Three Essays in Microeconometrics

by

Xinwei Ma

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in The University of Michigan
2019

Doctoral Committee:

Professor Matias D. Cattaneo, Chair
Assistant Professor Andreas Hagemann
Professor Lutz Kilian
Professor Rocío Titiunik

Xinwei Ma

xinweima@umich.edu

ORCID iD: 0000-0001-8827-9146

To my parents, 马践原 and 其泊热

ACKNOWLEDGMENTS

First and foremost, I am deeply indebted to Matias Cattaneo. As an advisor, he inspired me to study econometric theory: “I see you are usually thinking and asking the ‘right’ type of questions. You should think seriously whether you are interested in doing some econometrics research.” These “right questions” shaped my doctoral dissertation, and turned me into the scholar I am today. As a colleague, we started working together since the summer of 2014, and I have continuously benefited from his insightful comments and endless patience. As a friend, he constantly reminds me: “Are you not working!”

I have been fortunate to be surrounded by amazing faculty at the University of Michigan. It has been a source of encouragement and good advice. I am especially grateful for my dissertation committee members, Andreas Hagemann, Lutz Kilian and Rocío Titiunik. They have been critical and truly dedicated, and spent tremendous effort in polishing my research writing and presentation skill.

I would also like to express my gratitude to the many of my friends. It is them who make foreign cities home. It was a pleasure to share an office with Xing Guo, Kenichi Nagasawa and Elchin Suleymanov. They made Lorch 107 the most hard-working and productive office.

Last but not least, I thank my parents. They are never able to understand what econometrics is and why robust inference is important. However, it was them who taught me love learning in the first place and kept doing so for the past 25 years. This dissertation would not have been possible without their support and sacrifices – It has not always been easy to send their only child overseas, and I could have spent much more time with them.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

ABSTRACT

Traditional econometric methods can perform poorly in applications. The poor performance is usually due to challenges faced by researchers conducting empirical data analysis, yet overlooked by large sample reasonings that depend on stringent conditions. Such lack of robustness can be detrimental to economic decision making and prescribing policy recommendations. This dissertation consists of three connected chapters on important issues in microeconometric theory, with a particular emphasis on developing robust inference procedures in program evaluation and other microeconomic settings.

The first chapter discusses the implications of small probability weights entering the inverse probability weighting estimator, and proposes an inference procedure that is robust to not only small probability weights but also a wide range of trimming choices. Robustness is achieved by combining resampling techniques with a novel bias correction method. This chapter is based on the working paper "Robust Inference Using Inverse Probability Weighting" (Ma and Wang, 2019).

In an important class of two-step semi-parametric models, the second chapter provides estimation and inference procedures that are robust to including high-dimensional covariates in the first-step estimation. Robustness is achieved by the jackknife bias correction, and the bootstrap is employed for statistical inference. This chapter is based on the paper "Two-Step Estimation and Inference with Possibly Many Included Covariates" (Cattaneo, Jansson and Ma, 2018d).

The third chapter develops a non-parametric estimator of probability density functions based on local polynomial techniques. The proposed estimator is easy to implement and is robust to discontinuities in the underlying density – an important concern in empirical research. This chapter is based on the working paper "Simple Local Polynomial Density Estimators" (Cattaneo, Jansson and Ma, 2019b).

CHAPTER I

Robust Inference Using Inverse Probability Weighting

Abstract. *Inverse Probability Weighting (IPW) is widely used in program evaluation and other empirical economics applications. As Gaussian approximations perform poorly in the presence of "small denominators," trimming is routinely employed as a regularization strategy. However, ad hoc trimming of the observations renders usual inference procedures invalid for the target estimand, even in large samples. This chapter proposes an inference procedure that is robust not only to small probability weights entering the IPW estimator, but also to a wide range of trimming threshold choices. Our inference procedure employs resampling with a novel bias correction technique. Specifically, we show that both the IPW and trimmed IPW estimators can have different (Gaussian or non-Gaussian) limiting distributions, depending on how "close to zero" the probability weights are and on the trimming threshold. Our method provides more robust inference for the target estimand by adapting to these different limiting distributions. This robustness is partly achieved by correcting a non-negligible trimming bias. We demonstrate the finite-sample accuracy of our method in a simulation study, and we illustrate its use by revisiting a dataset from the National Supported Work program.*

I.1 Introduction

Inverse Probability Weighting (IPW) is widely used in program evaluation settings, such as instrumental variables, difference-in-differences and counterfactual analysis. Other applications of IPW include survey adjustment, data combination, and models involving missing data or measurement error. In practice, it is common to observe small probability weights entering the IPW estimator. This renders inference based on standard Gaussian approximations invalid, even in large samples, because these approximations rely crucially on the probability weights being well-separated from zero. In a recent study, Busso, DiNardo and McCrary (2014) investigated the finite sample performance of commonly used IPW treat-

This chapter is based on the working paper "Robust Inference Using Inverse Probability Weighting" (Ma and Wang, 2019).

ment effect estimators, and documented that small probability weights can be detrimental to statistical inference. In response to this problem, observations with probability weights below a certain threshold are often excluded from subsequent statistical analysis. The exact amount of trimming, however, is usually ad hoc and will affect the performance of the IPW estimator and the corresponding confidence interval in nontrivial ways.

In this chapter, we show that both the IPW and trimmed IPW estimators can have different (Gaussian or non-Gaussian) limiting distributions, depending on how "close to zero" the probability weights are and on how the trimming threshold is specified. We propose an inference procedure that adapts to these different limiting distributions, making it robust not only to small probability weights, but also to a wide range of trimming threshold choices. To achieve this "two-way robustness," our method employs a resampling technique combined with a novel bias correction, which remains valid for the target estimand even when trimming induces a non-negligible bias. In addition, we propose an easy-to-implement method for choosing the trimming threshold by minimizing an empirical analogue of the asymptotic mean squared error.

To understand why standard inference procedures are not robust to small probability weights, we first consider the large-sample properties of the IPW estimator

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{\hat{e}(X_i)}, \tag{I.1}$$

where $D_i \in \{0,1\}$ is binary, Y_i is the outcome of interest, and $e(X_i) = \mathbb{P}[D_i = 1|X_i]$ is the probability weight conditional on the covariates, with $\hat{e}(X_i)$ being its estimate. The asymptotic framework we employ is general and allows, but does not require that the probability weights have a heavy tail near zero. If the probability weights are bounded away from zero, the IPW estimator is $\sqrt{n}$-consistent with a limiting Gaussian distribution. Otherwise, a slower-than-$\sqrt{n}$ convergence rate and a non-Gaussian limiting distribution can emerge, for which regular large-sample approximation no longer applies. Specifically, in the latter case,

$$\frac{n}{a_n}\Big(\hat{\theta}_n - \theta_0\Big) \xrightarrow{\mathrm{d}} \mathcal{L}(\gamma_0, \alpha_+(0), \alpha_-(0)), \tag{I.2}$$

where θ_0 is the parameter of interest and $a_n \to \infty$ is a sequence of normalizing factors. The limiting distribution, $\mathcal{L}(\cdot)$, depends on three parameters. The first parameter γ_0 is related to the "tail behavior" of the probability weights near zero. Only if the tail is relatively thin, the limiting distribution will be Gaussian; otherwise it will be a Lévy stable distribution. In the non-Gaussian case, the limiting distribution does not need to be symmetric, with its two tails characterized by $\alpha_+(0)$ and $\alpha_-(0)$. Another complication in the non-Gaussian case is that the convergence rate, n/a_n, is typically unknown, and depends again on how "close to zero" the probability weights are.

In an effort to circumvent this problem, practitioners typically use trimming as a regularization strategy. The idea is to exclude observations with small probability weights from the analysis. However, the performance of standard inference procedures is sensitive to the amount of trimming. We study the trimmed IPW estimator

$$\hat{\theta}_{n,b_n} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{\hat{e}(X_i)} \mathbb{1}_{\hat{e}(\mathbf{x}_i) \geq b_n}. \tag{I.3}$$

The large-sample properties of this estimator depend heavily on the choice of the trimming threshold, b_n. In particular,

$$\frac{n}{a_{n,b_n}} \left(\hat{\theta}_{n,b_n} - \theta_0 - \mathsf{B}_{n,b_n} \right) \xrightarrow{\text{d}} \mathcal{L}(\gamma_0, \alpha_+(\cdot), \alpha_-(\cdot)). \tag{I.4}$$

Compared to (I.2), the most noticeable change is that a trimming bias B_{n,b_n} emerges. This bias has order $\mathbb{P}[e(X) \leq b_n]$, hence it will vanish asymptotically if the trimming threshold shrinks to zero. However, the trimming bias can still contribute to the mean squared error of the estimator nontrivially. Furthermore, it can be detrimental to statistical inference, since the limiting distribution is shifted away from the target estimand by $\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}$, which may not vanish even in large samples. Indeed, in a simple simulation setting with sample size $n = 2,000$ and a trimming threshold $b_n = 0.036$, the bias B_{n,b_n} is already quite severe (three times as large as the variability of the point estimate). Another noticeable change with trimming is that the normalizing factor, a_{n,b_n}, can depend on the trimming threshold. As a result, the trimmed IPW estimator may have a different convergence rate compared to the untrimmed estimator. An extreme case is fixed trimming ($b_n = b > 0$), which forces the probability weights to be well-separated from zero. In this case, the trimmed estimator converges to a pseudo-true parameter at the usual parametric rate $n/a_{n,b_n} = \sqrt{n}$. Finally, the form of the limiting distribution also changes and can depend on two infinite dimensional objects, $\alpha_+(\cdot)$ and $\alpha_-(\cdot)$, making inference based on the estimated limiting distribution prohibitively difficult.

As the large-sample properties of both the IPW and trimmed IPW estimators are sensitive to small probability weights and to the amount of trimming, it is important to develop an inference procedure that automatically adapts to the relevant limiting distributions. However, it is difficult to base inference on estimates of the nuisance parameters in (I.2) or (I.4), and the standard nonparametric bootstrap is known to fail in our setting (Athreya, 1987; Knight, 1989). We instead propose the use of subsampling (Politis and Romano, 1994). In particular, we show that subsampling provides valid approximations to the limiting distribution in (I.2) for the IPW estimator, and automatically adapts to the distribution in (I.4) under trimming. With self-normalization (i.e., subsampling a Studentized statistic), it also overcomes the difficulty of having a possibly unknown convergence rate.

Subsampling alone does not suffice for valid inference due to the bias induced by trimming. A desirable inference procedure should be valid even when the trimming bias is nonnegligible. That is, it should be robust not only to small probability weights but also to a wide range of trimming threshold choices. To achieve this "two-way robustness," we combine subsampling with a novel bias correction method based on local polynomial regression. Specifically, our method regresses the outcome variable on a polynomial of the probability weight in a region local to 0, and estimates the trimming bias with the regression coefficients. In the current context, however, local polynomial regressions cannot be analyzed with standard techniques available in the literature (Fan and Gijbels, 1996), as the density of the probability weights can be arbitrarily close to zero in the subsample $D = 1$. Both the variance and bias of the local polynomial regression change considerably.

Finally, we address the question of how to choose the trimming threshold. One extreme possibility is fixed trimming ($b_n = b > 0$). Although fixed trimming helps restore asymptotic Gaussianity by forcing the probability weights to be bounded away from zero, this practice is difficult to justify, unless one is willing to re-interpret the estimation and inference result completely (Crump, Hotz, Imbens and Mitnik, 2009). We instead propose to determine the trimming threshold by taking into consideration both the bias and variance of the trimmed IPW estimator. We suggest an easy-to-implement method to choose the trimming threshold by minimizing an empirical analogue of the asymptotic mean squared error.

From a practical perspective, results in this chapter relate to the large literature on program evaluation and causal inference (Imbens and Rubin, 2015; Abadie and Cattaneo, 2018; Hernán and Robins, 2018). Inverse weighting type estimators are widely used in missing data models (Robins, Rotnitzky and Zhao, 1994; Wooldridge, 2007) and for estimating treatment effects (Hirano, Imbens and Ridder, 2003; Cattaneo, 2010). They also feature in settings such as instrumental variables (Abadie, 2003), difference-in-differences (Abadie, 2005), counterfactual analysis (DiNardo, Fortin and Lemieux, 1996) and survey sampling adjustment (Wooldridge, 1999). From a theoretical perspective, the IPW estimator is known to behave poorly when the probability weights are close to zero (Khan and Tamer, 2010). Some attempts have been made to deal with this problem. Heiler and Kazak (2018) also consider how to conduct inference when the probability weights can be arbitrarily close to zero. They establish a stable convergence result for the (untrimmed) IPW estimator, a conclusion similar to (I.2), and propose the use of subsampling for inference. However, they do not address the issue of trimming, nor do they discuss how the trimming threshold should be chosen in practice. Chaudhuri and Hill (2016) propose a trimming strategy based on the absolute magnitude of $|DY/e(X)|$. However, their method only allows the trimming of a few observations. Moreover, both inference and bias correction rely on estimates of certain tail features, which can be difficult to obtain. Hong, Leung and Li (2018) consider a setting where observations fall into finitely many strata, and propose to measure the severity of limited overlap by how fast the propensity score approaches an extreme. To conduct inference for moments of ratios, Sasaki and Ura (2018) propose a trimming method and a companion sieve-based bias correction technique.

Trimming has also been studied in the literature on heavy-tailed random variables. As in our setting, different limiting distributions can emerge (Csörgő, Haeusler and Mason, 1988; Hahn and Weiner, 1992; Berkes, Horváth and Schauer, 2012). However, the focus in that literature has been almost exclusively on extreme order statistics. Hence, the results do not apply to the trimming strategy which practitioners use. Crump, Hotz, Imbens and Mitnik (2009) and Yang and Ding (2018) are two exceptions. They consider the probability weight based trimming, as we do in this chapter, but both studies assume that the probability weights are already bounded away from zero. Trimming is not unique to the inverse probability weighting framework. Hill and Renault (2012) propose tail trimming for the variance targeting estimator. It turns out that tail trimming is crucial to establish asymptotic normality, as Vaynman and Beare (2014) show that stable convergence may arise for the untrimmed variance targeting estimator.

With the IPW estimator as a special case, Cattaneo and Jansson (2018) and Cattaneo, Jansson and Ma (2018d) show how an asymptotic bias can arise in a two-step semiparametric setting where the first step employs small bandwidths, which corresponds to undersmoothing, or many covariates, which corresponds to overfitting. Along another direction, Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018b) develop robust inference procedures against oversmoothing bias. The first-order bias we document in this chapter is both qualitatively and quantitatively different, as it emerges due to trimming and will be present even when the probability weights are directly observed (making the estimator a one-step procedure), and certainly will not disappear with model selection or machine learning methods (Athey, Imbens and Wager, 2018; Belloni, Chernozhukov, Chetverikov, Hansen and Kato, 2018; Farrell, 2015; Farrell, Liang and Misra, 2018).

In Section I.2, we study the large-sample properties of the IPW estimator, and show that subsampling provides valid distributional approximations. In Section I.3, we extend our analysis to the trimmed IPW estimator, for which we discuss in detail the bias correction required for our robust inference procedure. A data-driven method to choose the trimming threshold is also proposed. Section I.4 shows how our framework can be extended to provide robust inference for treatment effects and parameters defined through a nonlinear moment condition. Section I.5 provides numerical evidence from a wide array of simulation designs and an empirical example. Section I.6 concludes. Additional results, preliminary lemmas and all proofs are collected in Section I.7 and I.8.

I.2 The IPW Estimator

Let (Y_i, D_i, X_i), $i = 1, 2, \cdots, n$ be a random sample from $Y \in \mathbb{R}$, $D \in \{0, 1\}$ and $X \in \mathbb{R}^{d_x}$. Recall that the probability weight is defined as $e(X) = \mathbb{P}[D = 1|X]$. Define the conditional moments of the outcome variable as

$$\mu_s(e(X)) = \mathbb{E}[Y^s|e(X), D = 1], \quad s > 0,$$

then the parameter of interest is $\theta_0 = \mathbb{E}[DY/e(X)] = \mathbb{E}[\mu_1(e(X))]$. At this level of generality, we do not attach specific interpretations to the parameter and the random variables in our model. To facilitate understanding, one can think of Y as an observed outcome variable and D as an indicator of treatment status, hence the parameter is the population average of one potential outcome (see Section I.4.1 for a treatment effect setting).

As previewed in Section I.1, the large-sample properties of the IPW estimator $\hat{\theta}_n$ depend on the tail behavior of the probability weight near zero. If $e(X)$ is bounded away from zero, the IPW estimator is $\sqrt{n}$-consistent and asymptotically Gaussian. In the presence of small probability weights, however, a non-Gaussian limiting distribution can emerge. In this section, we first discuss the assumptions and formalize the notion of probability weights "being close to zero" or "having a heavy tail." Then we give precise statements on the large-sample properties of the IPW estimator, and propose an inference procedure that is robust to small probability weights.

I.2.1 Tail Behavior

For an estimator that takes the form of a sample average (or more generally can be linearized into such), distributional approximation based on the central limit theorem only requires a finite variance. The problem with inverse probability weighting with "small denominators," however, is that the estimator may not have a finite variance. In this case, distributional convergence relies on tail features, which we formalize in the following assumption.

Assumption I.1 (Regularly varying tail)
For some $\gamma_0 > 1$, the probability weight has a regularly varying tail with index $\gamma_0 - 1$ at zero:

$$\lim_{t\downarrow 0} \frac{\mathbb{P}[e(X) \leq tx]}{\mathbb{P}[e(X) \leq t]} = x^{\gamma_0 - 1}, \qquad \text{for all } x > 0. \qquad \|$$

Assumption I.1 only imposes a local restriction on the tail behavior of the probability weights, and is common when dealing with sums of heavy-tailed random variables. This assumption encompasses the special case that $\mathbb{P}[e(X) \leq x] = c(x)x^{\gamma_0-1}$ with $\lim_{x\downarrow 0} c(x) > 0$ (i.e., approximately polynomial tail).[1] To see how the tail index γ_0 features in data, Figure I.1 shows the distribution of the probability weights simulated with $\gamma_0 = 1.5$. There, it is clear that the probability weights exhibit a heavy tail near 0 (more precisely, the density of $e(X)$, if it exists, diverges to infinity). In Section I.5.2, we illustrate this point with estimated probability weights from an empirical example, and a similar pattern emerges. Later in Theorem I.1, we show that $\gamma_0 = 2$ is the boundary case that separates the Gaussian and the non-Gaussian limiting distributions for the IPW estimator. With $\gamma_0 = 2$, the probability weight is approximately uniformly distributed, a fact that can be used in practice as a rough guidance on the magnitude of this tail index.

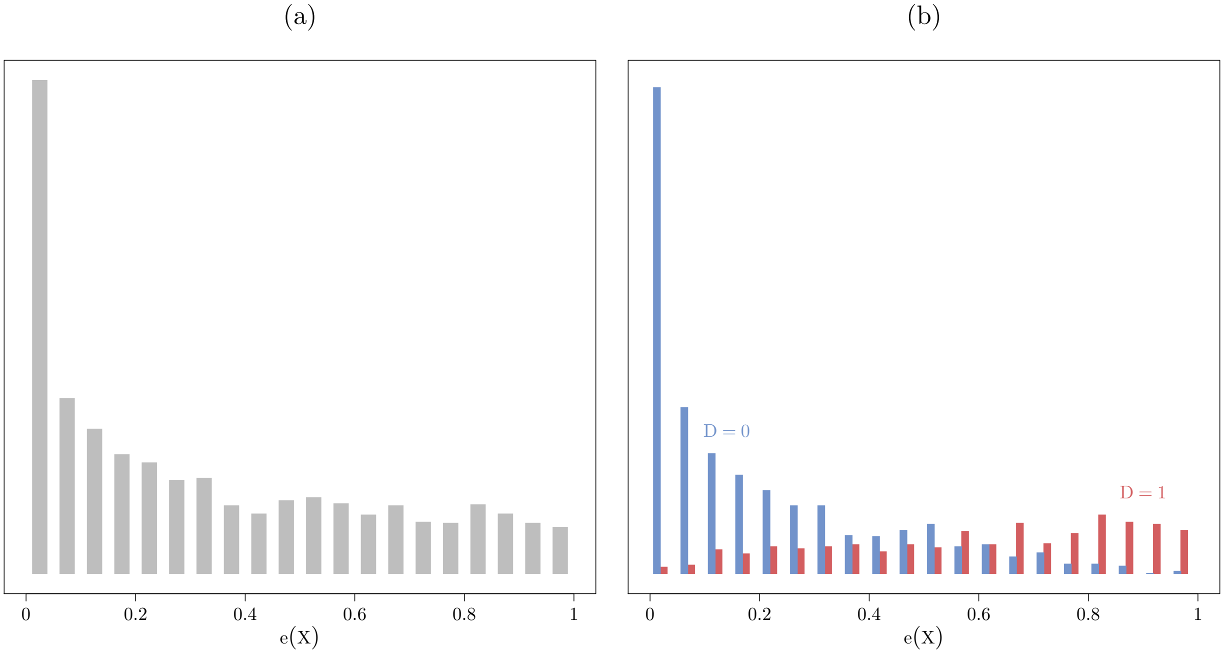
Remark I.1 (Identification) The requirement $\gamma_0 > 1$ ensures point identification of the parameter θ_0, as it implies $\mathbb{P}[e(X) = 0] = 0$. $\|$

Remark I.2 (Tail property of the inverse weight) Assumption I.1 can be equivalently rewritten as a tail condition of the inverse weight: $\mathbb{P}[D/e(X) \geq x] \approx x^{-\gamma_0}$, as $x \uparrow \infty$. (Precisely, $D/e(X)$ has a regularly varying tail at ∞ with index $-\gamma_0$.) Therefore, γ_0 determines what moments the inverse weight possesses. For our purpose, it is more instructive to have a result on the tail behavior of $DY/e(X)$. This is made precise in Lemma I.1, for which an additional assumption is needed. $\|$

Assumption I.1 characterizes the tail behavior of the probability weights. However, it alone does not suffice for the IPW estimator to have a limiting distribution. The reason is that, for sums of random variables without finite variance to converge in distribution, one needs not only a restriction on the shape of the tail, but also a "tail balance condition." This should be compared to the asymptotically Gaussian case, in which no tail restriction is necessary beyond a finite variance.

[1] Assumption I.1 is equivalent to $\mathbb{P}[e(X) \leq x] = c(x)x^{\gamma_0-1}$ with $c(x)$ being a slowly varying function. Because $c(x)$ does not need to have a well-defined limit as $x \downarrow 0$, Assumption I.1 is more general than assuming an approximately polynomial tail. See Section I.7 for more detail.

Figure I.1. Illustration of γ_0.



Note. Sample size: $n = 2,000$. $\mathbb{P}[e(X) \le x] = x^{\gamma_0 - 1}$ with $\gamma_0 = 1.5$. (a) Distribution of the probability weights. (b) Distribution of the probability weights, separately for subgroups $D = 1$ (red) and $D = 0$ (blue).

Assumption I.2 (Conditional distribution of Y)

(i) For some $\varepsilon > 0$, $\mathbb{E}\left[|Y|^{(\gamma_0 \vee 2)+\varepsilon}\middle|e(X) = x, D = 1\right]$ is uniformly bounded. (ii) There exists a probability distribution F, such that for all bounded and continuous $\ell(\cdot)$, $\mathbb{E}[\ell(Y)|e(X) = x, D = 1] \to \int_{\mathbb{R}} \ell(y)F(\mathrm{d}y)$ as $x \downarrow 0$. ‖

This assumption has two parts. The first part requires the tail of Y to be thinner than that of $D/e(X)$, therefore the tail behavior of $DY/e(X)$ is largely driven by the "small denominator $e(X)$." As our primary focus is the implication of small probability weights entering the IPW estimator rather than a heavy-tailed outcome variable, we maintain this assumption. The second part requires convergence of the conditional distribution of Y given $e(X)$ and $D = 1$. Together, they help characterize the tail behavior of $DY/e(X)$. Specifically, the two tails of $DY/e(X)$ are balanced.

Lemma I.1 (Tail property of $DY/e(X)$)

Under Assumptions I.1 and I.2,

$$\lim_{x\to\infty} \frac{x\mathbb{P}[DY/e(X) > x]}{\mathbb{P}[e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0}\alpha_+(0), \quad \lim_{x\to\infty} \frac{x\mathbb{P}[DY/e(X) < -x]}{\mathbb{P}[e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0}\alpha_-(0),$$

where

$$\alpha_+(x) = \lim_{t\to 0} \mathbb{E}\Big[|Y|^{\gamma_0}\mathbb{1}_{Y>x}\Big| e(X)=t, D=1\Big],\quad \alpha_-(x) = \lim_{t\to 0} \mathbb{E}\Big[|Y|^{\gamma_0}\mathbb{1}_{Y<x}\Big| e(X)=t, D=1\Big].$$

$\|$

Assuming the distribution of the outcome variable is nondegenerate conditional on the probability weights being small (i.e., $\alpha_+(0)+\alpha_-(0)>0$), Lemma I.1 shows that $DY/e(X)$ has regularly varying tails with index $-\gamma_0$. As a result, γ_0 determines which moment of the IPW estimator is finite: for $s<\gamma_0$, $\mathbb{E}[|DY/e(X)|^s]<\infty$, and for $s>\gamma_0$, the moment is infinite. Thanks to Assumption I.2(ii), Lemma I.1 also implies that $DY/e(X)$ has balanced tails: the ratio $\frac{\mathbb{P}[DY/e(X)>x]}{\mathbb{P}[|DY/e(X)|>x]}$ tends to a finite constant. It turns out that without a finite variance, the limiting distribution of the IPW estimator is non-Gaussian, and the limiting distribution depends on both the left and right tails of $DY/e(X)$. This should be compared to the asymptotically Gaussian case, where delicate tail properties do not feature in the asymptotic distribution beyond a finite second moment. Thus, tail balancing (and Assumption I.2(ii)) is indispensable for developing a large sample theory allowing small probability weights entering the IPW estimator.

Lemma I.1 also helps clarify different consequences of small probability weights/small denominators. If $\gamma_0>2$, the IPW estimator is asymptotically Gaussian: $\sqrt{n}(\hat{\theta}_n-\theta_0)\xrightarrow{\text{d}}\mathcal{N}(0,\mathbb{V}[DY/e(X)])$, although the probability weights can still be close to zero. The reason is that, with large $\gamma_0>2$, small denominators appear so infrequently that they will not affect the large-sample properties.

For $\gamma_0\in(1,2]$, the IPW estimator no longer has finite variance, and without further restrictions on the data generating process, the parameter is not $\sqrt{n}$-estimable. Since the distribution of $e(X)$ does not approach zero fast enough (or equivalently, the density of $e(X)$, if it exists, diverges to infinity), it represents the empirical difficulty of dealing with small probability weights entering the IPW estimator, for which regular asymptotic analysis no longer applies.

Remark I.3 (Limited overlap) When estimating treatment effects (see Section I.4.1 for a setup), it is possible that covariates are distributed very differently across the treatment and the control group. Even worse, for some region in the covariates distribution, one may observe abundant units from one group, yet units from the other group are scarce. This is commonly referred to as “limited overlap,” and is one instance in which extreme probability weights (propensity scores) can arise (Imbens and Rubin, 2015, Chapter 14).

Hong, Leung and Li (2018) consider a setting where observations fall into finitely many strata (hence the propensity score has a finite support), and propose to use the quantity

"$n \min_{1\leq i\leq n} e(X_i)$" as the measure of the effective sample size (severity of limited overlap). They require this measure to diverge in large samples, which is equivalent to $\gamma_0 > 2$ in our setting. To see this connection,

$$\mathbb{P}\Big[n \min_{1\leq i\leq n} e(X_i) > x\Big] = \Big(1 - \mathbb{P}[e(X) \leq n^{-1}x]\Big)^n \asymp \Big(1 - \big(n^{-1}x\big)^{\gamma_0 - 1}\Big)^n,$$

so that $n \min_{1\leq i\leq n} e(X_i) \xrightarrow{\mathrm{p}} \infty$ if and only if $\gamma_0 > 2$, which guarantees that the IPW estimator is $\sqrt{n}$-consistent and asymptotically Gaussian. ‖

Remark I.4 (Implied tail of X) To see how the tail behavior of the probability weights is related to that of the covariates X, we consider a Logit model:

$$e(X) = \exp(X^{\mathrm{T}}\pi_0)/(1 + \exp(X^{\mathrm{T}}\pi_0)).$$

Note that when the index $X^{\mathrm{T}}\pi_0$ approaches $-\infty$, the probability weight approaches zero, and

$$\mathbb{P}[e(X) \leq x] = \mathbb{P}\left[\frac{1}{1 + \exp(-X^{\mathrm{T}}\pi_0)} \leq x\right] = \mathbb{P}\left[X^{\mathrm{T}}\pi_0 < -\log(x^{-1} - 1)\right].$$

As a result, Assumption I.1 is equivalent to that, for all x large enough, $\mathbb{P}[X^{\mathrm{T}}\pi_0 < -x] \approx e^{-(\gamma_0 - 1)x}$, meaning that the (left) tail of $X^{\mathrm{T}}\pi_0$ is approximately sub-exponential. ‖

I.2.2 Large Sample Properties of the IPW Estimator

The following theorem characterizes the limiting distribution of the IPW estimator. To make the result concise, we assume the oracle (rather than estimated) probability weights are used, making the IPW estimator a one-step procedure. We extend the theorem to estimated probability weights in the next subsection.

Theorem I.1 (Large sample properties of the IPW estimator)
Assume Assumptions I.1 and I.2 hold with $\alpha_+(0) + \alpha_-(0) > 0$. Let a_n be defined from

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{DY}{e(X)} - \theta_0\right|^2 \mathbb{1}_{|DY/e(X)|\leq a_n}\right] \to 1.$$

Then (I.2) *holds with $\mathcal{L}(\gamma_0, \alpha_+(0), \alpha_-(0))$ being:*
(i) the standard Gaussian distribution if $\gamma_0 \geq 2$; and

(ii) the Lévy stable distribution if $\gamma_0 < 2$, with characteristic function:

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x)\right\},$$
$$\textit{where } M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(0)\mathbb{1}_{x\geq 0} + \alpha_-(0)\mathbb{1}_{x<0}\Big)\right]. \qquad \|$$

This theorem demonstrates how a non-Gaussian limiting distribution can emerge when the IPW estimator does not have a finite variance ($\gamma_0 < 2$). The limiting Lévy stable distribution is generally not symmetric (unless the outcome variable is conditionally symmetrically distributed), and has tails much heavier than that of a Gaussian distribution. As a result, inference procedures based on the standard Gaussian approximation perform poorly.

Theorem I.1 also shows how the convergence rate of the IPW estimator depends on the tail index γ_0. For $\gamma_0 > 2$, the IPW estimator converges at the usual parametric rate $n/a_n = \sqrt{n}$. This extends to the $\gamma_0 = 2$ case, except that an additional slowly varying factor is present in the convergence rate. For $\gamma_0 < 2$, a_n is only implicitly defined from a truncated second moment, and generally does not have an explicit formula. One can consider the special case that the probability weights have an approximately polynomial tail: $\mathbb{P}[e(X) \leq x] \asymp x^{\gamma_0 - 1}$, for which a_n can be set to n^{1/γ_0}. As a result, the IPW estimator will have a slower convergence rate if the probability weights have a heavier tail at zero (i.e., smaller γ_0). Fortunately, the (unknown) convergence rate is captured by self-normalization (Studentization), which we employ in our robust inference procedure.

As a technical remark, the characteristic function in Theorem I.1(ii) has an equivalent representation, from which we deduce several properties of the limiting Lévy stable distribution. In particular,

$$\psi(\zeta) = -|\zeta|^{\gamma_0}\frac{\Gamma(3-\gamma_0)}{\gamma_0(\gamma_0-1)}\left[-\cos\left(\frac{\gamma_0\pi}{2}\right) + i\frac{\alpha_+(0)-\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\mathrm{sgn}(\zeta)\sin\left(\frac{\gamma_0\pi}{2}\right)\right],$$

where $\Gamma(\cdot)$ is the gamma function and $\mathrm{sgn}(\cdot)$ is the sign function. First, this distribution is not symmetric unless $\alpha_+(0) = \alpha_-(0)$. Second, the characteristic function has a sub-exponential tail, meaning that the limiting stable distribution has a smooth density function (although in general it does not have a closed-form expression). Finally, the above characteristic function is continuous in γ_0, in the sense that as $\gamma_0 \uparrow 2$, it reduces to the standard Gaussian characteristic function.

I.2.3 Estimated Probability Weights

The probability weights are usually unknown and are estimated in a first step, which are then plugged into the IPW estimator, making it a two-step estimation problem. In this subsection, we discuss how estimating the probability weights in a first step will affect the results of Theorem I.1. To start, consider the following expansion:

$$\frac{n}{a_n}\left(\hat{\theta}_n - \theta_0\right) = \underbrace{\frac{1}{a_n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i)} - \theta_0\right)}_{\text{Theorem I.1}} + \underbrace{\frac{1}{a_n}\sum_{i=1}^{n}\frac{D_iY_i}{e(X_i)}\left(\frac{e(X_i)}{\hat{e}(X_i)} - 1\right)}_{\text{Proposition I.1}},$$

where the first term is already captured by Theorem I.1. At this level of generality, it is not possible to determine whether the second term in the above expansion has a nontrivial (first order) impact. In fact, nothing prevents the second term from being dominant in large samples, which happens, for example, when the probability weights are estimated at a rate slower than n/a_n. Even if the probability weights are estimated at the usual parametric rate, the difference between their inverses may not be small at all (due to the presence of "small estimated denominators"). In this subsection, we first impose high-level assumptions and discuss the impact of employing estimated probability weights. Then we specialize to generalized linear models, and verify the high-level assumptions for Logit and Probit models which are widely used in applied work.

Assumption I.3 (First step)
The probability weights are parametrized as $e(X,\pi)$ with $\pi \in \Pi$, and $e(\cdot)$ is continuously differentiable with respect to π. Let $e(X) = e(X,\pi_0)$ and $\hat{e}(X) = e(X,\hat{\pi}_n)$. Further,
(i) $\sqrt{n}(\hat{\pi}_n - \pi_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} h(D_i, X_i) + o_{\mathrm{p}}(1)$, where $h(D_i, X_i)$ is mean zero and has a finite variance.
(ii) For some $\varepsilon > 0$, $\mathbb{E}\left[\sup_{\pi:|\pi-\pi_0|\le\varepsilon}\left|\frac{e(X_i)}{e(X_i,\pi)^2}\frac{\partial e(X_i,\pi)}{\partial \pi}\right|\right] < \infty$. $\|$

Now we state the analogue of Theorem I.1 but with the probability weights estimated in a first step.

Proposition I.1 (IPW estimator with estimated probability weights)
Assume Assumptions I.1–I.3 hold with $\alpha_+(0) + \alpha_-(0) > 0$. Let a_n be defined from

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{DY}{e(X)} - \theta_0 - A_0 h(D,X)\right|^2 \mathbb{1}_{|DY/e(X) - A_0 h(D,X)| \le a_n}\right] \to 1,$$

where $A_0 = \mathbb{E}\left[\frac{\mu_1(e(X))}{e(X)}\frac{\partial e(X,\pi)}{\partial \pi}\bigg|_{\pi=\pi_0}\right]$. Then the IPW estimator has the following linear rep-

resentation:

$$\frac{n}{a_n}\left(\hat{\theta}_n - \theta_0\right) = \frac{1}{a_n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i)} - \theta_0 - A_0 h(D_i, X_i)\right) + o_{\mathrm{p}}(1),$$

and the conclusions of Theorem I.1 hold with estimated probability weights. ‖

To understand Proposition I.1, we again consider two cases. In the first case, the ratio has a finite variance: $\mathbb{V}[DY/e(X)] < \infty$, and estimating the probability weights in a first step will contribute to the asymptotic variance. The second case corresponds to $\mathbb{V}[DY/e(X)] = \infty$, implying that the final estimator, $\hat{\theta}_n$, has a slower convergence rate compared to the first-step estimated probability weights. As a result, the two definitions of the scaling factor a_n (in Theorem I.1 and in the above proposition) are asymptotically equivalent, and the limiting distribution will be the same regardless of whether the probability weights are known or estimated.

Now we consider generalized linear models (GLMs) for the probability weights, and show that Assumption I.3 holds under very mild primitive conditions.

Lemma I.2 (Primitive conditions for GLMs)
Assume Assumptions I.1 holds with $e(X, \pi_0) = \mathfrak{L}(X^{\mathrm{T}}\pi_0)$. Further,
(i) π_0 is the unique minimizer of $\mathbb{E}[|D - \mathfrak{L}(X^{\mathrm{T}}\pi)|^2]$ in the interior of the compact parameter space Π, and $\hat{\pi}_n = \mathrm{argmin}_{\pi\in\Pi}\sum_{i=1}^{n}|D_i - \mathfrak{L}(X_i^{\mathrm{T}}\pi)|^2$.
(ii) For some $\varepsilon > 0$, $\mathbb{E}\left[\sup_{\pi:|\pi-\pi_0|\leq\varepsilon}\left|\frac{\mathfrak{L}(X_i^{\mathrm{T}}\pi_0)\mathfrak{L}^{(1)}(X_i^{\mathrm{T}}\pi)}{\mathfrak{L}(X_i^{\mathrm{T}}\pi)^2}X\right|\right] < \infty$.
(iii) $\mathbb{E}[\mathfrak{L}^{(1)}(X^{\mathrm{T}}\pi_0)^2XX^{\mathrm{T}}]$ is nonsingular.
Then Assumption I.3 holds with $h(D_i, X_i) = \left(\mathbb{E}\left[\mathfrak{L}^{(1)}(X^{\mathrm{T}}\pi_0)^2XX^{\mathrm{T}}\right]\right)^{-1}(D_i - \mathfrak{L}(X_i^{\mathrm{T}}\pi_0))$ $\mathfrak{L}^{(1)}(X_i^{\mathrm{T}}\pi_0)X_i$. ‖

This lemma provides sufficient conditions to verify Assumption I.3 when the probability weight takes a generalized linear form, hence also justifies the result in Proposition I.1. Most of the conditions in Lemma I.2 are standard, except for part (ii). In the following remark we discuss in detail how this condition can be justified in Logit and Probit models.

Remark I.5 (Logit and Probit models) Assuming a Logit model for the probability weights: $e(X_i, \pi) = e^{X_i^{\mathrm{T}}\pi}/(1+e^{X_i^{\mathrm{T}}\pi})$, a sufficient condition for Lemma I.2(ii) is the covariates having a sub-exponential tail: $\mathbb{E}[e^{\varepsilon|X|}] < \infty$ for some (small) $\varepsilon > 0$. This should be compared to Remark I.4, where we show that for Assumption I.1 to hold in a Logit model, the index $X^{\mathrm{T}}\pi_0$ needs to have a sub-exponential left tail. Therefore, this sufficient condition is fully compatible with, and in a sense is "implied" by Assumption I.1.

As for the Probit model, condition (ii) in Lemma I.2 is implied by a sub-Gaussian tail of the covariates: $\mathbb{E}[e^{\varepsilon|X|^2}] < \infty$ for some (small) $\varepsilon > 0$. Again, it is possible to show that

Assumption I.1 implies a sub-Gaussian left tail for the index $X^{\mathrm{T}}\pi_0$. Thus, the requirement $\mathbb{E}[e^{\varepsilon|X|^2}] < \infty$ is fairly weak and does not contradict Assumption I.1. $\|$

I.2.4 Robust Inference

The limiting distribution of the IPW estimator can be quite complicated, and depends on multiple nuisance parameters which are usually difficult to estimate. In addition, the usual nonparametric bootstrap fails to provide a valid distributional approximation when $\gamma_0 < 2$ (Athreya, 1987; Knight, 1989). As a result, conducting statistical inference is particularly challenging. Subsampling is a powerful data-driven method to approximate the (limiting) distribution of a statistic. It draws samples of size $m \ll n$ and recomputes the statistic with each subsample. Therefore, subsampling provides distributional approximation as if many independent sets of random samples were available. Following is the detailed algorithm.

Algorithm I.1 (Robust inference using the IPW estimator)
Let $\hat{\theta}_n$ be defined as in (I.1), and

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\hat{e}(X_i)} - \hat{\theta}_n\right)^2}.$$

Step 1. Sample $m \ll n$ observations from the original data without replacement, denoted by $(Y_i^\star, D_i^\star, X_i^\star)$, $i = 1, 2, \cdots, m$.
Step 2. Construct the IPW estimator with the new subsample, and the self-normalized statistic as

$$T_m^\star = \frac{\hat{\theta}_m^\star - \hat{\theta}_n}{S_m^\star/\sqrt{m}}, \qquad S_m^\star = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{D_i^\star Y_i^\star}{\hat{e}^\star(X_i^\star)} - \hat{\theta}_m^\star\right)^2}.$$

Step 3. Repeat Step 1 and 2, and a $(1-\alpha)\%$-confidence interval can be constructed as

$$\left[\hat{\theta}_n - q_{1-\frac{\alpha}{2}}(T_m^\star)\frac{S_n}{\sqrt{n}} \quad , \quad \hat{\theta}_n - q_{\frac{\alpha}{2}}(T_m^\star)\frac{S_n}{\sqrt{n}}\right],$$

where $q_{(\cdot)}(T_m^\star)$ denotes the quantile of the statistic $T_m^\star$. $\|$

Subsampling validity typically relies on the existence of a limiting distribution (Politis and Romano, 1994; Romano and Wolf, 1999). We follow this approach, and justify our robust inference procedure by showing that the self-normalized statistic, $T_n = \sqrt{n}(\hat{\theta}_n - \theta_0)/S_n$, converges in distribution. Under $\gamma_0 > 2$, the term S_n in Algorithm I.1 converges in probability, and T_n converges to a Gaussian distribution by the Slutsky theorem. Asymptotic Gaussianity of T_n continues to hold for $\gamma_0 = 2$. Under $\gamma_0 < 2$, T_n still converges in distribution, although the limit is neither Gaussian nor Lévy stable. We characterize this limiting distribution in the proof of the following theorem.

Theorem I.2 (Validity of robust inference)
Under the assumptions of Theorem I.1 (or Proposition I.1 with estimated probability weights), and assume $m \to \infty$ and $m/n \to 0$. Then

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}[T_n \le t] - \mathbb{P}^\star[T^\star_m \le t] \right| \overset{\mathrm{p}}{\to} 0. \qquad \|$$

Before closing this section, we address several practical issues when applying the robust inference procedure. First, it is desirable to have an automatic and adaptive procedure to capture the possibly unknown convergence rate n/a_n, as the convergence rate depends on the tail index γ_0. In the subsampling algorithm, this is achieved by self-normalization (Studentization).

Second, one has to choose the subsample size m. Some suggestions have been made in the literature: Arcones and Giné (1991) suggest to use $m = \lfloor n/\log\log(n)^{1+\varepsilon}\rfloor$ for some $\varepsilon > 0$, although they consider the m-out-of-n bootstrap. Romano and Wolf (1999) propose a calibration technique. We use $m = \lfloor n/\log(n)\rfloor$ which performs quite well in our simulation study. Other choices such as $m = \lfloor n^{2/3}\rfloor$ and $\lfloor n^{1/2}\rfloor$ yield similar performance.

Finally, the denominator for self-normalization does not include all terms in the asymptotic linear representation stated in Proposition I.1. For example, with the probability weights estimated in a first step, an alternative is to use

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\hat{e}(X_i)} - \hat{\theta}_n - \hat{A}_n\hat{h}(D_i, X_i)\right)^2},$$

where $\hat{A}_n$ and $\hat{h}(\cdot)$ are plug-in estimates of A_0 and $h(\cdot)$. This alternative Studentization can be appealing for higher-order accuracy concerns (i.e., asymptotic refinements, Horowitz 2001; Politis, Romano and Wolf 1999). On the other hand, Algorithm I.1 is easier to implement since no additional estimation is needed.

I.3 Trimming

In response to small probability weights entering the IPW estimator, trimming is routinely employed as a regularization strategy. In this section, we first study the large-sample properties of the trimmed IPW estimator. It is shown that different limiting distributions can emerge, depending on how the trimming threshold is specified. Next, we study in detail the trimming bias, and show that for inference purpose it is typically nonnegligible or even explosive. These two findings explain why the point estimate is sensitive to the choice of the trimming threshold, and more importantly, why inference procedures based on the standard Gaussian approximation perform poorly. One extreme example is fixed trimming $b_n = b > 0$, with which the trimmed IPW estimator is $\sqrt{n}$-consistent and asymptotically Gaussian. However, it induces a bias that does not vanish even in large samples, forcing the researcher to change the target estimand and to re-interpret standard confidence intervals such as "point estimate ± 1.96×standard error."

As a remedy, we propose to combine resampling with a novel bias correction technique, where the latter employs local polynomial regression to approximate the trimming bias. Our inference procedure is robust not only to small probability weights but also to a wide range of trimming threshold choices. We also introduce a method to choose the trimming threshold by minimizing an empirical mean squared error, and discuss how our trimming threshold selector can be modified in a disciplined way if the researcher prefers to discard more observations.

