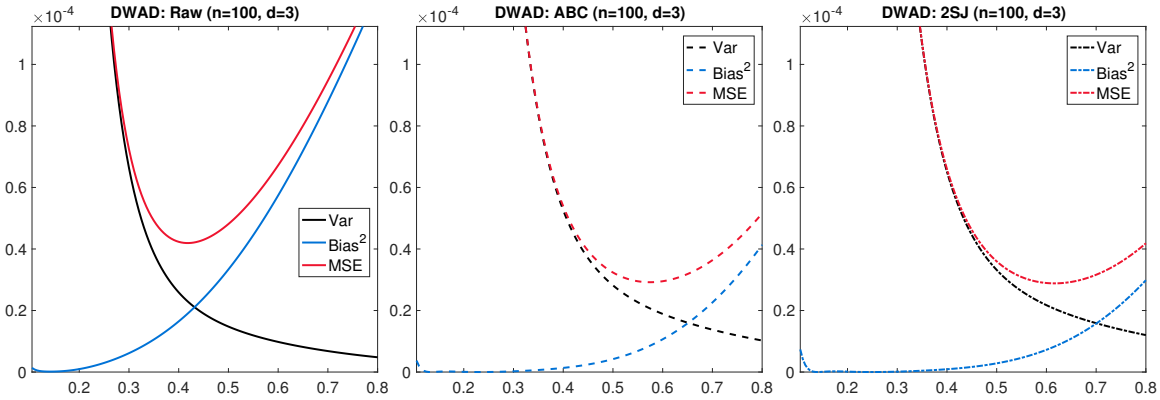

Figure 5: DWAD: Decomposition of Mean Squared Error

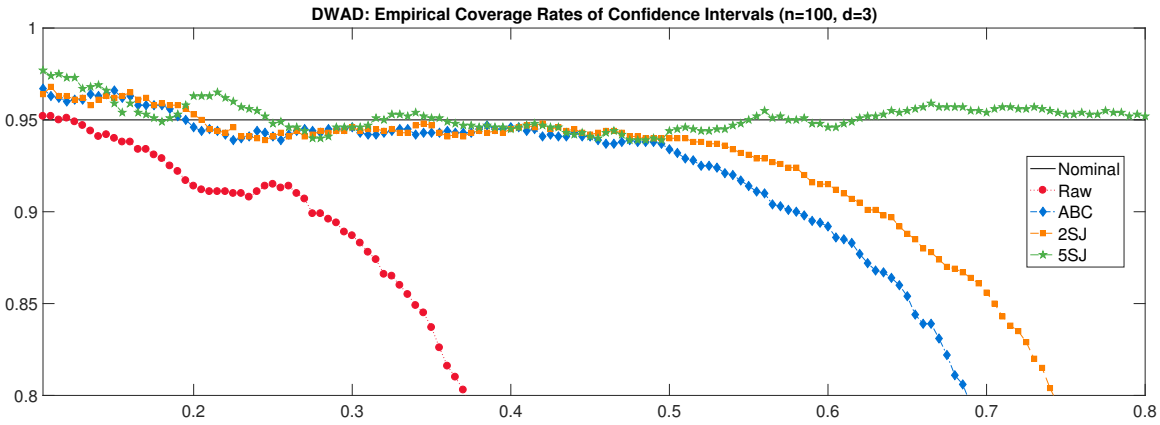

Figure 6: DWAD: Empirical Coverage Rates of Confidence Intervals

For the DWAD estimator, we present the results with $d = 3$, which is larger than the order of the Gaussian kernel ($m = 2$). The general patterns are the same as above. In this case, the MSE gains for the ABC and 2SJ estimators are more noticeable. When constructing the confidence intervals, we used the variance estimator proposed by (Cattaneo et al., 2014) (Case (b) of Theorem 2 therein), while the one considered by (Powell et al.,

1989) leads to over-coverage. The under-coverage of the CI based on the raw estimator is mainly due to the bias. In other cases, the coverage rates are pretty close to the nominal level, when the remaining biases are small. In particular, since the five-scale jackknife estimator successively removes bias for a large range of bandwidth, its CI continues to have good coverage rates across all the bandwidths considered in the simulation.

5 Conclusion

This paper extends the classic framework on semiparametric two-step models, which is developed by (Andrews, 1994), (Newey, 1994), and (Newey and McFadden, 1994), to allow for possibly low-precision nonparametric estimator. We have shown that there are two (or even more) different types of biases in the semiparametric estimator, when its nonparametric ingredient has a slower-than-$n^{1/4}$ convergence rate. We also have proposed two different methods to correct for these biases: one is multi-scale jackknife, the other is analytical-based bias correction. Our simulation study suggests that these bias-correction methods work quite well in finite samples for various kernel-based semiparametric two-step estimators.

References

Andrews, D. W. K. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62, 43–72.

Andrews, I. and A. Mikusheva (2016): "Conditional Inference with Functional Nuisance Parameter," *Econometrica*, 84, 1571–1612.

Bierens, H. J. (1987): "Kernel Estimators of Regression Functions," in *Advances in econometrics: Fifth World Congress*, vol. 1, 99–144.

Borovskikh, Y. V. (1996): *U-statistics in Banach Spaces*, V.S.P. Intl Science.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust nonparametric confidence intervals for regression-discontinuity designs," *Econometrica*, 82, 2295–2326.

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," *Journal of the American Statistical Association*, 108, 1243–1256.

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

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

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5549–5632.

Chen, X., O. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models When The Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017): "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review*, 107, 261–265.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018a): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2018b): "Locally Robust Semiparametric Estimation," Working paper.

Chernozhukov, V., W. K. Newey, and J. M. Robins (2018c): "Double/De-Biased Machine Learning Using Regularized Riesz Representers," Working paper.

Dehling, H. (2006): "Limit theorems for dependent U-statistics," in *Dependence in Probability and Statistics*, Springer, 65–86.

Hampel, F. R. (1974): "The influence curve and its role in robust estimation," *Journal of the American Statistical Association*, 69, 383–393.

Hoeffding, W. (1948): "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Statistics*, 19, 293–325.

Ichimura, H. and W. K. Newey (2017): "The Influence Function of Semiparametric Estimators," Working paper.

Ichimura, H. and P. E. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," in *Handbook of Econometrics*, ed. by J. J. Heckman and E. E. Leamer, Elsevier, vol. 6, 5549–5632.

Kollo, T. and D. von Rosen (2006): *Advanced multivariate statistics with matrices*, vol. 579, Springer Science & Business Media.

Korolyuk, V. S. and Y. V. Borovskich (1994): *Theory of U-statistics*, vol. 273 of *Mathematics and Its Applications*, Springer.

Li, J., Y. Liu, and D. Xiu (2019): "Efficient Estimation of Integrated Volatility Functionals via Multiscale Jackknife," *The Annals of Statistics*, 47, 156–176.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 1349–1382.

Newey, W. K., F. Hsieh, and J. M. Robins (2004): "Twicing Kernels and A Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947–962.

Newey, W. K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle and D. L. McFadden, Elsevier, vol. 4, 2111–2245.

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

Quenouille, M. H. (1949): "Problems in Plane Sampling," *The Annals of Mathematical Statistics*, 20, 335–375.

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

Stuetzle, W. and Y. Mittal (1979): "Some Comments on The Asymptotic Behavior of Robust Smoothers," in *Smoothing Techniques for Curve Estimation*, Springer, vol. 757 of *Lecture Notes in Mathematics*, 191–195.

Yang, X. (2020): "Semiparametric Estimation in Continuous-Time: Asymptotics for Integrated Volatility Functionals with Small and Large Bandwidths," *Journal of Business & Economic Statistics*, forthcoming.