I.3.1 Large Sample Properties of the trimmed IPW Estimator

If the untrimmed IPW estimator is already asymptotically Gaussian ($\gamma_0 \geq 2$, Theorem I.1(i)), so is the trimmed estimator. Therefore we restrict our attention to the $\gamma_0 < 2$ case. Also to make the result concise, we assume the probability weights are known, and postpone to the next subsection the impact of estimating the probability weights in a first step. Following is the main theorem characterizing the large-sample properties of the trimmed IPW estimator.

Theorem I.3 (Large sample properties of the trimmed IPW estimator)
Assume Assumptions I.1 and I.2 hold with $\gamma_0 < 2$ and $\alpha_+(0) + \alpha_-(0) > 0$. Further, let a_n be defined as in Theorem I.1.
(i) Light trimming: For $b_n a_n \to 0$, (I.4) *holds with $a_{n,b_n} = a_n$, and the limiting distribution is the Lévy stable distribution in Theorem I.1(ii).*
(ii) Heavy trimming: For $b_n a_n \to \infty$, (I.4) *holds with $a_{n,b_n} = \sqrt{n\mathbb{V}[DY/e(X)\mathbb{1}_{e(X)\geq b_n}]}$, and the limiting distribution is the standard Gaussian distribution.*
(iii) Moderate trimming: For $b_n a_n \to t \in (0,\infty)$, (I.4) *holds with $a_{n,b_n} = a_n$, and the limiting*

distribution is infinitely divisible with characteristic function:

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x)\right\},$$
$$\text{where } M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(tx)\mathbb{1}_{x\geq 0} + \alpha_-(tx)\mathbb{1}_{x<0}\Big)\right]. \qquad \|$$

For light trimming in part (i), b_n shrinks to zero fast enough so that asymptotically trimming becomes negligible, and the limiting distribution is Lévy stable as if there were no trimming. In part (ii), the trimming threshold shrinks to zero slowly, hence most of the small probability weights are excluded. This heavy trimming scenario leads to a Gaussian limiting distribution. Part (iii) lies between the two extremes. We refer to it as moderate trimming. On the one hand, a nontrivial number of small probability weights are discarded, making the limit no longer the Lévy stable distribution. On the other hand, the trimming is not heavy enough to restore asymptotic Gaussianity. The limiting distribution in this case is quite complicated, and depends on two (infinitely dimensional) nuisance parameters, $\alpha_+(\cdot)$ and $\alpha_-(\cdot)$. For this reason, inference is extremely challenging. As a technical remark, this limiting distribution is continuous in t, in the sense that as $t \to \infty$, it reduces to the standard Gaussian distribution; and as $t \downarrow 0$, it becomes the Lévy stable distribution.

Despite the limiting distribution taking on a complicated form, the trimming threshold choice in Theorem I.3(iii) is highly relevant, as it balances the bias and variance and leads to a mean squared error improvement over the untrimmed IPW estimator. In addition, unless one employs a very large trimming threshold, it is unclear how well the Gaussian approximation performs in samples of moderate size.

I.3.2 Estimated Probability Weights

Estimating the probability weights in a first step can affect the large-sample properties of the trimmed IPW estimator through two channels: the estimated weights enter the final estimator both through inverse weighting and through the trimming function. More precisely, we have the following expansion:

$$\frac{n}{a_{n,b_n}}\Big(\hat{\theta}_{n,b_n} - \theta_0 - \mathsf{B}_{n,b_n}\Big) = \underbrace{\frac{1}{a_n}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i)}\mathbb{1}_{e(X_i)\geq b_n} - \theta_0 - \mathsf{B}_{n,b_n}\right)}_{\text{Theorem I.3}}$$

$$+\underbrace{\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\frac{D_iY_i}{e(X_i)}\left(\frac{e(X_i)}{\hat{e}(X_i)}-1\right)\mathbb{1}_{\hat{e}(X_i)\geq b_n}}_{\text{Proposition I.1}}+\underbrace{\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\frac{D_iY_i}{e(X_i)}\left(\mathbb{1}_{\hat{e}(X_i)\geq b_n}-\mathbb{1}_{e(X_i)\geq b_n}\right)}_{\text{Proposition I.2}}.$$

Proposition I.2 shows that, despite the estimated probability weights entering both the denominator and the trimming function, the second channel is asymptotically negligible under an additional assumption, which turns out to be very mild in applications.

Assumption I.4 (Trimming threshold)
The trimming threshold satisfies $c_n\sqrt{b_n\mathbb{P}[e(X_i)\leq b_n]}\to 0$, where c_n is a positive sequence such that, for any $\varepsilon>0$,

$$c_n^{-1}\max_{1\leq i\leq n}\sup_{|\pi-\pi_0|\leq\varepsilon/\sqrt{n}}\left|\frac{1}{e(X_i)}\frac{\partial e(X_i,\pi)}{\partial\pi}\right|=o_{\mathrm{p}}(1).$$ ‖

Remark I.6 (Logit and Probit models) To verify Assumption I.4, it suffices to set $c_n=\log^2(n)$ for Logit and Probit models. Therefore, we only require the trimming threshold shrinking to zero faster than a logarithmic rate. ‖

Proposition I.2 (Trimmed IPW estimator with estimated probability weights)
Assume Assumptions I.1–I.4 hold with $\gamma_0<2$ and $\alpha_+(0)+\alpha_-(0)>0$, and let a_n be defined as in Proposition I.1. Then the conclusions of Theorem I.3 hold with estimated probability weights. ‖

From this proposition, estimating the probability weights in a first step does not lead to any first order impact beyond what has been stated in Proposition I.1. Equivalently, one can always assume that the true probability weights are used for trimming.

I.3.3 Balancing Bias and Variance

If the sole purpose of trimming is to stabilize the IPW estimator, one can argue that only a fixed trimming rule, $b_n=b\in(0,1)$, should be used. Such practice, however, completely ignores the bias introduced by trimming, forcing the researcher to change the target estimand and re-interpret the estimation/inference result (see, for example Crump, Hotz, Imbens and Mitnik 2009). Practically, the trimming threshold can be chosen by minimizing the asymptotic mean squared error. For this purpose, we characterize the bias and variance of the trimmed IPW estimator in the following lemma.

Lemma I.3 (Bias and variance of $\hat{\theta}_{n,b_n}$)
Assume Assumptions I.1 and I.2 hold with $\gamma_0 < 2$. Further, assume that $\mu_1(\cdot)$ and $\mu_2(\cdot)$ do not vanish near 0. Then the bias and variance of $\hat{\theta}_{n,b_n}$ are:

$$\begin{aligned}
\mathsf{B}_{n,b_n} &= -\mathbb{E}[\mu_1(e(X))\mathbb{1}_{e(X)\le b_n}] = -\mu_1(0)\mathbb{P}\left[e(X)\le b_n\right](1+o(1)),\\
\mathsf{V}_{n,b_n} &= \frac{1}{n}\mathbb{E}\left[\frac{\mu_2(e(X))}{e(X)}\mathbb{1}_{e(X)\ge b_n}\right](1+o(1)) = \mu_2(0)\frac{1}{n}\mathbb{E}\left[e(X)^{-1}\mathbb{1}_{e(X)\ge b_n}\right](1+o(1)).
\end{aligned}$$

In addition, $\mathsf{B}_{n,b_n}^2/\mathsf{V}_{n,b_n} \asymp nb_n\mathbb{P}[e(X)\le b_n]$. ‖

A natural question is how b_n can be chosen in practice. One possibility is to consider the leading mean squared error:

$$\begin{aligned}
\mathsf{B}_{n,b_n}^2 + \mathsf{V}_{n,b_n} &\approx \left[\mathbb{P}\left[e(X)\le b_n\right]\cdot\mu_1(0)\right]^2 + \frac{1}{n}\mathbb{E}\left[e(X)^{-1}\mathbb{1}_{e(X)\ge b_n}\right]\cdot\mu_2(0)\\
&= \left[\int_0^{b_n} \mathrm{d}\mathbb{P}[e(X)\le x]\cdot\mu_1(0)\right]^2 + \frac{1}{n}\int_{b_n}^1 x^{-1}\mathrm{d}\mathbb{P}[e(X)\le x]\cdot\mu_2(0),
\end{aligned}$$

and by taking derivative with respect to b_n, we have,

$$b_n^\dagger\cdot\mathbb{P}[e(X)\le b_n^\dagger] = \frac{1}{2n}\frac{\mu_2(0)}{\mu_1(0)^2}, \tag{I.5}$$

which gives the optimal trimming threshold.

The (mean squared error) optimal trimming $b_n^\dagger$ helps understand the three scenarios in Theorem I.3: light, moderate and heavy trimming. More importantly, it helps clarify whether (and when) the trimming bias features in the limiting distribution. (The trimming bias B_{n,b_n} vanishes as long as $b_n \to 0$. Scaled by the convergence rate, however, it may not be negligible even in large samples.) $b_n^\dagger$ corresponds to the moderate trimming scenario, and since it balances the leading bias and variance, the limiting distribution of the trimmed IPW estimator is not centered at the target estimand (i.e., it is asymptotically biased). A trimming threshold that shrinks more slowly than the optimal one corresponds to the heavy trimming scenario, where the bias dominates in the asymptotic distribution. The only scenario in which one can ignore the trimming bias for inference purposes is when light trimming is used. That is, the trimming threshold shrinks faster than $b_n^\dagger$. In large samples, however, no observation will be discarded. Overall, the trimming bias cannot be ignored if one wants to develop an inference procedure that is valid for the target estimand using the trimmed IPW estimator. In the next subsection, we propose an inference procedure that is valid for the target estimand under a range of trimming threshold choices. This is achieved by explicitly estimating and correcting the trimming bias with a novel application of local polynomial regression.

The following theorem shows that, under very mild regularity conditions, the optimal trimming threshold can be implemented in practice by solving the sample analogue of (I.5). In addition, it also provides a disciplined method for choosing the trimming threshold if the researcher prefers to employ a heavy trimming.

Theorem I.4 (Optimal trimming: implementation)
Assume Assumption I.1 holds, and $0 < \mu_2(0)/\mu_1(0)^2 < \infty$. For any $s > 0$, define b_n and $\hat{b}_n$ as:

$$b_n^s \mathbb{P}[e(X) \le b_n] = \frac{1}{2n}\frac{\mu_2(0)}{\mu_1(0)^2}, \qquad \hat{b}_n^s \left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{e(X)\le \hat{b}_n}\right) = \frac{1}{2n}\frac{\hat{\mu}_2(0)}{\hat{\mu}_1(0)^2},$$

where $\hat{\mu}_1(0)$ and $\hat{\mu}_2(0)$ are some consistent estimates of $\mu_1(0)$ and $\mu_2(0)$, respectively. Then $\hat{b}_n$ is consistent for b_n, in the sense that:

$$\frac{\hat{b}_n}{b_n} \xrightarrow{\mathrm{p}} 1.$$

Therefore, for $0 < s < 1$, $s = 1$ and $s > 1$, we have that $\hat{b}_n/b_n^\dagger$ converges in probability to 0, 1 and ∞, respectively.

If in addition Assumption I.3 holds, and for any $\varepsilon > 0$,

$$\max_{1\le i\le n} \sup_{|\pi-\pi_0|\le \varepsilon/\sqrt{n}} \left| \frac{1}{e(X_i)} \frac{\partial e(X_i,\pi)}{\partial \pi} \right| = o_{\mathrm{p}}\left(\sqrt{\frac{n}{\log(n)}}\right),$$

then $\hat{b}_n$ can be constructed with estimated probability weights. ‖

This theorem states that, as long as we can construct a consistent estimator for the ratio $\mu_2(0)/\mu_1(0)^2$, the optimal trimming threshold can be implemented in practice with the unknown distribution $\mathbb{P}[e(X) \le x]$ replaced by the standard empirical estimate. Although (I.5) and its sample analogue do not have closed-form solutions, finding $\hat{b}_n$ is quite easy, by first searching over the order statistics of the probability weights, and then performing a grid search in a interval with length of order n^{-1}.

In addition, Theorem I.4 allows the use of estimated probability weights for constructing $\hat{b}_n$. The extra condition turns out to be quite weak, and is easily satisfied if the probability weights are estimated in a Logit or Probit model.

Remark I.7 (Bias-variance trade-off when $\gamma_0 \ge 2$) The characterization of leading variance in Lemma I.3 only applies to $\gamma_0 < 2$. The trimming threshold in (I.5), however, remains to be mean squared error optimal even for $\gamma_0 \geq 2$. To show this, we need to characterize a higher order variance term. Assume for simplicity that $\gamma_0 > 2$, then the variance of the trimmed IPW estimator is

$$\begin{aligned}\frac{1}{n}\mathbb{V}\left[\frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n}\right] &= \frac{1}{n}\mathbb{E}\left[\frac{DY^2}{e(X)^2}\right] - \frac{1}{n}\mathbb{E}\left[\frac{DY^2}{e(X)^2}\mathbb{1}_{e(X)\leq b_n}\right] - \frac{1}{n}\left(\theta_0 + \mathsf{B}_{n,b_n}\right)^2 \\ &= \frac{1}{n}\mathbb{V}\left[\frac{DY}{e(X)}\right] - \frac{1}{n}\mathbb{E}\left[\frac{DY^2}{e(X)^2}\mathbb{1}_{e(X)\leq b_n}\right](1+o(1)),\end{aligned}$$

provided that $\mu_2(0) > 0$. In this case, the (asymptotic) mean squared error optimal trimming threshold is defined as the minimizer of:

$$\left[\int_0^{b_n} \mathrm{d}\mathbb{P}[e(X) \leq x]\cdot\mu_1(0)\right]^2 - \frac{1}{n}\int_0^{b_n} x^{-1}\mathrm{d}\mathbb{P}[e(X)\leq x]\cdot\mu_2(0),$$

which can be found by solving a first order condition and coincides with (I.5). The $\gamma_0 = 2$ case can be analyzed similarly, although one has to take extra care on a slowly varying term in the variance expansion. Finally, we note that Theorem I.4 remains valid and can be employed to estimate this optimal trimming threshold for $\gamma_0 \geq 2$. ‖

I.3.4 Bias Correction and Robust Inference

To motivate our bias correction technique, recall that the bias is $\mathsf{B}_{n,b_n} = -\mathbb{E}[\mu_1(e(X))\mathbb{1}_{e(X)\leq b_n}]$, where $\mu_1(\cdot)$ is the expectation of the outcome Y conditional on the probability weight and $D = 1$. Next, we replace the expectation by a sample average, and the unknown conditional expectation by a p-th order polynomial expansion, and the bias is approximated by

$$-\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=0}^{p}\frac{1}{j!}\mu_1^{(j)}(0)e(X_i)^j\right)\mathbb{1}_{e(X_i)\leq b_n}.$$

Here, $\mu_1^{(j)}(0)$ is the j-th derivative of $\mu_1(\cdot)$ evaluated at 0, and has to be estimated. Given that we do not impose parametric assumptions on the conditional expectation beyond certain degree of smoothness, we employ local polynomial regression (Fan and Gijbels, 1996).

Our procedure takes two steps. In the first step, one implements a p-th order local polynomial regression of the outcome variable on the probability weight using the $D = 1$ subsample in a region $[0, h_n]$, where $(h_n)_{n\geq 1}$ is a bandwidth sequence. In the second step, the estimated bias is constructed by replacing the unknown conditional expectation function and its derivatives by the first-step estimates. Following is the detailed algorithm, which is illustrated in Figure I.2.

Algorithm I.2 (Bias estimation)

Step 1. With the $D = 1$ subsample, regress the outcome variable Y_i on the (estimated) probability weight in a region $[0, h_n]$:

$$\Big[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\Big]' = \operatorname*{argmin}_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n} D_i \Big[Y_i - \sum_{j=0}^{p} \beta_j \hat{e}(X_i)^j\Big]^2 \mathbb{1}_{\hat{e}(X_i)\le h_n}.$$

Step 2. Construct the bias correction term as

$$\hat{\mathsf{B}}_{n,b_n} = -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=0}^{p}\hat{\beta}_j \hat{e}(X_i)^j\right)\mathbb{1}_{\hat{e}(X_i)\le b_n},$$

so that the bias-corrected estimator is $\hat{\theta}^{\mathsf{bc}}_{n,b_n} = \hat{\theta}_{n,b_n} - \hat{\mathsf{B}}_{n,b_n}$. ‖

By inspecting the bias-corrected estimator, our procedure can be understood as a "local regression adjustment," since we replace the trimmed observations by its conditional expectation, which is further approximated by a local polynomial. In the local polynomial regression step, it is possible to incorporate other kernel functions: we use the uniform kernel $\mathbb{1}_{\hat{e}(X_i)\le h_n}$ to avoid introducing additional notation, but all the main conclusions continue to hold with other commonly employed kernel functions, such as the triangular and Epanechnikov kernels. As for the order of local polynomial regression, common choices are $p = 1$ and 2, which reduce the bias to a satisfactory level without introducing too much additional variation.

Standard results form the local polynomial regression literature require the density of the design variable to be bounded away from zero, which is not satisfied in our context. When the probability weight is close to zero, it becomes very difficult to observe $D = 1$. Equivalently, in the subsample which we use for the local polynomial regression, the distribution of the probability weights quickly vanishes near the origin.[2] As a result, nonstandard scaling is needed to derive large-sample properties of $\hat{\mu}_1^{(j)}(0)$.

The following theorem shows the validity of our bias correction procedure.

Theorem I.5 (Large sample properties of the estimated bias)

Assume Assumptions I.1 and I.2 (and in addition Assumption I.3 and I.4 with estimated probability weights) hold. Further, assume (i) $\mu_1(\cdot)$ is $p+1$ times continuously differentiable; (ii) $\mu_2(0) - \mu_1(0)^2 > 0$; (iii) the bandwidth sequence satisfies $nh_n^{2p+3}\mathbb{P}[e(X) \le h_n] \asymp 1$; (iv) $nb_n^{2p+3}\mathbb{P}[e(X) \le b_n] \to 0$. Then the bias correction is valid, and does not affect the asymptotic

[2] More precisely, $\mathbb{P}[e(X) \le x|D = 1] \prec x$ as $x \downarrow 0$, meaning that in the $D = 1$ subsample, the density of the probability weights (if it exists) tends to zero: $f_{e(X)|D=1}(0) = 0$.

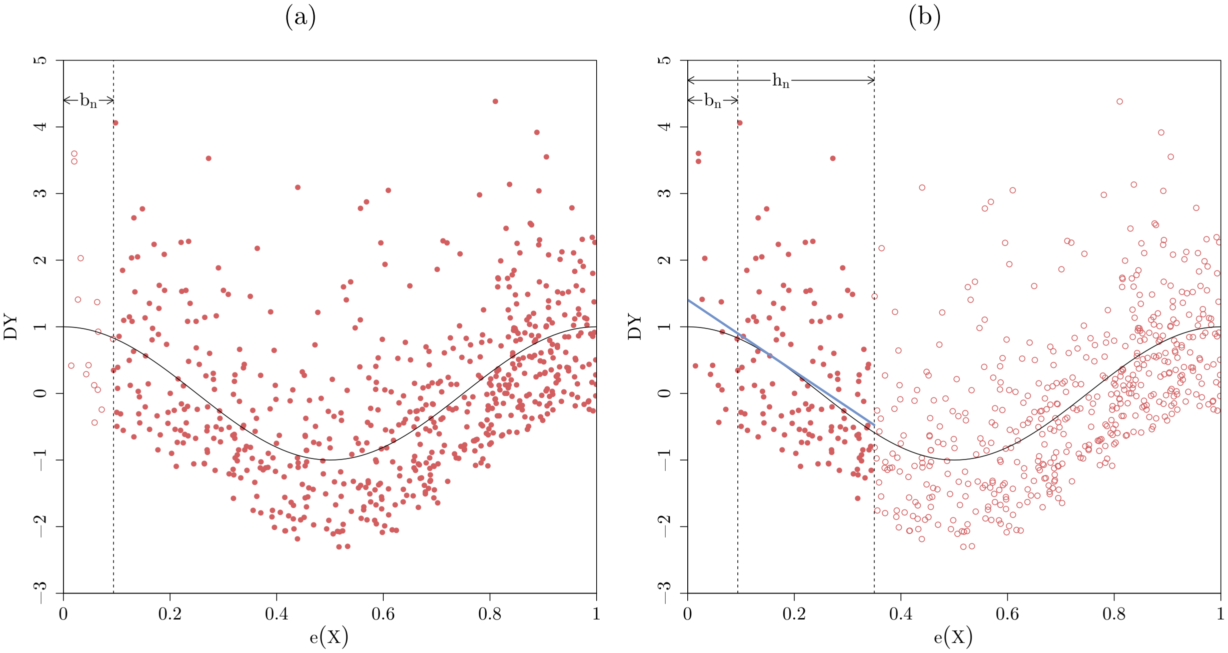
distribution:

$$\hat{\theta}^{\mathsf{bc}}_{n,b_n} - \theta_0 = \Big(\hat{\theta}_{n,b_n} - \mathsf{B}_{n,b_n} - \theta_0\Big)(1 + o_{\mathrm{p}}(1)). \qquad \|$$

Theorem I.5 has several important implications. First, our bias correction is valid for a wide range of trimming threshold choices, as long as the trimming threshold does not shrink to zero too slowly: $nb_n^{2p+3}\mathbb{P}[e(X) \le b_n] \to 0$. However, fixed trimming $b_n = b \in (0,1)$ is ruled out (except for the trivial case where the probability weight is already bounded away from zero). This is not surprising, since under fixed trimming the correct scaling is $\sqrt{n}$, and generally the bias cannot be estimated at this rate without additional parametric assumptions.

Second, it gives a guidance on how the bandwidth for the local polynomial regression can be chosen. In practice, this is done by solving $n\hat{h}_n^{2p+3}\hat{\mathbb{P}}[e(X) \le \hat{h}_n] = c$ for some $c > 0$, so that the resulting bandwidth makes the (squared) bias and variance of the local polynomial regression the same order. A simple strategy is to set $c = 1$. It is also possible to construct a

Figure I.2. Trimming and local polynomial bias correction.



Note. (a) Illustration of Trimming. Circles: trimmed observations. Solid dots: observations included in the estimator. Solid curve: conditional expectation function $\mathbb{E}[Y|e(X), D = 1]$. (b) Illustration of the local polynomial regression. Solid dots: observations used in the local polynomial regression. Solid straight line: local linear regression function.

bandwidth that minimizes the leading mean squared error of the local polynomial regression, for which c has to be estimated in a pilot step.

Finally, it shows how trimming and bias correction together can help improve the convergence rate of the (untrimmed) IPW estimator. From Theorem I.3(ii), we have $|\hat{\theta}_{n,b_n} - \theta_0 - \mathsf{B}_{n,b_n}| = O_{\mathrm{p}}((n/a_{n,b_n})^{-1})$, where the convergence rate $n/a_{n,b_n}$ is typically faster when a heavier trimming is employed. This, however, should not be interpreted as a real improvement, as the trimming bias can be so large that the researcher effectively changes the target estimand to $\theta_0 - \mathsf{B}_{n,b_n}$. with bias correction, it is possible to achieve a faster rate of convergence for the target estimand, since under the assumptions of Theorem I.5, one has $|\hat{\theta}^{\mathsf{bc}}_{n,b_n} - \theta_0| = O_{\mathrm{p}}((n/a_{n,b_n})^{-1})$, which is valid for a wide rage of trimming threshold choices.

Together with our bias correction technique, subsampling can be employed to conduct statistical inference and to construct confidence intervals that are valid for the target estimand. Although Theorem I.5 states that estimating the bias does not have a first order contribution to the limiting distribution, it may still introduce additional variability in finite samples (Calonico, Cattaneo and Farrell, 2018). Therefore, we recommend subsampling the bias-corrected statistic.

Algorithm I.3 (Robust inference using the trimmed IPW estimator)
Let $\hat{\theta}^{\mathsf{bc}}_{n,b_n}$ be defined as in Algorithm I.2, and

$$S_{n,b_n} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{D_iY_i}{\hat{e}(X_i)}\mathbb{1}_{\hat{e}(X_i)\geq b_n} - \hat{\theta}_{n,b_n}\right)^2}.$$

Step 1. Sample $m \ll n$ observations from the original data without replacement, denoted by $(Y_i^\star, D_i^\star, X_i^\star)$, $i = 1, 2, \cdots, m$.

Step 2. Construct the trimmed IPW estimator and the bias correction term from the new subsample, and the bias-corrected and self-normalized statistic as

$$T^\star_{m,b_m} = \frac{\hat{\theta}^{\star\mathsf{bc}}_{m,b_m} - \hat{\theta}^{\mathsf{bc}}_{n,b_n}}{S^\star_{m,b_m}/\sqrt{m}}, \qquad S^\star_{m,b_m} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{D_i^\star Y_i^\star}{\hat{e}^\star(X_i^\star)}\mathbb{1}_{\hat{e}^\star(X_i^\star)\geq b_m} - \hat{\theta}^\star_{m,b_m}\right)^2}.$$

Step 3. Repeat Step 1 and 2, and a $(1-\alpha)\%$-confidence interval can be constructed as

$$\left[\hat{\theta}^{\mathsf{bc}}_{n,b_n} - q_{1-\frac{\alpha}{2}}(T^\star_{m,b_m})\frac{S_{n,b_n}}{\sqrt{n}} \quad , \quad \hat{\theta}^{\mathsf{bc}}_{n,b_n} - q_{\frac{\alpha}{2}}(T^\star_{m,b_m})\frac{S_{n,b_n}}{\sqrt{n}}\right],$$

where $q_{(\cdot)}(T^\star_{m,b_m})$ denotes the quantile of the statistic $T^\star_{m,b_m}$. ‖

Same as Theorem I.2, the validity of our inference procedure relies on establishing a limiting distribution for the self-normalized statistic, $T_{n,b_n} = \sqrt{n}(\hat{\theta}^{\mathsf{bc}}_{n,b_n} - \theta_0)/S_{n,b_n}$. This is relatively easy if $\gamma_0 \geq 2$ or a heavy trimming is employed, in which case T_{n,b_n} is asymptotically Gaussian. With light or moderate trimming under $\gamma_0 < 2$, the limiting distribution of T_{n,b_n} depends on the trimming threshold and is quite complicated. This technical by-product generalizes Logan, Mallows, Rice and Shepp (1973).

Theorem I.6 (Validity of robust inference)
Under the assumptions of Theorem I.1 (or Proposition I.2 with estimated probability weights) and Theorem I.5, and assume $m \to \infty$ and $m/n \to 0$. Then

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}[T_{n,b_n} \leq t] - \mathbb{P}^\star[T^\star_{m,b_m} \leq t] \right| \overset{\mathrm{p}}{\to} 0. \qquad \|$$

I.4 Extensions

In this section, we discuss two extensions of the current IPW framework. In the first extension, we consider treatment effect estimation under selection on observables. In the second extension, we consider a general estimating equation where the parameter is defined by a possibly nonlinear moment condition, not necessarily a population mean.

I.4.1 Treatment Effect Estimation

Given the prominent role of treatment effect estimands in program evaluation, we extend the IPW framework along this direction. Let the binary indicator denote a treatment status, $D = 1$ for the treatment group and 0 for the control group. The corresponding potential outcomes are denoted by $Y(1)$ and $Y(0)$, respectively. The observed outcome is $Y = DY(1) + (1 - D)Y(0)$. Throughout this subsection, we maintain the selection on observables assumption that, conditional on the covariates X, D and $(Y(1), Y(0))$ are independent. Following the convention in the literature, we use the terminology "propensity score" rather than probability weight. We ignore the issue of using estimated propensity scores for ease of exposition (see Section I.2.3 and I.3.2 for discussions).

Treatment Effect on the Treated (ATT)

We first consider the treatment effect on the treated estimand: $\tau_0^{\mathsf{ATT}} = \mathbb{E}[Y(1)-Y(0)|D = 1]$. Both Assumption I.1 and I.2 can be modified in a straightforward way.

Assumption I.5 (ATT)
(i) For some $\gamma_0 > 1$, the propensity score has a regularly varying tail with index $\gamma_0 - 1$ at

one:

$$\lim_{t\downarrow 0}\frac{\mathbb{P}[1-e(X)\le tx]}{\mathbb{P}[1-e(X)\le t]}=x^{\gamma_0-1},\qquad \textit{for all } x>0.$$

(ii) For some $\varepsilon>0$, $\mathbb{E}\big[|Y(0)+Y(1)|^{(\gamma_0\vee 2)+\varepsilon}\big|e(X)=x\big]$ is uniformly bounded. There exists a probability distribution $F_{(0)}$, such that for all bounded and continuous $\ell(\cdot)$, $\mathbb{E}[\ell(Y(0))|e(X)=x]\to\int_{\mathbb{R}}\ell(y)F_{(0)}(\mathrm{d}y)$ as $x\uparrow 1$. ‖

Assumption I.5(i) suffices for identification, as it implies $\mathbb{P}[e(X)=1]=0$. Using inverse probability weighting, a natural estimator of $\tau_0^{\texttt{ATT}}$ is

$$\hat{\tau}_n^{\texttt{ATT}}=\frac{1}{n_1}\sum_{i=1}^{n}\left[D_iY_i-\frac{e(X_i)}{1-e(X_i)}(1-D_i)Y_i\right]=\frac{1}{n}\sum_{i=1}^{n}\frac{(D_i-e(X_i))Y_i}{\hat{\mathbb{P}}[D=1](1-e(X_i))},$$

where $n_1=\sum_{i=1}^{n}D_i$ is size of the treated group, and $\hat{\mathbb{P}}[D=1]=n_1/n$. It should be clear that propensity scores that are close to 1 will pose a challenge to both estimation and inference. The following proposition characterizes the large sample properties of $\hat{\tau}_n^{\texttt{ATT}}$.

Proposition I.3 (Large sample properties of the ATT estimator)
Assume Assumption I.5 holds with $\alpha_{(0),+}(0)+\alpha_{(0),-}(0)>0$, where

$$\alpha_{(0),+}(x)=\lim_{t\to 1}\mathbb{E}\left[|Y(0)|^{\gamma_0}\mathbb{1}_{Y(0)>x}\middle|e(X)=t\right],\ \alpha_{(0),-}(x)=\lim_{t\to 1}\mathbb{E}\left[|Y(0)|^{\gamma_0}\mathbb{1}_{Y(0)<x}\middle|e(X)=t\right].$$

Let a_n be defined from

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{(D-e(X))Y}{\mathbb{P}[D=0](1-e(X))}-\tau_0^{\texttt{ATT}}\right|^2\mathbb{1}_{\left|\frac{(D-e(X))Y}{\mathbb{P}[D=0](1-e(X))}\right|\le a_n}\right]\to 1.$$

Then $\frac{n}{a_n}(\hat{\tau}_n^{\texttt{ATT}}-\tau_0^{\texttt{ATT}})$ converges in distribution, with the limit being:
(i) the standard Gaussian distribution if $\gamma_0\ge 2$; and
(ii) the Lévy stable distribution if $\gamma_0<2$, with characteristic function:

$$\psi(\zeta)=\exp\left\{\int_{\mathbb{R}}\frac{e^{i\zeta x}-1-i\zeta x}{x^2}M(\mathrm{d}x)\right\},$$
$$\textit{where } M(\mathrm{d}x)=\mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_{(0),+}(0)+\alpha_{(0),-}(0)}|x|^{1-\gamma_0}\Big(\alpha_{(0),+}(0)\mathbb{1}_{x<0}+\alpha_{(0),-}(0)\mathbb{1}_{x\ge 0}\Big)\right].\quad \|$$

Proposition I.3 and Theorem I.1 share common features. The limiting distribution can be Gaussian or non-Gaussian, depending on the tail behavior of the propensity score near

1. In the latter case, the limiting distribution is smooth, heavy-tailed but not necessarily symmetric (and usually does not have a closed-form distribution or density function).

We also consider the trimmed ATT estimator, which takes the following form

$$\begin{aligned}\hat{\tau}^{\mathtt{ATT}}_{n,b_n} &= \frac{1}{n_1}\sum_{i=1}^{n}\left[D_iY_i - \frac{e(X_i)}{1-e(X_i)}(1-D_i)Y_i\mathbb{1}_{1-e(X_i)\geq b_n}\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{(D_i-e(X_i))Y_i}{\hat{\mathbb{P}}[D=1](1-e(X_i))}\mathbb{1}_{1-e(X_i)\geq(1-D_i)b_n}.\end{aligned}$$

That is, observations from the control group with propensity scores above $1-b_n$ are discarded. It can be shown that the trimming bias is

$$\mathsf{B}_{n,b_n} = \frac{1}{\mathbb{P}[D=1]}\mathbb{E}\Big[e(X)\mathbb{E}[Y(0)|e(X)]\mathbb{1}_{e(X)\geq 1-b_n}\Big].$$

To implement bias correction, one first regresses the outcome variable on a p-th polynomial of the propensity score, using only observations from the control group:

$$\Big[\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p\Big]' = \operatorname*{argmin}_{\beta_0,\beta_1,\cdots,\beta_p}\sum_{i=1}^{n}(1-D_i)\Big[Y_i-\sum_{j=0}^{p}\beta_je(X_i)^j\Big]^2\mathbb{1}_{e(X_i)\geq 1-h_n}.$$

Then the bias is estimated by

$$\hat{\mathsf{B}}_{n,b_n} = \frac{1}{n_1}\sum_{i=1}^{n}\sum_{j=0}^{p}\hat{\beta}_je(X_i)^{j+1}\mathbb{1}_{e(X_i)\geq 1-b_n}.$$

Next we discuss the large sample properties of the trimmed ATT estimator, for which we focus on the $\gamma_0 < 2$ case.

Proposition I.4 (Large sample properties of the trimmed ATT estimator) *Assume Assumption I.5 holds with $\gamma_0 < 2$ and $\alpha_{(0),+}(0)+\alpha_{(0),-}(0) > 0$. Further, let a_n be defined as in Proposition I.3.*

(i) For $b_na_n \to 0$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATT}}_{n,b_n} - \tau^{\mathtt{ATT}}_0 - \mathsf{B}_{n,b_n})$ converges to the Lévy stable distribution in Proposition I.3(ii).

(ii) For $b_na_n \to \infty$, let $a_{n,b_n} = \sqrt{n\mathbb{V}[\frac{(D-e(X))Y}{\mathbb{P}[D=1](1-e(X))}\mathbb{1}_{1-e(X)\geq(1-D)b_n}]}$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATT}}_{n,b_n} - \tau^{\mathtt{ATT}}_0 - \mathsf{B}_{n,b_n})$ converges to the standard Gaussian distribution.

(iii) For $b_na_n \to t \in (0,\infty)$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATT}}_{n,b_n} - \tau^{\mathtt{ATT}}_0 - \mathsf{B}_{n,b_n})$ converges to an

infinitely divisible distribution with characteristic function:

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x)\right\},$$

$$\textit{where } M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_{(0),+}(0) + \alpha_{(0),-}(0)} |x|^{1-\gamma_0} \Big(\alpha_{(0),+}(-tx)\mathbb{1}_{x<0} + \alpha_{(0),-}(-tx)\mathbb{1}_{x\geq 0}\Big)\right]. \qquad \|$$

Average Treatment Effect (ATE)

The average treatment effect, $\tau_0^{\mathtt{ATE}} = \mathbb{E}[Y(1) - Y(0)]$, is another commonly employed treatment effect estimand. Because both small and large propensity scores can lead to "small denominators," Assumptions I.1 and I.2 have to be properly modified. To be specific, we require

Assumption I.1 (ATE)
(i) For some $\gamma_0 > 1$ and $\omega \in [0,1]$,

$$\lim_{t\downarrow 0} \frac{\mathbb{P}[e(X)\leq t]}{\mathbb{P}[e(X)\leq t] + \mathbb{P}[1-e(X)\leq t]} = \omega,$$

$$\textit{and} \quad \lim_{t\downarrow 0} \frac{\mathbb{P}[e(X)\leq tx] + \mathbb{P}[1-e(X)\leq tx]}{\mathbb{P}[e(X)\leq t] + \mathbb{P}[1-e(X)\leq t]} = x^{\gamma_0 - 1}, \qquad \textit{for all } x > 0.$$

(ii) For some $\varepsilon > 0$, $\mathbb{E}[|Y(1)+Y(0)|^{(\gamma_0\vee 2)+\varepsilon}|e(X)=x]$ is uniformly bounded. Further, there exist probability distributions, $F_{(1)}$ and $F_{(0)}$, such that for all bounded and continuous $\ell(\cdot)$, $\mathbb{E}[\ell(Y(1))|e(X)=x] \to \int \ell(y)F_{(1)}(\mathrm{d}y)$ and $\mathbb{E}[\ell(Y(0))|e(X)=1-x] \to \int \ell(y)F_{(0)}(\mathrm{d}y)$ as $x\downarrow 0$. ‖

Note that in part (i), we do not require the two tails of the propensity score having the same index, since it is possible to have $\omega = 0$ or 1. Asymptotically, the heavier tail "wins." Part (i) also implies $\mathbb{P}[e(X)=0] = \mathbb{P}[e(X)=1] = 0$, meaning that the ATE is identified. Part (ii) takes into account that both potential outcomes can affect the tail behavior of the estimator. The following is a natural estimator of ATE using inverse probability weighting:

$$\hat{\tau}_n^{\mathtt{ATE}} = \frac{1}{n}\sum_{i=1}^n \left[\frac{D_iY_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right] = \frac{1}{n}\sum_{i=1}^n \frac{(2D_i-1)Y_i}{1-D_i+(2D_i-1)e(X_i)}.$$

Assumption I.1 suffices to characterize the tail of $\frac{(2D-1)Y}{(1-D+(2D-1)e(X))}$. For future reference, let

$$\alpha_{(1),+}(x) = \lim_{t\to 0}\mathbb{E}\Big[|Y(1)|^{\gamma_0}\mathbb{1}_{Y(1)>x}\Big|e(X)=t\Big], \quad \alpha_{(1),-}(x) = \lim_{t\to 0}\mathbb{E}\Big[|Y(1)|^{\gamma_0}\mathbb{1}_{Y(1)<x}\Big|e(X)=t\Big],$$

and re-define $\alpha_+(x)$ and $\alpha_-(x)$ as

$$\alpha_+(x) = \omega\alpha_{(1),+}(x) + (1-\omega)\alpha_{(0),-}(-x), \quad \alpha_-(x) = \omega\alpha_{(1),-}(x) + (1-\omega)\alpha_{(0),+}(-x).$$

The following proposition summarizes the large sample properties of the ATE estimator.

Proposition I.5 (Large sample properties of the ATE estimator)
Assume Assumption I.1 holds with $\alpha_+(0) + \alpha_-(0) > 0$. Let a_n be defined from

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{(2D-1)Y}{1-D+(2D-1)e(X)} - \theta_0\right|^2 \mathbb{1}_{\left|\frac{(2D-1)Y}{1-D+(2D-1)e(X)}\right|\le a_n}\right] \to 1.$$

Then $\frac{n}{a_n}(\hat{\tau}_n^{\mathtt{ATE}} - \tau_0^{\mathtt{ATE}})$ converges in distribution, with the limit being:
(i) the standard Gaussian distribution if $\gamma_0 \ge 2$; and
(ii) the Lévy stable distribution if $\gamma_0 < 2$, with characteristic function:

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x)\right\},$$
$$\text{where } M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(0)\mathbb{1}_{x\ge 0} + \alpha_-(0)\mathbb{1}_{x<0}\Big)\right]. \qquad \|$$

For ATE estimation, trimming can lead to further complications beyond affecting the limiting distribution and introducing a bias: different trimming thresholds can be applied to the treatment and control groups. For the treatment group ($D = 1$), it is natural to discard observations with small propensity scores, while for the control group ($D = 0$) observations with large propensity scores will be dropped. To see how having two trimming thresholds can complicate the asymptotic analysis, assume $\omega = 1$ so that the propensity score has a heavier left tail, and Proposition I.5 essentially reduces to Theorem I.1. When different trimming thresholds are applied to small and large propensity scores in the treatment and control groups, however, the relative magnitude of the two tails can be overturned. To see this, consider the extreme scenario where fixed trimming is applied to the treatment group but no trimming (or light trimming) for the control group. Then the trimmed ATE estimator will be greatly influenced by the relatively heavier right tail of the propensity score (i.e., "small denominators" in the $D = 0$ subsample). To avoid cumbersome notation and lengthy discussions on each possible scenarios, we instead focus on a concrete trimming strategy, which illuminates how trimming affects the IPW-based ATE estimator, yet does not complicate the analysis too much. We consider the following trimmed ATE estimator:

$$\begin{aligned}\hat{\tau}^{\mathtt{ATE}}_{n,b_n} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_iY_i}{e(X_i)}\mathbb{1}_{e(X_i)\geq b_n} - \frac{(1-D_i)Y_i}{1-e(X_i)}\mathbb{1}_{e(X_i)\leq 1-b_n}\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{(2D_i-1)Y_i}{1-D_i+(2D_i-1)e(X_i)}\mathbb{1}_{1-D_i+(2D_i-1)e(X_i)\geq b_n}. \end{aligned} \tag{I.6}$$

The above trimming strategy can be understood as "discarding observations with small denominators." It is different, however, from "discarding observations with small or large propensity scores," since an observation in the control group is never trimmed because of a small propensity score, and vice versa, an observation in the treatment group is not trimmed even if it has a large propensity score.

The "symmetric trimming" in (I.6) is easy to analyze and implement, but employing different trimming thresholds is also justified in practice. As discussed, trimming introduces a bias which is generally non-negligible. For estimating the ATE, however, it is possible to achieve "small bias" by choosing the two trimming thresholds appropriately. To see this, the trimming bias in (I.6) is $\mathsf{B}_{n,b_n} = \mathbb{E}[\mathbb{E}[Y(0)|e(X)]\mathbb{1}_{e(X)\geq 1-b_n} - \mathbb{E}[Y(1)|e(X)]\mathbb{1}_{e(X)\leq b_n}] \approx \mathbb{E}[Y(0)|e(X)=1]\mathbb{P}[e(X)\geq 1-b_n] - \mathbb{E}[Y(1)|e(X)=0]\mathbb{P}[e(X)\leq b_n]$. Assuming that the propensity score has similar tails at the two ends and that the two conditional expectations have the same sign and magnitude, then it is possible to use different trimming thresholds so that the two components in the bias formula cancel each other. However, this strategy is not always feasible, especially when the two tails behave very differently.

Proposition I.6 (Large sample properties of the trimmed ATE estimator) *Assume Assumption I.1 holds with $\gamma_0 < 2$ and $\alpha_+(0)+\alpha_-(0) > 0$. Further, let a_n be defined as in Proposition I.5.*

(i) For $b_n a_n \to 0$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATE}}_{n,b_n} - \tau^{\mathtt{ATE}}_0 - \mathsf{B}_{n,b_n})$ converges to the Lévy stable distribution in Proposition I.5(ii).

(ii) For $b_n a_n \to \infty$, let $a_{n,b_n} = \sqrt{n\mathbb{V}[\frac{(2D-1)Y}{(1-D+(2D-1)e(X))}\mathbb{1}_{1-D+(2D-1)e(X)\geq b_n}]}$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATE}}_{n,b_n} - \tau^{\mathtt{ATE}}_0 - \mathsf{B}_{n,b_n})$ converges to the standard Gaussian distribution.

(iii) For $b_n a_n \to t \in (0,\infty)$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATE}}_{n,b_n} - \tau^{\mathtt{ATE}}_0 - \mathsf{B}_{n,b_n})$ converges to an infinitely divisible distribution with characteristic function:

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}}\frac{e^{i\zeta x}-1-i\zeta x}{x^2}M(\mathrm{d}x)\right\},$$

$$\textit{where } M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(tx)\mathbb{1}_{x\geq 0} + \alpha_-(tx)\mathbb{1}_{x<0}\Big)\right]. \qquad \|$$

Bias correction can be implemented according to Algorithm I.2 with a straightforward modification: one first runs two local polynomial regressions, one for the treatment group and the other for the control group:

$$\begin{aligned}\left[\hat{\beta}^{\mathtt{l}}_0, \hat{\beta}^{\mathtt{l}}_1, \cdots, \hat{\beta}^{\mathtt{l}}_p\right]' &= \operatorname*{argmin}_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n} D_i \left[Y_i - \sum_{j=0}^{p} \beta_j e(X_i)^j\right]^2 \mathbb{1}_{e(X_i)\le h_n}\\ \left[\hat{\beta}^{\mathtt{r}}_0, \hat{\beta}^{\mathtt{r}}_1, \cdots, \hat{\beta}^{\mathtt{r}}_p\right]' &= \operatorname*{argmin}_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^{n} (1-D_i) \left[Y_i - \sum_{j=0}^{p} \beta_j e(X_i)^j\right]^2 \mathbb{1}_{e(X_i)\ge 1-h_n}.\end{aligned}$$

Then the bias is estimated by

$$\hat{\mathsf{B}}_{n,b_n} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{p}\left(\hat{\beta}^{\mathtt{r}}_j \mathbb{1}_{e(X_i)\ge 1-b_n} - \hat{\beta}^{\mathtt{l}}_j \mathbb{1}_{e(X_i)\le b_n}\right) e(X_i)^j.$$

We assume the same bandwidth h_n is used for the two local polynomial regressions for simplicity, although in practice different bandwidths can be employed.

I.4.2 General Estimating Equation

We employ the same notation used in Section I.1. Instead of focusing on a population mean, the parameter θ_0 is defined by a possibly nonlinear moment condition $\mathbb{E}[\mu_1(e(X), \theta_0)] = 0$, where $\mu_1(e(X), \theta) = \mathbb{E}[g(Y, X, \theta)|e(X), D = 1]$ and g is a known function. Alternatively, we have $\mathbb{E}[Dg(Y_i, X_i, \theta_0)/e(X)] = 0$. For ease of exposition, we assume that both the parameter and the moment condition are univariate. To estimate θ_0, one can solve the following sample analogue:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i g(Y_i, X_i, \hat{\theta}_n)}{e(X_i)}.$$

Consistency of $\hat{\theta}_n$ can be established with a uniform law of large numbers (see, for example, Newey and McFadden 1994). Given that $\hat{\theta}_n$ is consistent, it is possible to employ a Taylor expansion provided that $g(\cdot)$ is continuously differentiable in θ, and under mild regularity conditions one can show

$$\frac{n}{a_n}(\hat{\theta}_n - \theta_0) = \frac{\Sigma_0}{a_n}\sum_{i=1}^{n}\frac{D_i g(Y_i, X_i, \theta_0)}{e(X_i)} + o_{\mathrm{p}}(1), \qquad \Sigma_0 = \left(-\mathbb{E}\left[\frac{\partial}{\partial\theta}\mu_1(e(X), \theta_0)\right]\right)^{-1}, \quad \text{(I.7)}$$

where n/a_n is a normalizing sequence which we specify in Proposition I.7. Once the estimator has been linearized as above, we can prove a result similar to Theorem I.1. To economize notation, define the random variables $G_i(\theta) = g(Y_i, X_i, \theta)$ and $G_i = G_i(\theta_0)$. We make the following assumption.

Assumption I.1 (GEE)
(i) θ_0 *is the unique root of* $\mathbb{E}[\mu_1(e(X),\theta)] = 0$ *in the interior of a compact parameter space* Θ.
(ii) $g(Y,X,\theta)$ *is continuously differentiable in* θ, *and* $\mathbb{E}[\sup_{\theta\in\Theta} |g(Y_i,X_i,\theta)| \vee |\frac{\partial}{\partial\theta} g(Y_i,X_i,\theta)|] < \infty$.
(iii) For some $\varepsilon > 0$, $\mathbb{E}[|G|^{(\gamma_0\vee 2)+\varepsilon}|e(X) = x, D = 1]$ *is uniformly bounded. There exists a probability distribution* F, *such that for any bounded and continuous function* ℓ, $\mathbb{E}[\ell(G)|e(X) = x, D = 1] \to \int_{\mathbb{R}} \ell(y)F(\mathrm{d}y)$ *as* $x \downarrow 0$. ‖

The following proposition characterizes the large-sample properties of the (IPW-based) GEE estimator $\hat{\theta}_n$.

Proposition I.7 (Large sample properties of the GEE estimator)
Assume Assumptions I.1 and I.1 hold with $\alpha_{G,+}(0) + \alpha_{G,-}(0) > 0$, *where*

$$\alpha_{G,+}(x) = \lim_{t\to 0} \mathbb{E}\left[|G|^{\gamma_0}\mathbb{1}_{G>x}\middle| e(X) = t, D = 1\right], \quad \alpha_{G,-}(x) = \lim_{t\to 0} \mathbb{E}\left[|G|^{\gamma_0}\mathbb{1}_{G<x}\middle| e(X) = t, D = 1\right].$$

Let a_n *be such that*

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{DG}{e(X)}\right|^2 \mathbb{1}_{|DG/e(X)|\le a_n}\right] \to 1.$$

Then $\frac{n}{a_n}(\hat{\theta}_n - \theta_0)$ *converges in distribution, with the limit being:*
(i) $\mathcal{N}(0,\Sigma_0^2)$ *if* $\gamma_0 \ge 2$; *and*
(ii) the Lévy stable distribution if $\gamma_0 < 2$, *with characteristic function:*

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}} \frac{e^{i\Sigma_0\zeta x} - 1 - i\Sigma_0\zeta x}{x^2} M(\mathrm{d}x)\right\},$$
$$\textit{where } M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_{G,+}(0) + \alpha_{G,-}(0)}|x|^{1-\gamma_0}\Big(\alpha_{G,+}(0)\mathbb{1}_{x\ge 0} + \alpha_{G,-}(0)\mathbb{1}_{x<0}\Big)\right]. \quad \|$$

Trimming can be implemented in an obvious way:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i g(Y_i, X_i, \hat{\theta}_{n,b_n})}{e(X_i)}\mathbb{1}_{e(X_i)\ge b_n}.$$

As long as the trimming threshold b_n shrinks to zero as the sample size increases, the trimmed estimator $\hat{\theta}_{n,b_n}$ will be consistent for θ_0. Assuming this is the case, we can again employ a

Taylor expansion and linearize the estimator:

$$\frac{n}{a_{n,b_n}}(\hat{\theta}_{n,b_n} - \theta_0 - \Sigma_0 \mathsf{B}_{n,b_n}) = \frac{\Sigma_0}{a_{n,b_n}} \sum_{i=1}^{n} \left[\frac{D_i G_i}{e(X_i)} \mathbb{1}_{e(X_i) \geq b_n} - \mathsf{B}_{n,b_n} \right] + o_{\mathrm{p}}(1),$$

$$\text{where} \qquad \mathsf{B}_{n,b_n} = -\mathbb{E}[\mu_1(e(X), \theta_0) \mathbb{1}_{e(X) \leq b_n}]. \tag{I.8}$$

The bias term we recover only represents the leading bias in an asymptotic linear expansion, with higher order bias absorbed into the $o_{\mathrm{p}}(1)$ term. The bias arises because after trimming the estimating equation may not have a zero mean in finite samples. Assuming $\mu_1(\cdot)$ is continuous in its first argument, the bias can be further simplified as $\mathsf{B}_{n,b_n} = -\mu_1(0, \theta_0)\mathbb{P}[e(X) \leq b_n]$, which gives its precise order. From this, one can immediately see that if $\mu_1(x, \theta_0) = 0$ for all x small enough, trimming does not induce any bias, and at the same time can improve the performance of the IPW estimator. Such "small bias" scenario, however, is difficult to justify in practice because it requires that the information provided by observations with small probability weights does not feature in the estimating equation.

Proposition I.8 (Large sample properties of the trimmed GEE estimator) *Assume Assumptions I.1 and I.1 hold with $\gamma_0 < 2$ and $\alpha_{G,+}(0) + \alpha_{G,-}(0) > 0$. Let a_n be defined as in Proposition I.7.*

(i) For $b_n a_n \to 0$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat{\theta}_{n,b_n} - \theta_0 - \Sigma_0 \mathsf{B}_{n,b_n})$ converges to the Lévy stable distribution in Proposition I.7(ii).

(ii) For $b_n a_n \to \infty$, let $a_{n,b_n} = \sqrt{n\mathbb{V}[DG/e(X)\mathbb{1}_{e(X) \geq b_n}]}$, then $\frac{n}{a_{n,b_n}}(\hat{\theta}_{n,b_n} - \theta_0 - \Sigma_0 \mathsf{B}_{n,b_n})$ converges to the Gaussian distribution $\mathcal{N}(0, \Sigma_0^2)$.

(iii) For $b_n a_n \to t \in (0, \infty)$, let $a_{n,b_n} = a_n$. Then $\frac{n}{a_{n,b_n}}(\hat{\theta}_{n,b_n} - \theta_0 - \Sigma_0 \mathsf{B}_{n,b_n})$ converges to an infinitely divisible distribution, with characteristic function:

$$\psi(\zeta) = \exp\left\{ \int_{\mathbb{R}} \frac{e^{i\Sigma_0 \zeta x} - 1 - i\Sigma_0 \zeta x}{x^2} M(\mathrm{d}x) \right\},$$

$$\textit{where } M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2 - \gamma_0}{\alpha_{G,+}(0) + \alpha_{G,-}(0)} |x|^{1-\gamma_0} \Big(\alpha_{G,+}(tx) \mathbb{1}_{x \geq 0} + \alpha_{G,-}(tx) \mathbb{1}_{x < 0} \Big) \right]. \quad \|$$

Both Proposition I.7 and I.8 can be further generalized to a vector-valued parameter. As long as the moment condition permits identification (and consistent estimation), one can employ the Cramér-Wold device to characterize the limiting distribution.

Selecting the trimming threshold is more complicated, since now the conditional first and second moment cannot be estimated directly. It is possible to employ a three-step procedure. In the first step, one constructs a pilot point estimate. Next, one can estimate the conditional moments applying local polynomial regression, with either $G_i(\hat{\theta}_n)$ or $G_i(\hat{\theta}_{n,b_n})$

as the dependent variable. In the final step, the trimming threshold is chosen by plugging the second-step estimated conditional moments into the procedure of Theorem I.4.

As a final remark, bias correction is still feasible in this setting by exploiting the asymptotic linear representation in (I.8). To form the bias estimate, one can employ the local polynomial regression technique and regress $G_i(\hat{\theta}_{n,b_n})$ on the probability weights to form an estimate of the bias B_{n,b_n} (Algorithm I.2). Then a bias estimate can be constructed as $\hat{\Sigma}_n\mathsf{B}_{n,b_n}$, where $\hat{\Sigma}_n$ estimates Σ_0 as by sample average.

I.5 Numerical Evidence

This section studies the performance of our inference procedure with a Monte Carlo experiment. Due to the possibly non-Gaussian limiting distributions and the trimming bias documented in Section I.2 and I.3, conventional confidence intervals can exhibit severe undercoverage. (Alternatively, conventional t-tests over-reject the null hypothesis.) On the other hand, our procedure is robust to small probability weights and remains valid for a wide range of trimming threshold choices. Indeed, the robust confidence interval (Algorithm I.1 and I.3) has an empirical coverage very close to the nominal level. We also showcase our method with a dataset from the National Supported Work program.

I.5.1 Simulation Study

The probability weight is distributed according to $\mathbb{P}[e(X) \leq x] = x^{\gamma_0 - 1}$ with $\gamma_0 = 1.5$. A typical realization is given in Figure I.1, which resembles the distribution of the estimated probability weights in our empirical application (Figure I.3(a)). With $\gamma_0 = 1.5$, the convergence rate of the IPW estimator is $n^{1/3}$. Conditional on the weight and $D = 1$, the outcome variable is generated as $\mu_1(e(X)) + \eta$, where the mean equation is either $\cos(2\pi e(X))$ or $1 - e(X)$, and the error η follows a chi-square distribution with four degrees of freedom, centered and scaled to have a zero mean and unit variance. The first specification represents the empirical difficulty of "small denominators" combined with unrestricted conditional mean heterogeneity of the outcome variable, as the conditional mean function is nonlinear in the probability weight. A typical realization of the outcome variable is given in Figure I.2. In the second specification, the leading bias remains the same, but the conditional mean function is linear in the probability weight. Our bias correction technique is therefore expected to perform well. Throughout, we use 5,000 Monte Carlo repetitions, and for each repetition, 1,000 subsampling iterations are used with subsample size $m = \lfloor n/\log(n) \rfloor$, and the full sample size is $n \in \{2,000,\ 5,000,\ 10,000\}$. We follow Theorem I.4 to set the trimming threshold, by solving $\hat{b}_n^s \hat{\mathbb{P}}[e(X_i) \leq \hat{b}_n] = (2n)^{-1}$ with $s \in \{1,\ 1.5,\ 2,\ 3\}$. For $s = 1$, the trimming threshold is rate optimal (in terms of the leading mean squared error) and corresponds to moderate trimming. The other cases fall into the heavy trimming category. Bias correction is based on Algorithm I.2, for which we employ a local linear regression.

The simulation results are collected in Table I.1 and I.2. Under "**Conventional**" we report bias, standard deviation and root mean squared error of the IPW estimator, both untrimmed ($\hat{\theta}_n$) and trimmed ($\hat{\theta}_{n,b_n}$). Note that they have been scaled by $n^{1-1/\gamma_0} = n^{1/3}$. We also report empirical coverage of the conventional Gaussian-based confidence interval under "cov," $[\hat{\theta}_n \pm 1.96 \cdot S_n/\sqrt{n}]$ using the untrimmed estimator, and $[\hat{\theta}_{n,b_n} \pm 1.96 \cdot S_{n,b_n}/\sqrt{n}]$ using the trimmed estimator. (S_n and S_{n,b_n} are defined in Algorithm I.1 and I.3.) Average confidence interval length is reported under "|ci|," scaled by $n^{1-1/\gamma_0} = n^{1/3}$. Under "**Robust**" we report bias, standard deviation and root mean squared error of the trimmed and bias-corrected IPW estimator, $\hat{\theta}^{\mathsf{bc}}_{n,b_n}$ (Algorithm I.2). Under "cov" we report empirical coverage of the subsampling-based confidence interval, using either the untrimmed IPW estimator (Algorithm I.1) or the trimmed and bias-corrected estimator (Algorithm I.3). Also reported is the average length of the subsampling-based confidence interval under "|ci|." In the following, we highlight several observations from Table I.1.

First, inference based on the Gaussian approximation performs poorly, as predicted by our theoretical results. Without trimming, the limiting distribution of the IPW estimator is heavy-tailed (Theorem I.1), and hence using critical values computed from Gaussian quantiles leads to confidence intervals that are overly optimistic/narrow. Although heavy trimming can help restore asymptotic Gaussianity (Theorem I.3(ii)), it is unclear how well distributional approximation based on this result performs in samples of moderate size (Theorem I.3(iii)). In addition, trimming introduces a bias that can significantly shift the limiting distribution away from the target parameter (Theorem I.3 and Lemma I.3). Indeed, in a sample of size $2,000$, using 0.1 as the trimming threshold will lead to a bias that is so severe that a nominal 95% confidence interval will have practically zero coverage. This shows why it is important to combine bias correction with a disciplined method to choose the trimming threshold, and how ad hoc trimming can be detrimental for statistical inference: the researcher essentially changes the target estimand.

Second, it is not surprising that employing a larger trimming threshold can help stabilize the estimator, leading to a smaller empirical standard deviation. However, the mean squared error increases due to the trimming bias. Indeed, by comparing the scaled bias across the three panels in Table I.1, it is clear that the bias is explosive when heavy trimming is used.

Third, despite the fact that the conditional mean function is highly nonlinear, our bias correction procedure successfully removes most of the bias, making the subsampling-based confidence interval having an empirical coverage very close to the 95% nominal level. The performance of our inference procedure is quite robust across a range of trimming threshold choices. For the very heavy trimming case, under-coverage remains to be an issue even with bias correction, because it is quite difficult to estimate a nonlinear function local to a point where observations are scarce. In addition, bias correction may introduce extra variability in samples of moderate size. This is again confirmed by our simulation results, and is why we recommend to conduct bias correction not only for the main estimator, but also in each subsampling iteration.

Now we consider how the form of the conditional mean function affects the performance of our procedure. In Table I.2, the conditional mean is a linear function of the probability weight. If this is known a priori, a better estimation strategy is to fit a global linear regression and extrapolate to observations with small probability weights. Such regression-based estimator will converge at the $\sqrt{n}$-rate and be asymptotically Gaussian. In practice, however, the shape of the conditional mean function is rarely known, so the setting in Table I.2 is best understood as a favorable situation in which our bias correction and inference procedure are expected to perform well. Indeed, the remaining bias is almost zero, and the subsampling-based confidence interval has an empirical coverage very close to the nominal 95% level.

I.5.2 Empirical Application

In this section, we revisit a dataset from the National Supported Work (NSW) program. Our aim is neither to give a thorough evaluation of the program nor to discuss to what extent experimental estimates can be recovered by non-experimental methods. Rather, we use it to illustrate how small probability weights may affect the performance of the IPW estimator, and to showcase our robust inference procedure.

The NSW is a labor training program implemented in 1970's by providing work experience, from 6 to 18 months, to individuals who face social or economic difficulties. It has been analyzed in multiple studies and along different directions since LaLonde (1986). We use the same dataset employed in Dehejia and Wahba (1999), and refer interested readers to the original work for detailed discussion on institutional background, variable definition, and sample inclusion. Briefly, our sample consists of the treated individuals in the NSW experimental group (sample size 185), and a nonexperimental comparison group from the Panel Study of Income Dynamics (PSID, sample size $1,157$). Besides the binary treatment indicator ($D = 1$ for NSW treated units and 0 for PSID comparison units) and the main outcome variable (Y) of post-intervention earning measured in 1978, information on age, education,

Table I.1. Simulation: $\gamma_0 = 1.5$, $\mathbb{E}[Y|e(X), D = 1] = \cos(2\pi e(X))$.

(a) $n = 2,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.377$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	$\lvert\text{ci}\rvert$	bias	sd	rmse	cov	$\lvert\text{ci}\rvert$
–	–.	0.131	3.773	3.776	0.775	7.308				0.844	21.235
0.004	0.170	0.800	1.493	1.694	0.740	5.116	0.238	1.565	1.583	0.924	7.387
0.016	1.338	1.576	0.979	1.855	0.541	3.713	0.465	1.169	1.258	0.926	5.757
0.036	4.606	2.373	0.741	2.486	0.158	2.849	0.628	1.064	1.236	0.913	4.973
0.094	19.225	3.718	0.503	3.752	0.000	1.956	0.711	0.999	1.226	0.906	4.219

(b) $n = 5,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.319$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	$\lvert\text{ci}\rvert$	bias	sd	rmse	cov	$\lvert\text{ci}\rvert$
–	–	0.025	5.681	5.681	0.786	7.948				0.869	37.240
0.002	0.173	0.764	1.546	1.724	0.755	5.336	0.259	1.592	1.613	0.928	7.196
0.010	1.689	1.697	0.966	1.953	0.514	3.717	0.485	1.103	1.205	0.916	5.233
0.025	6.653	2.692	0.714	2.785	0.077	2.805	0.696	0.961	1.187	0.891	4.457
0.072	32.182	4.484	0.478	4.510	0.000	1.885	0.883	0.894	1.257	0.846	3.780

(c) $n = 10,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.281$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	$\lvert\text{ci}\rvert$	bias	sd	rmse	cov	$\lvert\text{ci}\rvert$
–	–	0.053	7.909	7.909	0.787	7.761				0.862	59.629
0.001	0.168	0.781	1.575	1.758	0.757	5.404	0.213	1.609	1.623	0.922	6.944
0.007	1.994	1.812	0.975	2.058	0.477	3.698	0.441	1.086	1.172	0.910	4.870
0.019	8.752	2.971	0.708	3.054	0.037	2.756	0.668	0.916	1.134	0.877	4.097
0.059	47.837	5.175	0.466	5.196	0.000	1.824	0.895	0.831	1.221	0.817	3.490

Note. (i) $\hat{b}_n$: trimming threshold. (ii) $n_{\leq \hat{b}_n}$: effective number of trimmed observations. (iii) bias: empirical bias, scaled by n^{1-1/γ_0}. (iv) sd: empirical standard deviation, scaled by n^{1-1/γ_0}. (v) rmse: empirical root mean squared error, scaled by n^{1-1/γ_0}. (vi) cov: coverage probability (nominal level 0.95). (vii) $|\text{ci}|$: average confidence interval length, scaled by n^{1-1/γ_0}. **Conventional**: bias, sd and rmse are calculated for both the untrimmed ($\hat{\theta}_n$) and the trimmed ($\hat{\theta}_{n,b_n}$) IPW estimators. Coverage is calculated for the Gaussian-based confidence interval, $[\hat{\theta}_n \pm 1.96 \cdot S_n/\sqrt{n}]$ without trimming, and $[\hat{\theta}_{n,b_n} \pm 1.96 \cdot S_{n,b_n}/\sqrt{n}]$ with trimming. **Robust**: bias, sd and rmse are calculated for the trimmed and bias-corrected IPW estimator ($\hat{\theta}^{\mathsf{bc}}_{n,b_n}$, Algorithm I.2). Coverage is calculated for the subsampling-based confidence interval, using either the untrimmed (Algorithm I.1) or the trimmed and bias-corrected (Algorithm I.3) IPW estimator. $\hat{h}_n$: bandwidth for local polynomial bias correction. Number of Monte Carlo repetitions: 5000. Number of subsampling iterations: 1000. Subsample size: $\lfloor n/\log(n) \rfloor$.

Table I.2. Simulation: $\gamma_0 = 1.5$, $\mathbb{E}[Y|e(X), D = 1] = 1 - e(X)$.

(a) $n = 2,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.377$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.132	3.771	3.773	0.774	7.295				0.864	22.017
0.004	0.170	0.800	1.490	1.691	0.742	5.105	0.012	1.569	1.569	0.939	7.755
0.016	1.338	1.569	0.977	1.849	0.543	3.716	0.003	1.172	1.172	0.957	6.029
0.036	4.606	2.357	0.747	2.472	0.165	2.875	0.001	1.063	1.063	0.964	5.228
0.094	19.225	3.730	0.510	3.764	0.000	2.005	0.017	0.984	0.984	0.967	4.530

(b) $n = 5,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.319$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.025	5.678	5.678	0.784	7.933				0.873	37.233
0.002	0.173	0.763	1.549	1.726	0.754	5.323	0.031	1.600	1.601	0.935	7.334
0.010	1.689	1.692	0.967	1.949	0.514	3.712	0.015	1.112	1.112	0.956	5.346
0.025	6.653	2.676	0.719	2.771	0.081	2.817	0.015	0.967	0.967	0.963	4.559
0.072	32.182	4.467	0.491	4.494	0.000	1.927	0.019	0.890	0.890	0.964	3.958

(c) $n = 10,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.281$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.045	7.909	7.909	0.790	7.747				0.863	59.692
0.001	0.168	0.773	1.571	1.751	0.760	5.391	0.019	1.609	1.609	0.928	7.017
0.007	1.994	1.801	0.973	2.047	0.477	3.689	0.005	1.092	1.092	0.952	4.943
0.019	8.752	2.949	0.710	3.033	0.040	2.760	0.003	0.923	0.923	0.958	4.152
0.059	47.837	5.136	0.474	5.158	0.000	1.856	0.006	0.829	0.829	0.964	3.588

Note. (i) $\hat{b}_n$: trimming threshold. (ii) $n_{\leq \hat{b}_n}$: effective number of trimmed observations. (iii) bias: empirical bias, scaled by n^{1-1/γ_0}. (iv) sd: empirical standard deviation, scaled by n^{1-1/γ_0}. (v) rmse: empirical root mean squared error, scaled by n^{1-1/γ_0}. (vi) cov: coverage probability (nominal level 0.95). (vii) |ci|: average confidence interval length, scaled by n^{1-1/γ_0}. **Conventional**: bias, sd and rmse are calculated for both the untrimmed ($\hat{\theta}_n$) and the trimmed ($\hat{\theta}_{n,b_n}$) IPW estimators. Coverage is calculated for the Gaussian-based confidence interval, $[\hat{\theta}_n \pm 1.96 \cdot S_n/\sqrt{n}]$ without trimming, and $[\hat{\theta}_{n,b_n} \pm 1.96 \cdot S_{n,b_n}/\sqrt{n}]$ with trimming. **Robust**: bias, sd and rmse are calculated for the trimmed and bias-corrected IPW estimator ($\hat{\theta}^{\mathsf{bc}}_{n,b_n}$, Algorithm I.2). Coverage is calculated for the subsampling-based confidence interval, using either the untrimmed (Algorithm I.1) or the trimmed and bias-corrected (Algorithm I.3) IPW estimator. $\hat{h}_n$: bandwidth for local polynomial bias correction. Number of Monte Carlo repetitions: 5000. Number of subsampling iterations: 1000. Subsample size: $\lfloor n/\log(n) \rfloor$.

marital status, ethnicity and earnings in 1974 and 1975 are also available as pre-intervention individual characteristics (X). We follow the literature and focus on the treatment effect on the treated (ATT), which requires weighting observations from the comparison group by $\hat{e}(X)/(1-\hat{e}(X))$. As a result, probability weights that are close to 1 can pose a challenge to both estimation and inference.

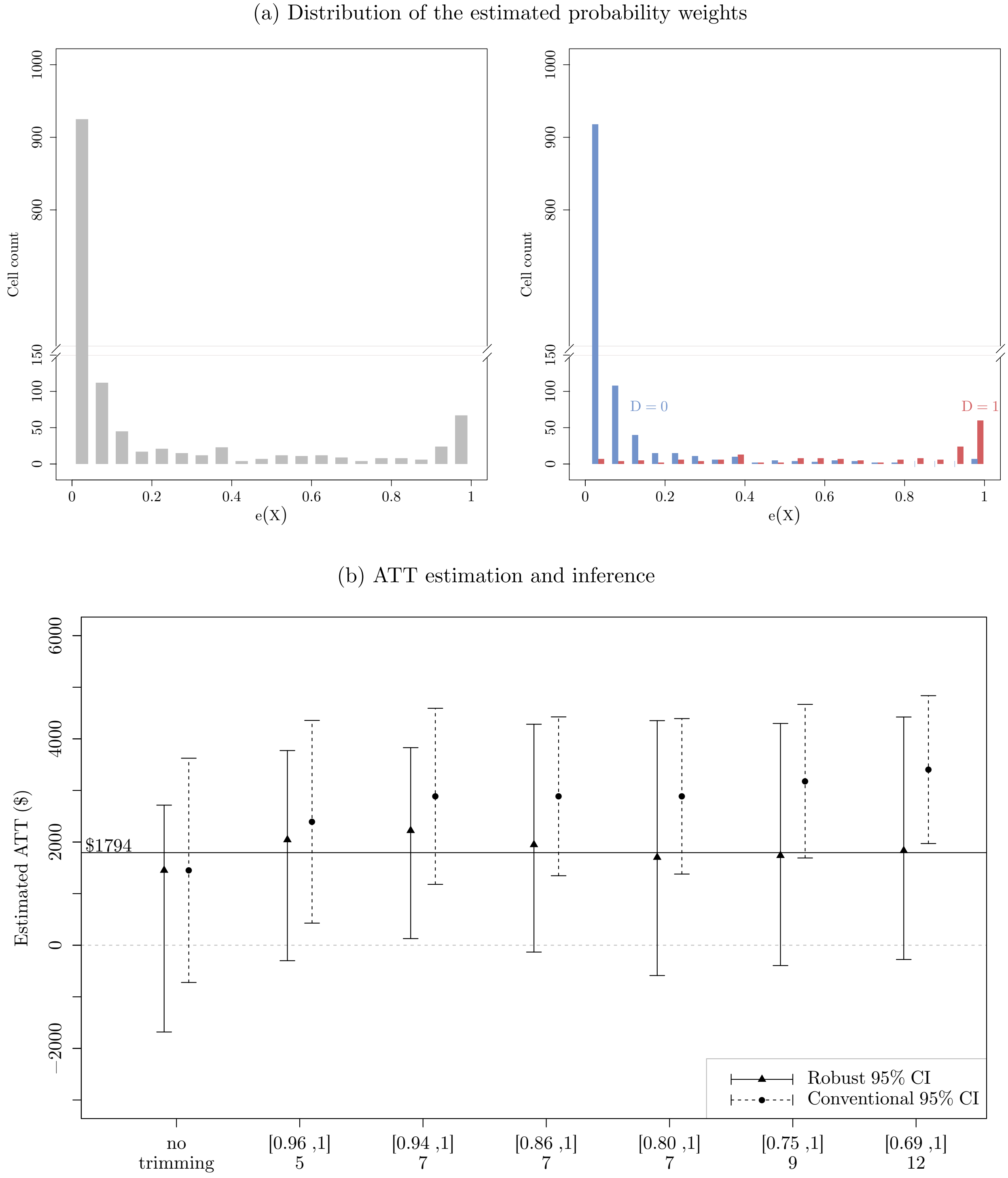
The probability weight is estimated in a Logit model with `age`, `education`, `earn1974`, `earn1975`, `age`^{2}, `education`^{2}, `earn1974`^{2}, `earn1975`^{2}, three indicators for `married`, `black` and `hispanic`, and an interaction term between `black` and unemployment status in 1974: `black` × `u74`. Figure I.3(a) plots the distribution of the estimated probability weights, which clearly exhibits a heavy tail near 1. Since $\gamma_0 = 2$ roughly corresponds to uniformly distributed probability weights, the tail index in this dataset should be well below 2, suggesting that standard inference procedures based on the Gaussian approximation may not perform well.

In Figure I.3(b), we plot the bias-corrected ATT estimates (solid triangles) and the robust 95% confidence intervals (solid vertical lines) with different trimming thresholds. For comparison, we also show conventional point estimates and confidence intervals (solid dots and dashed vertical lines, based on the Gaussian approximation) using the same trimming thresholds. Without trimming, the point estimate is \$1,451 with a confidence interval $[-1,763,\ 2,739]$. The robust confidence interval is asymmetric around the point estimate, a feature also predicted by our theory: probability weights that are close to 1 affect the estimation of $\mathbb{E}[Y(0)|D=1]$ and will subsequently contribute to a long left tail to the estimator, because the outcome variable is nonnegative.

For the trimmed IPW estimator, the trimming thresholds are chosen following Theorem I.4, and the region used for local polynomial bias estimation is $[0.71, 1]$, corresponding to a bandwidth $h_n = 0.29$. Under the mean squared error optimal trimming, units in the comparison group with probability weights above 0.96 (five observations) are discarded. Compared to the untrimmed case, the robust confidence interval becomes more symmetric.

In this empirical example, a noteworthy feature of our method is that both the bias-corrected point estimates and the robust confidence intervals remain quite stable for a range of trimming threshold choices, and the point estimates are very close to the experimental benchmark (\$1,794). This is in stark contrast to conventional confidence intervals that rely on Gaussian approximation. First, conventional confidence intervals fail to adapt to the non-Gaussian limiting distributions we documented in Theorem I.1 and I.3, and are overly optimistic/narrow. Second, by ignoring the trimming bias, they are only valid for a pseudo-true parameter implicitly defined by the trimming threshold. As a result, the researcher changes the target estimand each time a different trimming threshold is used, making conventional confidence intervals very sensitive to b_n.

Figure I.3. Empirical illustration: National Supported Work program.



(a) Distribution of the estimated probability weights

(b) ATT estimation and inference

Note. Panel (a): histogram of the estimated probability weights (propensity scores). Panel (b): estimated ATT for different trimming thresholds. Numbers below the horizontal axis show the trimming threshold/region and the effective number of observations trimmed from the comparison group. The experimental benchmark ($1, 794) is indicated by the solid horizontal line.

I.6 Conclusion

We study the large-sample properties of the Inverse Probability Weighting (IPW) estimator. We show that, in the presence of small probability weights, this estimator may have a slower-than-$\sqrt{n}$ convergence rate and a non-Gaussian limiting distribution. We also study the effect of discarding observations with small probability weights, and show that such trimming not only complicates the limiting distribution, but also causes a non-negligible bias. As a consequence, inference based on the standard Gaussian approximation can be highly unreliable when ad hoc trimming rules are used. We consider two extensions of our basic framework, one for treatment effect estimands and the other for parameters defined by a nonlinear estimating equation, and show that the aforementioned conclusions continue to hold more generally.

We propose an inference procedure that is robust not only to small probability weights entering the IPW estimator but also to a range of trimming threshold choices. The "two-way robustness" is achieved by combining resampling with a novel local polynomial-based bias-correction technique. We also propose a method to choose the trimming threshold by minimizing an empirical analogue of the asymptotic mean squared error. Implementation of our robust inference procedure and trimming threshold selector is straightforward. As the probability weights are typically unknown in applications, we allow the probability weights to be estimated in a first step. In particular, we show that the two workhorse models, Logit and Probit, can be employed under mild regularity conditions.

More generally, our results shed light on the reliability of conventional inference procedures using inverse weighting type estimators. One important insight is that with "small denominators," conventional inference procedures can be unreliable regardless of whether trimming is employed or not. It will be interesting to explore the possibility of estimating the denominator in a first step, perhaps with a nonparametric method or a high-dimensional model. The problem is considerably more challenging, because in both cases the estimated denominator can be highly volatile and its tail behavior can deviate significantly from the regular variation setting.

I.7 Additional Results and Preliminary Lemmas

For ease of reference, we collect some facts from Feller (1991) on regularly varying functions and distributional convergence of sums of random variables. We also provide preliminary lemmas for establishing the main results.

I.7.1 Regular Variation

In this subsection, we take X and Y as some generic univariate random variables, not necessarily the same as in the previous sections.

With finite second moments, weak convergence is not sensitive to delicate tail features. This is captured by the central limit theorem. However, weak convergence of sums of random variables without finite variance relies on additional tail properties. The appropriate notion is regular variation.

Definition I.1
A random variable X has regularly varying tail at ∞ with index $-\gamma < 0$, if for all $x > 0$, $\mathbb{P}[X > tx]/\mathbb{P}[X > t] \to x^{-\gamma}$ as $t \to \infty$. Similarly, X has regularly varying tail at $-\infty$ if for all $x > 0$, $\mathbb{P}[X < tx]/\mathbb{P}[X < t] \to x^{-\gamma}$ as $t \to -\infty$.

Assume $\mathbb{P}[X > 0] = 1$, then it has regularly varying tail at 0 with index γ if $1/X$ has regularly varying tail at ∞ with index $-\gamma$. ||

One special example of regular variation is "approximately polynomial tail": Assume $\mathbb{P}[X > x] = c(x)x^{-\gamma}$ with $\gamma > 0$ and $c(x)$ tending to a strictly positive constant, then X has regularly varying tail at ∞ with index $-\gamma$. Following is a complete characterization of regular variation.

Lemma I.4
Assume X has regularly varying tail at ∞ with index $-\gamma$, then for all x large enough,

$$\mathbb{P}[X > x] = x^{-\gamma}c(x), \qquad \text{with } c(x) = L(x)\exp\left\{\int_s^x \frac{R(t)}{t}\mathrm{d}t\right\}, \tag{I.9}$$

where $L(x)$ tends to a strictly positive constant, $\lim_{x\to\infty} R(x) = 0$, and s is some strictly positive constant. ||

If X has a regularly varying right tail with index $-\gamma$, then it is clear that $\mathbb{E}[X^\alpha \mathbb{1}_{X>0}]$ exists and is finite for any $\alpha < \gamma$. However, the expectation will be infinite for all $\alpha > \gamma$. For the purpose of studying distributional convergence of sums of heavy-tailed random variables, a more thorough characterization of the truncated moment $\mathbb{E}[X^\alpha \mathbb{1}_{0<X<x}]$ is necessary.

Lemma I.5
Assume X has a regularly varying right tail at ∞ with index $-\gamma$, then for any $\alpha > \gamma$,

$$\frac{\mathbb{E}[X^\alpha \mathbb{1}_{0<X<x}]}{x^\alpha \mathbb{P}[X > x]} \to \frac{\gamma}{\alpha - \gamma}, \qquad \text{as } x \to \infty.$$ ||

In previous sections, we take X to be the inverse probability weight multiplied by the binary indicator. However, the primary quantity of interest involves the outcome variable, and it is unclear how multiplication affects the tail behavior. The following lemma gives sufficient conditions under which the product XY has the same tail index as X. Despite being intuitive, it doesn't seem to be available in the literature.

Lemma I.6
Assume X is nonnegative and has a regularly varying tail with index $-\gamma$. Further assume (i) $\mathbb{E}[|Y|^\alpha|X = x]$ is uniformly bounded for some $\alpha > \gamma$, and (ii) there exists a distribution F, such that for all bounded and continuous $\ell(\cdot)$, $\mathbb{E}[\ell(Y)|X = x] \to \int \ell(y)F(\mathrm{d}y)$ as $x \to \infty$. Then

$$\lim_{x\to\infty} \frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \lim_{x\to\infty} \mathbb{E}[|Y|^\gamma \mathbb{1}_{Y>0}|X = x],$$
$$\lim_{x\to\infty} \frac{\mathbb{P}[XY < -x]}{\mathbb{P}[X > x]} = \lim_{x\to\infty} \mathbb{E}[|Y|^\gamma \mathbb{1}_{Y<0}|X = x].$$

Therefore the product XY has a regularly varying right (resp. left) tail with index $-\gamma$, if $\lim_{x\to\infty} \mathbb{P}[Y > 0|X = x] > 0$ (resp. $\lim_{x\to\infty} \mathbb{P}[Y < 0|X = x] > 0$). ∥

The first condition that $\mathbb{E}[|Y|^\alpha|X = x]$ is uniformly bounded is intuitive. To ensure the product that XY has the same tail behavior as X, one needs to assume that the tail of Y is uniformly thin enough. In general, it is not possible to drop the second requirement that $Y|X = x$ converges in distribution, unless one is willing to impose additional structures on the conditional distribution. Following is a example, which shows that when the conditional distribution of Y "oscillates" as X tends to infinity, the product XY does not have a regularly varying tail even when Y is bounded.

Example I.1 Assume $Y = 1$ for $X \in (2^j, 2^{j+1}]$ for $j = 1, 3, 5, \cdots$, and equals 0 otherwise, then on the grid $(2^j)_{j\geq 1}$, XY has right tail:

$$\mathbb{P}[XY > 2^j] = \sum_{k=j,\ k \text{ odd}}^{\infty} F_X(2^{k+1}) - F(2^k).$$

Now we take limit $j \to \infty$ along the sequence of odd numbers,

$$\lim_{j\to\infty,\ j \text{ odd}} \frac{\mathbb{P}[XY > 2^j]}{\mathbb{P}[X > 2^j]} = \lim_{j\to\infty,\ j \text{ odd}} \sum_{k=j,k \text{ odd}}^{\infty} \frac{F_X(2^{k+1}) - F(2^k)}{\mathbb{P}[X > 2^j]}$$
$$= \left(1 - 2^{-\gamma}\right) \sum_{k=0}^{\infty} 2^{-2k\gamma} = \frac{1 - 2^{-\gamma}}{1 - 2^{-2\gamma}}.$$

If we take the limit along the sequence of even numbers,

$$\lim_{j\to\infty,\ j \text{ even}} \frac{\mathbb{P}[XY > 2^j]}{\mathbb{P}[X > 2^j]} = \left(1 - 2^{-\gamma}\right) \sum_{k=1}^{\infty} 2^{-2k\gamma} = 2^{-2\gamma} \frac{1 - 2^{-\gamma}}{1 - 2^{-2\gamma}}.$$

Since X has regularly varying tail and the ratio $\mathbb{P}[XY > x]/\mathbb{P}[X > x]$ oscillates between two numbers, we conclude XY does not have regularly varying tail. ‖

I.7.2 Distributional Convergence

Assume $(X_{i,n})_{1\le i\le n, n\ge 1}$ is a triangular array, such that for each n, $(X_{i,n})_{1\le i\le n}$ are independently and identically distributed. The following lemma characterizes the asymptotic distribution of the sum $\sum_{i=1}^n X_{i,n}$, if exists.

Lemma I.7
Assume $\mathbb{E}[X_{i,n}] = 0$ for all n, and that the sum $\sum_{i=1}^n X_{i,n}$ converges in distribution. Then the limiting distribution has a characteristic function given by the canonical form:

$$\psi(\zeta) = \exp \int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x),$$

where M is a nonnegative measure satisfying (i) $M(I) < \infty$ for all bounded intervals I, and (ii) the integrals $\int_c^\infty x^{-1} M(\mathrm{d}x)$ and $\int_{-\infty}^{-c} x^{-1} M(\mathrm{d}x)$ are finite for all $c > 0$. ‖

The next lemma gives conditions under which the distributional convergence of the partial sum, $\sum_{i=1}^n X_{i,n}$, happens.

Lemma I.8
Assume $\mathbb{E}[X_{i,n}] = 0$ for all n, and let F_n be the distribution function of $X_{i,n}$. Then the sum $\sum_{i=1}^n X_{i,n}$ converges in distribution if and only if, for some measure M,

$$n\mathbb{E}\left[X_{i,n}^2 \mathbb{1}_{X_{i,n\in I}}\right] \to M(I)$$

for all compact intervals with $M(\partial I) = 0$; and

$$n(1 - F_n(c)) \to \int_c^\infty x^{-2} M(\mathrm{d}x), \qquad nF_n(-c) \to \int_{-\infty}^{-c} x^{-2} M(\mathrm{d}x),$$

for all $c > 0$ with $M(\{c\}) = 0$. In this case, the limiting distribution is infinitely divisible, and its characteristic function is given by the form in Lemma I.7. ‖

To understand the previous lemma, assume $X_{i,n} = Y_i/\sqrt{n}$ with $(Y_i)_{i\geq 1}$ being iid and $\mathbb{V}[Y_i] = \sigma^2$. Then it is quite easy to show that $M(I) = \sigma^2\mathbb{1}_{0\in I}$. That is, M is a point mass of size σ^2 at the origin. The integrand is $-\zeta^2/2$ at the origin by l'Hospital's rule, meaning that $\psi(\zeta) = e^{-\zeta^2\sigma^2/2}$, which is the characteristic function of the centered Gaussian distribution with variance σ^2. The situation becomes much more delicate if $X_{i,n}$ does not have a finite variance, and/or if it involves trimming that depends on the sample size. We will be using this lemma repeatedly in order to derive the asymptotic distributions of the IPW and the trimmed IPW estimators.

I.7.3 Local Polynomial Regression

Local polynomial regression is employed for estimating the trimming bias. To be more specific, the outcome variable is regressed on the probability weight in a region local to the origin. That is,

$$\hat{\boldsymbol{\beta}} = \Big[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\Big]' = \underset{\beta_0,\beta_1,\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} D_i \Big[Y_i - \sum_{j=0}^{p} \beta_j e(X_i)^j\Big]^2 \mathbb{1}_{e(X_i)\leq h_n},$$

where for ease of exposition we assume that the true probability weights are used. The following lemma characterizes the properties of the local polynomial estimates.

Lemma I.9

Assume Assumption I.1 and I.2 hold. In addition, assume (i) $\mu_1(\cdot)$ is $p+1$ times continuously differentiable; (ii) $\mu_2(0)-\mu_1(0)^2 > 0$; and (iii) the bandwidth sequence satisfies $nh_n\mathbb{P}[e(X) \leq h_n] \to \infty$ and $nh_n^{2p+3}\mathbb{P}[e(X) \leq h_n] = O(1)$. Let $\boldsymbol{\beta} = \Big[\mu_1(0), \mu_1^{(1)}(0), \cdots, \frac{1}{p!}\mu_1^{(p)}(0)\Big]'$ and $\hat{\boldsymbol{\beta}}$ be defined in the above, then

$$\sqrt{nh_n\mathbb{P}[e(X) \leq h_n]}\mathbf{H}_n \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} - h_n^{p+1}\mathbf{H}_n^{-1}\frac{\mu_1^{(p+1)}(0)}{(p+1)!}\mathbf{S}^{-1}\mathbf{R}\right) \rightsquigarrow \mathcal{N}\Big(\mathbf{0},\ (\mu_2(0) - \mu_1(0)^2)\mathbf{S}^{-1}\Big),$$

where $\mathbf{H}_n = \text{diagonal}(1, h_n, h_n^2, \cdots, h_n^p)$, $\mathbf{S} = (s_{ij})_{1\leq i,j\leq p}$ with $s_{ij} = (\gamma_0 - 1)/(\gamma_0 + i + j - 2)$, and $\mathbf{R} = (r_i)_{1\leq i\leq p}$ with $r_i = (\gamma_0 - 1)/(\gamma_0 + i + p)$. ‖

I.8 Proof

I.8.1 Proof of Lemma I.4

See Theorem VIII.9.1 and the corresponding corollary in Feller (1991). ■

I.8.2 Proof of Lemma I.5

See Theorem VIII.9.2 in Feller (1991). ■

I.8.3 Proof of Lemma I.6

We split the proof into three parts.

Part 1

We first assume X and Y are independent. For simplicity, we denote by F_X and F_Y the distribution functions of X and Y, and $\varepsilon = \alpha - \gamma > 0$. Define $a(y,x)$ be

$$a(y,x) = \frac{1 - F_X(x/y)}{1 - F_X(x)}.$$

Then from the definition of regular variation, one has $\lim_{x\to\infty} a(x,y) = y^\gamma$ for all $y > 0$. Consider the following limit:

$$\begin{aligned}\lim_{x\to\infty} \frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} &= \lim_{x\to\infty} \int_0^\infty a(y,x) F_Y(\mathrm{d}y) \\ &= \underbrace{\lim_{x\to\infty} \int_0^{b(x)^{1/(\gamma+\varepsilon)}} a(y,x) F_Y(\mathrm{d}y)}_{\text{(I)}} + \underbrace{\lim_{x\to\infty} \int_{b(x)^{1/(\gamma+\varepsilon)}}^\infty a(y,x) F_Y(\mathrm{d}y)}_{\text{(II)}},\end{aligned}$$

where $b(x)$ satisfies $\lim_{x\to\infty} b(x)(1 - F_X(x)) = \infty$ and $\lim_{x\to\infty} b(x)/x^{\gamma+\varepsilon} = 0$. We first show that the second limit is zero:

$$\begin{aligned}\text{(II)} &= \lim_{x\to\infty} \int_{b(x)^{1/(\gamma+\varepsilon)}}^\infty \frac{1 - F_X(x/y)}{1 - F_X(x)} F_Y(\mathrm{d}y) \leq \lim_{x\to\infty} \int_{b(x)^{1/(\gamma+\varepsilon)}}^\infty \frac{1}{1 - F_X(x)} F_Y(\mathrm{d}y) \\ &\leq \lim_{x\to\infty} \int_{b(x)^{1/(\gamma+\varepsilon)}}^\infty \frac{y^{\gamma+\varepsilon}}{(1 - F_X(x))b(x)} F_Y(\mathrm{d}y) \\ &\leq \lim_{x\to\infty} \frac{1}{(1 - F_X(x))b(x)} \mathbb{E}[|Y|^{\gamma+\varepsilon}] = 0.\end{aligned}$$

Now we consider (I), and show that for all x large enough, the integrand is bounded by an integrable function (of y), hence dominated convergence can be applied. First, we note that for $y \in (0,1)$, $a(y,x) \leq 1$ for all x. Therefore we only need to consider $y \in [1, b(x)^{1/(\gamma+\varepsilon)}]$.

Since $y \leq b(x)^{1/(\gamma+\varepsilon)}$, we have

$$\frac{x}{y} \geq \left(\frac{x^{\gamma+\varepsilon}}{b(x)}\right)^{\frac{1}{\gamma+\varepsilon}},$$

which can be made arbitrarily large for all x large enough. Also note that

$$a(y,x) = y^{\gamma}\frac{L(x/y)}{L(x)}\exp\left\{\int_x^{x/y}\frac{R(t)}{t}\mathrm{d}t\right\},$$

where the ratio $|L(x/y)/L(x)|$ is bounded by a constant for all x large enough, uniformly in y. Similarly, $|R(t)|$ can be chosen to be arbitrarily small, which means the exponential term is bounded by y^{ε}. Hence, for $y \in [1, b(x)^{1/(\gamma+\varepsilon)}]$,

$$a(y,x) \leq Cy^{\gamma+\varepsilon},$$

which is integrable with respect to the distribution F_Y. Applying the dominated convergence, one concludes that

$$\lim_{x\to\infty}\frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \int_0^{\infty} y^{\gamma}F_Y(\mathrm{d}y) = \mathbb{E}[Y^{\gamma}\mathbb{1}_{Y>0}],$$

so that the product XY also has regularly varying tail with index γ, provided that $\mathbb{P}[Y > 0] > 0$. Similar argument can be applied to analyze the left tail of XY.

Part 2

Now we drop the independence assumption, and assume instead that Y is bounded by a constant C. For simplicity, we use F to denote the limit of the conditional distribution $F_{Y|X=x}$ as $x \to \infty$. Same as before, $\varepsilon = \alpha - \gamma > 0$. First,

$$\frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \int_0^{\infty}\frac{\mathbb{P}[Y > x/y|X = y]}{\mathbb{P}[X > x]}F_X(\mathrm{d}y) = \int_{x/C}^{\infty}\frac{\mathbb{P}[Y > x/y|X = y]}{\mathbb{P}[X > x]}F_X(\mathrm{d}y).$$

Further, let $U \perp\!\!\!\perp (X,Y)$ be distributed according to F. Since the conditional distribution $Y|X = x$ converges weakly to that of U as $x \to \infty$, one has for all x large enough,

$$\Big|\mathbb{P}[Y > x|X = y] - \mathbb{P}[U > x]\Big| \leq \eta + \mathbb{1}_{x\in A(y)},$$

where $\eta > 0$ is arbitrary, and for fixed η, the set $A(x)$ takes the form

$$A(y) = \bigcup_{j=1}^{J} \Big(a_j - \delta(y), a_j + \delta(y)\Big),$$

with $\delta(y)$ monotonically decreases to zero as $y \to \infty$. Then we have

$$\begin{aligned}&\int_{x/C}^{\infty} \left| \frac{\mathbb{P}[Y > x/y|X = y] - \mathbb{P}[U > x/y]}{\mathbb{P}[X > x]} \right| F_X(\mathrm{d}y) \\ &\leq \eta \frac{\mathbb{P}[X > x/C]}{\mathbb{P}[X > x]} + \sum_{1 \leq j \leq J:\ 0 \leq a_j \leq C} \frac{F_X(x/(a_j - \delta(x/c))) - F_X(x/(a_j + \delta(x/c)))}{\mathbb{P}[X > x]},\end{aligned}$$

where the right-hand-side has limit ηC^{γ}. Since η is arbitrary, the left-hand-side tends to zero as $x \to \infty$. As a result, we have

$$\lim_{x \to \infty} \frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \lim_{x \to \infty} \int_{x/C}^{\infty} \frac{\mathbb{P}[U > x/y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y) = \lim_{x \to \infty} \frac{\mathbb{P}[XU > x]}{\mathbb{P}[X > x]}.$$

Since we have $U \perp\!\!\!\perp X$, Part 1 of this proof can be applied to obtain the desired result.

Part 3

Now we drop the boundedness condition on Y. For this purpose, we only need to show that the following

$$\int_0^{x/C} \frac{\mathbb{P}[Y > x/y|X = y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y), \qquad \int_0^{x/C} \frac{\mathbb{P}[U > x/y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y),$$

can be made arbitrarily small by choosing C large enough. We only show for the first term. By Markov's inequality and the assumption that $\mathbb{E}[|Y|^{\gamma+\varepsilon}|X = x]$ is uniformly bounded, we have

$$\begin{aligned}\int_0^{x/C} \frac{\mathbb{P}[Y > x/y|X = y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y) &\leq \Big(\sup_x \mathbb{E}[|Y|^{\gamma+\varepsilon}|X = x]\Big) \int_0^{x/C} \frac{y^{\gamma+\varepsilon}}{x^{\gamma+\varepsilon}\mathbb{P}[X > x]} F_X(\mathrm{d}y) \\ &\to \Big(\sup_x \mathbb{E}[|Y|^{\gamma+\varepsilon}|X = x]\Big) C^{-\varepsilon} \frac{\gamma}{\varepsilon},\end{aligned}$$

where the last convergence follows from Lemma I.5. ■

I.8.4 Proof of Lemma I.7 and I.8

See Section XVII.2 in Feller (1991). ■

I.8.5 Proof of Lemma I.9

We take $p = 1$ (i.e. local linear regression) for the proof, which allows us to show explicitly the form of various matrices. The general case can be proven similarly, although the notation becomes much more cumbersome. Define $\mathbf{r}(x) = [1, x]'$, then the estimator can be rewritten as

$$\left[\sum_{i=1}^{n} \mathbf{r}(e(X_i))\mathbf{r}(e(X_i))' w_i\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{r}(e(X_i)) Y_i w_i\right],$$

where $w_i = \mathbb{1}_{e(X_i)\le h_n, D_i=1}$. We use $F_{e(X)}$ to denote the distribution function of the probability weight. We first analyze the "denominator" term. Consider the following:

$$\mathbf{S}_n = \frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)} \sum_{i=1}^{n} \mathbf{r}(e(X_i)/h_n)\mathbf{r}(e(X_i)/h_n)' w_i,$$

whose expectation is given by

$$\begin{aligned}
\mathbb{E}[\mathbf{S}_n] &= \frac{1}{F_{e(X)}(h_n)} \int_0^{h_n} \mathbf{r}(x/h_n)\mathbf{r}(x/h_n)' x/h_n F_{e(X)}(\mathrm{d}x) \\
&= \frac{1}{F_{e(X)}(h_n)} \left[\mathbf{r}(1)\mathbf{r}(1)' F_{e(X)}(h_n) - \int_0^1 \begin{bmatrix} 1 & 2x \\ 2x & 3x^2 \end{bmatrix} F_{e(X)}(x h_n)\mathrm{d}x\right] \\
&= \left[\mathbf{r}(1)\mathbf{r}(1)' - \int_0^1 \begin{bmatrix} 1 & 2x \\ 2x & 3x^2 \end{bmatrix} x^{\gamma_0 - 1}\mathrm{d}x\right](1 + o(1)) = \begin{bmatrix} \frac{\gamma_0-1}{\gamma_0} & \frac{\gamma_0-1}{\gamma_0+1} \\ \frac{\gamma_0-1}{\gamma_0+1} & \frac{\gamma_0-1}{\gamma_0+2} \end{bmatrix}(1 + o(1)),
\end{aligned}$$

which is always invertible. Next we show that $\mathbf{S}_n$ converges to the expectation computed above. For this purpose, we consider the variance of individual terms in $\mathbf{S}_n$, which is bounded by

$$\begin{aligned}
&\frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)^2} \int_0^{h_n} (x/h_n)^{j+1} F_{e(X)}(\mathrm{d}x) \\
&= \frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)^2} \left[F_{e(X)}(h_n) - \int_0^1 (j+1)x^j F_{e(X)}(x h_n)\mathrm{d}x\right] \\
&= \frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)} \left[1 - \int_0^1 (j+1)x^j x^{\gamma_0 - 1}\mathrm{d}x\right](1 + o(1)) \\
&= \frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)} \frac{\gamma_0 - 1}{\gamma_0 + j}(1 + o(1)),
\end{aligned}$$

which shrinks to zero provided that $n h_n F_{e(X)}(h_n) \to \infty$.

Now we consider the "numerator" term. First ignore the expectation, and let $\eta_i = Y_i - \mathbb{E}[Y_i(1)|e(X_i)]$ be the residual from conditional expectation projection. Then the following

$$\mathbf{L}_n = \sqrt{\frac{1}{nh_nF_{e(X)}(h_n)}}\sum_{i=1}^{n}\mathbf{r}(e(X_i)/h_n)\eta_i w_i$$

has variance:

$$\begin{aligned}
\mathbb{V}[\mathbf{L}_n] &= \frac{1}{F_{e(X)}(h_n)}\int_0^{h_n} x/h_n\mathbf{r}(x/h_n)\mathbf{r}(x/h_n)'\mathbb{V}[Y|e(X)=x, D=1]F_{e(X)}(\mathrm{d}x)\\
&= (\mu_2(0) - \mu_1(0)^2)\frac{1}{F_{e(X)}(h_n)}\int_0^{h_n} x/h_n\mathbf{r}(x/h_n)\mathbf{r}(x/h_n)'F_{e(X)}(\mathrm{d}x)(1+o(1))\\
&= (\mu_2(0) - \mu_1(0)^2)\mathbb{E}[\mathbf{S}_n](1+o(1)).
\end{aligned}$$

The Lindeberg condition can easily be verified by calculating higher moments, and $\mathbf{L}_n$ will be asymptotically Gaussian provided that $nh_nF_{e(X)}(h_n) \to \infty$. We do not elaborate the argument here.

Next we consider the bias. Assuming μ_1 is twice continuously differentiable, then one has

$$\mu_1(x) = \mu_1(0) + \mu_1^{(1)}(0)x + \frac{1}{2}\mu_1^{(2)}(\tilde{x})x^2,$$

where $\tilde{x} \in [0, x]$. Now we rewrite the estimator as follows:

$$\begin{aligned}
&\left[\sum_{i=1}^{n}\mathbf{r}(e(X_i))\mathbf{r}(e(X_i))'w_i\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{r}(e(X_i))Y_iw_i\right] - \begin{bmatrix}\mu_1(0)\\ \mu_1^{(1)}(0)\end{bmatrix}\\
&= \mathbf{H}_n^{-1}\mathbf{S}_n^{-1}\left[\sqrt{\frac{1}{nh_nF_{e(X)}(h_n)}}\mathbf{L}_n + h_n^2\mathbf{R}_n\right],
\end{aligned}$$

where $\mathbf{H}_n$ is diagonal with elements 1 and h_n, and $\mathbf{R}_n$ is

$$\mathbf{R}_n = \frac{1}{nh_n^3F_{e(X)}(h_n)}\sum_{i=1}^{n}\mathbf{r}(e(X_i)/h_n)\mu_1^{(2)}(\lambda_ie(X_i)^2)e(X_i)^2w_i/2,$$

with $\lambda_i \in [0, 1]$. With the same technique applied to $\mathbf{S}_n$, one can show that

$$\left|\mathbf{R}_n - \mathbb{E}[\mathbf{R}_n]\right|^2 = \left|\mathbf{R}_n - \frac{\mu_1^{(2)}(0)}{2}\begin{bmatrix}\frac{\gamma_0-1}{\gamma_0+2}\\ \frac{\gamma_0-1}{\gamma_0+3}\end{bmatrix}(1+o(1))\right|^2 = O_{\mathrm{p}}\left(\frac{1}{nh_nF_{e(X)}(h_n)}\right).$$

■

I.8.6 Proof of Lemma I.1

Let $F_{1/e(X)}$ be the distribution function of the inverse probability weight $1/e(X)$. First consider the tail probability $\mathbb{P}[D/e(X) > x]$:

$$\begin{aligned}
&\mathbb{P}[D/e(X) > x] = \mathbb{E}[e(X)\mathbb{1}_{e(X)^{-1}>x}] = \int_x^\infty y^{-1}F_{1/e(X)}(\mathrm{d}y)\\
&= \int_x^\infty y^{-2}F_{1/e(X)}(y)\mathrm{d}y - x^{-1}F_{1/e(X)}(x)\\
&= x^{-1}\Big(1 - F_{1/e(X)}(x)\Big) - \int_x^\infty y^{-2}\Big(1 - F_{1/e(X)}(y)\Big)\mathrm{d}y\\
&= x^{-1}\Big(1 - F_{1/e(X)}(x)\Big) - \int_1^\infty x^{-1}y^{-2}\Big(1 - F_{1/e(X)}(xy)\Big)\mathrm{d}y.
\end{aligned}$$

Hence

$$\begin{aligned}
\lim_{x\to\infty}\frac{x\mathbb{P}[D/e(X) > x]}{\mathbb{P}[e(X) < x^{-1}]} &= \lim_{x\to\infty}\frac{x\mathbb{P}[D/e(X) > x]}{\mathbb{P}[e(X)^{-1} > x]} = 1 - \lim_{x\to\infty}\int_1^\infty y^{-2}\frac{1 - F_{1/e(X)}(xy)}{1 - F_{1/e(X)}(x)}\mathrm{d}y\\
&= 1 - \int_1^\infty y^{-2}y^{1-\gamma_0}\mathrm{d}y = \frac{\gamma_0 - 1}{\gamma_0}.
\end{aligned}$$

For the second line, interchanging integration and limit is permitted since the integrand is bounded by y^{-2}, which is integrable. Therefore $D/e(X)$ has regularly varying tail with index $-\gamma_0$. The rest follows from Lemma I.6. ■

I.8.7 Proof of Theorem I.1

Part (i)

We first assume $\gamma_0 > 2$ so that $DY/e(X)$ has finite variance, which is also nonzero since $\alpha_+(0) + \alpha_-(0) > 0$. Then we set $a_n = \sqrt{n\mathbb{V}[DY/e(X)]}$, which satisfies the requirement of the theorem. Then asymptotic Gaussianity follows from the central limit theorem.

Next we consider the $\gamma_0 = 2$ case, for which we compute the limits in Lemma I.8 and show that M is a point mass at the origin. Let

$$W_n = \frac{Z}{a_n}, \qquad Z = \frac{DY}{e(X)} - \theta_0,$$

and F_Z be the distribution function of Z. Without loss of generality, we assume $\alpha_+(0) > 0$, so that $DY/e(X)$ has a regularly varying right tail with index -2. First, note that for any

$0 < \eta < c$,

$$\int_0^{a_n c} x \frac{1-F_Z(x)}{a_n^2(1-F_Z(a_n))} \mathrm{d}x \geq \int_{a_n \eta}^{a_n c} x \frac{1-F_Z(x)}{a_n^2(1-F_Z(a_n))} \mathrm{d}x = \int_\eta^c x \frac{1-F_Z(a_n x)}{1-F_Z(a_n)} \mathrm{d}x \to \ln c - \ln \eta.$$

As a result, the left-hand-side diverges as $\eta > 0$ is arbitrary. Then one has $\int_0^{a_n c} y(1 - F_Z(y))\mathrm{d}y \succ a_n^2(1-F_Z(a_n))$ for any $c > 0$. Using a similar argument, we have $\int_{-a_n c}^0 y(1 - F_Z(y))\mathrm{d}y \succ a_n^2 F_Z(-a_n)$ for any $c > 0$. Now take a_n such that $n\mathbb{E}[W_n^2 \mathbb{1}_{|W_n| \leq 1}] \to 1$, then for any $c > 0$,

$$\begin{aligned} n\mathbb{E}[W_n^2 \mathbb{1}_{|W_n| \leq c}] &= \frac{n}{a_n^2} \mathbb{E}\left[Z^2 \mathbb{1}_{Z/a_n \leq c}\right] = \frac{n}{a_n^2} \int_{-a_n c}^{a_n c} x^2 F_Z(\mathrm{d}x) \\ &= n \left[c^2 F_Z(a_n c) - c^2 F_Z(-a_n c) - 2 \int_{-a_n c}^{a_n c} \frac{x}{a_n^2} F_Z(x) \mathrm{d}x\right] \\ &= n \left[-c^2(1-F_Z(a_n c)) + c^2 F_Z(-a_n c) + 2 \int_{-a_n c}^{a_n c} \frac{x}{a_n^2}(1-F_Z(x))\mathrm{d}x\right] \\ &= \left[2n \int_{-a_n c}^{a_n c} \frac{x}{a_n^2} F_Z(x) \mathrm{d}x\right](1+o(1)) \to 1. \end{aligned}$$

Therefore, we showed that for any compact interval I containing 0 in its interior, it satisfies $n\mathbb{E}[X_n \mathbb{1}_{|X_n| \leq c}] \to 1$. As a byproduct, $n(1-F_Z(a_n c)) \to 0$ and $nF_Z(-a_n c) \to 0$ for any $c > 0$. Hence the measure as in Lemma I.8 concentrates at the origin, showing that the limiting distribution is standard Gaussian.

Part (ii)

Again we assume, without loss of generality, that $\alpha_+(0) > 0$, so that $DY/e(X)$ has regularly varying right tail with index $-\gamma_0$. For $c > 0$, we compute the following:

$$\begin{aligned} n\Big(1 - F_Z(a_n c)\Big) &= \frac{1-F_Z(a_n c)}{1-F_{|Z|}(a_n)} n\Big(1 - F_{|Z|}(a_n)\Big) \\ &= \frac{1-F_Z(a_n c)}{1-F_{|Z|}(a_n)} \frac{a_n^2(1-F_{|Z|}(a_n))}{\mathbb{E}[|Z|^2 \mathbb{1}_{|Z| \leq a_n}]} \frac{n}{a_n^2} \mathbb{E}[|Z|^2 \mathbb{1}_{|Z| \leq a_n}] \\ &\to \frac{\alpha_+(0)}{\alpha_+(0)+\alpha_-(0)} \frac{2-\gamma_0}{\gamma_0} c^{-\gamma_0} = \int_c^\infty \frac{1}{x^2}\Big(\frac{(2-\gamma_0)\alpha_+(0)}{\alpha_+(0)+\alpha_-(0)} x^{1-\gamma_0}\Big)\mathrm{d}x. \end{aligned}$$

Similarly, we compute for the left tail:

$$nF_Z(-a_n c) \to \frac{\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)} \frac{2-\gamma_0}{\gamma_0} c^{-\gamma_0} = \int_c^\infty \frac{1}{x^2}\Big(\frac{(2-\gamma_0)\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)} x^{1-\gamma_0}\Big)\mathrm{d}x.$$

Therefore, we conjecture the measure M to be of the form:

$$M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(0)\mathbb{1}_{x\geq 0} + \alpha_-(0)\mathbb{1}_{x<0}\Big)\right].$$

Indeed, this is confirmed by computing the other condition in Lemma I.8. We verify for intervals $I = [c_1, c_2]$ with $c_1 > 0$,

$$\begin{aligned}
&n\mathbb{E}[X_n^2 \mathbb{1}_{|W_n|\in I}] = \frac{n}{a_n^2}\int_{a_n c_1}^{a_n c_2} x^2 F_Z(\mathrm{d}x) = n\left[c_2^2 F_Z(a_n c_2) - c_1^2 F_Z(a_n c_1) - 2\int_{c_1}^{c_2} x F_Z(a_n x)\mathrm{d}x\right] \\
&= n\left[-c_2^2\Big(1-F_Z(a_n c_2)\Big) + c_1^2\Big(1-F_Z(a_n c_1)\Big) + 2\int_{c_1}^{c_2} x\Big(1-F_Z(a_n x)\Big)\mathrm{d}x\right] \\
&= (1+o(1))\frac{2-\gamma_0}{\gamma_0}\left[-c_2^2\frac{1-F_Z(a_n c_2)}{1-F_{|Z|}(a_n)} + c_1^2\frac{1-F_Z(a_n c_1)}{1-F_{|Z|}(a_n)} + 2\int_{c_1}^{c_2} x\frac{1-F_Z(a_n x)}{1-F_{|Z|}(a_n)}\mathrm{d}x\right] \\
&\to \frac{2-\gamma_0}{\gamma_0}\frac{\alpha_+(0)}{\alpha_+(0)+\alpha_-(0)}\left[-c_2^{2-\gamma_0} + c_1^{2-\gamma_0} + 2\int_{c_1}^{c_2} x^{1-\gamma_0}\mathrm{d}x\right] \\
&= \frac{\alpha_+(0)}{\alpha_+(0)+\alpha_-(0)}\Big(c_2^{2-\gamma_0} - c_1^{2-\gamma_0}\Big) = M(I).
\end{aligned}$$

Given the measure M, the characteristic function can be found by evaluating the integral in lemma I.7, yielding

$$\begin{aligned}
&\int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x) \\
&= -|\zeta|^{\gamma_0}\frac{\Gamma(3-\gamma_0)}{\gamma_0(\gamma_0-1)}\cos\left(\frac{\gamma_0\pi}{2}\right)\left[i\frac{\alpha_+(0)-\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\mathrm{sgn}(\zeta)\tan\left(\frac{\gamma_0\pi}{2}\right) - 1\right].
\end{aligned}$$

■

I.8.8 Proof of Proposition I.1

To start,

$$\begin{aligned}
\frac{n}{a_n}\Big(\hat\theta_n - \theta_0\Big) &= \frac{1}{a_n}\sum_{i=1}^n\left(\frac{D_i Y_i}{e(X_i)} - \theta_0\right) + \frac{1}{a_n}\sum_{i=1}^n \frac{D_i Y_i}{e(X_i)}\left(\frac{e(X_i)}{\hat e(X_i)} - 1\right) \\
&= \frac{1}{a_n}\sum_{i=1}^n\left(\frac{D_i Y_i}{e(X_i)} - \theta_0\right) + \left(-\frac{1}{n}\sum_{i=1}^n \frac{D_i Y_i}{e(X_i, \tilde\pi_n)^2}\frac{\partial e(X_i, \tilde\pi_n)}{\partial \pi}\right)\frac{n}{a_n}\left(\hat\pi_n - \pi_0\right),
\end{aligned}$$

where $\tilde{\pi}_n$ is some convex combination of $\hat{\pi}_n$ and π_0, hence $|\tilde{\pi}_n - \pi_0| = O_{\mathrm{p}}(1/\sqrt{n})$. By Assumption I.3, the class

$$\left\{ \frac{D_i Y_i}{e(X_i, \pi)^2} \frac{\partial e(X_i, \pi)}{\partial \pi} \ : \ |\pi - \pi_0| \le \varepsilon \right\}$$

is Glivenko-Cantelli, hence

$$\frac{1}{n} \sum_{i=1}^{n} \frac{D_i Y_i}{e(X_i, \tilde{\pi}_n)^2} \frac{\partial e(X_i, \tilde{\pi}_n)}{\partial \pi} \xrightarrow{\mathrm{p}} \mathbb{E}\left[\frac{DY}{e(X_i)^2} \frac{\partial e(X_i, \pi_0)}{\partial \pi} \right].$$

Therefore, we have

$$\frac{n}{a_n}\left(\hat{\theta}_n - \theta_0\right) = \frac{1}{a_n} \sum_{i=1}^{n} \left(\frac{D_i Y_i}{e(X_i)} - \theta_0 - \mathbb{E}\left[\frac{\mu_1(e(X_i))}{e(X_i)} \frac{\partial e(X_i, \pi_0)}{\partial \pi} \right] h(D_i, X_i) \right) + o_{\mathrm{p}}(1).$$

For $\gamma_0 > 2$, we have $n/a_n \asymp \sqrt{n}$, and the above is asymptotically Gaussian. For the other case, the additional term in the summand is asymptotically negligible. ■

I.8.9 Proof of Lemma I.2

Consider the first step estimation problem, where the parameter π_0 is estimated by the nonlinear least squares:

$$\hat{\pi}_n = \operatorname*{argmin}_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^{n} \left| D_i - \mathfrak{L}(X_i^{\mathrm{T}} \pi) \right|^2,$$

where $\mathfrak{L}$ is the link function. Since Π is compact and $\mathfrak{L}$ is continuous in π, the class $\{ \left| D_i - \mathfrak{L}(X_i^{\mathrm{T}} \pi) \right|^2 \ : \ \pi \in \Pi \}$ is Glivenko-Cantelli with an finite envelop. Together with the assumption that π_0 is the unique minimizer of $\mathbb{E}[\left| D - \mathfrak{L}(X^{\mathrm{T}} \pi) \right|^2]$, $\hat{\pi}_n$ will be consistent for π_0. For simplicity, define

$$V = D - e(X) = D - \mathfrak{L}(X^{\mathrm{T}} \pi_0).$$

Then by a standard Taylor expansion argument,

$$\sqrt{n}\,(\hat{\pi}_n - \pi_0) = \left(\mathbb{E}\left[\mathfrak{L}^{(1)}(X^{\mathrm{T}} \pi_0)^2 X X^{\mathrm{T}} \right] \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \mathfrak{L}^{(1)}(X_i^{\mathrm{T}} \pi_0) X_i + o_{\mathrm{p}}(1),$$

provided that the inverse matrix is well-defined, and that the covariates have finite second moment $\mathbb{E}[|X|^2] < \infty$. This justifies Assumption I.3(i), with

$$h(D_i, X_i) = \left(\mathbb{E}\left[\mathfrak{L}^{(1)}(X^{\mathrm{T}}\pi_0)^2 X X^{\mathrm{T}}\right]\right)^{-1} V_i \mathfrak{L}^{(1)}(X_i^{\mathrm{T}}\pi_0) X_i.$$

■

I.8.10 Omitted Details of Remark I.5

Assumption I.3(ii) in Logit models

Note that

$$\begin{aligned}
\frac{\mathfrak{L}(X^{\mathrm{T}}\pi_0)}{\mathfrak{L}(X^{\mathrm{T}}\pi)^2}\frac{\partial}{\partial\pi}\mathfrak{L}(X^{\mathrm{T}}\pi) &= \frac{\mathfrak{L}(X^{\mathrm{T}}\pi_0)}{\mathfrak{L}(X^{\mathrm{T}}\pi)}\Big(1 - \mathfrak{L}(X^{\mathrm{T}}\pi)\Big)X \\
&= \frac{e^{X^{\mathrm{T}}\pi_0}}{e^{X^{\mathrm{T}}\pi_0}+1}\frac{e^{X^{\mathrm{T}}\pi}+1}{e^{X^{\mathrm{T}}\pi}}\frac{1}{e^{X^{\mathrm{T}}\pi}+1}X \\
&\le e^{X^{\mathrm{T}}(\pi_0-\pi)}X.
\end{aligned}$$

Then

$$\mathbb{E}\left[\sup_{|\pi-\pi_0|\le\varepsilon}\left|\frac{\mathfrak{L}(X^{\mathrm{T}}\pi_0)}{\mathfrak{L}(X^{\mathrm{T}}\pi)^2}\frac{\partial}{\partial\pi}\mathfrak{L}(X^{\mathrm{T}}\pi)\right|\right] \le \mathbb{E}\left[e^{\varepsilon|X|}X\right] \le \sqrt{\mathbb{E}[e^{2\varepsilon|X|}]\mathbb{E}[|X|^2]},$$

which will be finite if we can show that, for some small $\varepsilon > 0$, $\mathbb{E}[e^{2\varepsilon|X|}] < \infty$.

Assumption I.3(ii) in Probit models

The same argument can be applied here to show that the first step estimate has an asymptotic linear expansion. Hence we only verify Assumption I.3(ii). Note that

$$\begin{aligned}
\frac{\mathfrak{L}(X^{\mathrm{T}}\pi_0)}{\mathfrak{L}(X^{\mathrm{T}}\pi)^2}\frac{\partial}{\partial\pi}\mathfrak{L}(X^{\mathrm{T}}\pi) &= \frac{\Phi(X^{\mathrm{T}}\pi_0)\phi(X^{\mathrm{T}}\pi)}{\Phi(X^{\mathrm{T}}\pi)^2}X \\
&= \mathbb{1}_{X^{\mathrm{T}}\pi\ge-2}\frac{\Phi(X^{\mathrm{T}}\pi_0)\phi(X^{\mathrm{T}}\pi)}{\Phi(X^{\mathrm{T}}\pi)^2}X + \mathbb{1}_{X^{\mathrm{T}}\pi\le-2}\frac{\Phi(X^{\mathrm{T}}\pi_0)\phi(X^{\mathrm{T}}\pi)}{\Phi(X^{\mathrm{T}}\pi)^2}X \\
&\le \Phi(-2)^{-2}\Phi(X^{\mathrm{T}}\pi_0)\phi(X^{\mathrm{T}}\pi)X + \mathbb{1}_{X^{\mathrm{T}}\pi\le-2}\frac{\Phi(X^{\mathrm{T}}\pi_0)\phi(X^{\mathrm{T}}\pi)}{\Phi(X^{\mathrm{T}}\pi)^2}X \\
&\le \underbrace{\Phi(-2)^{-2}\Phi(X^{\mathrm{T}}\pi_0)\phi(X^{\mathrm{T}}\pi)X}_{\text{(I)}} + \underbrace{\mathbb{1}_{X^{\mathrm{T}}\pi\le-2}\frac{\phi(X^{\mathrm{T}}\pi_0)}{\phi(X^{\mathrm{T}}\pi)}\left(\frac{|X^{\mathrm{T}}\pi|^3}{|X^{\mathrm{T}}\pi|^2-1}\right)^2 X}_{\text{(II)}},
\end{aligned}$$

where for the last line, see Proposition 2.1.2 of Vershynin (2018). Term (I) is easily bounded by

$$\mathbb{E}\left[\sup_{|\pi-\pi_0|\leq\varepsilon} |(\mathrm{I})|\right] \leq \Phi(-2)^{-2}\phi(0)\mathbb{E}[|X|].$$

We can further bound (II) by

$$(\mathrm{II}) \leq 4\mathbb{1}_{X^{\mathrm{T}}\pi\leq -2}\exp\left\{\frac{1}{2}|X|^2|\pi+\pi_0||\pi-\pi_0|\right\}|X^{\mathrm{T}}\pi|^2 X,$$

Hence

$$\mathbb{E}\left[\sup_{|\pi-\pi_0|\leq\varepsilon} |(\mathrm{II})|\right] \leq 4(|\pi_0|+\varepsilon)^2\mathbb{E}\left[\exp\left\{\frac{1}{2}|X|^2\varepsilon(2|\pi_0|+\varepsilon)\right\}|X|^3\right],$$

which is finite if $\mathbb{E}[e^{\varepsilon(2|\pi_0|+\varepsilon)|X|^2}] < \infty$ for some small $\varepsilon > 0$. ■

I.8.11 Proof of Theorem I.2

Define:

$$Z = \frac{DY}{e(X)} - \theta_0, \qquad U_n = \frac{1}{a_n}\sum_{i=1}^n Z_i, \qquad V_n = \sqrt{\frac{1}{a_n^2}\sum_{i=1}^n Z_i^2}.$$

We first establish the joint limiting distribution of (U_n, V_n^2) under $\gamma_0 < 2$, which is the only interesting case. (Otherwise the self-normalized statistic is asymptotically Gaussian). The argument relies on a modification of the method in Feller 1991, Chapter XVII. To start, consider the characteristic function:

$$\begin{aligned}\mathbb{E}\left[e^{i(\zeta_1 U_n+\zeta_2 V_n^2)}\right] &= \left(\mathbb{E}\left[e^{i(\zeta_1 W_n+\zeta_2 W_n^2)}\right]\right)^n \\ &= \left(1+\frac{1}{n}\int_{\mathbb{R}}\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}nx^2F_{W_n}(\mathrm{d}x)\right)^n,\end{aligned}$$

where

$$W_n = \frac{Z}{a_n}.$$

Let $K : \mathbb{R} \to (0,\infty)$ be an auxiliary function which is smooth, symmetric, and satisfies $\lim_{x\to\infty} xK(x) = 1$.

Take $I = [c_1, c_2]$ to be a compact interval with $0 \leq c_1 < c_2$, following the same argument used to prove Theorem I.1(ii),

$$
\begin{aligned}
&\int_I K(x) n x^2 F_{W_n}(\mathrm{d}x) \\
&= n\mathbb{E}[K(W_n) W_n^2 \mathbb{1}_{W_n \in I}] \\
&= \frac{n}{a_n^2} \mathbb{E}\left[K\left(\frac{Z}{a_n}\right) Z^2 \mathbb{1}_{Z/a_n \in I} \right] \\
&= \frac{n}{a_n^2} \int_{a_n c_1}^{a_n c_2} K\left(\frac{x}{a_n}\right) x^2 \mathrm{d}F_Z(x) \\
&= n \left[K(c_2) c_2^2 F_Z(a_n c_2) - K(c_1) c_1^2 F_Z(a_n c_1) - \int_{c_1}^{c_2} \left(2xK\left(x\right) + x^2 K^{(1)}\left(x\right)\right) F_Z(a_n x) \mathrm{d}x \right] \\
&= n \Big[- K(c_2) c_2^2 \Big(1 - F_Z(a_n c_2)\Big) + K(c_1) c_1^2 \Big(1 - F_Z(a_n c_1)\Big) \\
&\qquad + \int_{c_1}^{c_2} \left(2xK\left(x\right) + x^2 K^{(1)}\left(x\right)\right) \Big(1 - F_Z(a_n x)\Big) \mathrm{d}x \Big] \\
&\to \frac{2-\gamma_0}{\gamma_0} \frac{\alpha_+(0)}{\alpha_+(0) + \alpha_-(0)} \left[-K(c_2) c_2^{2-\gamma_0} + K(c_1) c_1^{2-\gamma_0} + \int_{c_1}^{c_2} \left(2xK\left(x\right) + x^2 K^{(1)}\left(x\right)\right) x^{-\gamma_0} \mathrm{d}x \right] \\
&= M^\dagger(I),
\end{aligned}
$$

where the measure $M^\dagger(\mathrm{d}x)$ is defined as

$$
M^\dagger(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_+(0) + \alpha_-(0)} K(x) |x|^{1-\gamma_0} \Big(\alpha_+(0) \mathbb{1}_{x \geq 0} + \alpha_-(0) \mathbb{1}_{x < 0} \Big) \right].
$$

The same convergence holds for compact intervals $[c_1, c_2]$ with $c_2 \leq 0$. Finally, we note that

$$
\int_{\mathbb{R}} K(x) n x^2 F_{W_n}(\mathrm{d}x) \to M^\dagger(\mathbb{R}) \in (0, \infty).
$$

Therefore, we have the following distributional convergence:

$$
\frac{K(x) n x^2 F_{W_n}(\mathrm{d}x)}{\int_{\mathbb{R}} K(x) n x^2 F_{W_n}(\mathrm{d}x)} \overset{\mathrm{d}}{\to} \frac{M^\dagger(\mathrm{d}x)}{M^\dagger(\mathbb{R})}.
$$

Since the following is bounded and continuous of x

$$
\frac{e^{i(\zeta_1 x + \zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2 K(x)}
$$

for any $\zeta_1, \zeta_2 \in \mathbb{R}$, we have

$$\int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} n x^2 F_{W_n}(\mathrm{d}x) = \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2 K(x)} K(x) n x^2 F_{W_n}(\mathrm{d}x)$$
$$\to \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2 K(x)} M^{\dagger}(\mathrm{d}x) = \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} M(\mathrm{d}x),$$

where $M(\mathrm{d}x)$ is defined in Theorem I.1(ii). To summarize, we showed:

$$\mathbb{E}\left[e^{i(\zeta_1 U_n + \zeta_2 V_n^2)}\right] \to \exp\left\{\int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} M(\mathrm{d}x)\right\}.$$

A similar result was derived in Logan, Mallows, Rice and Shepp (1973). However, our argument only relies on the fact that Z has a regularly varying tail, while they impose the stronger assumption that Z follows a Lévy stable distribution. Given the joint limiting characteristic function, Logan, Mallows, Rice and Shepp (1973) showed that the limiting distribution does not have positive mass on $\mathbb{R} \times \{0\}$, implying that U_n/V_n has a well-defined limiting distribution. Further, the limiting distribution has a smooth density function.

For the self-normalized statistic T_n in Theorem I.2, we rely on Proposition I.1, which claims that estimating the probability weights in a first step does not contribute to the limiting distribution when $\gamma_0 < 2$. Then with simple algebra,

$$T_n = \frac{U_n}{V_n}\sqrt{\frac{n-1}{n - V_n^2}}.$$

As a result, T_n has the same limiting distribution as U_n/V_n. Therefore, subsampling is valid by standard arguments in Politis and Romano (1994) (or Romano and Wolf 1999). ■

I.8.12 Proof of Theorem I.3

Part (i)

Take $c > 0$ and first consider the following probability:

$$\int_0^{b_n} x \mathbb{P}[Y > a_n c x | e(X) = x, D = 1] F_{e(X)}(\mathrm{d}x) \leq \int_0^{b_n} x F_{e(X)}(\mathrm{d}x) = \mathbb{P}\left[\frac{D}{e(X)} > b_n^{-1}\right].$$

If $a_n b_n \to 0$, the right-hand-side will be asymptotically negligible compared to $\mathbb{P}[D/e(X) > a_n c]$ for any $c > 0$. As a result, we have for $a_n b_n \to 0$,

$$\frac{\mathbb{P}\left[\frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n} > a_n c\right]}{\mathbb{P}\left[\frac{D}{e(X)} > a_n c\right]} = \frac{1}{\mathbb{P}\left[\frac{D}{e(X)} > a_n c\right]}\int_{b_n}^{1} x\mathbb{P}[Y > a_n cx | e(X) = x, D = 1]F_{e(X)}(\mathrm{d}x)$$

$$= \frac{1}{\mathbb{P}\left[\frac{D}{e(X)} > a_n c\right]}\Big[\int_{0}^{1} x\mathbb{P}[Y > a_n cx | e(X) = x, D = 1]F_{e(X)}(\mathrm{d}x)$$

$$- \int_{0}^{b_n} x\mathbb{P}[Y > a_n cx | e(X) = x, D = 1]F_{e(X)}(\mathrm{d}x)\Big]$$

$$= o(1) + \frac{\mathbb{P}\left[\frac{DY}{e(X)} > a_n c\right]}{\mathbb{P}\left[\frac{D}{e(X)} > a_n c\right]} \to \alpha_+(0),$$

as claimed in Lemma I.1. Therefore, the same Lévy stable limiting distribution emerges under light trimming.

Part (ii)

First note that $nb_n^2\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq b_n^{-1}}] \to \infty$ and that S_n has unit variance. Hence we only need to verify the Lindeberg condition.

$$\frac{n}{a_{n,b_n}^{2+\eta}}\mathbb{E}\left[\frac{DY^{2+\eta}}{e(X)^{2+\eta}}\mathbb{1}_{e(X)\geq b_n}\right] \leq C\frac{n}{a_{n,b_n}^{2+\eta}}\mathbb{E}\left[\frac{1}{e(X)^{1+\eta}}\mathbb{1}_{e(X)\geq b_n}\right] = C\frac{n}{a_{n,b_n}^{2+\eta}}\int_1^{1/b_n} x^{1+\eta}F_{1/e(X)}(\mathrm{d}x)$$

$$\leq C' n^{-\eta/2}\left[\int_1^{1/b_n} x^{1+\eta}F_{1/e(X)}(\mathrm{d}x)\right]\left[\int_1^{1/b_n} xF_{1/e(X)}(\mathrm{d}x)\right]^{-2-\eta}$$

$$= C'\left[\frac{\int_1^{1/b_n} x^{1+\eta}F_{1/e(X)}(\mathrm{d}x)}{b_n^{-1-\eta}\mathbb{P}[e(X)\leq b_n]}\right]\left[\frac{\int_1^{1/b_n} xF_{1/e(X)}(\mathrm{d}x)}{b_n^{-1}\mathbb{P}[e(X)\leq b_n]}\right]^{-2-\eta}\frac{n\mathbb{P}[|DY/e(X)|\geq b_n^{-1}]}{nb_n\mathbb{P}[e(X)\leq b_n]}$$

$$\frac{\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq b_n^{-1}}]}{b_n^{-2}\mathbb{P}[|DY/e(X)|\geq b_n^{-1}]}\frac{1}{nb_n^2\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq b_n^{-1}}]} \to 0,$$

by Lemma I.5.

Part (iii)

Again we ignore the centering, since it is irrelevant for computing the tail probabilities or truncated moments. Let F_U be the limiting distribution of $F_{Y|e(X)=x,D=1}$ as $x \to 0$, $U \perp\!\!\!\perp (X, Y)$ be distributed according to F_U, and $c > 0$. We first compute the following limit:

$$
\begin{aligned}
&\lim_{n\to\infty} n\mathbb{P}\left[\frac{DU}{e(X)}\mathbb{1}_{e(X)\geq ta_n^{-1}} > a_n c\right] = n\int_0^\infty \mathbb{P}\left[\frac{D}{e(X)}\mathbb{1}_{e(X)\geq ta_n^{-1}} > \frac{a_n c}{x}\right] F_U(\mathrm{d}x) \\
&= \lim_{n\to\infty} n\int_0^\infty \int_{t/a_n}^{x/(a_n c)} yF_{e(X)}(\mathrm{d}y)F_U(\mathrm{d}x) = \lim_{n\to\infty} n\int_{ct}^\infty \int_{t/a_n}^{x/(a_n c)} yF_{e(X)}(\mathrm{d}y)F_U(\mathrm{d}x) \\
&= \lim_{n\to\infty} n\int_{ct}^\infty \left[\frac{x}{a_n c}F_{e(X)}\left(\frac{x}{a_n c}\right) - \frac{t}{a_n}F_{e(X)}\left(\frac{t}{a_n}\right) - \int_{t/a_n}^{x/(a_n c)} F_{e(X)}(y)\mathrm{d}y\right] F_U(\mathrm{d}x) \\
&= \lim_{n\to\infty} n\int_{ct}^\infty \left[\frac{x}{a_n c}F_{e(X)}\left(\frac{x}{a_n c}\right) - \frac{t}{a_n}F_{e(X)}\left(\frac{t}{a_n}\right) - \frac{1}{a_n}\int_{t}^{x/c} F_{e(X)}\left(\frac{y}{a_n}\right)\mathrm{d}y\right] F_U(\mathrm{d}x) \\
&= \lim_{n\to\infty} \left[\frac{nF_{e(X)}(a_n)}{a_n}\right] \\
&\quad \left[\int_{ct}^\infty \left[\frac{x}{c}\frac{F_{e(X)}\left(x/(a_n c)\right)}{F_{e(X)}\left(1/a_n\right)} - t\frac{F_{e(X)}\left(t/a_n\right)}{F_{e(X)}\left(1/a_n\right)} - \int_t^{x/c}\frac{F_{e(X)}(y/a_n)}{F_{e(X)}(1/a_n)}\mathrm{d}y\right] F_U(\mathrm{d}x)\right] \\
&= \lim_{n\to\infty} \left[\frac{nF_{e(X)}(a_n)}{a_n}\right] \left[\int_{ct}^\infty \left[\left(\frac{x}{c}\right)^{\gamma_0} - t^{\gamma_0} - \int_t^{x/c} y^{\gamma_0 - 1}\mathrm{d}y\right] F_U(\mathrm{d}x)\right] \\
&= \frac{\gamma_0 - 1}{\gamma_0}\lim_{n\to\infty} \left[\frac{nF_{e(X)}(a_n)}{a_n}\right] \left[\int_{ct}^\infty \left[\left(\frac{x}{c}\right)^{\gamma_0} - t^{\gamma_0}\right] F_U(\mathrm{d}x)\right].
\end{aligned}
$$

Finally we note that

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{nF_{e(X)}(a_n)}{a_n} = \lim_{n\to\infty} n\mathbb{P}[|DY/e(X)| > a_n]\frac{F_{e(X)}(a_n)}{a_n\mathbb{P}[|DY/e(X)| > a_n]} \\
&= \lim_{n\to\infty} \frac{n}{a_n^2}\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq a_n}]\frac{a_n^2\mathbb{P}[|DY/e(X)| > a_n]}{\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq a_n}]}\frac{F_{e(X)}(a_n)}{a_n\mathbb{P}[|DY/e(X)| > a_n]} \\
&= \frac{2-\gamma_0}{\gamma_0 - 1}\frac{1}{\alpha_+(0) + \alpha_-(0)}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\lim_{n\to\infty} n\mathbb{P}\left[\frac{DU}{e(X)}\mathbb{1}_{e(X)\geq ta_n^{-1}} > a_n c\right] &= \frac{2-\gamma_0}{\gamma_0}\frac{1}{\alpha_+(0) + \alpha_-(0)}\left[\int_{ct}^\infty \left[\left(\frac{x}{c}\right)^{\gamma_0} - t^{\gamma_0}\right] F_U(\mathrm{d}x)\right] \\
&= \int_c^\infty \frac{1}{x^2}\left[\frac{2-\gamma_0}{\alpha_+(0) + \alpha_-(0)}x^{1-\gamma_0}\alpha_+(tx)\right]\mathrm{d}x.
\end{aligned}
$$

Similarly, we can obtain, for the left tail, that

$$
\lim_{n\to\infty} n\mathbb{P}\left[\frac{DU}{e(X)}\mathbb{1}_{e(X)\geq ta_n^{-1}} < -a_n c\right] = \int_c^\infty \frac{1}{x^2}\left[\frac{2-\gamma_0}{\alpha_+(0) + \alpha_-(0)}x^{1-\gamma_0}\alpha_-(tx)\right]\mathrm{d}x,
$$

where F_{-U} is the distribution function of $-U$. Define a measure M as

$$M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)} |x|^{1-\gamma_0} \Big(\alpha_+(tx) \mathbb{1}_{x\geq 0} + \alpha_-(tx) \mathbb{1}_{x<0} \Big) \right],$$

and we verify the other condition in Lemma I.8. For simplicity, take $I = [c_1, c_2]$ with $0 < c_1 < c_2$ and $t = 1$. Then the truncated second moment is

$$\begin{aligned}
&\frac{n}{a_n^2} \mathbb{E} \left[\frac{DU^2}{e(X)^2} \mathbb{1}_{e(X)\geq a_n^{-1}} \mathbb{1}_{DU/e(X)\mathbb{1}_{e(X)\geq a_n^{-1}} \in [a_n c_1, a_n c_2]} \right] \\
&= \frac{n}{a_n^2} \int_{-\infty}^{\infty} \int_0^1 \mathbb{1}_{x\geq a_n^{-1}} \mathbb{1}_{x\in[u/(a_n c_2), u/(a_n c_1)]} \frac{u^2}{x} F_{e(X)}(\mathrm{d}x) F_U(\mathrm{d}u) \\
&= \frac{n}{a_n^2} \int_{c_1}^{\infty} \int_{((u/c_2)\vee 1)/a_n}^{u/(a_n c_1)} \frac{u^2}{x} F_{e(X)}(\mathrm{d}x) F_U(\mathrm{d}u) \\
&= \frac{n}{a_n^2} \int_{c_1}^{\infty} u^2 \left[\frac{F_{e(X)}(u/(a_n c_1))}{u/(a_n c_1)} - \frac{F_{e(X)}(((u/c_2)\vee 1)/a_n)}{((u/c_2)\vee 1)/a_n} + \int_{((u/c_2)\vee 1)/a_n}^{u/(a_n c_1)} \frac{1}{x^2} F_{e(X)}(x)\mathrm{d}x \right] F_U(\mathrm{d}u) \\
&= n \int_{c_1}^{\infty} u^2 \Big[\frac{F_{e(X)}(u/(a_n c_1))}{a_n u/c_1} - \frac{F_{e(X)}(((u/c_2)\vee 1)/a_n)}{a_n((u/c_2)\vee 1)} \\
&\qquad\qquad + \int_{((u/c_2)\vee 1)}^{u/c_1} \frac{1}{a_n x^2} F_{e(X)}(x/a_n)\mathrm{d}x \Big] F_U(\mathrm{d}u) \\
&= (1+o(1)) \frac{2-\gamma_0}{\gamma_0 - 1} \frac{1}{\alpha_+ + \alpha_-} \int_{c_1}^{\infty} u^2 \Big[\frac{1}{u/c_1} \frac{F_{e(X)}(u/(a_n c_1))}{F_{e(X)}(1/a_n)} - \frac{1}{(u/c_2)\vee 1} \frac{F_{e(X)}(((u/c_2)\vee 1)/a_n)}{F_{e(X)}(1/a_n)} \\
&\qquad\qquad + \int_{((u/c_2)\vee 1)}^{u/c_1} \frac{1}{x^2} \frac{F_{e(X)}(x/a_n)}{F_{e(X)}(1/a_n)} \mathrm{d}x \Big] F_U(\mathrm{d}u) \\
&\to \frac{2-\gamma_0}{\gamma_0 - 1} \frac{1}{\alpha_+(0) + \alpha_-(0)} \int_{c_1}^{\infty} u^2 \left[(u/c_1)^{\gamma_0 - 2} - ((u/c_2)\vee 1)^{\gamma_0 - 2} + \int_{((u/c_2)\vee 1)}^{u/c_1} x^{\gamma_0 - 3} \mathrm{d}x \right] F_U(\mathrm{d}u) \\
&= -\frac{1}{\alpha_+(0) + \alpha_-(0)} \int_{c_1}^{\infty} u^2 \left[(u/c_1)^{\gamma_0 - 2} - ((u/c_2)\vee 1)^{\gamma_0 - 2} \right] F_U(\mathrm{d}u) \\
&= -\frac{1}{\alpha_+(0) + \alpha_-(0)} \left[\int_{c_1}^{c_2} \frac{u^{\gamma_0}}{c_1^{\gamma_0 - 2}} - u^2 F_U(\mathrm{d}u) + \int_{c_2}^{\infty} \frac{u^{\gamma_0}}{c_1^{\gamma_0 - 2}} - \frac{u^{\gamma_0}}{c_2^{\gamma_0 - 2}} F_U(\mathrm{d}u) \right],
\end{aligned}$$

which, by simple algebra, can be shown to be the same as $M([c_1, c_2])$. The next step is to replace $DU/e(X)$ by $DY/e(X)$. The same argument used to proved Lemma I.6 applies here, which we do not repeat. ■

I.8.13 Proof of Proposition I.2

To start,

$$
\begin{aligned}
&\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i,\hat{\pi}_n)}\mathbb{1}_{e(X_i,\hat{\pi}_n)\geq b_n}-\theta_0-\mathsf{B}_{n,b_n}\right)\\
&=\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i,\pi_0)}\mathbb{1}_{e(X_i,\pi_0)\geq b_n}-\theta_0-\mathsf{B}_{n,b_n}\right) && \text{(I)}\\
&+\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i,\hat{\pi}_n)}\mathbb{1}_{e(X_i,\hat{\pi}_n)\geq b_n}-\frac{D_iY_i}{e(X_i,\pi_0)}\mathbb{1}_{e(X_i,\pi_0)\geq b_n}\right), && \text{(II)}
\end{aligned}
$$

where asymptotic properties of (I) is discussed in Theorem I.3. For (II), we further expand it as

$$
\begin{aligned}
\text{(II)}=&\underbrace{\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i,\hat{\pi}_n)}-\frac{D_iY_i}{e(X_i,\pi_0)}\right)\mathbb{1}_{e(X_i,\hat{\pi}_n)\geq b_n}}_{\text{(II.1)}}\\
&+\underbrace{\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\frac{D_iY_i}{e(X_i,\pi_0)}\left(\mathbb{1}_{e(X_i,\hat{\pi}_n)\geq b_n}-\mathbb{1}_{e(X_i,\pi_0)\geq b_n}\right)}_{\text{(II.2)}}.
\end{aligned}
$$

By the same argument used in Proposition I.1, it satisfies

$$
\text{(II.1)}=-\frac{1}{a_n}\sum_{i=1}^{n}A_0h(D_i,X_i)+o_{\mathrm{p}}(1).
$$

For (II.2), we first make some auxiliary calculations. Take π be a generic element in the parameter space Π,

$$
\frac{e(X_i,\pi)}{e(X_i,\pi_0)}-1=\frac{1}{e(X_i,\pi_0)}\frac{\partial e(X_i,\tilde{\pi})}{\partial\pi}(\pi-\pi_0),
$$

where $\tilde{\pi}$ is some convex combination of π and π_0. Next define

$$
Z_i(\varepsilon)=\sup_{|\pi-\pi_0|\leq\varepsilon}\left|\frac{1}{e(X_i,\pi_0)}\frac{\partial e(X_i,\pi)}{\partial\pi}\right|.
$$

Then we have

$$
\left|\mathbb{1}_{e(X_i,\pi)\geq b_n}-\mathbb{1}_{e(X_i,\pi_0)\geq b_n}\right|\leq\mathbb{1}_{\frac{b_n}{1+Z_i(\varepsilon)\varepsilon}\leq e(X_i,\pi_0)\leq\frac{b_n}{1-Z_i(\varepsilon)\varepsilon}}+\mathbb{1}_{|\pi-\pi_0|>\varepsilon}.
$$

Now fix some $K > 0$ and let $\varepsilon = K/\sqrt{n}$ in the above, we have

$$|(\mathrm{II}.2)| \leq \underbrace{\frac{1}{a_{n,b_n}} \sum_{i=1}^{n} \frac{D_i|Y_i|}{e(X_i,\pi_0)} \mathbb{1}_{\frac{b_n}{1+Z_i(\frac{K}{\sqrt{n}})\frac{K}{\sqrt{n}}} \leq e(X_i,\pi_0) \leq \frac{b_n}{1-Z_i(\frac{K}{\sqrt{n}})\frac{K}{\sqrt{n}}}}}_{(\mathrm{II}.2.1)} + \underbrace{(\mathrm{II}.2)\mathbb{1}_{|\hat{\pi}_n - \pi_0| > \frac{K}{\sqrt{n}}}}_{(\mathrm{II}.2.2)}.$$

Now take a sequence c_n, we expand (II.2.1) as

$$|(\mathrm{II}.2.1)| \leq \underbrace{\frac{1}{a_{n,b_n}} \sum_{i=1}^{n} \frac{D_i|Y_i|}{e(X_i,\pi_0)} \mathbb{1}_{\frac{b_n}{1+\frac{K}{\sqrt{n}}c_n} \leq e(X_i,\pi_0) \leq \frac{b_n}{1-\frac{K}{\sqrt{n}}c_n}}}_{(\mathrm{II}.2.1.1)} + \underbrace{(\mathrm{II}.2.1)\mathbb{1}_{\max_{1\leq i\leq n} Z_i(\frac{K}{\sqrt{n}}) > c_n}}_{(\mathrm{II}.2.1.2)}.$$

Further,

$$\begin{aligned}
\mathbb{E}[|(\mathrm{II}.2.1.1)|] &\precsim \frac{n}{a_{n,b_n}} \left[F_{e(X)}\left(\frac{b_n}{1 - \frac{K}{\sqrt{n}}c_n} \right) - F_{e(X)}\left(\frac{b_n}{1 + \frac{K}{\sqrt{n}}c_n} \right) \right] \\
&\precsim \frac{n}{a_{n,b_n}} F_{e(X)}(b_n) \left[\left(1 + 2\frac{\frac{K}{\sqrt{n}}c_n}{1 - \frac{K}{\sqrt{n}}c_n} \right)^{\gamma_0 - 1} - 1 \right] \asymp \frac{n}{a_{n,b_n}} F_{e(X)}(b_n) \frac{K}{\sqrt{n}} c_n.
\end{aligned}$$

From Lemma I.3, the above becomes

$$K\sqrt{b_n \mathbb{P}[e(X) \leq b_n]}c_n \to 0.$$

Put all pieces together, we have for any $\varrho > 0$

$$\limsup_n \mathbb{P}\left[|(\mathrm{II}.2)| > \varrho \right] \to \limsup_n \mathbb{P}\left[|\hat{\pi}_n - \pi_0| > \frac{K}{\sqrt{n}} \right],$$

since only (II.2.2) can be non-degenerate. The left-hand-side is independent of K and the right-hand-side decreases to 0 as $K \uparrow \infty$, we have that (II.2) converges in probability to zero. ■

I.8.14 Omitted Details of Remark I.6

Bounding c_n in Logit models

Let

$$Z_i(\varepsilon) = \sup_{|\pi - \pi_0| \leq \varepsilon} \left| \frac{e(X_i,\pi)}{e(X_i,\pi_0)}(1 - e(X_i,\pi))X_i \right|,$$

for which it suffices to consider (see the proof of Proposition I.2)

$$Z_i(\varepsilon) = \sup_{|\pi-\pi_0|\le\varepsilon} e^{|X_i|\cdot|\pi-\pi_0|}|X_i| = e^{\varepsilon|X_i|}|X_i|.$$

By our assumption, X_i is sub-exponential, hence

$$\begin{aligned}\max_{1\le i\le n} Z_i\left(\frac{K}{\sqrt{n}}\right) &\le \left(\max_{1\le i\le n} e^{\frac{K}{\sqrt{n}}|X_i|}\right)\left(\max_{1\le i\le n}|X_i|\right)\\ &= O_{\mathrm{p}}\left(n^{\frac{K}{\varepsilon\sqrt{n}}}\right) O_{\mathrm{p}}\left(\log(n)\right) = O_{\mathrm{p}}(\log(n)).\end{aligned}$$

Bounding c_n in Probit models

Let

$$Z_i(\varepsilon) = \sup_{|\pi-\pi_0|\le\varepsilon} \frac{\phi(X_i^{\mathrm{T}}\pi)\phi(X_i^{\mathrm{T}}\pi_0)}{\phi(X_i^{\mathrm{T}}\pi_0)\Phi(X_i^{\mathrm{T}}\pi_0)}\left|X_i\right|.$$

Again for our purposes, it suffices to consider $X_i^{\mathrm{T}}\pi_0 \ll 0$, hence

$$Z_i(\varepsilon) = \sup_{|\pi-\pi_0|\le\varepsilon} \frac{\phi(X_i^{\mathrm{T}}\pi)}{\phi(X_i^{\mathrm{T}}\pi_0)}|X_i|^3 = e^{\frac{1}{2}|X_i|^2\varepsilon(2|\pi_0|+\varepsilon)}|X_i|^3.$$

By our assumption, X_i is sub-Gaussian, hence

$$\max_{1\le i\le n} Z_i\left(\frac{K}{\sqrt{n}}\right) \le \left(\max_{1\le i\le n} e^{\frac{1}{2}|X_i|^2\frac{K}{\sqrt{n}}(2|\pi_0|+\frac{K}{\sqrt{n}})}\right)\left(\max_{1\le i\le n}|X_i|^3\right) = O_{\mathrm{p}}\left(\log(n)^{\frac{3}{2}}\right).$$

■

I.8.15 Proof of Lemma I.3

The bias of $\hat{\theta}_{n,b_n}$ is quite easy to derive. Note that the IPW estimator $\hat{\theta}_n$ is unbiased for θ_0, hence the bias can be written as the following expectation:

$$\begin{aligned}\mathsf{B}_{n,b_n} = \mathbb{E}[\hat{\theta}_{n,b_n}] - \theta_0 &= -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{e(X_i)}\mathbb{1}_{e(X_i)\le b_n}\right]\\ &= -\mathbb{E}\Big[\mathbb{E}[Y|e(X), D=1]\mathbb{1}_{e(X)<b_n}\Big] \approx -\mu_1(0)\cdot\mathbb{P}[e(X)\le b_n],\end{aligned}$$

so that the leading bias vanishes at the rate $\mathbb{P}[e(X) \leq b_n]$, unless the data generating process is that the conditional mean shrinks as the probability weight approaches zero[3].

For the variance of $DY/e(X)\mathbb{1}_{e(X)\geq b_n}$, we note that when $\gamma_0 \in (1,2)$ and $b_n \to 0$, it diverges to infinity. As a result,

$$\begin{aligned}\mathsf{V}_{n,b_n} &= \frac{1}{n}\mathbb{V}\left[\frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n}\right] \approx \frac{1}{n}\mathbb{E}\left[\frac{DY^2}{e(X)^2}\mathbb{1}_{e(X)\geq b_n}\right]\\ &= \frac{1}{n}\int_{b_n}^{1}\frac{\mathbb{E}[Y^2|e(X)=x, D=1]}{x}\mathrm{d}\mathbb{P}[e(X)\leq x].\end{aligned}$$

As one may suspect, the behavior of the above integral is not “sensitive” to the conditional second moment of Y, since what matters is the tail behavior of the probability weight.

To simplify notation, let $a = \lim_{y\to 0}\mathbb{E}[Y^2|e(X)=y, D=1]$. Choose $c > 0$ small enough so that

$$\sup_{x\leq c}\Big|\mathbb{E}[Y^2|e(X)=x, D=1] - a\Big| \leq \eta.$$

Then

$$\frac{\int_{b_n}^{1} a x^{-1}F_{e(X)}(\mathrm{d}x)}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} = 1 + \frac{A+B-C}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)},$$

where

$$\begin{aligned}A &= \int_{c}^{1} a x^{-1}F_{e(X)}(\mathrm{d}x)\\ B &= \int_{b_n}^{c}\Big(a - \mathbb{E}[Y^2|e(X)=x, D=1]\Big)x^{-1}F_{e(X)}(\mathrm{d}x)\\ C &= \int_{c}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x).\end{aligned}$$

Note that

$$\frac{A}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} \to 0, \quad \frac{C}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} \to 0.$$

[3] In $\hat{\theta}_{n,b_n}$ we use the entire sample size n for normalization, rather than the effective number of observations $n_{b_n} = \sum_{i=1}^{n}\mathbb{1}_{e(X)\geq b_n}$. We note that even when n_{b_n} is used, the order of bias does not change, unless one has $\lim_{x\to 0}\mathbb{E}[Y|e(X)=x, D=1] = \theta_0$, so that the limiting conditional expectation equals exactly the target parameter.

For B, we have

$$\frac{B}{\int_{b_n}^{1} \mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} \le \frac{\eta}{\inf_{x\in[0,c]} \mathbb{E}[Y^{2,D=1}|e(X)=x]},$$

which can be made arbitrarily small. Hence

$$\frac{\int_{b_n}^{1} ax^{-1}F_{e(X)}(\mathrm{d}x)}{\int_{b_n}^{1} \mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} \to 1.$$

For the final claim, we first note, by a slight modification of Lemma I.5,

$$\frac{b_n^{-1}\mathbb{P}[e(X)\le b_n]}{\mathbb{E}[e(X)^{-1}\mathbb{1}_{e(X)\ge b_n}]} \to \frac{2-\gamma_0}{\gamma_0 - 1},$$

as $b_n \to 0$, from which the desired result follows. ■

I.8.16 Proof of Theorem I.4

Let $\hat{F}_{e(X)}(x) = \sum_{i=1}^{n} \mathbb{1}_{e(X)\le x}/n$. We first consider the behavior of $b^s\hat{F}_{e(X)}(b)$ at b_n (defined in the theorem), which is given by the following probability bound (Markov's inequality):

$$\begin{aligned}
\mathbb{P}\left[n\left|b_n^s\hat{F}_{e(X)}(b_n) - b_n^sF_{e(X)}(b_n)\right| > \delta\right] &\le n^2\left(\frac{b_n^s}{\delta}\right)^2 \mathbb{E}\left|\hat{F}_{e(X)}(b_n) - F_{e(X)}(b_n)\right|^2 \\
&= n\left(\frac{b_n^s}{\delta}\right)^2 \mathbb{V}\left[\mathbb{1}_{e(X)\le b_n}\right] \\
&= n\left(\frac{b_n^s}{\delta}\right)^2 F_{e(X)}(b_n)(1-F_{e(X)}(b_n)) \\
&= \frac{c_0}{\delta^2}b_n^s\Big(1+o(1)\Big),
\end{aligned}$$

which implies

$$n\left|b_n^s\hat{F}_{e(X)}(b_n) - b_n^sF_{e(X)}(b_n)\right| \xrightarrow{\mathrm{p}} 0.$$

To complete the proof, take some constant $a\in(0,1)$, and define $b_{l,n}$ and $b_{r,n}$ as:

$$b_{l,n}^sF_{e(X)}(b_{l,n}) = \frac{ac_0}{n}, \quad b_{r,n}^sF_{e(X)}(b_{r,n}) = \frac{c_0}{an}.$$

Then it is easy to see that

$$\begin{aligned}
\mathbb{P}\left[\hat{b}_n \leq b_{l,n}\right] &\leq \mathbb{P}\left[b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) \geq \hat{b}_n^s \hat{F}_{e(X)}(\hat{b}_n)\right] = \mathbb{P}\left[b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) \geq \frac{\hat{c}_n}{n}\right] \\
&= \mathbb{P}\left[b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) - b_{l,n}^s F_{e(X)}(b_{l,n}) \geq \frac{(1-a)c_0 + (\hat{c}_n - c_0)}{n}\right] \\
&= \mathbb{P}\Big[n\left(b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) - b_{l,n}^s F_{e(X)}(b_{l,n})\right) \geq \underbrace{(1-a)c_0 + (\hat{c}_n - c_0)}_{\xrightarrow{\mathrm{P}}(1-a)c_0>0}\Big] \to 0,
\end{aligned}$$

since the first term $n\left(b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) - b_{l,n}^s F_{e(X)}(b_{l,n})\right)$ is $o_p(1)$. Using a similar technique, we can show that $\mathbb{P}[\hat{b}_n \geq b_{r,n}] \to 0$. Therefore,

$$\mathbb{P}\left[b_{l,n} \leq \hat{b}_n \leq b_{r,n}\right] = \mathbb{P}\left[\frac{b_{l,n}}{b_n} \leq \frac{\hat{b}_n}{b_n} \leq \frac{b_{r,n}}{b_n}\right] \to 1.$$

Since the choice of a is arbitrary, we only need to show that both $b_{l,n}/b_n$ and $b_{r,n}/b_n$ are arbitrarily close to 1 for all a close to 1. To see this, note that since $b_n \to 0$, one has

$$a = \frac{b_{l,n}^s F_{e(X)}(b_{l,n})}{b_n^s F_{e(X)}(b_n)} = \underbrace{\frac{b_{l,n}^s}{b_n^s} \frac{F_{e(X)}((b_{l,n}/b_n)b_n)}{F_{e(X)}(b_n)}}_{\to(b_{l,n}/b_n)^{\gamma_0-1}} = \left(\frac{b_{l,n}}{b_n}\right)^{\gamma_0-1+s}(1+o(1)).$$

and the same argument applies to $b_{r,n}$.

To show that estimated probability weights can be employed, we only need to show that for all $\delta > 0$,

$$\mathbb{P}\left[n\left|b_n^s \hat{F}_{\hat{e}(X)}(b_n) - b_n^s \hat{F}_{e(X)}(b_n)\right| > \delta\right] \to 0,$$

where again b_n is defined in the theorem. From the proof of Proposition I.2, we have, for any $|\pi - \pi_0| \leq \varepsilon$,

$$\left|\mathbb{1}_{e(X_i,\pi)\geq b_n} - \mathbb{1}_{e(X_i,\pi_0)\geq b_n}\right| \leq \mathbb{1}_{\frac{b_n}{1+Z_i(\varepsilon)\varepsilon} \leq e(X_i,\pi_0) \leq \frac{b_n}{1-Z_i(\varepsilon)\varepsilon}},$$

and

$$Z_i(\varepsilon) = \sup_{|\pi-\pi_0|\leq\varepsilon}\left|\frac{1}{e(X_i,\pi_0)}\frac{\partial e(X_i,\pi)}{\partial \pi}\right|.$$

Therefore, for any $K > 0$,

$$\begin{aligned}
&\mathbb{P}\left[n\left|b_n^s\hat{F}_{\hat{e}(X)}(b_n) - b_n^s\hat{F}_{e(X)}(b_n)\right| > \delta\right] \\
&\leq \mathbb{P}\left[b_n^s\sum_{i=1}^{n}\mathbb{1}_{\frac{b_n}{1+Z_i(\frac{K}{\sqrt{n}})\frac{K}{\sqrt{n}}}\leq e(X_i,\pi_0)\leq\frac{b_n}{1-Z_i(\frac{K}{\sqrt{n}})\frac{K}{\sqrt{n}}}} > \delta\right] + \mathbb{P}\left[|\hat{\pi}_n - \pi_0| \geq \frac{K}{n}\right] \\
&\leq \mathbb{P}\left[b_n^s\sum_{i=1}^{n}\mathbb{1}_{\frac{b_n}{1+\frac{K}{\sqrt{n}}c_n}\leq e(X_i,\pi_0)\leq\frac{b_n}{1-\frac{K}{\sqrt{n}}c_n}} > \delta\right] + \mathbb{P}\left[\max_{1\leq i\leq n} Z_i(\frac{K}{\sqrt{n}}) > c_n\right] + \mathbb{P}\left[|\hat{\pi}_n - \pi_0| \geq \frac{K}{n}\right],
\end{aligned}$$

and c_n is to be specified. For the first term, one has

$$\begin{aligned}
&\mathbb{E}\left[b_n^s\sum_{i=1}^{n}\mathbb{1}_{\frac{b_n}{1+\frac{K}{\sqrt{n}}c_n}\leq e(X_i,\pi_0)\leq\frac{b_n}{1-\frac{K}{\sqrt{n}}c_n}}\right] \\
&= nb_n^s\left[F_{e(X)}\left(\frac{b_n}{1-\frac{K}{\sqrt{n}}c_n}\right) - F_{e(X)}\left(\frac{b_n}{1+\frac{K}{\sqrt{n}}c_n}\right)\right] \\
&\precsim nb_n^sF_{e(X)}(b_n)\left[\left(1+2\frac{\frac{K}{\sqrt{n}}c_n}{1-\frac{K}{\sqrt{n}}c_n}\right)^{\gamma_0-1} - 1\right] \\
&\asymp nb_n^sF_{e(X)}(b_n)\frac{K}{\sqrt{n}}c_n \asymp \frac{K}{\sqrt{n}}c_n \to 0,
\end{aligned}$$

which holds if $c_n = \sqrt{n/\log(n)}$. By our assumption,

$$\mathbb{P}\left[\max_{1\leq i\leq n} Z_i\left(\frac{K}{\sqrt{n}}\right) > c_n\right] \to 0.$$

Finally,

$$\mathbb{P}\left[|\hat{\pi}_n - \pi_0| \geq \frac{K}{n}\right]$$

can be made arbitrarily small by taking K large. Since

$$\mathbb{P}\left[n\left|b_n^s\hat{F}_{\hat{e}(X)}(b_n) - b_n^s\hat{F}_{e(X)}(b_n)\right| > \delta\right]$$

does not depend on K, this probability converges to 0 for all $\delta > 0$. ■

I.8.17 Proof of Theorem I.5

We assume the true probability weights are used in the local polynomial regression, as estimating the probability weights in a first step does not have a first order contribution to the local polynomial regression. We first consider a (trivial) situation where $nF_{e(X)}(b_n) \to 0$. This clearly falls into the light trimming scenario of Theorem I.3(i). To show that our bias correction does not contribute to the limiting distribution, note that

$$\frac{n}{a_{n,b_n}}|\hat{\mathsf{B}}_{n,b_n} - \mathsf{B}_{n,b_n}| \le \underbrace{\frac{n}{a_{n,b_n}}|\mathsf{B}_{n,b_n}|}_{o_{\mathrm{p}}(1),\text{ due to light trimming}} + \frac{n}{a_{n,b_n}}\left|\hat{\mathsf{B}}_{n,b_n}\right|.$$

The second term has expansion

$$\frac{n}{a_{n,b_n}}\left|\hat{\mathsf{B}}_{n,b_n}\right| \le \underbrace{\left|\sum_{j=0}^{p}\hat{\beta}_j\right|}_{O_{\mathrm{p}}(1),\text{ Lemma I.9}} \frac{n}{a_{n,b_n}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{e(X_i)\le b_n},$$

where by Markov's inequality,

$$\frac{n}{a_{n,b_n}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{e(X_i)\le b_n} = O_{\mathrm{p}}\left(\frac{n}{a_{n,b_n}}\mathbb{E}[\mathbb{1}_{e(X_i)\le b_n}]\right) = O_{\mathrm{p}}\left(\frac{n}{a_{n,b_n}}F_{e(X)\le b_n}\right) = o_{\mathrm{p}}(1),$$

since we assumed $nF_{e(X)}(b_n) \to 0$ and for all cases we consider, $a_{n,b_n} \to \infty$.

Now we proceed to prove the theorem assuming $nF_{e(X)}(b_n) \gtrsim 1$. Note that the true bias B_{n,b_n} has order $F_{e(X)}(b_n)$, hence we consider the relative accuracy:

$$\frac{n}{a_{n,b_n}}|\hat{\mathsf{B}}_{n,b_n} - \mathsf{B}_{n,b_n}| \sim \left(\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}\right)\frac{|\hat{\mathsf{B}}_{n,b_n} - \mathsf{B}_{n,b_n}|}{F_{e(X)}(b_n)} \le \left(\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}\right)\Big((\mathrm{I}) + (\mathrm{II}) + (\mathrm{III})\Big),$$

where

$$\begin{aligned}
(\mathrm{I}) &= \sum_{j=0}^{p}(\mathrm{I})_j = \sum_{j=0}^{p}\left(\frac{|\hat{\mu}_1^{(j)}(0) - \mu_1^{(j)}(0)|}{j!F_{e(X)}(b_n)}\frac{1}{n}\sum_{i=1}^{n}e(X_i)^j\mathbb{1}_{e(X_i)\le b_n}\right)\\
(\mathrm{II}) &= \frac{1}{(p+1)!F_{e(X)}(b_n)}\frac{1}{n}\sum_{i=1}^{n}\mu_1^{(p+1)}(\lambda_i e(X_i))e(X_i)^{p+1}\mathbb{1}_{e(X_i)\le b_n}\\
(\mathrm{III}) &= \frac{1}{nF_{e(X)}(b_n)}\sum_{i=1}^{n}\Big(\mu_1(e(X_i))\mathbb{1}_{e(X_i)\le b_n} - \mathbb{E}[\mu_1(e(X_i))\mathbb{1}_{e(X_i)\le b_n}]\Big),
\end{aligned}$$

with $\lambda_i \in [0,1]$, by a $(p+1)$-th order Taylor expansion. For $0 \leq j \leq p$,

$$\begin{aligned}\mathbb{E}[e(X)^j \mathbb{1}_{e(X)\leq b_n}] &= \int_0^{b_n} x^j F_{e(X)}(\mathrm{d}x) = b_n^j F_{e(X)}(b_n) - \int_0^{b_n} j x^{j-1} F_{e(X)}(x)\mathrm{d}x \\ &= (1+o(1))\frac{\gamma_0 - 1}{\gamma_0 + j - 1} F_{e(X)}(b_n) b_n^j.\end{aligned}$$

Similarly, its variance has order:

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n e(X_i)^j \mathbb{1}_{e(X)\leq b_n}\right] \leq \frac{1}{n}\mathbb{E}[e(X_i)^{2j}\mathbb{1}_{e(X)\leq b_n}] = (1+o(1))\frac{1}{n}\frac{\gamma_0 - 1}{\gamma_0 + 2j - 1} F_{e(X)}(b_n) b_n^{2j}.$$

Hence, we have

$$\frac{1}{F_{e(X)}(b_n)}\frac{1}{n}\sum_{i=1}^n e(X_i)^j \mathbb{1}_{e(X)\leq b_n} = O_{\mathrm{p}}\left(b_n^j + b_n^j\sqrt{\frac{1}{nF_{e(X)}(b_n)}}\right),$$

which implies that (II) has order:

$$\text{(II)} = O_{\mathrm{p}}\left(b_n^{p+1} + b_n^{p+1}\sqrt{\frac{1}{nF_{e(X)}(b_n)}}\right)$$

By Lemma I.9, term (I) has order:

$$\begin{aligned}\text{(I)} &= O_{\mathrm{p}}\left[\left(\sqrt{\frac{1}{nh_nF_{e(X)}(h_n)}} + h_n^{p+1}\right)\cdot\left(\sum_{j=0}^p \frac{b_n^j}{h_n^j}\right)\cdot\left(1+\sqrt{\frac{1}{nF_{e(X)}(b_n)}}\right)\right] \\ &= O_{\mathrm{p}}\left[\left(\sqrt{\frac{1}{nh_nF_{e(X)}(h_n)}} + h_n^{p+1}\right)\cdot\left(1\vee\frac{b_n^p}{h_n^p}\right)\cdot\left(1+\sqrt{\frac{1}{nF_{e(X)}(b_n)}}\right)\right].\end{aligned}$$

Now we consider some concrete situations. First, assume $b_n = b_n^\star$ being the optimal trimming threshold. Then we know that $a_{n,b_n} \sim b_n^{-1}$. Then for bias correction to be successful, we need $b_n = o(h_n)$. Therefore, h_n should be chosen such that the bias and variance in Lemma I.9 is balanced, which requires $nh_n^{2p+3}F_{e(X)}(h_n) \sim 1$. To explicitly calculate the order of the remaining bias, we assume for simplicity that $F_{e(X)}(h_n) \sim h_n^{\gamma_0 - 1}$, which gives $n^{-(p+1)/(2p+\gamma_0+2)}$, and for $p=1$ (i.e. local linear regression), it becomes $n^{-2/(\gamma_0+4)}$. Since we assumed $\gamma_0 < 2$, the remaining bias, after normalization, is at most $n^{-1/3}$.

For light trimming, the previous discussion continues to hold, although it is not necessary to conduct bias correction in this case.

The heavy trimming case is delicate. We know that in the extreme case where fixed trimming is employed $b_n = b \in (0,1)$, no bias correction will be satisfactory in the sense that the remaining bias, after normalization, either explodes or the that the noise from bias correction will contribute to asymptotic variance. The reason is simple: under fixed trimming, the correct normalization is $\sqrt{n}$, but bias cannot be estimated at a faster rate unless parametric assumption is imposed on the conditional mean function. As a result, one has to rule out fixed trimming unless the researcher is willing to reinterpret the target estimand.

For heavy trimming, we first note that $n/a_{n,b_n} = \sqrt{\mathsf{V}_{n,b_n}}$, hence by Lemma I.3,

$$\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n} = \sqrt{nb_n F_{e(X)}(b_n)}.$$

Therefore to make sure term (II) is negligible, we need

$$\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}(\mathrm{II}) = o_{\mathrm{p}}(1), \qquad \Rightarrow \qquad nb_n^{2p+3}F_{e(X)}(b_n) \to 0.$$

We can still employ the MSE optimal h_n, and the bias correction will be successful.

Finally, we note that term (III) has mean zero and variance of order $1/(nF_{e(X)}(b_n))$, so that

$$\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}(\mathrm{III}) \precsim \sqrt{nb_n F_{e(X)}(b_n)}\sqrt{\frac{1}{nF_{e(X)}(b_n)}} = \sqrt{b_n},$$

which is negligible. ■

I.8.18 Proof of Theorem I.6

Define:

$$Z = \frac{DY}{e(X)}\mathbb{1}_{e(X)\ge b_n} - \mathbb{E}\left[\frac{DY}{e(X)}\mathbb{1}_{e(X)\ge b_n}\right], \qquad U_n = \frac{1}{a_{n,b_n}}\sum_{i=1}^{n} Z_i, \qquad V_n = \sqrt{\frac{1}{a_{n,b_n}^2}\sum_{i=1}^{n} Z_i^2}.$$

Similar as the proof of Theorem I.2, we first establish the joint limiting distribution of (U_n, V_n^2). Consider the characteristic function:

$$\begin{aligned}\mathbb{E}\left[e^{i(\zeta_1 U_n + \zeta_2 V_n^2)}\right] &= \left(\mathbb{E}\left[e^{i(\zeta_1 W_n + \zeta_2 W_n^2)}\right]\right)^n \\ &= \left(1 + \frac{1}{n}\int_{\mathbb{R}} \frac{e^{i(\zeta_1 x + \zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} nx^2 F_{W_n}(\mathrm{d}x)\right)^n,\end{aligned}$$

where $W_n = Z/a_{n,b_n}$. Let $K : \mathbb{R} \to (0, \infty)$ be an auxiliary function that is smooth, symmetric, and satisfies $\lim_{x\to\infty} xK(x) = 1$.

Light trimming

The proof is essentially the same as that of Theorem I.2. Take $I = [c_1, c_2]$ to be a compact interval with $0 \leq c_1 < c_2$, then

$$\begin{aligned}
&\int_I K(x) n x^2 F_{W_n}(\mathrm{d}x) \\
&= n\mathbb{E}[K(W_n) W_n^2 \mathbb{1}_{W_n \in I}] \\
&= \frac{n}{a_{n,b_n}^2} \mathbb{E}\left[K\left(\frac{Z}{a_{n,b_n}}\right) Z^2 \mathbb{1}_{Z/a_{n,b_n} \in I}\right] \\
&= \frac{n}{a_{n,b_n}^2} \int_{a_{n,b_n}c_1}^{a_{n,b_n}c_2} K\left(\frac{x}{a_{n,b_n}}\right) x^2 \mathrm{d}F_Z(x) \\
&= n\left[K(c_2)c_2^2 F_Z(a_{n,b_n}c_2) - K(c_1)c_1^2 F_Z(a_{n,b_n}c_1) - \int_{c_1}^{c_2} \left(2xK(x) + x^2 K^{(1)}(x)\right) F_Z(a_{n,b_n}x)\mathrm{d}x\right] \\
&= n\Big[-K(c_2)c_2^2\Big(1 - F_Z(a_{n,b_n}c_2)\Big) + K(c_1)c_1^2\Big(1 - F_Z(a_{n,b_n}c_1)\Big) \\
&\qquad\quad + \int_{c_1}^{c_2} \left(2xK(x) + x^2 K^{(1)}(x)\right)\Big(1 - F_Z(a_{n,b_n}x)\Big)\mathrm{d}x\Big].
\end{aligned}$$

The tail probabilities can be calculated as in the proof of Theorem I.3(i), implying

$$\begin{aligned}
&\int_I K(x) n x^2 F_{W_n}(\mathrm{d}x) \\
&\to \frac{2-\gamma_0}{\gamma_0} \frac{\alpha_+(0)}{\alpha_+(0) + \alpha_-(0)} \left[-K(c_2)c_2^{2-\gamma_0} + K(c_1)c_1^{2-\gamma_0} + \int_{c_1}^{c_2} \left(2xK(x) + x^2 K^{(1)}(x)\right) x^{-\gamma_0}\mathrm{d}x\right] \\
&= M^\dagger(I),
\end{aligned}$$

where the measure $M^\dagger(\mathrm{d}x)$ is defined as

$$M^\dagger(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0) + \alpha_-(0)} K(x)|x|^{1-\gamma_0}\Big(\alpha_+(0)\mathbb{1}_{x\geq 0} + \alpha_-(0)\mathbb{1}_{x<0}\Big)\right].$$

The same convergence holds for compact intervals $[c_1, c_2]$ with $c_2 \leq 0$. Finally, we note that

$$\int_{\mathbb{R}} K(x) n x^2 F_{W_n}(\mathrm{d}x) \to M^\dagger(\mathbb{R}) \in (0, \infty).$$

Therefore, we have the following distributional convergence:

$$\frac{K(x)nx^2F_{W_n}(\mathrm{d}x)}{\int_{\mathbb{R}} K(x)nx^2F_{W_n}(\mathrm{d}x)} \xrightarrow{\mathrm{d}} \frac{M^\dagger(\mathrm{d}x)}{M^\dagger(\mathbb{R})}.$$

Since the following is bounded and continuous,

$$\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2K(x)},$$

we have

$$\int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}nx^2F_{W_n}(\mathrm{d}x) = \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2K(x)}K(x)nx^2F_{W_n}(\mathrm{d}x)$$
$$\to \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2K(x)}M^\dagger(\mathrm{d}x) = \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}M(\mathrm{d}x),$$

where

$$M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(0)\mathbb{1}_{x\geq 0}+\alpha_-(0)\mathbb{1}_{x<0}\Big)\right],$$

as defined in Theorem I.1(ii). To summarize, we showed:

$$\mathbb{E}\left[e^{i(\zeta_1 U_n+\zeta_2 V_n^2)}\right] \to \exp\left\{\int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}M(\mathrm{d}x)\right\},$$

which defines the joint limiting distribution of (U_n, V_n^2).

Moderate trimming

We do not repeat the lengthy argument. With the tail probability calculations used for Theorem I.3(iii), one has

$$\mathbb{E}\left[e^{i(\zeta_1 U_n+\zeta_2 V_n^2)}\right] \to \exp\left\{\int_{\mathbb{R}} \frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}M(\mathrm{d}x)\right\},$$

where

$$M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(tx)\mathbb{1}_{x\geq 0}+\alpha_-(tx)\mathbb{1}_{x<0}\Big)\right],$$

as defined in Theorem I.3(iii).

Heavy trimming

This case is much easier, and one can directly show that U_n/V_n converges to the standard Gaussian distribution.

For all three cases, U_n/V_n has a well-defined limiting distribution. And since we focus on $\gamma_0 < 2$, the impact of estimating the probability weights can be ignored. Therefore, the self-normalized statistic T_{n,b_n} has the same limiting distribution as U_n/V_n, and subsampling is valid by standard arguments in Politis and Romano (1994) (or Romano and Wolf 1999). ■

I.8.19 Proof of Proposition I.3

Rewrite the estimator as

$$\hat{\tau}_n^{\texttt{ATT}} = \frac{c_0}{\hat{c}_n}\frac{1}{n}\sum_{i=1}^{n}\frac{(D_i - e(X_i))Y_i}{c_0(1-e(X_i))},$$

where $c_0 = \mathbb{P}[D=1]$, and $\hat{c}_n = n^{-1}\sum_{i=1}^n D_i$. We first consider the tail behavior of $(D - e(X))Y/(c_0(1-e(X))$. Note that

$$\begin{aligned}\mathbb{P}\left[\frac{(D-e(X))Y}{c_0(1-e(X))} > x\right] &= \mathbb{P}[D=1]\mathbb{P}\left[\frac{Y(1)}{c_0} > x\middle| D=1\right]\\ &\quad + \mathbb{P}[D=0]\mathbb{P}\left[\frac{e(X)Y(0)}{c_0(1-e(X))} < -x\middle| D=0\right],\end{aligned}$$

where we take $x > 0$. To proceed, let $F_{1-e(X)}$ be the distribution function of $1-e(X)$, then

$$\begin{aligned}&\lim_{x\downarrow 0}\frac{\mathbb{P}[1-e(X)\le x|D=0]}{x\mathbb{P}[1-e(X)\le x]} = \lim_{x\downarrow 0}\frac{\mathbb{P}[D=0|1-e(X)\le x]}{x\mathbb{P}[D=0]}\\ &= \lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[1-e(X)\le x]\mathbb{P}[D=0]}\int_0^x yF_{1-e(X)}(\mathrm{d}y)\\ &= \lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[1-e(X)\le x]\mathbb{P}[D=0]}\left(xF_{1-e(X)}(x) - \int_0^x F_{1-e(X)}(y)\mathrm{d}y\right)\\ &= \lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[1-e(X)\le x]\mathbb{P}[D=0]}\left(xF_{1-e(X)}(x) - \int_0^1 xF_{1-e(X)}(xy)\mathrm{d}y\right)\\ &= \lim_{x\downarrow 0}\frac{1}{\mathbb{P}[D=0]}\left(1 - \int_0^1\frac{F_{1-e(X)}(xy)}{F_{1-e(X)}(x)}\mathrm{d}y\right)\\ &= \frac{1}{\mathbb{P}[D=0]}\left(1-\int_0^1 y^{\gamma_0-1}\mathrm{d}y\right) = \frac{\gamma_0-1}{\gamma_0}\frac{1}{\mathbb{P}[D=0]}.\end{aligned}$$

Applying the same argument used to prove Lemma I.1, one has

$$\begin{aligned}
&\lim_{x\to\infty} \frac{\mathbb{P}[D=0]\mathbb{P}\left[\frac{e(X)Y(0)}{c_0(1-e(X))} < -x \middle| D=0\right]}{x^{-1}\mathbb{P}[1-e(X) < x^{-1}]} \\
&= \lim_{x\to\infty} \frac{\mathbb{P}[D=0]\mathbb{P}[1-e(X) < x^{-1}|D=0]}{x^{-1}\mathbb{P}[1-e(X) < x^{-1}]} \frac{\mathbb{P}\left[\frac{e(X)Y(0)}{c_0(1-e(X))} < -x \middle| D=0\right]}{\mathbb{P}[1-e(X) < x^{-1}|D=0]} \\
&= \frac{\gamma_0 - 1}{\gamma_0} c_0^{-\gamma_0} \alpha_{(0),-}(0),
\end{aligned}$$

where

$$\alpha_{(0),-}(x) = \lim_{t\to 1} \mathbb{E}\left[|Y(0)|^{\gamma_0} \mathbb{1}_{Y(0)<x} \middle| e(X) = t\right].$$

Therefore,

$$\lim_{x\to\infty} \frac{\mathbb{P}\left[\frac{(D-e(X))Y}{c_0(1-e(X))} > x\right]}{x^{-1}\mathbb{P}[1-e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0} c_0^{-\gamma_0} \alpha_{(0),-}(0).$$

Similarly, we have

$$\lim_{x\to\infty} \frac{\mathbb{P}\left[\frac{(D-e(X))Y}{c_0(1-e(X))} < -x\right]}{x^{-1}\mathbb{P}[1-e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0} c_0^{-\gamma_0} \alpha_{(0),+}(0).$$

As a result, $(D-e(X))Y/(c_0(1-e(X))$ has regularly varying tails with index $-\gamma_0$ if $\alpha_{(0),+}(0)+\alpha_{(0),-}(0) > 0$. The rest of the proof employs the same argument used for Theorem I.1. ■

I.8.20 Proof of Proposition I.4

This employs the same argument used for Theorem I.3 and Proposition I.3. ■

I.8.21 Proof of Proposition I.5

We first consider the tail behavior of $(2D-1)Y/(1-D+(2D-1)e(X))$. For this, we note that

$$\begin{aligned}
\mathbb{P}\left[\frac{(2D-1)Y}{1-D+(2D-1)e(X)} > x\right] &= \mathbb{P}[D=1]\mathbb{P}\left[\frac{Y(1)}{e(X)} > x \middle| D=1\right] \\
&\quad + \mathbb{P}[D=0]\mathbb{P}\left[\frac{Y(0)}{1-e(X)} < -x \middle| D=0\right],
\end{aligned}$$

where we take $x > 0$. Then if $\omega > 0$,

$$\begin{aligned}
&\lim_{x\downarrow 0} \frac{\mathbb{P}[e(X) \leq x | D = 1]}{x\mathbb{P}[e(X) \leq x]} = \lim_{x\downarrow 0} \frac{\mathbb{P}[D = 1 | e(X) \leq x]}{x\mathbb{P}[D = 1]} \\
&= \lim_{x\downarrow 0} \frac{1}{x\mathbb{P}[e(X) \leq x]\mathbb{P}[D = 1]} \int_0^x y F_{e(X)}(\mathrm{d}y) \\
&= \lim_{x\downarrow 0} \frac{1}{x\mathbb{P}[e(X) \leq x]\mathbb{P}[D = 1]} \left(x\mathbb{P}[e(X) \leq x] - \int_0^x F_{e(X)}(y)\mathrm{d}y \right) \\
&= \lim_{x\downarrow 0} \frac{1}{x\mathbb{P}[e(X) \leq x]\mathbb{P}[D = 1]} \left(x\mathbb{P}[e(X) \leq x] - \int_0^1 xF_{e(X)}(xy)\mathrm{d}y \right) \\
&= \lim_{x\downarrow 0} \frac{1}{\mathbb{P}[D = 1]} \left(1 - \int_0^1 \frac{F_{e(X)}(xy)}{F_{e(X)}(x)}\mathrm{d}y \right) = \frac{1}{\mathbb{P}[D = 1]} \left(1 - \int_0^1 y^{\gamma_0 - 1}\mathrm{d}y \right) = \frac{\gamma_0 - 1}{\gamma_0} \frac{1}{\mathbb{P}[D = 1]}.
\end{aligned}$$

Therefore, conditional on $D = 1$, the probability weight has regularly varying left tail with index γ_0. Applying the same argument used to prove Lemma I.1, one has

$$\begin{aligned}
&\lim_{x\to\infty} \frac{\mathbb{P}[D = 1]\mathbb{P}\left[\frac{Y(1)}{e(X)} > x \middle| D = 1 \right]}{x^{-1}\mathbb{P}[e(X) < x^{-1}]} \\
&= \lim_{x\to\infty} \frac{\mathbb{P}[D = 1]\mathbb{P}[e(X) < x^{-1} | D = 1]}{x^{-1}\mathbb{P}[e(X) < x^{-1}]} \frac{\mathbb{P}\left[\frac{Y(1)}{e(X)} > x \middle| D = 1 \right]}{\mathbb{P}[e(X) < x^{-1} | D = 1]} \\
&= \frac{\gamma_0 - 1}{\gamma_0} \alpha_{(1),+}(0).
\end{aligned}$$

Similarly, we can show that if $\omega < 1$,

$$\lim_{x\downarrow 0} \frac{\mathbb{P}[D = 0]\mathbb{P}\left[\frac{Y(0)}{1-e(X)} < -x \middle| D = 0 \right]}{x^{-1}\mathbb{P}[1 - e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0} \alpha_{(0),-}(0).$$

Together, they imply

$$\lim_{x\to\infty} \frac{x\mathbb{P}\left[\frac{(2D-1)Y}{1-D+(2D-1)e(X)} > x \right]}{\mathbb{P}[e(X) < x^{-1}] + \mathbb{P}[1 - e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0} \Big(\omega\alpha_{(1),+}(0) + (1 - \omega)\alpha_{(0),-}(0) \Big).$$

By the same argument,

$$\lim_{x\to\infty} \frac{x\mathbb{P}\left[\frac{(2D-1)Y}{1-D+(2D-1)e(X)} < -x \right]}{\mathbb{P}[e(X) < x^{-1}] + \mathbb{P}[1 - e(X) < x^{-1}]} = \frac{\gamma_0 - 1}{\gamma_0} \Big(\omega\alpha_{(1),-}(0) + (1 - \omega)\alpha_{(0),+}(0) \Big).$$

As a result, $(2D-1)Y/(1-D+(2D-1)e(X))$ has regularly varying tail with index $-\gamma_0$ if

$$\omega\Big(\alpha_{(1),+}(0)+\alpha_{(1),-}(0)\Big)+(1-\omega)\Big(\alpha_{(0),+}(0)+\alpha_{(0),-}(0)\Big)>0.$$

The rest of the proof employs the same argument used for Theorem I.1. ■

I.8.22 Proof of Proposition I.6

This employs the same argument used for Theorem I.3 and Proposition I.5. ■

I.8.23 Proof of Proposition I.7

This employs the same argument used for Theorem I.1. ■

I.8.24 Proof of Proposition I.8

This employs the same argument used for Theorem I.3. ■

CHAPTER II

Two-Step Estimation and Inference with Possibly Many Included Covariates

Abstract. *This chapter studies the implications of including many covariates in a first-step estimate entering a two-step estimation procedure. We find that a first order bias emerges when the number of included covariates is "large" relative to the square-root of sample size, rendering standard inference procedures invalid. We show that the jackknife is able to estimate this "many covariates" bias consistently, thereby delivering a new automatic bias-corrected two-step point estimator. The jackknife also consistently estimates the standard error of the original two-step point estimator. For inference, we develop a valid post-bias-correction bootstrap approximation that accounts for the additional variability introduced by the jackknife bias-correction. We find that the jackknife bias-corrected point estimator and the bootstrap post-bias-correction inference perform excellent in simulations, offering important improvements over conventional two-step point estimators and inference procedures, which are not robust to including many covariates. We apply our results to an array of distinct treatment effect, policy evaluation, and other applied microeconomics settings. In particular, we discuss production function and marginal treatment effect estimation in detail.*

II.1 Introduction

Two-step estimators are very important and widely used in empirical work in economics and other disciplines. This approach involves two estimation steps: first an unknown quantity is estimated, and then this estimate is plugged in a moment condition to form the second and final point estimator of interest. For example, inverse probability weighting (IPW) and generated regressors methods fit naturally into this framework, both used routinely in treatment effect and policy evaluation settings. In practice, researchers often include many

This chapter is based on the paper "Two-Step Estimation and Inference with Possibly Many Included Covariates" (Cattaneo, Jansson and Ma, 2018d).

covariates in the first-step estimation procedure in an attempt to flexibly control for as many confounders as possible, even after model selection or model shrinking has been used to select out some of all available covariates. Conventional (post-model selection) estimation and inference results in this context, however, assume that the number of covariates included in the estimation is "small" relative to the sample size, and hence the effect of overfitting in the first estimation step is ignored in current practice.

We show that two-step estimators can be severely biased when too many covariates are included in a linear-in-parameters first-step, a fact that leads to invalid inference procedures even in large samples. This crucial, but often overlooked fact implies that many empirical conclusions will be incorrect whenever many covariates are used. For example, we find from a very simple simulation setup with a first step estimated with 80 i.i.d. variables, sample size of $2,000$, and even no misspecification bias, that a conventional 95% confidence interval covers the true parameter with probability 76% due to the presence of the many covariates bias we highlight in this chapter (Table II.1 below).[1] This result is not specific to our simulation setting, as our general results apply broadly to many other treatment effect, policy evaluation, and applied microeconomics settings: IPW estimation under unconfoundedness, semiparametric difference-in-differences, local average response function estimation, marginal treatment effects, control function methods, and production function estimation, just to mention a few other popular examples. We illustrate the usefulness of our results by considering the estimation and inference for the marginal treatment effect (Heckman and Vytlacil, 2005) when possibly many covariates/instruments are present. This offers new estimation and inference results in instrumental variable (IV) settings allowing for treatment effect heterogeneity and many covariates/instruments.

The presence of the generic many covariates bias we highlight implies that developing more robust procedures accounting for possibly many covariates entering the first step estimation is highly desirable. Such robust methods would give more credible empirical results, thereby providing more plausible testing of substantive hypotheses as well as more reliable policy prescriptions. With this goal in mind, we show that jackknife bias-correction is able to remove the many covariates bias we uncover in a fully automatic way. Under mild conditions on the design matrix, we prove consistency of the jackknife bias and variance estimators, even when many covariates are included in the first-step estimation. Indeed, our simulations in the context of MTE estimation show that jackknife bias-correction is quite effective in removing the many covariates bias, exhibiting roughly a 50% bias reduction (Table II.1

[1] Including 80 regressors is quite common in empirical work: e.g., settings with 50 residential dummy indicators, a few covariates entering linearly and quadratically, and perhaps some interactions among these variables.

below). We also show that the mean squared error of the jackknife bias-corrected estimator is substantially reduced whenever many covariates are included. More generally, our results give a new, fully automatic, jackknife bias-corrected two-step estimator with demonstrably superior properties to use in applications.

For inference, while the jackknife bias correction and variance estimation deliver a valid Gaussian distributional approximation in large samples, we find in our simulations that the associated inference procedures do not perform as well in small samples. As discussed in Calonico, Cattaneo and Farrell (2018) in the context of kernel-based nonparametric inference, a crucial underlying issue is that bias correction introduces additional variability not accounted for in samples of moderate size (we confirm this finding in our simulations). Therefore, to develop better inference procedures in finite samples, we also establish validity of a bootstrap method applied to the jackknife-based bias-corrected Studentized statistic, which can be used to construct valid confidence intervals and conduct valid hypothesis tests in a fully automatic way. This procedure is a hybrid of the wild bootstrap (first-step estimation) and the multiplier bootstrap (second-step estimation), which is fast and easy to implement in practice because it avoids recomputing the relatively high-dimensional portion of the first estimation step. Under generic regularity conditions, we show that this bootstrap procedure successfully approximates the finite sample distribution of the bias-corrected jackknife-based Studentized statistic, a result that is also borne out in our simulation study.

Put together, our results not only highlight the important negative implications of overfitting the first-step estimate in generic two-step estimation problems, which leads to a first order many covariates bias in the distributional approximation, but also provide fully automatic resampling methods to construct more robust estimators and inference procedures. Furthermore, because our results remain asymptotically valid when only a few covariates are used, they provide strict asymptotic improvement over conventional methods currently used in practice. All our results are fully automatic and do not require additional knowledge about the data generating process, which implies that they can be easily implemented in empirical work using straightforward resampling methods on any computing platform.

Our work is related to several interconnected literatures in econometrics and statistics. From a classical semiparametric perspective, when the many covariates included in the first-step are taken as basis expansions of some underlying fixed dimension regressor, our final estimator becomes a two-step semiparametric estimator with a nonparametric series-based preliminary estimate. Conventional large sample approximations in this case are well known (e.g., Newey and McFadden, 1994; Chen, 2007; Ichimura and Todd, 2007, and references therein). From this perspective, our result contributes not only to this classical semiparametric literature, but also to the more recent work in the area, which has developed distributional approximations that are more robust to tuning parameter choices and underlying assumptions (e.g., smoothness). In particular, first, Cattaneo, Crump and Jansson (2013) and Cattaneo and Jansson (2018) develop approximations for two-step non-linear kernel-based semiparametric estimators when possibly a "small" bandwidth is used, which leads to a first-order bias due to undersmoothing the preliminary kernel-based nonparametric estimate, and show that inference based on the nonparametric bootstrap automatically accounts for the small bandwidth bias explicitly, thereby offering more robust inference procedures in that context.[2] Second, Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018b) study the complementary issue of "large" bandwidth or "small" number of series terms, and develop more robust inference procedures in that case. Their approach is to modify the estimating equation so that the resulting new two-step estimator is less sensitive to oversmoothing (i.e., underfitting) the first-step nonparametric estimator. Our result complements this recent literature by offering new inference procedures with demonstrably more robust properties to undersmoothing (i.e., overfitting) a first step series-based estimator, results that are not currently available in the semiparametrics literature. See Section II.3 for more details.

Our results go beyond semiparametrics because we do not assume (but allow for) the first-step estimate to be a nonparametric series-based estimator. In fact, we do not rely on any specific structure of the covariates in the first step, nor do we rely on asymptotic linear representations. Thus, our results also contribute to the literature on high-dimensional models in statistics and econometrics (e.g., Mammen, 1989, 1993; El Karoui, Bean, Bickel, Lim and Yu, 2013; Cattaneo, Jansson and Newey, 2018f; Li and Müller, 2017, and references therein) by developing generic distributional approximations for two-step estimators where the first-step estimator is possibly high-dimensional. See also Fan, Lv and Qi (2011) for a survey and discussion on high-dimensional and ultra-high-dimensional models.[3] A key distinction here is that the class of estimators we consider is defined through a moment condition that is non-linear in the first step estimate (e.g., propensity score, generated regressor, etc.). Previous work on high-dimensional models has focused exclusively on either linear least squares regression or one-step (possibly non-linear) least squares regression. In contrast, this chapter covers a large class of two-step non-linear procedures, going well be-

[2] A certain class of *linear* semiparametric estimators has a very different behavior when undersmoothing the first step nonparametric estimator; see Cattaneo, Crump and Jansson (2010, 2014a,b) and Cattaneo, Jansson and Newey (2018e) for discussion and references. In particular, their results show that undersmoothing leads to an additional variance contribution (due to the underlying linearity of the model), while in the present chapter we find a bias contribution instead (due to the non-linearity of the models considered).

[3] We call models high-dimensional when the number of available covariates is at most a fraction of the sample size and ultra-high-dimensional when the number of available covariates is larger than the sample size.

yond least squares regression for the second step estimation procedure. Most interestingly, our results show formally that when many covariates are included in a first-step estimation the resulting two-step estimator exhibits a bias of order $k/\sqrt{n}$ in the distributional approximation, where k denotes the number of included covariates and n denotes the sample size. This finding contrasts sharply with previous results for high-dimensional linear regression models with many covariates, where it has been found that including many covariates leads to a variance contribution (not a bias contribution as we find herein) in the distributional approximation, which is of order k/n (not $k/\sqrt{n}$ as we find herein). By implication, the many covariates bias we uncover in this chapter will have a first-order effect on inference when fewer covariates are included relative to the case of high-dimensional linear regression models.

Our results also have implications for the recent and rapidly growing literature on inference after covariate/model selection in ultra-high-dimensional settings under sparsity conditions (e.g., Belloni, Chernozhukov and Hansen, 2014; Farrell, 2015; Belloni, Chernozhukov, Fernández-Val and Hansen, 2017, and references therein). In this literature, the total number of available covariates/instruments is allowed to be much larger than the sample size, but the final number of included covariates/instruments is much smaller than the sample size, as most available covariates are selected out by some penalization or model selection method (e.g., LASSO) employing some form of a sparsity assumption. This implies that the number of included covariates/instruments effectively used for estimation and inference (k in our notation) is much smaller than the sample size, as the underlying distribution theory in that literature requires $k/\sqrt{n} = o(1)$. Therefore, because $k/\sqrt{n} = O(1)$ is the only restriction assumed in this chapter, our results shed new light on situations where the number of selected or included covariates, possibly after model selection, is not "small" relative to the sample size. We formally show that valid inference post-model selection requires that a relatively small number of covariates enter the final specification, since otherwise a first order bias will be present in the distributional approximations commonly employed in practice, thereby invalidating the associated inference procedures. Our results do not employ any sparsity assumption and allow for any kind of regressors, including many fixed effects, provided the first-step estimate can be computed.

Our findings are also qualitatively connected to the literature on non-linear panel models with fixed effects (Fernandez-Val and Weidner, 2199, and references therein) in at least two ways. First, in that context a first-order bias arises when the number of time periods (T) is proportional to the number of entities (N), just like we uncover a first-order bias when $k \propto \sqrt{n}$, and in both cases this bias can be heuristically attributed to an incidental parameters/overfitting problem. Second, in that literature jackknife bias correction was shown to be able to remove the large-(N, T) bias, just like we establish a similar result in this chapter for a class of two-step estimators with high-dimensional first-step. Beyond these two superficial connections, however, our findings are both technically and conceptually quite different from the results already available in the large-(N, T) non-linear panel fixed effects literature.

Section II.2 introduces the setup and gives an overview of our results. Section II.3 gives details on the main properties of the two-step estimator, in particular characterizing the non-vanishing bias due to many covariates entering the first-step estimate. Section II.4 establishes validity of the jackknife bias and variance estimator, and therefore presents our proposed bias-corrected two-step estimator, while Section II.5 establishes valid distributional approximations for the jackknife-based bias-corrected Studentized statistic using a carefully modified bootstrap method. Section II.6 applies our main results to the marginal treatment effect (Heckman and Vytlacil, 2005) estimation with a Monte Carlo experiment and an empirical illustration building on the work of Carneiro, Heckman and Vytlacil (2011). Finally, Section II.7 concludes. Additional results, preliminary lemmas and all proofs are collected in Section II.8 and II.9.

II.2 Setup and Overview of Results

We consider a two-step GMM setting where $\mathbf{w}_i = (\mathbf{y}_i^{\mathrm{T}}, r_i, \mathbf{z}_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, 2, \ldots, n$, denotes an observed random sample, and the finite dimensional parameter of interest $\boldsymbol{\theta}_0$ solves uniquely the (possibly over-identified) vector-valued moment condition $\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)] = \mathbf{0}$ with $\mu_i = \mathbb{E}[r_i|\mathbf{z}_i]$. Thus, we specialize the general two-step GMM approach in that we view the unknown scalar μ_i as a "generated regressor" depending on possibly many covariates $\mathbf{z}_i \in \mathbb{R}^k$, which we take as the included variables entering the first-step specification. Our results extend immediately to vector-valued unknown $\boldsymbol{\mu}_i$, albeit with cumbersome notation.

Given a first-step estimate $\hat{\mu}_i$ of μ_i, which we construct by projecting r_i on the possibly high-dimensional covariate $\mathbf{z}_i$ with least squares, as discussed further below, we study the two-step estimator:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \left| \boldsymbol{\Omega}_n^{1/2} \sum_{i=1}^{n} \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}) \right|, \tag{II.1}$$

where $|\cdot|$ denotes the Euclidean norm, $\Theta \subseteq \mathbb{R}^{d_\theta}$ is the parameter space, and $\boldsymbol{\Omega}_n$ is a (possibly random) positive semi-definite conformable weighting matrix with positive definite probability limit $\boldsymbol{\Omega}_0$. Regularity conditions on the known moment function $\mathbf{m}(\cdot)$ are given in the next section.

When the dimension of the included variables $\mathbf{z}_i$ is "small" relative to the sample size, $k = o(\sqrt{n})$, textbook large sample theory is valid, and hence estimation and inference can be conducted in the usual way (e.g., Newey and McFadden, 1994). However, when the dimension of the included covariates used to approximate the unknown component μ_i is "large" relative to the sample size, $k = O(\sqrt{n})$, standard distribution theory fails. To be more specific, under fairly general regularity conditions, we show in Section II.3 that:

$$\mathscr{V}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \mathscr{B}) \overset{\mathrm{d}}{\to} \mathcal{N}(\mathbf{0},\ \mathbf{I}), \tag{II.2}$$

where $\overset{\mathrm{d}}{\to}$ denotes convergence in distribution, with limits always taken as $n \to \infty$ and $k = O(\sqrt{n})$, and $\mathscr{V}$ and $\mathscr{B}$ denoting, respectively, the approximate variance and bias of the estimator $\hat{\boldsymbol{\theta}}$. This result has a key distinctive feature relative to classical textbook results: a first-order bias $\mathscr{B}$ emerges whenever "many" covariates are included, that is, whenever k is "large" relatively to n in the sense that $k/\sqrt{n} \not\to 0$. A crucial practical implication of this finding is that conventional inference procedures that disregard the presence of the first-order bias will be incorrect even asymptotically, since $\mathscr{V}^{-1/2}\mathscr{B} = O_{\mathrm{p}}(k/\sqrt{n})$ is non-negligible. For example, non-linear treatment effect, instrumental variables and control function estimators employing "many" included covariates in a first-step estimation will be biased, thereby giving over-rejection of the null hypothesis of interest. In Section II.6 we illustrate this problem using simulated data in the context of instrumental variable models with many instruments/covariates, where we find that typical hypothesis tests over-reject the null hypothesis four times as often as they should in practically relevant situations.

Putting aside the bias issue when many covariates are used in the first-step estimation, another important issue regarding (II.2) is the characterization and estimation of the variance $\mathscr{V}$. Because the possibly high-dimensional covariates $\mathbf{z}_i$ are not necessarily assumed to be a series expansion, or other type of convergent sequence of covariates, the variance $\mathscr{V}$ is harder to characterize and estimate. In fact, our distributional approximation leading to (II.2) is based on a quadratic approximation of $\hat{\boldsymbol{\theta}}$, as opposed to the traditional linear approximation commonly encountered in the semiparametrics literature (Newey, 1994; Chen, 2007; Hahn and Ridder, 2013), thereby giving a more general characterization of the variability of $\hat{\boldsymbol{\theta}}$ with potentially better finite sample properties.

Nevertheless, our first main result (II.2) suggests that valid inference in two-step GMM settings is possible even when many covariates are included in the first-step estimation, if consistent variance and bias estimators are available. Our second main result (in Section II.4) shows that the jackknife offers an easy-to-implement and automatic way to approximate both the variance and the bias:

$$\mathscr{T} \;=\; \hat{\mathscr{V}}^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \hat{\mathscr{B}}) \;\overset{\mathrm{d}}{\to}\; \mathcal{N}(\mathbf{0},\ \mathbf{I}), \tag{II.3}$$

To implement the jackknife method, one first constructs $\hat{\boldsymbol{\theta}}^{(\ell)}$, for which the ℓ^{th} observation is deleted and then both steps are re-estimated using the remaining observations. Denote by $\hat{\boldsymbol{\theta}}^{(\cdot)} = n^{-1}\sum_{\ell=1}^{n}\hat{\boldsymbol{\theta}}^{(\ell)}$ the average of the leave-one-observation-out estimators, then

$$\hat{\mathscr{B}} = (n-1)(\hat{\boldsymbol{\theta}}^{(\cdot)} - \hat{\boldsymbol{\theta}}), \quad \hat{\mathscr{V}} = \frac{n-1}{n}\sum_{\ell=1}^{n}(\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\cdot)})(\hat{\boldsymbol{\theta}}^{(\ell)} - \hat{\boldsymbol{\theta}}^{(\cdot)})^{\mathrm{T}}. \tag{II.4}$$

Simulation evidence reported in Section II.6 confirms that the jackknife provides an automatic data-driven method able to approximate quite well both the bias and the variance of the estimator $\hat{\boldsymbol{\theta}}$, even when many covariates are included in the first-step estimation procedure. An important virtue of the jackknife is that it can be implemented very fast in special settings, which is particularly important in high-dimensional situations. Indeed, our first-step estimator will be constructed using least-squares, a method that is particularly amenable to jackknifing.

While result (II.3) could be used for inference in large samples, a potential drawback is that the jackknife bias-correction introduces additional variability not accounted for in samples of moderate size. Therefore, to improve inference further in applications, we develop a new, specifically tailored bootstrap-based distributional approximation to the jackknife-based bias-corrected and Studentized statistic. Our method combines the wild bootstrap (first-step) and the multiplier bootstrap (second-step), while explicitly taking into account the effect of jackknifing under the multiplier bootstrap law (see Section II.5 for more details). To be more specific, our third and final main result is:

$$\sup_{t\in\mathbb{R}^{d_\theta}}\Big|\mathbb{P}[\mathscr{T} \le t] - \mathbb{P}^\star[\mathscr{T}^\star \le t]\Big| \overset{\mathrm{p}}{\to} 0, \qquad \mathscr{T}^\star \;=\; \hat{\mathscr{V}}^{\star\,-1/2}(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \hat{\mathscr{B}}^\star), \tag{II.5}$$

where $\hat{\boldsymbol{\theta}}^\star$ is a bootstrap counterpart of $\hat{\boldsymbol{\theta}}$, $\hat{\mathscr{B}}^\star$ and $\hat{\mathscr{V}}^\star$ are properly weighted jackknife bias and variance estimators under the bootstrap distribution, respectively, and $\mathbb{P}^\star$ is the bootstrap probability law conditional on the data. Our bootstrap approach is fully automatic and captures explicitly the distributional effects of estimating the bias (and variance) using the jackknife, and hence delivers a better finite sample approximation. Simulation evidence reported in Section II.6 supports this result.

In sum, valid and more robust inference in two-step GMM settings with possibly many covariates entering the first-step estimate can be conducted by combining results (II.3) and

(II.5). Specifically, our approach requires three simple and automatic stages: (i) constructing the two-step estimator $\hat{\boldsymbol{\theta}}$, (ii) constructing the jackknife bias and variance estimators $\hat{\mathscr{B}}$ and $\hat{\mathscr{V}}$, and finally (iii) conducting inference as usual but employing bootstrap quantiles obtained from (II.5) instead of those from the normal approximation. In the remainder of this chapter we formalize these results and illustrate them using simulated as well as real data.

II.3 The Effect of Including Many Covariates

In this section we formalize the implications of overfitting the first-step estimate entering (II.1), and show that under fairly general conditions the estimator $\hat{\boldsymbol{\theta}}$, and transformations thereof, exhibit a first-order bias whenever k is "large", that is, whenever $k \propto \sqrt{n}$. The results in this section justify, in particular, the distributional approximation in (II.2).

We first present some regularity conditions maintained throughout this chapter. A random variable is said to be in BM_ℓ (bounded moments) if its ℓ^{th} moment is finite, and in BCM_ℓ (bounded conditional moments) if its ℓ^{th} conditional on $\mathbf{z}_i$ moment is bounded uniformly by a finite constant. In addition, define the transformation

$$\mathcal{H}_i^{\alpha,\delta}(\mathbf{m}) = \sup_{(|\mu-\mu_i|+|\boldsymbol{\theta}-\boldsymbol{\theta}_0|)^\alpha \leq \delta} \frac{|\mathbf{m}(\mathbf{w}_i, \mu, \boldsymbol{\theta}) - \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|}{(|\mu - \mu_i| + |\boldsymbol{\theta} - \boldsymbol{\theta}_0|)^\alpha}.$$

The following assumption collects some basic notation and regularity conditions.

Assumption II.1 (Regularity conditions)
Let $0 < \delta,\ \alpha,\ C < \infty$ be some fixed constants. (i) $\mathbf{m}$ is twice continuously differentiable in μ with derivatives denoted by $\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0) = \frac{\partial}{\partial \mu}\mathbf{m}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)$ and $\ddot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0) = \frac{\partial^2}{\partial \mu^2}\mathbf{m}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)$. In addition, $\mathbf{m}$ and $\dot{\mathbf{m}}$ are continuously differentiable in $\boldsymbol{\theta}$. (ii) $\mathcal{H}_i^{\alpha,\delta}(\mathbf{m})$, $\mathcal{H}_i^{\alpha,\delta}(\frac{\partial \mathbf{m}}{\partial \boldsymbol{\theta}})$, $\mathcal{H}_i^{\alpha,\delta}(\frac{\partial \dot{\mathbf{m}}}{\partial \boldsymbol{\theta}}) \in \mathsf{BM}_1$. (iii) $\mathbf{m}_i$, $\dot{\mathbf{m}}_i$, $\ddot{\mathbf{m}}_i$, $\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}})$, ε_i^3, $|\dot{\mathbf{m}}_i\varepsilon_i|$, $|\ddot{\mathbf{m}}_i|\varepsilon_i^2$, $|\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}})|\varepsilon_i^2 \in \mathsf{BCM}_2$, where $\mathbf{m}_i = \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta})$, $\dot{\mathbf{m}}_i = \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta})$, $\ddot{\mathbf{m}}_i = \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta})$, and $\varepsilon_i = r_i - \mu_i$. (iv) $\mathbf{M}_0 = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\right]$ has full (column) rank d_θ. ‖

These conditions are standard in the literature. They require smoothness of $\mathbf{m}(\mathbf{w}, \mu, \boldsymbol{\theta})$ with respect to both μ and $\boldsymbol{\theta}$, and boundedness of (conditional) moments of various orders. In future work we plan to extend our results to non-differentiable second-step estimating equations.

II.3.1 First-Step Estimation

We are interested in understanding the effects of including possibly many covariates $\mathbf{z}_i$, that is, in cases where its dimension k is possibly "large" relative to the sample size. For tractability and simplicity, we consider linear approximations to the unknown component:

$$\mu_i = \mathbb{E}[r_i|\mathbf{z}_i] = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + \eta_i, \qquad \mathbb{E}[\mathbf{z}_i\eta_i] = \mathbf{0}, \tag{II.6}$$

for a non-random vector $\boldsymbol{\beta}$, where η_i represents the error in the best linear approximation. This motivates the least-squares first-step estimate:

$$\hat{\mu}_i = \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^k} \sum_{i=1}^{n}(r_i - \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})^2, \tag{II.7}$$

which is quite common in empirical work. It is possible to allow for non-linear models, but such methods are harder to handle mathematically and usually do not perform well numerically when $\mathbf{z}_i$ is of large dimension. Furthermore, a non-linear approach will be computationally more difficult, as we discuss in more detail below. Our proofs explicitly exploit the linear regression representation of $\hat{\mu}_i$ to scale down the already quite involved technical work. Nevertheless, we also conducted preliminary theoretical work to verify that the main results presented below carry over to non-linear least-squares estimators (e.g., logistic regression when r_i is binary).

Using the first-step estimate $\hat{\mu}_i$ in (II.7), we investigate the implications of introducing possibly many covariates $\mathbf{z}_i$, and thus our approximations allow for (but do not require that) k being "large" relative to the sample size. Specifically, we show that when $k \propto \sqrt{n}$ conventional inference procedures become invalid due to a new bias term in the asymptotic approximations.

In some settings, the covariates $\mathbf{z}_i$ can have approximation power beyond the first-step estimation, as it occurs for instance when these covariates are basis expansions. To allow for this possibility, we also define, for a non-random matrix $\boldsymbol{\Gamma}$,

$$\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i] = \boldsymbol{\Gamma}\mathbf{z}_i + \boldsymbol{\zeta}_i, \qquad \mathbb{E}[\mathbf{z}_i\boldsymbol{\zeta}_i^{\mathrm{T}}] = \mathbf{0}, \tag{II.8}$$

where $\boldsymbol{\zeta}_i$ is the error from the best linear approximation of $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ based on $\mathbf{z}_i$. This approximation error will not be small in general, because our result allows for generic high-dimensional first-step covariates. However, in some special cases it can be small as we discuss further below.

The following assumption collects the key restrictions we impose on the first-step procedure.

Assumption II.2 (First-step)
(i) $\max_{1\le i\le n}|\hat{\mu}_i - \mu_i| = o_{\mathrm{p}}(1)$. *(ii)* $\mathbb{E}[|\eta_i|^2] = o(n^{-1/2})$ *and* $\mathbb{E}[|\eta_i|^2]\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = o(n^{-1})$. ‖

This assumption imposes high-level conditions on the covariates $\mathbf{z}_i$ entering the first-step estimate (II.7), covering both series-based nonparametric estimation and, more generally, many covariates settings. Assumption II.2(i) requires uniform consistency of $\hat{\mu}_i$ for μ_i only, without a convergence rate. Primitive conditions can be found in the vast literatures on nonparametric sieve estimation and high-dimensional models. Underlying this assumption is the implicit requirement that the best linear approximation of μ_i based on $\mathbf{z}_i$ in (II.6) should vanish asymptotically.

Assumption II.2(ii) concerns the approximation power of the covariates $\mathbf{z}_i$ explicitly, measured in terms of the mean squared error of best linear approximations. It requires, at least, that the best linear approximation error in (II.6) is sufficiently small relative to the sample size in mean square. The condition $\mathbb{E}[|\eta_i|^2] = o(n^{-1/2})$ cannot be dropped without affecting the interpretation of the final estimand $\boldsymbol{\theta}_0$ because the first-step best linear approximation error will affect (in general) the probability limit of the resulting two-step estimator. In other words, either the researcher assumes that the best linear approximation is approximately exact in large samples, or needs to change the interpretation of the probability limit of the two-step estimator because of the misspecification introduced in the first step. The latter approach is common in empirical work, where researchers often employ a "flexible" parametric model, such as linear regression, Probit or Logit, all of which are misspecified in general.

Furthermore, the exact quality of approximation for the first-step estimate required in Assumption II.2(ii) depends on the quality of approximation in (II.8). At one extreme, the covariates $\mathbf{z}_i$ may not offer any approximation of $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ in mean square, in which case $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = O(1)$, and hence the relevant restriction becomes $\mathbb{E}[|\eta_i|^2] = o(n^{-1})$. This corresponds to the case of many generic covariates $\mathbf{z}_i$ and non-linear $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i]$, that is, cases where $\mathbf{z}_i$ are not basis of approximation and/or $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ can not be well approximated by a linear combination of $\mathbf{z}_i$.

At the other extreme, if $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ can be well approximated by the best linear mean square prediction based on $\mathbf{z}_i$ so that, at least, $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = O(n^{-1/2})$, then the relevant restriction on the first-step estimate becomes $\mathbb{E}[|\eta_i|^2] = o(n^{-1/2})$. This case encompasses the standard two-step semiparametric setup, where the covariates $\mathbf{z}_i$ include basis expansions able to approximate $\mu_i = \mathbb{E}[r_i|\mathbf{z}_i]$ and $\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu, \boldsymbol{\theta}_0)|\mathbf{z}_i]$ accurately enough in mean square (usually justified by smoothness of these conditional expectations). From this perspective, the sufficient conditions $\mathbb{E}[|\eta_i|^2] = o(n^{-1/2})$ and $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = O(n^{-1/2})$ reassemble the usual requirement of better than $n^{1/4}$-consistency of first-step nonparametric estimators in two-step semiparametrics (see Cattaneo and Jansson, 2018, and references therein), but this is imposed only on best linear approximation errors (i.e., misspecification/smoothing bias), which are exacerbated for small k and not for large k, the latter being the main focus of the present chapter.

II.3.2 Distribution Theory

It is not difficult to establish $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{p}} \boldsymbol{\theta}_0$, even when $k/\sqrt{n} = O(1)$. Thus, we impose the following high-level assumption.

Assumption II.3 (Consistency)
(i) $\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{p}} \boldsymbol{\theta}_0$ *the unique solution of* $\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta})] = \mathbf{0}$ *and an interior point of* Θ. *(ii)* $\boldsymbol{\Omega}_n \xrightarrow{\mathrm{p}} \boldsymbol{\Omega}_0$ *positive definite.* ‖

On the other hand, the $\sqrt{n}$-scaled mean squared error and distributional properties of the estimator $\hat{\boldsymbol{\theta}}$ will change depending on whether k is "small" or "large" relative to the sample size. To describe heuristically the result, consistency of $\hat{\boldsymbol{\theta}}$ and a second-order Taylor series expansion give:

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx \quad & \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) && \text{(II.9)}\\
+ \quad & \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big) && \text{(II.10)}\\
+ \quad & \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0 \sum_{i=1}^{n} \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big)^2, && \text{(II.11)}
\end{aligned}$$

where $\boldsymbol{\Sigma}_0 = -(\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{M}_0)^{-1}\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0$.

Term (II.9) will be part of the influence function. Using conventional large sample approximations (i.e., k fixed or at most $k/\sqrt{n} \to 0$), term (II.10) contributes to the variability of $\hat{\boldsymbol{\theta}}$ as a result of estimating the first step, and term (II.11) will be negligible. Here, however, we show that under the many covariates assumption $k/\sqrt{n} \not\to 0$, both (II.10) and (II.11) will deliver nonvanishing bias terms. The main intuition is as follows: as the number of covariates increases relative to the sample size, the error in $\hat{\mu}_i - \mu_i$ also increases and features in terms (II.10) and (II.11). This, in turn, affects the finite sample performance of the usual asymptotic approximations, delivering unreliable results in applications. To be specific, the term (II.10) contributes a leave-in bias arising from using the same observation to estimate μ_i and later the parameter $\boldsymbol{\theta}_0$, while the term (II.11) contributes with a bias arising from averaging (non-linear) squared errors in the estimation of μ_i.

The following theorem formalizes our main finding. The proof relies on several preliminary results given in Section II.8. Let $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n]^{\mathrm{T}}$ be the first step included covariates and $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-}\mathbf{Z}^{\mathrm{T}}$ be the projection matrix with elements $\{\pi_{ij} : 1 \leq i, j \leq n\}$.

Theorem II.1 (Asymptotic normality)
Suppose Assumption II.1, II.2 and II.3 hold. If $k = O(\sqrt{n})$, then (II.2) holds with

$$\mathscr{B} = \mathbf{\Sigma}_0 \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbf{B}_i|\mathbf{Z}], \quad \mathscr{V} = \frac{1}{n}\mathbf{\Sigma}_0 \left(\mathbb{V}[\mathbb{E}[\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{Z}]] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}[\mathbf{\Psi}_i|\mathbf{Z}] \right) \mathbf{\Sigma}_0,$$

where

$$\begin{aligned} \mathbf{B}_i &= \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(r_i - \mu_i)\pi_{ii} + \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\sum_{j=1}^{n}(r_j - \mu_j)^2\pi_{ij}^2, \\ \mathbf{\Psi}_i &= \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) + \Big(\sum_{j=1}^{n} \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)|\mathbf{Z}]\pi_{ij} \Big)(r_i - \mu_i). \end{aligned}$$

‖

Using well known properties of projection matrices, it follows that $\mathscr{B} = O_{\mathrm{p}}(k/n)$ and non-zero in general, and thus the distributional approximation in Theorem II.1 will exhibit a first-order asymptotic bias $\mathscr{V}^{-1/2}\mathscr{B}$ whenever k is "large" relative to the sample size (e.g., $k \propto \sqrt{n}$). In turn, this result implies that conventional inference procedures ignoring this first-order distributional bias will be invalid, leading to over-rejection of the null hypothesis of interest and under-coverage of the associated confidence intervals. Section II.6 presents simulation evidence capturing this phenomena.

To understand the implications of the above theorem, we discuss the two terms in $\mathbf{B}_i$. The first term corresponds to the contribution from (II.10), because a first order approximation gives $\mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \approx \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(\hat{\mu}_i - \mu_i) \approx \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(\sum_j \pi_{ij}(r_j - \mu_j))$. Because $\mathbb{E}[r_j - \mu_j|\mathbf{z}_j] = 0$, this bias is proportional to the sample average of $\mathbb{C}\mathsf{ov}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0), r_i - \mu_i|\mathbf{z}_i]\pi_{ii}$. Hence the bias, due to the linear contribution of $\hat{\mu}_i$, will be zero if there is no residual variation in the sensitivity measure $\dot{\mathbf{m}}$ (i.e., $\mathbb{V}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i] = \mathbf{0}$) or, more generally, the residual variation in the sensitivity measure $\dot{\mathbf{m}}$ is uncorrelated to the first step error term (i.e., $\mathbb{C}\mathsf{ov}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0), r_i - \mu_i|\mathbf{z}_i] = \mathbf{0}$).

The second term in $\mathbf{B}_i$ captures the quadratic dependence of the estimating equation on the unobserved μ_i, coming from (II.11). Because of the quadratic nature, this bias represents the accumulated estimation error when $\hat{\mu}_i$ is overfitted. When $i \neq j$, which is the main part of the bias, $\mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)(r_j - \mu_j)^2|\mathbf{z}_i, \mathbf{z}_j] = \mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i]\mathbb{E}[(r_j - \mu_j)^2|\mathbf{z}_j]$, and hence this portion of the bias will be non-zero unless an estimating equation linear in μ_i is considered or, slightly more generally, $\mathbb{E}[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i] = \mathbf{0}$. Intuitively, overfitting the first step does not give a quadratic contribution if the estimating equation is not sensitive to the first step on average to the second order.

The first bias can be manually removed by employing a leave-one-out estimator of μ_i. However, the second bias cannot be removed this way. Furthermore, the leave-one-out estimator $\hat{\mu}_i^{(i)}$ usually has higher variability compared with $\hat{\mu}_i$, hence the second bias will be amplified, which is confirmed by our simulations.

Chernozhukov, Escanciano, Ichimura, Newey and Robins (2018b) introduced the class of locally robust estimators, which are a generalization of doubly robust estimators (e.g., Bang and Robins, 2005) and the efficient influence function estimators (e.g., Cattaneo, 2010, p. 142). These estimators can offer demonstrable improvements in terms of smoothing/approximation bias rate restrictions and, consequently, they offer robustness to "small" k (underfitting). See also Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2018a) and Newey and Robins (2018) for related approaches. This type of estimators are carefully constructed so that (II.10) is removed, but they do not account for (II.11). Because the "large" k bias is in part characterized by (II.11), locally robust estimators cannot (in general) reduce the bias we uncover in this chapter. Therefore, our methods complement locally robust estimation by offering robustness to overfitting, that is, situations where the first step estimate includes possibly many covariates. Cattaneo and Jansson (2018) illustrate this fact in the context of kernel-based estimation.

Consider next the variance and distributional approximation. Theorem II.1 shows that the distributional properties of $\hat{\boldsymbol{\theta}}$ are based on a double sum in general, and hence it does not have an "influence function" or asymptotically linear representation. Nevertheless, after proper Studentization, asymptotic normality holds as in (II.2). The following remark summarizes the special case when the estimator, after bias correction, does have an asymptotic linear representation.

Remark II.1 (Asymptotic linear representation) Suppose the conditions of Theorem II.1 hold. If, in addition, $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = o(1)$, then

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \mathscr{B}) = \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) + \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|\mathbf{z}_i](r_i - \mu_i) \Big\} + o_{\mathrm{p}}(1),$$

hence $\hat{\boldsymbol{\theta}}$ is asymptotically linear after bias correction even when $k/\sqrt{n} \not\to 0$. However, $\hat{\boldsymbol{\theta}}$ is asymptotically linear if and only if $k/\sqrt{n} \to 0$ in general. See Newey (1994) and Hahn and Ridder (2013) for more discussion on asymptotic linearity and variance calculations. ‖

In practice one needs to estimate both the bias and the variance to conduct valid statistical inference. Plug-in estimators could be constructed to this end, though additional unknown functions would need to be estimated (e.g., conditional expectations of derivatives of the estimating equation). Under regularity conditions, these estimators would be consistent for the bias and variance terms. As a practically relevant alternative, we show in the upcoming sections that the jackknife can be used to estimate both the bias and variance, and that a carefully crafted resampling method can be used to conduct inference. The key advantage of these results is that they are fully automatic, and therefore can be used for any model considered in practice without having to re-derive and plug-in for the exact expressions each time.

Remark II.2 (Delta method) Our results apply directly to many other estimands via the so-called delta method. Let $\boldsymbol{\varphi}(\cdot)$ be a possibly vector-valued continuously differentiable function of the parameter $\boldsymbol{\theta}_0$ with gradient $\dot{\boldsymbol{\varphi}}(\cdot)$. Then, under the conditions of Theorem II.1,

$$\Big(\dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\mathscr{V}\dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)^{\mathrm{T}}\Big)^{-1/2}\Big(\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}})-\boldsymbol{\varphi}(\boldsymbol{\theta}_0)-\dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)\mathscr{B}\Big)\ \overset{\mathrm{d}}{\to}\ \mathrm{Normal}(\mathbf{0},\ \mathbf{I}),$$

provided that $\dot{\boldsymbol{\varphi}}(\boldsymbol{\theta}_0)$ is full rank. Hence, the usual delta method can be used for estimation and inference in our setting, despite the presence of potentially many covariates entering the first-step estimate. ‖

Plug-in consistent estimation of the appropriate GMM efficient weighting matrix is also possible given our regularity conditions, but we do not give details here to conserve space.

II.4 Jackknife Bias Correction and Variance Estimation

We show that the jackknife is able to estimate consistently the many covariate bias and the asymptotic variance of $\hat{\boldsymbol{\theta}}$, even when $k = O(\sqrt{n})$, and without assuming a valid asymptotic linear representation for $\hat{\boldsymbol{\theta}}$.

The jackknife estimates are constructed by simply deleting one observation at the time and then re-estimating both the first and second steps. To be more specific, let $\hat{\mu}_i^{(\ell)}$ denote the first-step estimate after the ℓ^{th} observation is removed from the dataset. Then, the leave-ℓ-out two-step estimator is

$$\hat{\boldsymbol{\theta}}^{(\ell)} = \arg\min_{\boldsymbol{\theta}} \Big|\boldsymbol{\Omega}_n^{1/2} \sum_{i=1,i\neq\ell}^{n} \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^{(\ell)}, \boldsymbol{\theta})\Big|, \qquad \ell = 1, 2, \ldots, n.$$

Finally, the bias and variance estimates are constructed as in (II.4). This approach is fully data-driven and automatic. In addition, another appealing feature of the jackknife in our case is that it is possible to exploit the specific structure of the problem to reduce computational burden. Specifically, because we consider a linear regression fit for the first step, the leave-ℓ-out estimate $\hat{\mu}_i^{(\ell)}$ can easily be obtained by

$$\hat{\mu}_i^{(\ell)} = \hat{\mu}_i + \frac{\hat{\mu}_\ell - r_\ell}{1 - \pi_{\ell\ell}} \cdot \pi_{i\ell}, \qquad 1 \leq i \leq n,$$

where recall that $\pi_{i\ell}$ is the $(i,\ell)^{\text{th}}$ element of the projection matrix for the first step $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-}\mathbf{Z}^{\mathrm{T}}$. Since recomputing the first-step estimate can be time-consuming when k is large, the above greatly simplifies the algorithm and reduces computing time.

To show the validity of the jackknife, we impose the following additional mild assumptions on the possibly large dimensional covariates $\mathbf{z}_i$, captured through the projection matrix of the first-step estimate.

Assumption II.4 (Jackknife)
(i) $\sum_{1\leq i\leq n} \pi_{ii}^2 = o_{\mathrm{p}}(k)$. *(ii)* $\max_{1\leq i\leq n} 1/(1-\pi_{ii}) = O_{\mathrm{p}}(1)$. ‖

The first two conditions together correspond to "design balance", which states that asymptotically the projection matrix is not "concentrated" on a few observations. They are slightly weaker than $\max_{1\leq i\leq n} \pi_{ii} = o_{\mathrm{p}}(1)$, which is commonly assumed in the literature on high-dimensional statistics. For more discussion on design balance in linear least squares models see, e.g., Chatterjee and Hadi (1988). With these conditions, we obtain the following result.

Theorem II.2 (Jackknife-based valid inference)
Suppose Assumption II.1, II.2, II.3 and II.4 hold. If $k = O(\sqrt{n})$, then (II.3) holds. ‖

By showing the validity of the jackknife, one can construct confidence intervals and conduct hypothesis tests using the jackknife bias and variance estimators, and the normal approximation. In particular, bias correction will not affect the variance of the asymptotic distribution. On the other hand, any bias correction technique is likely to introduce additional variability, which can be nontrivial in finite samples. This is indeed confirmed by our simulation studies. In the next section, we introduce a carefully crafted fully automatic bootstrap method that can be applied to the bias-corrected Studentized statistic to obtain better finite sample distributional approximations.

Remark II.3 (Delta method) Consider the setup of Remark II.2, where the goal is to conduct estimation and inference for a (smooth) function of $\boldsymbol{\theta}_0$. In this case, the estimator is $\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}})$. There are at least three ways to conduct bias correction: (i) plug-in method leading to $\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}} - \hat{\mathscr{B}})$, (ii) linearization-based method leading to $\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}}) - \dot{\boldsymbol{\varphi}}(\hat{\boldsymbol{\theta}})\hat{\mathscr{B}}$, and (iii) direct jackknife of $\boldsymbol{\varphi}(\hat{\boldsymbol{\theta}})$. The three methods are asymptotically equivalent, and can be easily implemented in practice. The same argument applies to the variance estimator when $\boldsymbol{\varphi}(\boldsymbol{\theta}_0)$ is the target parameter. ‖

II.5 Bootstrap Inference after Bias Correction

In this section we develop a fast, automatic and specifically tailored bootstrap-based approach to conducting post-bias-correction inference in our setting. The method combines the wild bootstrap (first-step estimation) and the multiplier bootstrap (second-step estimation) to give an easy-to-implement valid distributional approximation to the finite sample distribution of the jackknife-based bias-corrected Studentized statistic in (II.3). See Mammen (1993) for a related result in the context of a high-dimensional one-step linear regression model without any bias-correction, and Kline and Santos (2012) for some recent higher-order results in the context of parametric low-dimensional linear regression models.

Let $\omega_i^\star$, $i = 1, 2, \cdots, n$ be i.i.d. bootstrap weights with $\mathbb{E}[\omega_i^\star] = 1$, $\mathbb{V}[\omega_i^\star] = 1$, $\mathbb{E}[(\omega_i^\star - 1)^3] = 0$ and finite fourth moment. First, we describe the bootstrap construction of $\hat{\boldsymbol{\theta}}^\star$. We employ the wild bootstrap to obtain $\hat{\mu}_i^\star$, mimicking the first-step estimate (II.7): we regress $r_i^\star$ on $\mathbf{z}_i$, where $r_i^\star = \hat{\mu}_i + (\omega_i^\star - 1)(r_i - \hat{\mu}_i)$. Then, we employ the multiplier bootstrap to obtain $\hat{\boldsymbol{\theta}}^\star$, mimicking the second-step estimate (II.1):

$$\hat{\boldsymbol{\theta}}^\star = \arg\min_{\boldsymbol{\theta}} \Big| \boldsymbol{\Omega}_n^{1/2} \sum_{i=1}^n \omega_i^\star \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^\star, \boldsymbol{\theta}) \Big|. \tag{II.12}$$

Second, we describe the bootstrap construction of $\hat{\mathscr{B}}$ and $\hat{\mathscr{V}}$; that is, the implementation of the jackknife bias and variance estimators under the bootstrap. Because we employ a multiplier bootstrap, the jackknife estimates need to be adjusted to account for the effective number of observations under the bootstrap law. Thus, we have:

$$\hat{\mathscr{B}}^\star = (n-1)(\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}}^\star), \qquad \hat{\mathscr{V}}^\star = \frac{n-1}{n} \sum_{\ell=1}^n \omega_\ell^\star (\hat{\boldsymbol{\theta}}^{\star,(\ell)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)})(\hat{\boldsymbol{\theta}}^{\star,(\ell)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)})^{\mathrm{T}},$$

where $\hat{\boldsymbol{\theta}}^{\star,(\cdot)} = n^{-1} \sum_{\ell=1}^n \omega_\ell^\star \hat{\boldsymbol{\theta}}^{\star,(\ell)}$, and

$$\hat{\boldsymbol{\theta}}^{\star,(\ell)} \;=\; \arg\min_{\boldsymbol{\theta}} \Big| \boldsymbol{\Omega}_n^{1/2} \Big\{ \sum_{i=1}^n \big[\omega_i^\star - \mathbb{1}_{(i=\ell)}\big] \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i^{\star,(\ell)}, \boldsymbol{\theta}) \Big\} \Big|, \qquad \ell = 1, 2, \dots, n.$$

Here $\hat{\mu}_i^{\star,(\ell)}$ is obtained by regressing $r_i^\star$ on $\mathbf{z}_i$, without using the ℓ^{th} observation. Equivalently, the jackknife deletes the ℓ^{th} observation in the first step wild bootstrap, and reduces the ℓ^{th}

weight $\omega_\ell^\star$ by 1 in the second step multiplier bootstrap.

Our resampling approach employs the wild bootstrap to form $\hat{\mu}_i^\star$, which is very easy and fast to implement and does not require recomputing the possibly high-dimensional projection matrix $\mathbf{\Pi}$, and then uses the same bootstrap weights to construct $\hat{\boldsymbol{\theta}}^\star$ via a multiplier resampling approach. It is possible to use the multiplier bootstrap for both estimation steps, which would give a more unified treatment, but such an approach is harder to implement and does not utilize efficiently (from a computational point of view) the specific structure of the first-step estimate. To be more specific, employing the multiplier bootstrap in the first-step estimation leads to $\hat{\mu}_i^\star = \mathbf{z}_i^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{W}^\star\mathbf{Z})^{-}\mathbf{Z}^{\mathrm{T}}\mathbf{W}^\star\mathbf{R}$, where $\mathbf{R} = [r_1, r_2, \ldots, r_n]^{\mathrm{T}}$ and $\mathbf{W}^\star$ is a diagonal matrix with diagonal elements $\{\omega_i^\star\}_{1\leq i\leq n}$, which requires recomputing the projection matrix for each bootstrap replication. In contrast, our bootstrap approach leads to $\hat{\mu}_i^\star = \mathbf{z}_i^{\mathrm{T}}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-}\mathbf{Z}^{\mathrm{T}}\mathbf{R}^\star$, where $\mathbf{R}^\star = [r_1^\star, r_2^\star, \ldots, r_n^\star]^{\mathrm{T}}$. As discussed before, this important practical simplification also occurs because we are employing a linear regression fit in the first step. Employing the standard nonparametric bootstrap may also be possible, but additional (stronger) regularity conditions would be required. Last but not least, we note that combining the jackknife with the multiplier bootstrap naïvely (that is, deleting the ℓ^{th} observation with its weight $\omega_\ell^\star$ altogether in the second step) does not deliver a consistent variance estimate.

Only two additional mild, high-level conditions on the bootstrap analogue first-step and second-step estimators are imposed as follows.

Assumption II.5 (Bootstrap)
(i) $\max_{1\leq i\leq n}|\hat{\mu}_i^\star - \hat{\mu}_i| = o_{\mathrm{p}}(1)$. *(ii)* $|\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}}| = o_{\mathrm{p}}(1)$. ‖

The following theorem summarizes our main result for inference employing the bootstrap after jackknife bias and variance estimation.

Theorem II.3 (Bootstrap validity)
Suppose Assumption II.1, II.2, II.3, II.4 and II.5 hold. If $k = O(\sqrt{n})$, then (II.5) holds. ‖

It is common to assume the bootstrap weights $\omega_i^\star$ to have mean 1 and variance 1. For the jackknife bias and variance estimator to be consistent under the bootstrap distribution, we also need that the third central moment of $\omega_i^\star$ is zero. Examples include $\omega_i^\star = 1 + e_i^\star$ with $e_i^\star$ following the Rademacher distribution or the six-point distribution proposed in Webb (2014).

For inference, consider for example the one dimensional case: $\dim(\boldsymbol{\theta}_0) = 1$. The bootstrap percentile-t bias-corrected (equal tail) confidence interval for θ_0 is

$$\left[\ \hat{\theta} - \hat{\mathscr{B}} - \hat{q}_{1-\alpha/2}\cdot\sqrt{\hat{\mathscr{V}}}\quad,\quad \hat{\theta} - \hat{\mathscr{B}} - \hat{q}_{\alpha/2}\cdot\sqrt{\hat{\mathscr{V}}}\ \right],$$

where $\hat{q}_\alpha = \inf\{t \in \mathbb{R} : \hat{F}(t) \geq \alpha\}$ is the empirical α^{th} quantile of $\{\mathscr{T}_b^\star : 1 \leq b \leq B\}$, with $\hat{F}(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}[\mathscr{T}_b^\star \leq t]$ and $\mathscr{T}_b^\star$ denoting the bootstrap statistic in (II.5) in b^{th} simulation.

II.6 Numerical Evidence

We provide numerical evidence for the methods developed in this chapter. First, we offer a short introduction to the marginal treatment effect, and then present a Monte Carlo experiment constructed in the context of MTE estimation, which highlights the role of the many covariates bias and showcases the role of jackknife bias correction and bootstrap approximation for estimation and inference. Second, also in the context of MTE estimation and inference, we offer an empirical illustration following the work of Carneiro, Heckman and Vytlacil (2011).

II.6.1 Marginal Treatment Effect

Originally proposed by Björklund and Moffitt (1987), and later developed and popularized by Heckman and Vytlacil (2005) and Heckman, Urzua and Vytlacil (2006), the marginal treatment effect (MTE) is an important parameter of interest in program evaluation and causal inference. Not only it can be viewed as a limiting version of the local average treatment effect (LATE) of Imbens and Angrist (1994) for continuous instrumental variables (c.f. Angrist, Graddy and Imbens, 2000), but also it can be used to unify and interpret many other treatment effects parameters such as the average treatment effect or the treatment effect on the treated. Another appealing feature of the MTE is that it provides a description of treatment effect heterogeneity.

To describe the MTE, we adopt a potential outcomes framework under random sampling. Suppose $(Y_i, T_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, 2, \ldots, n$, is i.i.d., where Y_i is the outcome of interest, T_i is a treatment status indicator, $\mathbf{X}_i \in \mathbb{R}^{d_x}$ is a d_x-variate vector of observable characteristics, and $\mathbf{Z}_i \in \mathbb{R}^k$ is k-variate vector of "instruments" (which may include $\mathbf{X}_i$ and transformations thereof). The observed data is generated according to the following switching regression model, also known as potential outcomes or the Roy model,

$$\begin{aligned} Y_i &= T_i Y_i(1) + (1 - T_i) Y_i(0), \quad Y_i(1) = g_1(\mathbf{X}_i) + U_{1i},\ Y_i(0) = g_0(\mathbf{X}_i) + U_{0i}, && \text{(II.13)} \\ T_i &= \mathbb{1}[P_i \geq V_i], \quad P_i = P(\mathbf{Z}_i) = \mathbb{E}[T_i|\mathbf{Z}_i],\ V_i|\mathbf{X}_i \sim \text{Uniform}[0,1], && \text{(II.14)} \end{aligned}$$

where $Y_i(1)$ and $Y_i(0)$ are the potential outcomes when an individual receives the treatment or not, (U_{1i}, U_{0i}, V_i) are unobserved error terms, and P_i is the propensity score or probability of selection. The selection equation (II.14) is taken essentially without loss of generality to be of the single threshold-crossing form (see Vytlacil, 2002, for more discussion), though this representation may affect the interpretation of the unobserved heterogeneity.

The (conditional on $\mathbf{X}_i$) MTE at level a is defined as

$$\tau_{\texttt{MTE}}(a|\mathbf{x}) = \mathbb{E}[Y_i(1) - Y_i(0)|V_i = a, \mathbf{X}_i = \mathbf{x}].$$

The MTE will be constant in a if either (i) the individual treatment effect $Y_i(1) - Y_i(0)$ is constant, or (ii) there is no selection on unobservables, that is, the error terms of the outcome equation (II.13) are unrelated to that of the selection equation (II.14). The parameter $\tau_{\texttt{MTE}}(a|\mathbf{x})$ is understood as the treatment effect for the subpopulation where an infinitesimal increase in the propensity score leads to a change in participation status. Note that for a close to 1, the MTE measures the treatment effect in a subpopulation that is very unlikely to be treated. Other treatment and policy effects can be recovered using the MTE.

Two assumptions are made to facilitate identification. First, the collection of instruments $\mathbf{Z}_i$ is nondegenerate and independent of the error terms (U_{1i}, U_{0i}, V_i) conditional on the covariates $\mathbf{X}_i$. Second, $0 < \mathbb{P}[T_i = 1|\mathbf{X}_i] < 1$, so that conditional on the covariates, both treated and untreated individuals are observable in the population. It can then be shown that, for any limit point a in the support of the propensity score, $\tau_{\texttt{MTE}}(a|\mathbf{x})$ is

$$\tau_{\texttt{MTE}}(a|\mathbf{x}) = \frac{\partial}{\partial a}\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}].$$

This representation shows that the MTE is identifiable, and could in principle be estimated by standard nonparametric techniques (once P_i is estimated). In practice, however, nonparametric methods for estimating $\tau_{\texttt{MTE}}(a|\mathbf{x})$ and functionals thereof are often avoided because of the curse of dimensionality, the negative impact of smoothing and tuning parameters, and efficiency considerations. A flexible parametric functional form can be used instead: $\mathbb{E}[Y_i|P_i, \mathbf{X}_i] = e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)$, where $e(\cdot)$ is a known function up to some finite dimensional parameter $\boldsymbol{\theta}_0$.

Therefore, the MTE estimator is often constructed as follows:

$$\begin{aligned}
\hat{\tau}_{\texttt{MTE}}(a|\mathbf{x}) &= \frac{\partial}{\partial a}e(\mathbf{x}, a, \hat{\boldsymbol{\theta}}), \qquad \hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \Big(Y_i - e(\mathbf{X}_i, \hat{P}_i, \boldsymbol{\theta})\Big)^2, \\
\hat{P}_i &= \mathbf{Z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \Big(T_i - \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\beta}\Big)^2,
\end{aligned}$$

Identification and estimation of the MTE, as well as other policy-relevant parameters based on it, require exogenous variation in the treatment equation (II.14) induced by instrumental variables. In practice, researchers induce this variation by (i) employing many instruments, possibly generating them using power expansions and interactions, and (ii) including interactions with the "raw" or expanded instruments. Employing a flexible, high-dimensional specification for the probability of selection is also useful to mitigate misspecification errors. These observations have led researchers to employ many covariates/instruments in the probability of selection, that is, have a "large" k relative to the sample size. In this paper, we show that flexibly modeling the probability of selection can lead to a first-order bias in the estimation of the MTE and related policy-relevant estimands, even when the outcome equation is modeled parametrically and low-dimensional. Furthermore, we provide automatic bias-correction and inference procedures based on resampling methods.

The following result characterizes the asymptotic properties of the estimated MTE.

Corollary II.1 (Asymptotic Normality: MTE)
Suppose the assumptions of Theorem II.1 hold. Then, for $\hat{\boldsymbol{\theta}}$,

$$\begin{aligned}\mathbf{B}_i &= \frac{\partial^2 e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial P_i \partial \boldsymbol{\theta}} \Big[(1-P_i)\cdot \mathbb{E}[T_i Y_i(1)|\mathbf{Z}_i] - P_i \cdot \mathbb{E}[(1-T_i)Y_i(0)|\mathbf{Z}_i]\Big]\pi_{ii} \\ &\quad + \frac{1}{2}\sum_{j=1}^{n}\left[\frac{\partial^2 e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial P_i \partial \boldsymbol{\theta}}\tau_{\mathtt{MTE}}(P_i|\mathbf{X}_i) + \frac{1}{2}\frac{\partial e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial \tau_{\mathtt{MTE}}(P_i|\mathbf{X}_i)}{\partial P_i}\right] P_j(1-P_j)\pi_{ij}^2, \\ \boldsymbol{\Psi}_i &= \frac{\partial e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\Big(Y_i - e(\mathbf{X}_i, P_i, \boldsymbol{\theta}_0)\Big) - \left(\sum_{j=1}^{n}\frac{\partial e(\mathbf{X}_j, P_j, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\tau_{\mathtt{MTE}}(P_j|\mathbf{X}_j)\pi_{ij}\right)(T_i - P_i). \quad \|\end{aligned}$$

The above result gives a precise characterization of the asymptotic possibly first-order bias and variance of $\hat{\boldsymbol{\theta}}$ via the results in Theorem II.1. To obtain the corresponding result for the estimated MTE, $\hat{\tau}_{\mathtt{MTE}}(a|\mathbf{x})$, the delta method is employed and an extra multiplicative factor $\partial^2 e(\mathbf{x}, a, \boldsymbol{\theta}_0)/\partial a \partial \boldsymbol{\theta}^{\mathrm{T}}$ shows up. As a result, both the bias and variance for the estimated MTE will depend on the evaluation point $(\mathbf{x}|a)$.

To understand the implications of the above corollary, we consider the bias terms. Note that the factor associated with π_{ii} essentially captures treatment effect heterogeneity (in the outcome equation) and self-selection. To make it zero, one needs to assume there is no heterogeneous treatment effect and that the agents do not act on idiosyncratic characteristics that are unobservable to the analyst. For the second bias term associated with π_{ij}^2, note that it involves both the level of the MTE and its curvature. Hence the second bias is related not only to treatment effect heterogeneity captured through the shape of the MTE, but also to the magnitude of the treatment effect. Thus, aside from the off chance of these terms canceling each other, the many instruments bias will be zero only when there is neither heterogeneity nor self-selection, and the treatment effect is zero. Since these conditions are unlikely to hold in empirical work, even in randomized controlled trials, we expect the many instruments bias to have a direct implication in most practical cases. Therefore, conventional estimation and inference methods that do not account for the many instruments bias will be invalid, even in large samples, when many instruments are included in the estimation.

II.6.2 Simulation Study

We set the potential outcomes to $Y_i(0) = U_{0i}$ and $Y_i(1) = 0.5 + U_{1i}$. We assume there are many potential instruments $\mathbf{Z}_i = [1, Z_{1,i}, Z_{2,i}, \ldots, Z_{199,i}]$, with $Z_{\ell,i} \sim \text{Uniform}[0,1]$ independent across $\ell = 1, 2, \ldots, 199$. The selection equation is assumed to take a very parsimonious form: $T_i = \mathbb{1}\left[0.1 + Z_{1,i} + Z_{2,i} + Z_{3,i} + Z_{4,i} \geq V_i\right]$. In this case Assumption II.2 holds automatically without misspecification error. Finally, the error terms are distributed as $V_i|\mathbf{Z}_i \sim \text{Uniform}[0,1]$, $U_{0i}|\mathbf{Z}_i, V_i \sim \text{Uniform}[-1,1]$ and $U_{1i}|\mathbf{Z}_i, V_i \sim \text{Uniform}[-0.5, 1.5 - 2V_i]$. Because additional covariates $\mathbf{X}_i$ do not feature in this data generating process, the treatment effect heterogeneity and self-selection are captured by the correlation between U_{1i} and V_i.

It follows that $\mathbb{E}[Y_i|P_i = a] = a - \frac{a^2}{2}$, and the MTE is $\tau_{\texttt{MTE}}(a) = 1 - a$. Given a random sample index by $i = 1, 2, \ldots, n$, the second-step regression model is set to $\mathbb{E}[Y_i|P_i] = \theta_1 + \theta_2 \cdot P_i + \theta_3 \cdot P_i^2$ and therefore the estimated MTE is $\hat{\tau}_{\texttt{MTE}}(a) = \hat{\theta}_2 + 2a \cdot \hat{\theta}_3$ with $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)'$ denoting the least-squares estimators of $(\theta_1, \theta_2, \theta_3)'$. We consider the quantity $\sqrt{n}\,(\hat{\tau}_{\texttt{MTE}}(a) - \tau_{\texttt{MTE}}(a))$ at $a = 0.5$, with and without bias correction, for two sample sizes $n = 1,000$ and $n = 2,000$, and across $2,000$ simulation repetitions. To estimate the propensity score, we regress T_i on a constant term and $\{Z_{\ell,i}\}$ for $1 \leq \ell \leq k - 1$, where the number of covariates k ranges from 5 to 200. Note that $k = 5$ corresponds to the most parsimonious model which is correctly specified.

For inference, we consider two approaches. In the conventional approach, the many instruments bias is ignored, and hypothesis testing is based on normal approximation to the t-statistic, where the standard error comes from the simulated sampling variability of the estimator (i.e. the oracle standard error, which is infeasible). That is, this benchmark approach considers the infeasible statistic $(\hat{\tau}_{\texttt{MTE}} - \tau_{\texttt{MTE}})/\sqrt{\mathbb{V}[\hat{\tau}_{\texttt{MTE}}]}$, with $\mathbb{V}[\hat{\tau}_{\texttt{MTE}}]$ denoting the simulation variance of $\hat{\tau}_{\texttt{MTE}}$, and employs standard normal quantiles. The other approach, which follows the results in this chapter, utilizes both the jackknife and the bootstrap: the feasible statistic $(\hat{\tau}_{\texttt{MTE}} - \hat{\mathscr{B}} - \tau_{\texttt{MTE}})/\sqrt{\hat{\mathscr{V}}}$ is constructed as in Section II.4 and inference is conducted using the bootstrap approximation as in Section II.5.

The results are collected in Table II.1. The bias is small with small k, as the most parsimonious model is correctly specified. With more instruments added to the propensity score estimation, the many instruments bias quickly emerges, and without bias correction, it leads to severe empirical undercoverage (conventional 95% confidence is used). Interestingly, the finite sample variance shrinks at the same time. Therefore for this particular DGP, incorporating many instruments not only leads to biased estimates, but also gives the illusion that the parameter is estimated precisely. With jackknife bias correction, there is much less empirical size distortion, and the empirical coverage rate remains well-controlled even with 200 instruments used in the first step. Moreover, the jackknife bias correction also (partially) restores the true variability of the estimator.

Although the focus here is on inference and, in particular, empirical coverage of associated testing procedures, it is also important to know how the bias correction will affect the Standard Deviation (sd) and the Mean Squared Error (MSE) of the point estimators. Recall that the model is correctly specified with 5 instruments, hence it should not be surprising that incorporating bias correction there increases the variability of the estimator and the MSE – although the impact is very small. As more instruments are included, however, the MSE increases rapidly without bias correction, while the MSE of the bias corrected estimator remains relatively stable. In particular, this finding is driven by a sharp reduction in bias that more than compensates the increase in variability of the estimator. A larger variance of the bias-corrected estimator is expected, as additional sampling variability is introduced by the bias correction. All in all, the bias-corrected estimator seems to be appealing not only for inference, but also for point estimation because it performs better in terms of MSE when the number of instruments is moderate or large.

II.6.3 Empirical Illustration

To illustrate our procedure, we consider estimating the marginal returns to college education following the work of Carneiro, Heckman and Vytlacil (2011, CHV hereafter) with MTE methods. The data consists of a subsample of white males from the 1979 National Longitudinal Survey of Youth (NLSY79), and the sample size is $n = 1,747$. The outcome variable, Y_i, is the log wage in 1991, and the sample is split according to the treatment variable $T_i = 0$ (high school dropouts and high school graduates), and $T_i = 1$ (with some college education or college graduates). The dataset includes covariates on individual and family background information, and four "raw" instrumental variables: presence of four-year college, average tuition, local unemployment and wage rate, measured at age 17 of the survey participants.[4]

We normalize the estimates by the difference of average education level between the two groups, so that the estimates are interpreted as the return to per year of college education.

[4] Source: National Longitudinal Surveys, Bureau of Labor Statistics. Disclaimer: This research was conducted with restricted access to Bureau of Labor Statistics (BLS) data. The views expressed here do not necessarily reflect the views of the BLS.

Table II.1. Simulation: marginal treatment effects.

			Conventional					**Bias-Corrected**				
k	k/n	$k/\sqrt{n}$	bias	sd	$\sqrt{\text{mse}}$	coverage	length	bias	sd	$\sqrt{\text{mse}}$	coverage	length
Panel (a) $n = 1000$												
5	0.00	0.16	0.14	4.72	4.73	0.95	18.51	−0.21	4.93	4.93	0.93	18.28
20	0.02	0.63	1.73	4.11	4.46	0.93	16.11	0.18	5.26	5.27	0.94	19.81
40	0.04	1.26	3.08	3.54	4.69	0.86	13.88	1.03	5.11	5.22	0.94	19.67
60	0.06	1.90	3.96	3.22	5.11	0.77	12.63	1.75	5.02	5.32	0.93	19.27
80	0.08	2.53	4.61	3.00	5.50	0.66	11.76	2.28	4.91	5.41	0.92	18.67
100	0.10	3.16	5.10	2.83	5.83	0.56	11.08	2.65	4.78	5.46	0.90	18.28
120	0.12	3.79	5.55	2.67	6.16	0.46	10.48	2.96	4.66	5.51	0.89	17.80
140	0.14	4.43	5.97	2.54	6.49	0.35	9.98	3.24	4.57	5.60	0.87	17.46
160	0.16	5.06	6.35	2.45	6.81	0.26	9.59	3.46	4.43	5.62	0.86	17.15
180	0.18	5.69	6.69	2.33	7.09	0.18	9.13	3.58	4.35	5.63	0.86	16.97
200	0.20	6.32	7.03	2.23	7.38	0.12	8.75	3.81	4.22	5.69	0.84	16.75
Panel (b) $n = 2000$												
5	0.00	0.11	0.13	4.85	4.85	0.95	19.00	−0.12	4.95	4.95	0.93	18.21
20	0.01	0.45	1.42	4.47	4.69	0.94	17.51	0.06	5.16	5.16	0.94	19.31
40	0.02	0.89	2.73	4.17	4.99	0.90	16.36	0.54	5.35	5.38	0.94	19.72
60	0.03	1.34	3.78	3.95	5.47	0.84	15.47	1.18	5.44	5.57	0.93	19.75
80	0.04	1.79	4.62	3.74	5.95	0.76	14.67	1.82	5.43	5.73	0.91	19.59
100	0.05	2.24	5.27	3.55	6.35	0.68	13.91	2.33	5.37	5.86	0.90	19.31
120	0.06	2.68	5.77	3.37	6.68	0.59	13.22	2.74	5.27	5.94	0.90	19.04
140	0.07	3.13	6.27	3.20	7.03	0.49	12.53	3.21	5.11	6.04	0.88	18.85
160	0.08	3.58	6.67	3.07	7.35	0.41	12.03	3.53	5.05	6.16	0.87	18.66
180	0.09	4.02	7.07	2.95	7.65	0.32	11.54	3.87	4.95	6.28	0.85	18.40
200	0.10	4.47	7.42	2.83	7.94	0.26	11.11	4.13	4.84	6.36	0.85	18.22

Note. The marginal treatment effect is evaluated at $a = 0.5$. Panel (a) and (b) correspond to sample size $n = 1000$ and 2000, respectively. Statistics are centered at the true value. $k = 5$ is the correctly specified model. (i) k: number of instruments used for propensity score estimation. (ii) bias: empirical bias (scaled by $\sqrt{n}$). (iii) sd: empirical standard deviation (scaled by $\sqrt{n}$). (iv) $\sqrt{\text{mse}}$: empirical root-MSE (scaled by $\sqrt{n}$). (v) coverage: empirical coverage of a 95% confidence interval. Without bias correction, it is based on normal approximation and simulated sampling variability of the estimator (i.e. the oracle standard error). With bias correction, the test is based on the percentile-t method, where the bias-corrected and Studentized statistic is bootstrapped 500 times (Rademacher weights). (vi) length: the average confidence interval length (scaled by $\sqrt{n}$).

We make the same assumption as in CHV that the error terms are jointly independent of the covariates and the instruments. Then, $\tau_{\texttt{MTE}}(a|\mathbf{x}) = \partial \mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}]/\partial a$ with

$$\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}] = \mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}_0 + a \cdot \mathbf{x}^{\mathrm{T}}\boldsymbol{\delta}_0 + \boldsymbol{\phi}(a)^{\mathrm{T}}\boldsymbol{\theta}_0,$$

where $P_i = \mathbb{P}[T_i = 1|\mathbf{Z}_i]$ is the propensity score, and $\boldsymbol{\phi}$ is some fixed transformation. The covariates $\mathbf{X}_i$ include (i) linear and square terms of corrected AFQT score, education of mom, number of siblings, permanent average local unemployment rate and wage rate at age 17; (ii) indicator of urban residency at age 14; (iii) cohort dummy variables; and (iv) average local unemployment rate and wage rate in 1991, and linear and square terms of work experience in 1991. For the selection equation, the instruments $\mathbf{Z}_i$ include (i), (ii) and (iii) described earlier, as well as (v) the four raw instruments as well as their interactions with corrected AFQT score, education of mom and number of siblings. To make the functional form of the propensity score flexible, we also include interactions among the variables described in (i), and interactions between the cohort dummies and corrected AFQT score, education of mom and number of siblings.

We are employing the same covariates, instruments, and modeling assumptions as in CHV, but our estimation strategy is different than theirs. For the first step, the selection equation (propensity score) is estimated using a linear probability model with $k = 66$ as more interaction terms are included (which implies $k/\sqrt{n} = 1.58$), while CHV employ a Logit model with $k = 35$. Thus, our estimation approach reflects Assumption II.2 in the sense that we assume away misspecification errors from using a flexible (high-dimensional) linear probability model, while CHV assume away misspecification errors from using a lower dimensional Logit model. For the second step, while the specification of $\mathbb{E}[Y_i|P_i = a, \mathbf{X}_i = \mathbf{x}]$ coincides, we estimate the partially linear model (that is, the $\boldsymbol{\phi}(a)$ component) using a flexible polynomial in P_i while CHV employ a kernel local polynomial approach with a bandwidth of about 0.30 over the support $[0, 1]$. To be specific, we implement the second step estimation by using least-squares regression with a fourth-order polynomial of the estimated propensity score $\boldsymbol{\phi}(\hat{P}_i) = [\hat{P}_i, \hat{P}_i^2, \hat{P}_i^3, \hat{P}_i^4]^{\mathrm{T}}$. Here the dimension of $\mathbf{X}_i$ is 23, so the second step model can be regarded as either "flexible" parametric or high-dimensional.

We summarize the empirical findings in Figure II.1, where we plot the estimated MTE evaluate at the sample average of $\mathbf{X}_i$. In the upper panel of this figure, we plot the estimated MTE together with 95% confidence intervals (solid and dashed blue line), using conventional two-step estimation methods (i.e., without bias correction and employing the standard normal approximation). These empirical results are quite similar to those presented by CHV, both graphically and numerically. In particular, for individuals who are very likely to enroll in college, the per year return can be as high as 30%, while the return to college can also be as low as -20% for people who are very unlikely to enroll. Integrating the estimated MTE gives an estimator of the average treatment effect, which is roughly 9%.

The upper panel of Figure II.1, also depicts the bias-corrected MTE estimator (dashed red line). The average treatment effect corresponding to the bias-corrected MTE is 8%, quite close to the previous estimate. On the other hand, the bias-corrected MTE curve has much steeper slope, implying a wider range of heterogeneity for returns to college education. This bias-corrected MTE curve lies close to the boundary of the confidence intervals constructed using the conventional two-step method, hinting at the possibility of a many instruments/covariate bias in the conventional estimate (blue line).

The lower panel of Figure II.1 plots the bias corrected MTE estimator, together with the confidence intervals constructed using our proposed bootstrap-based method, which takes into account the extra variability introduced by bias correction. Not surprisingly, the new confidence intervals are wider than the conventional ones.

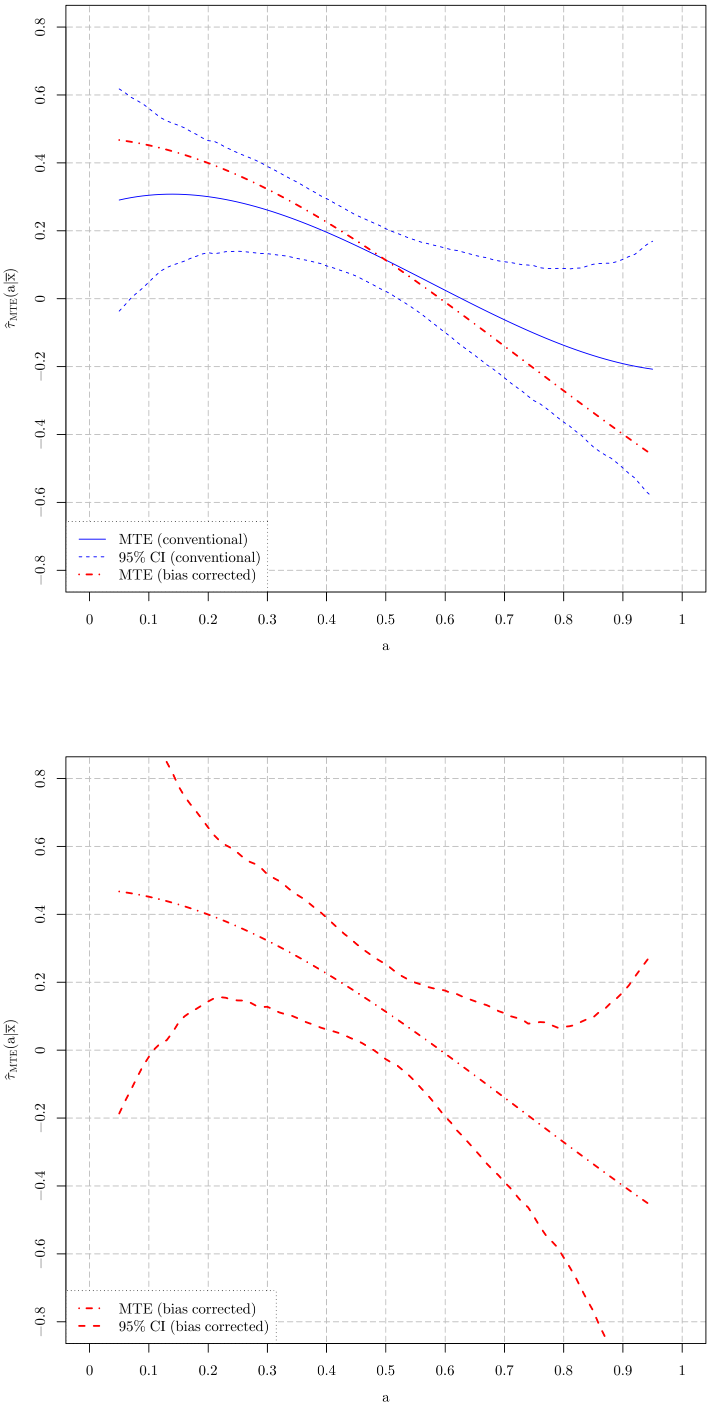
II.7 Conclusion

We studied the distributional properties of two-step estimators, and functionals thereof, when possibly many covariates are used to fit the first-step estimate (e.g., a propensity score, generated regressors or control functions). We show that overfitting in the first step estimation leads to a first-order bias in the distributional approximation of the two-step estimator. As a consequence, the limiting distribution is no longer centered at zero and usual inference procedures become invalid, possibly exhibiting severe empirical size distortions in finite samples.

As a remedy for the many covariates bias we uncover, we develop bias correction methods using the jackknife. Importantly, this approach is data-driven and fully automatic, and does not require additional resampling beyond what would be needed to compute the jackknife standard error, which we show is also consistent in our setting even when many covariates are used. Therefore, implementation is straightforward and is available in any statistical computing software. Furthermore, to improve finite sample inference after bias-correction, we also establish validity of an appropriately modified bootstrap for the jackknife-based bias-corrected Studentized statistic. We demonstrate the performance of our estimation and inference procedures in a comprehensive simulation study and an empirical illustration.

From a more general perspective, our main results give one additional contribution. They shed new light on the ultra-high-dimensional literature: one important implication is that typical sparsity assumptions imposed in that literature cannot be dropped in the context of non-linear models, since otherwise the effective number of included covariates will remain large after model selection, which in turn will lead to a non-vanishing first-order bias in the distributional approximation for the second-step estimator. It would be interesting to explore whether resampling methods are able to successfully remove this many selected or included covariates bias in ultra-high-dimensional settings, where model selection techniques are also used as a first-step estimation device.

Figure II.1. Estimated marginal treatment effects.



Note. The marginal treatment effect, $\hat{\tau}_{\mathtt{MTE}}(a|\bar{\mathbf{X}})$, is evaluated at mean value of the covariates. Bootstrap is used to construct the confidence interval, with 500 repetitions. *Top*: Estimated MTE without bias correction (solid blue line), together with 95% confidence interval (dashed blue line). Also included is the bias-corrected MTE (dashed red line). *Bottom:* Bias-corrected MTE, together with 95% confidence interval, taking into account the effect of bias correction.

II.8 Additional Results and Preliminary Lemmas

In this section we collect some preliminary results which are used to establish the main results in the earlier sections. Proof of these results and the main theorems are given in the next section.

II.8.1 Properties of Projection Matrices

Recall that $\mathbf{\Pi} = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-}\mathbf{Z}^{\mathrm{T}}$ is the projection matrix, with its entries denoted by π_{ij}. Then the first conclusion is that

$$\mathrm{tr}[\mathbf{\Pi}] = k.$$

And since $\mathbf{\Pi}$ is a projection matrix, one has $\mathbf{\Pi\Pi} = \mathbf{\Pi}$, which means

$$\pi_{ij} = \sum_{\ell} \pi_{i\ell}\pi_{j\ell}.$$

Also, $\pi_{ij} = \pi_{ji}$ (i.e., $\mathbf{\Pi}$ is symmetric), and $0 \leq \pi_{ii} \leq 1$ from the idempotency of the projection matrix.

Next consider the trace of $\mathbf{\Pi\Pi} = \mathbf{\Pi}^2$:

$$k = \mathrm{tr}[\mathbf{\Pi}^2] = \sum_i \sum_j \pi_{ij}^2 = \sum_i \pi_{ii}^2 + \sum_{i,j,j\neq i} \pi_{ij}^2,$$

which implies that

$$\sum_i \pi_{ii}^2 \leq k, \qquad \sum_{i,j} \pi_{ij}^2 \leq k.$$

Next we replace π_{ii} by $\sum_j \pi_{ij}^2$, which gives

$$k \geq \sum_i \pi_{ii}^2 = \sum_i \pi_{ii}\left(\sum_j \pi_{ij}^2\right) = \sum_i \sum_j \pi_{ii}\pi_{ij}^2,$$

hence

$$\sum_i \pi_{ii}^3 \leq k, \qquad \sum_{i,j} \pi_{ii}\pi_{ij}^2 \leq k.$$

Now make a further replacement,

$$k \geq \sum_i \pi_{ii}^2 = \sum_i \left(\sum_j \pi_{ij}^2 \right)^2 = \sum_i \pi_{ii}^4 + \sum_{i,j,i \neq j} \pi_{ij}^4 + \sum_{i,j,\ell,j \neq \ell} \pi_{ij}^2 \pi_{i\ell}^2.$$

One direct consequence is that

$$\sum_i \pi_{ii}^4 \leq k, \qquad \sum_{i,j} \pi_{ij}^4 \leq k, \qquad \sum_{i,j,\ell} \pi_{ij}^2 \pi_{i\ell}^2 \leq k.$$

We summarize the above in the following lemma:

Lemma II.1

Let $\mathbf{\Pi}$ *be a projection matrix with rank at most* k*, then:*
(i) $\mathbf{\Pi}$ *is symmetric, nonnegative definite, and* $\mathbf{\Pi}^2 = \mathbf{\Pi}$*, which implies* $\pi_{ij} = \sum_\ell \pi_{i\ell}\pi_{j\ell}$.
(ii) The diagonal elements satisfy

$$0 \leq \pi_{ii} \leq 1 \;\forall i, \qquad \textit{and} \qquad \sum_i \pi_{ii} = \mathrm{tr}[\mathbf{\Pi}] \leq k. \tag{II.15}$$

(iii) The following higher order summations hold:

$$\sum_i \pi_{ii}^2 \leq k, \qquad \sum_{i,j} \pi_{ij}^2 \leq k, \tag{II.16}$$

$$\sum_i \pi_{ii}^3 \leq \sum_i \pi_{ii}^2 \leq k, \qquad \sum_{i,j} \pi_{ii}\pi_{ij}^2 \leq \sum_i \pi_{ii}^2 \leq k, \tag{II.17}$$

$$\sum_i \pi_{ii}^4 \leq \sum_i \pi_{ii}^2 \leq k, \qquad \sum_{i,j} \pi_{ij}^4 \leq \sum_i \pi_{ii}^2 \leq k, \qquad \sum_{i,j,\ell} \pi_{ij}^2 \pi_{i\ell}^2 \leq \sum_i \pi_{ii}^2 \leq k. \tag{II.18}$$

$\|$

II.8.2 Summation Expansion

We first consider the expansion of $(\sum_{i,j,i \neq j} a_{ij})^2$, where $a_{ij} \neq a_{ji}$.

$$\left(\sum_{i,j,i \neq j} a_{ij} \right)^2 = \sum_{\substack{i,j,i',j' \\ i \neq j,\ i' \neq j'}} a_{ij} a_{i'j'} = \sum_{\substack{i,j,i',j' \\ \text{distinct}}} a_{ij} a_{i'j'} + \sum_{\substack{i,j,j' \\ \text{distinct}}} a_{ij} a_{ij'} + \sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij} a_{i'i} + \sum_{\substack{i,j,j' \\ \text{distinct}}} a_{ij} a_{jj'}$$
$$+ \sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij} a_{i'j} + \sum_{\substack{i,j \\ i \neq j}} a_{ij}^2 + \sum_{\substack{i,j \\ i \neq j}} a_{ij} a_{ji}.$$

Note that the two terms $\sum_{\substack{i,j,i' \\ \text{distinct}}} a_{ij}a_{i'i}$ and $\sum_{\substack{i,j,j' \\ \text{distinct}}} a_{ij}a_{jj'}$ are identical by relabeling, hence

Lemma II.2

$$\left(\sum_{i,j,i\neq j} a_{ij}\right)^2 = \sum_{\substack{i,j,i',j' \\ distinct}} a_{ij}a_{i'j'} + \sum_{\substack{i,j,j' \\ distinct}} a_{ij}a_{ij'} + 2\sum_{\substack{i,j,i' \\ distinct}} a_{ij}a_{i'i} + \sum_{\substack{i,j,i' \\ distinct}} a_{ij}a_{i'j} + \sum_{\substack{i,j \\ i\neq j}} a_{ij}^2 + \sum_{\substack{i,j \\ i\neq j}} a_{ij}a_{ji}. \qquad \text{(II.19)}$$

‖

A special case is when $a_{ij} = a_{ji}$ so that the two indices are exchangeable. Then

Lemma II.3

$$(i,j)\text{-}exchangeable \qquad \left(\sum_{i,j,i\neq j} a_{ij}\right)^2 = \sum_{\substack{i,j,i',j' \\ distinct}} a_{ij}a_{i'j'} + 4\sum_{\substack{i,j,i' \\ distinct}} a_{ij}a_{ii'} + 2\sum_{\substack{i,j \\ i\neq j}} a_{ij}^2. \qquad \text{(II.20)}$$

‖

Next we consider $(\sum_{\substack{i,j,\ell \\ \text{distinct}}} a_i b_{ij\ell})^2$, where $b_{ij\ell} = b_{i\ell j}$, i.e. for b the last two indices are exchangeable. For convenience define the following

$$d_i = \sum_{j,\ell,j\neq\ell} b_{ij\ell}, \qquad c_i = \sum_{\substack{j,\ell \\ j\neq i,\ell\neq i,j\neq\ell}} b_{ij\ell}.$$

Then

$$c_i = d_i - 2\sum_j b_{iij} + 2b_{iii} = d_i - 2\sum_{j,j\neq i} b_{iij}.$$

And the decomposition becomes

$$\left(\sum_{\substack{i,j,\ell \\ \text{distinct}}} a_i b_{ij\ell}\right)^2 = \left(\sum_i a_i c_i\right)^2 = \sum_i a_i^2 c_i^2 + \sum_{i,i',i\neq i'} a_i a_{i'} c_i c_{i'}.$$

To make further progress, consider

$$\begin{aligned}
c_i^2 &= \left(d_i - 2\sum_{j,j\neq i} b_{iij}\right)^2 = \left(\sum_{j,\ell,j\neq\ell} b_{ij\ell}\right)^2 + 4\left(\sum_{j,j\neq i} b_{iij}\right)^2 - 4\left(\sum_{j,\ell,j\neq\ell} b_{ij\ell}\right)\left(\sum_{\ell',\ell'\neq i} b_{ii\ell'}\right)\\
&= \sum_{\substack{j,\ell,j',\ell'\\ \text{distinct}}} b_{ij\ell}b_{ij'\ell'} + 4\sum_{\substack{j,\ell,j'\\ \text{distinct}}} b_{ij\ell}b_{ijj'} + 2\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}^2\\
&\qquad + 4\sum_{j,j\neq i} b_{iij}^2 + 4\sum_{\substack{j,\ell\\ j\neq i,\ell\neq i,j\neq\ell}} b_{iij}b_{ii\ell} - 4\sum_{\substack{j,\ell,\ell'\\ j\neq\ell,\ell'\neq i}} b_{ij\ell}b_{ii\ell'},
\end{aligned}$$

and

$$c_ic_{i'} = \left(\sum_{j,\ell,j\neq\ell} b_{ij\ell}\right)\left(\sum_{j,\ell,j\neq\ell} b_{i'j\ell}\right) = \sum_{\substack{j,\ell,j',\ell'\\ \text{distinct}}} b_{ij\ell}b_{i'j'\ell'} + 4\sum_{\substack{j,\ell,\ell'\\ \text{distinct}}} b_{ij\ell}b_{i'j\ell'} + 2\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}b_{i'j\ell}.$$

Therefore we have the following

Lemma II.4

$$\begin{aligned}
&(j,\ell)\textit{-exchangeable}\left(\sum_{\substack{i,j,\ell\\ \textit{distinct}}} a_ib_{ij\ell}\right)^2 = \sum_i a_i^2\left[\sum_{\substack{j,\ell,j',\ell'\\ \textit{distinct}}} b_{ij\ell}b_{ij'\ell'} + 4\sum_{\substack{j,\ell,j'\\ \textit{distinct}}} b_{ij\ell}b_{ijj'} + 2\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}^2\right]\\
&+ 4\sum_i a_i^2\left[\sum_{j,j\neq i} b_{iij}^2 + \sum_{\substack{j,\ell\\ j\neq i,\ell\neq i,j\neq\ell}} b_{iij}b_{ii\ell}\right]\\
&- 4\sum_i a_i^2\left[\sum_{\substack{j,\ell,\ell'\\ j\neq\ell,\ell'\neq i}} b_{ij\ell}b_{ii\ell'}\right] + \sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell,j',\ell'\\ \textit{distinct}}} b_{ij\ell}b_{i'j'\ell'}\right] + 4\sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell,\ell'\\ \textit{distinct}}} b_{ij\ell}b_{i'j\ell'}\right]\\
&+ 2\sum_{i,i',i\neq i'} a_ia_{i'}\left[\sum_{\substack{j,\ell\\ j\neq\ell}} b_{ij\ell}b_{i'j\ell}\right]. \qquad \text{(II.21)}
\end{aligned}$$

$\|$

II.8.3 Preliminary Lemmas

The following lemma justifies an expansion of the estimator.

Lemma II.5
If Assumption II.1, II.2 and II.3 hold, and $k = O(\sqrt{n})$, then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \boldsymbol{\Sigma}_0 \left[\frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\right] \Big(1 + o_{\mathrm{p}}(1)\Big), \tag{II.22}$$

where $\boldsymbol{\Sigma}_0 = -(\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{M}_0)^{-1}\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0$. ‖

A Taylor expansion with respect to the first-step estimate, $\hat{\mu}_i$, gives

$$\begin{aligned}
\frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) = {} & \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) && \text{(II.9)}\\
& + \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big) && \text{(II.10)}\\
& + \frac{1}{\sqrt{n}} \sum_i \frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i - \mu_i\Big)^2 && \text{(II.11)}\\
& + o_{\mathrm{p}}(1).
\end{aligned}$$

The following lemma shows that (II.10) contributes to not only the asymptotic variance, but also the asymptotic bias.

Lemma II.6
If Assumption II.1, II.2 and II.3 hold, and $k = O(\sqrt{n})$, then

$$\text{(II.10)} = \frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right) \cdot \varepsilon_i + \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + o_{\mathrm{p}}(1),$$

where $\mathbf{b}_{1,i} = \mathbb{E}\Big[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \cdot \varepsilon_i\ \Big|\mathbf{z}_i\Big]$. *If, in addition,* $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = o(1)$, *then*

$$\frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right) \cdot \varepsilon_i = \frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\ |\mathbf{z}_i] \cdot \varepsilon_i + o_{\mathrm{p}}(1). \quad \|$$

Inspection of the proof of this lemma shows that only $\mathbb{E}[\eta_i^2] = o(1)$ and $\mathbb{E}[|\boldsymbol{\zeta}_i|^2]\mathbb{E}[\eta_i^2] = o(n^{-1})$ is required; the stronger assumption $\mathbb{E}[\eta_i^2] = o(n^{-1/2})$ will be used when studying the quadratic term (II.11) in the expansion. Furthermore, when $\mathbb{E}[|\boldsymbol{\zeta}_i|^2] = o(1)$, this lemma shows that it is possible to drop the double sum as well as the projection matrix in the variance component, leading to an asymptotic linear representation.

The following lemma shows that the quadratic term (II.11) also contributes a bias.

Lemma II.7
If Assumption II.1, II.2 and II.3 hold, and $k = O(\sqrt{n})$, then

$$\text{(II.11)} = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + O_{\mathrm{p}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathrm{p}}(1),$$

where $\mathbf{b}_{2,ij} = \frac{1}{2}\mathbb{E}\left[\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \cdot \varepsilon_j^2 \;\middle|\mathbf{z}_i, \mathbf{z}_j\right]$. ∥

The following proposition combines the previous lemmas, and gives the asymptotic representation of the estimator $\hat{\boldsymbol{\theta}}$ when $k = O(\sqrt{n})$

Proposition II.1 (Asymptotic representation)
If Assumption II.1, II.2 and II.3 hold, and $k = O(\sqrt{n})$, then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 - \frac{\boldsymbol{\mathcal{B}}}{\sqrt{n}}\right) = \bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 + o_{\mathrm{p}}(1),$$

where

$$\boldsymbol{\mathcal{B}} = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\left[\sum_i \mathbf{b}_{1,i}\pi_{ii} + \sum_{i,j} \mathbf{b}_{2,ij}\pi_{ij}^2\right]$$

$$\bar{\boldsymbol{\Psi}}_1 = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\left[\sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\right],\ \bar{\boldsymbol{\Psi}}_2 = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0\left[\sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right) \cdot \varepsilon_i\right].\ \|$$

Here we use $\boldsymbol{\mathcal{B}}$ to denote the bias term. Note that $\boldsymbol{\mathcal{B}} = O_{\mathrm{p}}(k/\sqrt{n})$ hence is non-vanishing under the assumption that $k \propto \sqrt{n}$. The term $\boldsymbol{\mathcal{B}}$ can be viewed as the bias of the limiting distribution. In the earlier sections, we use $\mathscr{B}$ to denote the bias of $\hat{\boldsymbol{\theta}}$. The two terms are connected through the $\sqrt{n}$-scaling: $\boldsymbol{\mathcal{B}} = \sqrt{n}\mathscr{B}$. In addition, for the asymptotic representation, we use

$$\boldsymbol{\Psi}_i = \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) + \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right) \cdot \varepsilon_i,$$

and therefore $\bar{\boldsymbol{\Psi}}_1 + \bar{\boldsymbol{\Psi}}_2 = \boldsymbol{\Sigma}_0 \sum_i \boldsymbol{\Psi}_i/\sqrt{n}$.

We also consider an asymptotic representation for the bootstrap, implemented without jackknifing.

Proposition II.2 (Asymptotic representation: bootstrap)

Assume II.1, II.2, II.3 and II.5 hold, and $k = O(\sqrt{n})$. Then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \frac{\boldsymbol{\mathcal{B}} + \boldsymbol{\mathcal{B}}'}{\sqrt{n}}\right) = \bar{\boldsymbol{\Psi}}^\star_1 + \bar{\boldsymbol{\Psi}}^\star_2 + o_{\mathrm{p}}(1),$$

where $\boldsymbol{\mathcal{B}}$ is given in Proposition II.1, and

$$\begin{aligned}
\boldsymbol{\mathcal{B}}' &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\left[\sum_i \mathbf{b}_{2,ii} \cdot \pi_{ii}^2 \cdot \mathbb{E}[e_i^{\star 3}]\right] \\
\bar{\boldsymbol{\Psi}}^\star_1 &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\left[\sum_i \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \cdot e_i^\star\right] \\
\bar{\boldsymbol{\Psi}}^\star_2 &= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}}\left[\sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right)\varepsilon_i \cdot e_i^\star\right].
\end{aligned} \qquad \|$$

II.9 Proof

II.9.1 Proof of Lemma II.5

We apply Taylor expansion to the GMM problem, which gives

$$\begin{aligned}
o_{\mathrm{p}}(1) &= \left[\frac{1}{n}\sum_i \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\right]^{\mathrm{T}} \boldsymbol{\Omega}_n \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}) \\
&= \left[\frac{1}{n}\sum_i \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}})\right]^{\mathrm{T}} \boldsymbol{\Omega}_n \\
&\qquad \left(\frac{1}{\sqrt{n}}\sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) + \left[\frac{1}{n}\sum_i \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\right] \sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\right),
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ is (possibly random) convex combination of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Then we have

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= -(\hat{\mathbf{M}}_n^{\mathrm{T}} \boldsymbol{\Omega}_n \tilde{\mathbf{M}}_n)^{-1} \hat{\mathbf{M}}_n^{\mathrm{T}} \boldsymbol{\Omega}_n \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) + o_{\mathrm{p}}(1) \\
&= -(\mathbf{M}_0^{\mathrm{T}} \boldsymbol{\Omega}_0 \mathbf{M}_0)^{-1} \mathbf{M}_0^{\mathrm{T}} \boldsymbol{\Omega}_0 \frac{1}{\sqrt{n}} \sum_i \mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) + o_{\mathrm{p}}(1),
\end{aligned}$$

where

$$\hat{\mathbf{M}}_n = \frac{1}{n}\sum_i \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \hat{\boldsymbol{\theta}}), \qquad \tilde{\mathbf{M}}_n = \frac{1}{n}\sum_i \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\mathbf{m}(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}}).$$

In the above, we used the fact that both $\hat{\mathbf{M}}_n$ and $\tilde{\mathbf{M}}_n$ converge in probability to $\mathbf{M}_0$. This is easily shown by noting that (c.f. Assumption II.1)

$$\begin{aligned}
&\left|\hat{\mathbf{M}}_n - \frac{1}{n}\sum_i \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\right| \\
&\le \Big(\frac{1}{n}\sum_i \mathcal{H}_i^{\alpha,\delta}(\partial \mathbf{m}/\partial \boldsymbol{\theta})\Big) \cdot \Big(\max_{1\le i\le n} |\hat{\mu}_i - \mu_i| + |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|\Big)^{\alpha} = o_{\mathrm{p}}(1),
\end{aligned}$$

since $\hat{\mu}_i$ is uniformly consistent and $\hat{\boldsymbol{\theta}}$ is consistent. And note that $n^{-1}\sum_i \partial \mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}^{\mathrm{T}}$ $\xrightarrow{\mathrm{p}}$ $\mathbf{M}_0$ by the law of large numbers. The same argument applies to $\tilde{\mathbf{M}}_n$. ■

II.9.2 Proof Lemma II.6

Approximation Bias

For simplicity, let $\dot{\mathbf{m}}_i = \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)$, then

$$\begin{aligned}
\left|\frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}_i\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right| &\le \left|\frac{1}{\sqrt{n}}\sum_i \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i]\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right| \\
&+ \left|\frac{1}{\sqrt{n}}\sum_i (\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i])\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right|,
\end{aligned}$$

and we call the two terms (I) and (II) respectively. For term (I), we use projection matrix property, which implies

$$\begin{aligned}
\text{(I)} &= \left|\frac{1}{\sqrt{n}}\sum_i \eta_i\Big(\mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i] - \sum_j \pi_{ij}\mathbb{E}[\dot{\mathbf{m}}_j|\mathbf{z}_j]\Big)\right| \\
&\le \sqrt{n}\sqrt{\frac{1}{n}\sum_i \eta_i^2}\sqrt{\frac{1}{n}\sum_i \left|\mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i] - \sum_j \pi_{ij}\mathbb{E}[\dot{\mathbf{m}}_j|\mathbf{z}_j]\right|^2}.
\end{aligned}$$

By further splitting the conditional expectation $\mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i]$ into a linear projection and an error term,

$$
\begin{aligned}
\text{(I)} &\leq \sqrt{n}\sqrt{\frac{1}{n}\sum_i \eta_i^2}\sqrt{\frac{1}{n}\sum_i \Big|\boldsymbol{\zeta}_i - \sum_j \pi_{ij}\boldsymbol{\zeta}_j\Big|^2} \leq \sqrt{n}\sqrt{\frac{1}{n}\sum_i \eta_i^2}\sqrt{\frac{1}{n}\sum_i \Big|\boldsymbol{\zeta}_i\Big|^2} \\
&= O_{\mathrm{p}}\left(\sqrt{n\mathbb{E}[\eta_i^2]\mathbb{E}[|\boldsymbol{\zeta}_i|^2]}\right) = o_{\mathrm{p}}(1).
\end{aligned}
$$

The second term (II) can be bounded with conditional expectation and variance calculations. First note that since η_i is the error from linear approximation, this term has zero conditional mean:

$$
\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_i (\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i])\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right] \\
&= \frac{1}{\sqrt{n}}\sum_i \mathbb{E}[\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i]|\mathbf{z}_i]\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big) = \mathbf{0}.
\end{aligned}
$$

Next we consider the conditional second moment:

$$
\begin{aligned}
&\left|\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_i (\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i])\Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)\right]\right| \\
&\precsim \frac{1}{n}\sum_i \Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)^2 \mathbb{E}[|\dot{\mathbf{m}}_i - \mathbb{E}[\dot{\mathbf{m}}_i|\mathbf{z}_i]|^2|\mathbf{z}_i] \\
&\precsim \frac{1}{n}\sum_i \Big(\eta_i - \sum_j \pi_{ij}\eta_j\Big)^2 \leq \frac{1}{n}\sum_i \eta_i^2 = O_{\mathrm{p}}\left(\mathbb{E}[\eta_i^2]\right) = o_{\mathrm{p}}(1),
\end{aligned}
$$

where for the second line, we used the assumption that $\dot{\mathbf{m}}_i$ has uniformly bounded conditional variance.

Influence Function and Asymptotic Bias

The conclusion will be self-evident after two decompositions. First rewrite $\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) = \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) - \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\ |\mathbf{z}_i\right] + \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\ |\mathbf{z}_i\right]$ as the conditional expectation decomposition. Then

$$
\text{(II.10)} = \frac{1}{\sqrt{n}}\sum_i \left(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)\ |\mathbf{z}_j]\pi_{ij}\right)\cdot \varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i,j} \mathbf{u}_i\varepsilon_j\pi_{ij},
$$

where we use $\mathbf{u}_i = \dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) - \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\ |\mathbf{z}_i\right]$ to save notation. Then

$$\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij} = \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right] + O_{\mathrm{p}}\left(\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right]^{1/2}\right),$$

where we use $\mathbb{E}_{[\cdot|\mathbf{Z}]}$ and $\mathbb{V}_{[\cdot|\mathbf{Z}]}$ to denote the expectation and variance conditional on $\{\mathbf{z}_i, \mu_i\}_{1\le i\le n}$, respectively. Then

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right] = \frac{1}{\sqrt{n}}\sum_{i}\mathbf{b}_{1,i}\pi_{ii},$$

with $\mathbf{b}_{1,i} = \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{u}_i\varepsilon_i] = \mathbb{E}_{[\cdot|\mathbf{Z}]}[\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\varepsilon_i]$, since

$$i \neq j \qquad \Rightarrow \qquad \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\varepsilon_j\right] = \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{u}_i]\cdot\mathbb{E}_{[\cdot|\mathbf{Z}]}[\varepsilon_j] = \mathbf{0}.$$

Next we estimate the order of the conditional variance. To this end, consider

$$\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right)\left(\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right)^{\mathrm{T}}\right]\\
&= \frac{1}{n}\sum_{i,j,i',j'}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i'}^{\mathrm{T}}\varepsilon_j\varepsilon_{j'}\pi_{ij}\pi_{i'j'}\right]\\
&= \frac{1}{n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i'}^{\mathrm{T}}\varepsilon_i\varepsilon_{i'}\pi_{ii}\pi_{i'i'}\right] && (i=j,\ i'=j')\\
&+ \frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i}^{\mathrm{T}}\varepsilon_j\varepsilon_j\pi_{ij}\pi_{ij}\right] && (i=i',\ j=j')\\
&+ \frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{j}^{\mathrm{T}}\varepsilon_j\varepsilon_i\pi_{ij}\pi_{ij}\right] && (i=j',\ j=i')\\
&+ \frac{1}{n}\sum_{i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_{i}^{\mathrm{T}}\varepsilon_i\varepsilon_i\pi_{ii}\pi_{ii}\right]. && (i=j=i'=j')
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right]\\
&=\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right)\left(\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right)^{\mathrm{T}}\right]\\
&-\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right]\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{u}_i\varepsilon_j\pi_{ij}\right]^{\mathrm{T}}\\
&=\frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\varepsilon_j\varepsilon_j\pi_{ij}\pi_{ij}\right]+\frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_j^{\mathrm{T}}\varepsilon_j\varepsilon_i\pi_{ij}\pi_{ij}\right]+\frac{1}{n}\sum_{i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\varepsilon_i\varepsilon_i\pi_{ii}\pi_{ii}\right]\\
&-\frac{1}{n}\sum_i\mathbf{b}_{1,i}\mathbf{b}_{1,i}^{\mathrm{T}}\pi_{ii}^2.
\end{aligned}$$

Due to Assumption II.1, the above terms are easily bounded by

$$\begin{aligned}
\left|\frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\varepsilon_j\varepsilon_j\pi_{ij}\pi_{ij}\right]\right| &\precsim \frac{1}{n}\sum_{i,j}\pi_{ij}^2\leq\frac{k}{n}\\
\left|\frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_j^{\mathrm{T}}\varepsilon_j\varepsilon_i\pi_{ij}\pi_{ij}\right]\right| &\precsim \frac{1}{n}\sum_{i,j}\pi_{ij}^2\leq\frac{k}{n}\\
\left|\frac{1}{n}\sum_{i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\varepsilon_i\varepsilon_i\pi_{ii}\pi_{ii}\right]\right| &\precsim \frac{1}{n}\sum_{i}\pi_{ii}^2\leq\frac{k}{n}\\
\left|\frac{1}{n}\sum_{i}\mathbf{b}_{1,i}\mathbf{b}_{1,i}^{\mathrm{T}}\pi_{ii}^2\right| &\precsim \frac{1}{n}\sum_{i}\pi_{ii}^2\leq\frac{k}{n},
\end{aligned}$$

which closes the proof.

Variance Simplification

For notational convenience, denote $\mathbf{a}_i=\mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\mathbf{z}_i]$. Then it suffices to give conditions such that

$$\frac{1}{\sqrt{n}}\sum_i\left[\mathbf{a}_i-\sum_j\mathbf{a}_j\pi_{ij}\right]\varepsilon_i=o_{\mathrm{p}}(1).$$

Note that the conditional variance of the LHS is (use Assumption II.1)

$$\begin{aligned}
&\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\left|\frac{1}{\sqrt{n}}\sum_i\left[\mathbf{a}_i-\sum_j\mathbf{a}_j\pi_{ij}\right]\varepsilon_i\right|\right]\precsim\frac{1}{n}\sum_i\left|\mathbf{a}_i-\sum_j\mathbf{a}_j\pi_{ij}\right|^2\\
&=\frac{1}{n}\sum_i\left|\mathbf{a}_i-\boldsymbol{\Gamma}\mathbf{z}_i+\boldsymbol{\Gamma}\mathbf{z}_i-\sum_j\mathbf{a}_j\pi_{ij}\right|^2\\
&\leq\frac{2}{n}\sum_i\left(\left|\mathbf{a}_i-\boldsymbol{\Gamma}\mathbf{z}_i\right|^2+\left|\boldsymbol{\Gamma}\mathbf{z}_i-\sum_j\mathbf{a}_j\pi_{ij}\right|^2\right)=\frac{2}{n}\sum_i\left|\mathbf{a}_i-\boldsymbol{\Gamma}\mathbf{z}_i\right|^2+\frac{2}{n}\sum_i\left|\sum_j(\mathbf{a}_j-\boldsymbol{\Gamma}\mathbf{z}_j)\pi_{ij}\right|^2\\
&\leq\frac{4}{n}\sum_i\left|\mathbf{a}_i-\boldsymbol{\Gamma}\mathbf{z}_i\right|^2 \qquad\qquad \text{(Projection)}\\
&=o_{\mathrm{p}}(1),
\end{aligned}$$

where the last line shows why the assumption in Lemma II.6 is sufficient. Note that by projection, $\boldsymbol{\Gamma}\mathbf{z}_i=\sum_j\boldsymbol{\Gamma}\mathbf{z}_j\pi_{ij}$. ■

II.9.3 Proof of Lemma II.7

Approximation Error

For the current proof, we use $\ddot{\mathbf{m}}_i=\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)$ for notational convenience. Then recall that $\hat{\mu}_i-\mu_i=\sum_j\pi_{ij}\varepsilon_j-(\eta_i-\sum_j\pi_{ij}\eta_j)$. Then

$$\begin{aligned}
\text{(II.11)}&=\frac{1}{2\sqrt{n}}\sum_i\ddot{\mathbf{m}}_i\left(\sum_j\pi_{ij}\varepsilon_j-(\eta_i-\sum_j\pi_{ij}\eta_j)\right)^2\\
&=\underbrace{\frac{1}{2\sqrt{n}}\sum_i\ddot{\mathbf{m}}_i\left(\sum_j\pi_{ij}\varepsilon_j\right)^2}_{\text{(I)}}+\underbrace{\frac{1}{2\sqrt{n}}\sum_i\ddot{\mathbf{m}}_i\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2}_{\text{(II)}}\\
&\quad\underbrace{-\frac{1}{\sqrt{n}}\sum_i\ddot{\mathbf{m}}_i\left(\sum_j\pi_{ij}\varepsilon_j\right)\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)}_{\text{(III)}}.
\end{aligned}$$

We first deal with (II). Again we make a conditional expectation expansion of $\ddot{\mathbf{m}}_i$, which implies

$$|(\mathrm{II})| \le \underbrace{\left|\frac{1}{2\sqrt{n}}\sum_i \mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i]\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2\right|}_{(\mathrm{II}.1)} + \underbrace{\left|\frac{1}{2\sqrt{n}}\sum_i(\ddot{\mathbf{m}}_i-\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i])\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2\right|}_{(\mathrm{II}.2)}.$$

(II.1) has the simple bound:

$$\begin{aligned}(\mathrm{II}.1) &\le \frac{1}{2\sqrt{n}}\sum_i |\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i]|\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2 \precsim \frac{1}{\sqrt{n}}\sum_i\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2 \le \frac{1}{\sqrt{n}}\sum_i\eta_i^2\\ &= O_{\mathrm{p}}(\sqrt{n}\mathbb{E}[\eta_i^2]) = o_{\mathrm{p}}(1),\end{aligned}$$

where we used the assumption that $\ddot{\mathbf{m}}_i$ has uniformly bounded conditional expectation.

For (II.2), we employ conditional expectation and variance calculation. Note that it has zero conditional expectation:

$$\begin{aligned}&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{2\sqrt{n}}\sum_i(\ddot{\mathbf{m}}_i-\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i])\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2\right]\\ &=\frac{1}{2\sqrt{n}}\sum_i\mathbb{E}[\ddot{\mathbf{m}}_i-\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i]|\mathbf{z}_i]\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2=\mathbf{0}.\end{aligned}$$

The conditional variance is bounded by the following:

$$\begin{aligned}&\left|\mathbb{V}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{2\sqrt{n}}\sum_i(\ddot{\mathbf{m}}_i-\mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i])\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2\right]\right| \precsim \frac{1}{n}\sum_i\mathbb{E}[|\ddot{\mathbf{m}}_i|^2|\mathbf{z}_i]\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^4\\ &\precsim \frac{1}{n}\sum_i\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^4=\frac{1}{n}\sum_i\check{\eta}_i^4,\end{aligned}$$

where in the second line we used the assumption that $\ddot{\mathbf{m}}_i$ has uniformly bounded conditional second moment, and we use $\check{\eta}_i=\eta_i-\sum_j\pi_{ij}\eta_j$ for simplicity. Next, note that

$$\frac{1}{n}\left(\sum_i\check{\eta}_i^4+\sum_{i,j,i\neq j}\check{\eta}_i^2\check{\eta}_j^2\right)=\left(\frac{1}{\sqrt{n}}\sum_i\check{\eta}_i^2\right)^2\le\left(\frac{1}{\sqrt{n}}\sum_i\eta_i^2\right)^2=o_{\mathrm{p}}(1),$$

so that we conclude the previous conditional variance is asymptotically negligible.

For term (III), we first compute its conditional expectation:

$$\begin{aligned}
&\left|\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{\sqrt{n}}\sum_i \ddot{\mathbf{m}}_i\left(\sum_j \pi_{ij}\varepsilon_j\right)\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)\right]\right| = \left|\frac{1}{\sqrt{n}}\sum_i \pi_{ii}\mathbb{E}[\ddot{\mathbf{m}}_i\varepsilon_i|\mathbf{z}_i]\left(\eta_i - \sum_j \pi_{ij}\eta_j\right)\right| \\
&\le \sqrt{n}\sqrt{\frac{1}{n}\sum_i \pi_{ii}^2|\mathbb{E}[\ddot{\mathbf{m}}_i\varepsilon_i|\mathbf{z}_i]|^2}\sqrt{\frac{1}{n}\sum_i\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2} \\
&\lesssim \sqrt{n}\sqrt{\frac{1}{n}\sum_i \pi_{ii}^2}\sqrt{\frac{1}{n}\sum_i\left(\eta_i-\sum_j\pi_{ij}\eta_j\right)^2} \\
&= o_{\mathrm{p}}\left(\sqrt{n}\sqrt{\frac{k}{n}}\frac{1}{n^{1/4}}\right) = o_{\mathrm{p}}\left(\sqrt{\frac{k}{\sqrt{n}}}\right) = o_{\mathrm{p}}(1).
\end{aligned}$$

Here for the second line, we use the assumption that $\mathbb{E}[\ddot{\mathbf{m}}_i\varepsilon_i|\mathbf{z}_i]$ is uniformly bounded. Hence to bound (III), it suffices to consider the conditional second moment, which is bounded by the following (where $\check{\eta}_i = \eta_i - \sum_j \pi_{ij}\eta_j$):

$$\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\frac{1}{n}\sum_{i,j,k,\ell}|\ddot{\mathbf{m}}_i||\ddot{\mathbf{m}}_j|\varepsilon_k\varepsilon_\ell\pi_{ik}\pi_{j\ell}\check{\eta}_i\check{\eta}_j\right] \\
&= \frac{1}{n}\sum_{\substack{i,j,\ell\\ \text{distinct}}}\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\mathbb{E}[|\ddot{\mathbf{m}}_j||\mathbf{z}_j]\mathbb{E}[\varepsilon_\ell^2|\mathbf{z}_\ell]\pi_{i\ell}\pi_{j\ell}\check{\eta}_i\check{\eta}_j && \text{(III.1: } k=\ell) \\
&+ \frac{1}{n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\mathbb{E}[|\ddot{\mathbf{m}}_j|\varepsilon_j^2|\mathbf{z}_j]\pi_{ij}\pi_{jj}\check{\eta}_i\check{\eta}_j && \text{(III.2: } j=k=\ell) \\
&+ \frac{1}{n}\sum_{\substack{i,\ell\\ \text{distinct}}}\mathbb{E}[|\ddot{\mathbf{m}}_i|^2|\mathbf{z}_i]\mathbb{E}[\varepsilon_\ell^2|\mathbf{z}_\ell]\pi_{i\ell}^2\check{\eta}_i^2 && \text{(III.3: } i=j, k=\ell) \\
&+ o_{\mathrm{p}}(1), && (i=k, j=\ell)
\end{aligned}$$

where the last $o_{\mathrm{p}}(1)$ is the squared conditional expectation, and has been handled earlier. (III.3) is the simplest, which has bound

$$\text{(III.3)} \lesssim \frac{1}{n}\sum_{i,\ell}\pi_{i\ell}^2\check{\eta}_i^2 = \frac{1}{n}\sum_i \pi_{ii}\check{\eta}_i^2 \le \frac{1}{n}\sum_i \check{\eta}_i^2 \le \frac{1}{n}\sum_i \eta_i^2 = o_{\mathrm{p}}(1).$$

(III.1) is also easy, since by projection property, one has (it is easier to write it into a quadratic matrix form)

$$\begin{aligned}
\text{(III.1)} &\lesssim \frac{1}{n}\sum_{i,j,\ell} \mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\mathbb{E}[|\ddot{\mathbf{m}}_j||\mathbf{z}_j]\pi_{i\ell}\pi_{j\ell}\check{\eta}_i\check{\eta}_j \\
&\le \frac{1}{n}\sum_{i} |\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]|^2|\check{\eta}|_i^2 \lesssim \frac{1}{n}\sum_{i} \check{\eta}_i^2 \le \frac{1}{n}\sum_{i} \eta_i^2 = o_{\mathrm{p}}(1).
\end{aligned}$$

(III.2) is bounded by the following:

$$\begin{aligned}
\text{(III.2)} &\lesssim \left|\frac{1}{n}\sum_{j} \mathbb{E}[|\ddot{\mathbf{m}}_j|\varepsilon_j^2|\mathbf{z}_j]\pi_{jj}\check{\eta}_j \sum_{i} \mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\pi_{ij}\check{\eta}_i\right| \\
&\le \sqrt{\frac{1}{n}\sum_{j}(\mathbb{E}[|\ddot{\mathbf{m}}_j|\varepsilon_j^2|\mathbf{z}_j])^2\pi_{jj}^2\check{\eta}_j^2}\sqrt{\frac{1}{n}\sum_{j}\left(\sum_{i}\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\pi_{ij}\check{\eta}_i\right)^2} \\
&\lesssim \sqrt{\frac{1}{n}\sum_{j}\eta_j^2}\sqrt{\frac{1}{n}\sum_{j,i,\ell}\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\pi_{ij}\check{\eta}_i\mathbb{E}[|\ddot{\mathbf{m}}_\ell||\mathbf{z}_\ell]\pi_{\ell j}\check{\eta}_\ell} \\
&= \sqrt{\frac{1}{n}\sum_{j}\eta_j^2}\sqrt{\frac{1}{n}\sum_{i,\ell}\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\check{\eta}_i\pi_{i\ell}\mathbb{E}[|\ddot{\mathbf{m}}_\ell||\mathbf{z}_\ell]\check{\eta}_\ell} \\
&\le \sqrt{\frac{1}{n}\sum_{j}\eta_j^2}\sqrt{\frac{1}{n}\sum_{i}(\mathbb{E}[|\ddot{\mathbf{m}}_i||\mathbf{z}_i]\check{\eta}_i)^2} \lesssim \sqrt{\frac{1}{n}\sum_{j}\eta_j^2}\sqrt{\frac{1}{n}\sum_{i}\check{\eta}_i^2} \le \frac{1}{n}\sum_{j}\eta_j^2 = o_{\mathrm{p}}(1),
\end{aligned}$$

which concludes the proof.

Asymptotic Bias

Again we define $\ddot{\mathbf{m}}_i = \ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)$ to save notation. For the proof again we consider the expansion

$$\begin{aligned}
&\frac{1}{2\sqrt{n}}\sum_{i}\ddot{\mathbf{m}}_i\left(\sum_{j}\pi_{ij}\varepsilon_j\right)^2 = \frac{1}{2\sqrt{n}}\sum_{i,j,\ell}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell \\
&= \underbrace{\frac{1}{2\sqrt{n}}\sum_{\substack{i,j,\ell \\ \text{distinct}}}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell}_{\text{(I)}} + \underbrace{\frac{1}{2\sqrt{n}}\sum_{i,j,i\neq j}\ddot{\mathbf{m}}_i\pi_{ij}^2\varepsilon_j^2}_{(II)} + \underbrace{\frac{2}{2\sqrt{n}}\sum_{i,j,i\neq j}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{ii}\varepsilon_i\varepsilon_j}_{(III)} + \underbrace{\frac{1}{2\sqrt{n}}\sum_{i}\ddot{\mathbf{m}}_i\pi_{ii}^2\varepsilon_i^2}_{(Iv)}.
\end{aligned}$$

Expectation

It is easy to see that both (I) and (III) have zero conditional expectation. Hence we consider (II) and (IV).

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\mathrm{II})\right] = \frac{1}{2\sqrt{n}} \sum_{i,j,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i \pi_{ij}^2 \varepsilon_j^2\right] = \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_{2,ij}\pi_{ij}^2.$$

where the last line uses (II.16). And

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(\mathrm{IV})\right] = \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{2,ii}\pi_{ii}^2.$$

Variance, Term (I)

First for (I) we use (II.21) with $a_i = \ddot{\mathbf{m}}_i$ and (ignore the 1/2 in front) $b_{ij\ell} = \pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell$, and

$$\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}} \sum_{\substack{i,j,\ell \\ \text{distinct}}} \ddot{\mathbf{m}}_i \pi_{ij}\pi_{i\ell}\varepsilon_j\varepsilon_\ell\right)^2\right] \\
&= \underbrace{\frac{2}{n}\sum_i \sum_{j,\ell,j\neq\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_i b_{ij\ell}^2\right]}_{(\mathrm{I}.1)} + \underbrace{\frac{4}{n}\sum_i \sum_{j,j\neq i} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_i b_{iij}^2\right]}_{(\mathrm{I}.2)} \\
&\underbrace{+ \frac{2}{n}\sum_{i,i',i\neq i'} \sum_{j,\ell,j\neq\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_{i'} b_{ij\ell} b_{i'jl}\right]}_{(\mathrm{I}.3)}.
\end{aligned}$$

Next by (II.18) and (II.17), respectively,

$$(\mathrm{I}.1) \precsim \frac{1}{n}\sum_i \sum_{j,\ell,j\neq\ell} \pi_{ij}^2\pi_{i\ell}^2 \leq \frac{1}{n}\sum_i \pi_{ii}^2 \leq \frac{k}{n}.\ (\mathrm{I}.2) \precsim \frac{1}{n}\sum_{i,j} \pi_{ii}^2\pi_{ij}^2 \leq \frac{1}{n}\sum_i \pi_{ii}^3 \leq \frac{k}{n}.$$

And

$$\begin{aligned}
(\mathrm{I}.3) &= \frac{2}{n}\sum_{i,i',i\neq i'} \sum_{j,j',j\neq j'} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_{i'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\varepsilon_j^2\varepsilon_{j'}^2\right] \\
&= \frac{2}{n}\sum_{\substack{i,i',j,j' \\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_{i'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\varepsilon_j^2\varepsilon_{j'}^2\right] \\
&\quad + \frac{4}{n}\sum_{\substack{i,i' \\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_{i'} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\varepsilon_i^2\varepsilon_{i'}^2\right] + \frac{8}{n}\sum_{\substack{i,i',j \\ \text{distinct}}} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_{i'} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\varepsilon_i^2\varepsilon_j^2\right].
\end{aligned}$$

Define $\mathbf{c}_i = \mathbb{E}[\ddot{\mathbf{m}}_i|\mathbf{z}_i]$, $d_j = \mathbb{E}[\varepsilon_j^2|\mathbf{z}_j]$, and $\mathbf{e}_i = \mathbb{E}[\varepsilon_i^2\ddot{\mathbf{m}}_i|\mathbf{z}_i]$, and with (II.21) the above becomes

$$\begin{aligned}
\text{(I.3)} &= \frac{2}{n}\sum_{\substack{i,i',j,j'\\ \text{distinct}}} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}d_jd_{j'} + \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_{i'} + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\mathbf{e}_i^{\mathrm{T}}\mathbf{c}_{i'}d_j \\
&= \frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j,j',j\neq j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}d_jd_{j'} \\
&+ \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\left(\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_{i'} - \mathbf{c}_i^{\mathrm{T}}d_i\mathbf{c}_{i'}d_{i'}\right) + \frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\left(\mathbf{e}_i - \mathbf{c}_id_i\right)^{\mathrm{T}}\mathbf{c}_{i'}d_j \\
&= \underbrace{\frac{2}{n}\sum_{i,i',j,j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}d_jd_{j'}}_{\text{(I.3.1)}} + \underbrace{\frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\left(\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_{i'} - \mathbf{c}_i^{\mathrm{T}}d_i\mathbf{c}_{i'}d_{i'}\right)}_{\text{(I.3.2)}} \\
&+ \underbrace{\frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\left(\mathbf{e}_i - \mathbf{c}_id_i\right)^{\mathrm{T}}\mathbf{c}_{i'}d_j}_{\text{I.3.3}} - \underbrace{\frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j}\pi_{ij}^2\pi_{i'j}^2\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}d_j^2}_{\text{(I.3.4)}} \\
&- \underbrace{\frac{2}{n}\sum_{i}\sum_{j,j'}\pi_{ij}^2\pi_{ij'}^2|\mathbf{c}_i|^2d_jd_{j'}}_{\text{(I.3.5)}}\,.
\end{aligned}$$

Then use (II.16)

$$\begin{aligned}
|\text{(I.3.1)}| &= \left|\frac{2}{n}\sum_{i,i',j,j'} \pi_{ij}\pi_{ij'}\pi_{i'j}\pi_{i'j'}\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}d_jd_{j'}\right| \\
&= \left|\frac{2}{n}\sum_{i,i'}\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}\left(\sum_j \pi_{ij}\pi_{i'j}d_j\right)^2\right| \leq \max_{1\leq i,i'\leq n}|\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}|\frac{2}{n}\sum_{i,i'}\left(\sum_j \pi_{ij}\pi_{i'j}d_j\right)^2 \\
&= \max_{1\leq i,i'\leq n}|\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}|\frac{2}{n}\sum_{i,i',j,j'}\pi_{ij}\pi_{ij'}d_jd_{j'} = \max_{1\leq i,i'\leq n}|\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}|\frac{2}{n}\sum_{j,j'}d_jd_{j'}\left(\sum_i \pi_{ij}\pi_{ij'}\right)^2 \\
&\leq \max_{1\leq i,i',j,j'\leq n}|\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}d_jd_{j'}|\frac{2}{n}\sum_{j,j'}\left(\sum_i \pi_{ij}\pi_{ij'}\right)^2 \leq \max_{1\leq i,i',j,j'\leq n}|\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'}d_jd_{j'}|\frac{2}{n}\sum_{j,j'}\pi_{jj'}^2 \precsim \frac{k}{n}.
\end{aligned}$$

And by (II.17)

$$
\begin{aligned}
|(\mathrm{I.3.2})| &= \left|\frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'}\left(\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_{i'} - \mathbf{c}_i^{\mathrm{T}} d_i \mathbf{c}_{i'} d_{i'}\right)\right| \leq \max_{1\leq i,i'\leq n} |\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_{i'} - \mathbf{c}_i^{\mathrm{T}} d_i \mathbf{c}_{i'} d_{i'}| \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii}\pi_{ii'}^2\pi_{i'i'} \\
&\leq \max_{1\leq i,i'\leq n} |\mathbf{e}_i^{\mathrm{T}}\mathbf{e}_{i'} - \mathbf{c}_i^{\mathrm{T}} d_i \mathbf{c}_{i'} d_{i'}| \frac{4}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \pi_{ii'}^2\pi_{i'i'} \lesssim \frac{k}{n}.
\end{aligned}
$$

And by (II.16) and (II.18)

$$
\begin{aligned}
|(\mathrm{I.3.3})| &= \left|\frac{8}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \pi_{ii}\pi_{ij}\pi_{ii'}\pi_{i'j}\left(\mathbf{e}_i - \mathbf{c}_i d_i\right)^{\mathrm{T}}\mathbf{c}_{i'} d_j\right| \\
&\lesssim \frac{1}{n}\sum_{i',j,i'\neq j} |\mathbf{c}_{i'} d_j||\pi_{i'j}| \left|\sum_{\substack{i\\ i\neq i', i\neq j}} (\mathbf{e}_i - \mathbf{c}_i d_i)\pi_{ii}\pi_{ij}\pi_{ii'}\right| \\
&\lesssim \frac{1}{n}\sqrt{\sum_{i',j}\pi_{i'j}^2}\sqrt{\sum_{i',j}\left|\sum_{\substack{i\\ i\neq i', i\neq j}} (\mathbf{e}_i - \mathbf{c}_i d_i)\pi_{ii}\pi_{ij}\pi_{ii'}\right|^2} \\
&\lesssim \frac{\sqrt{k}}{n}\sqrt{\sum_{i,i',j,j'} (\mathbf{e}_i - \mathbf{c}_i d_i)^{\mathrm{T}}(\mathbf{e}_{j'} - \mathbf{c}_{j'} d_{j'})\pi_{ii}\pi_{ij}\pi_{ii'}\pi_{j'j'}\pi_{jj'}\pi_{i'j'}} \\
&= \frac{\sqrt{k}}{n}\sqrt{\sum_{i,j'} (\mathbf{e}_i - \mathbf{c}_i d_i)^{\mathrm{T}}(\mathbf{e}_{j'} - \mathbf{c}_{j'} d_{j'})\pi_{ii}\pi_{ij'}^2\pi_{j'j'}} \\
&\lesssim \frac{\sqrt{k}}{n}\sqrt{\sum_{i,j'}\pi_{ii}\pi_{ij'}^2\pi_{j'j'}} \leq \frac{\sqrt{k}}{n}\sqrt{\sum_{i,j'}\pi_{ii}\pi_{ij'}^2} \leq \frac{\sqrt{k}}{n}\sqrt{\sum_{i}\pi_{ii}^2} \leq \frac{k}{n}.
\end{aligned}
$$

And by (II.18)

$$
|(\mathrm{I.3.4})| = \left|\frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j}\pi_{ij}^2\pi_{i'j}^2\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'} d_j^2\right| \leq \max_{1\leq i,i',j\leq n}|\mathbf{c}_i^{\mathrm{T}}\mathbf{c}_{i'} d_j^2|\frac{2}{n}\sum_{i,i',i\neq i'}\sum_{j}\pi_{ij}^2\pi_{i'j}^2 \lesssim \frac{k}{n}.
$$

And by (II.18)

$$
|(\mathrm{I.3.5})| = \left|\frac{2}{n}\sum_{i}\sum_{j,j'}\pi_{ij}^2\pi_{ij'}^2|\mathbf{c}_i|^2 d_j d_{j'}\right| \leq \max_{1\leq i,j,j'\leq n}||\mathbf{c}_i|^2 d_j d_{j'}|\frac{2}{n}\sum_{i}\sum_{j,j'}\pi_{ij}^2\pi_{ij'}^2 \lesssim \frac{k}{n}.
$$

Variance, Term (II)

Then for (II), one has (by using (II.19))

$$\begin{aligned}
&\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\ddot{\mathbf{m}}_i\pi_{ij}^2\varepsilon_j^2\right)^2\right]-\left(\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\pi_{ij}^2\varepsilon_j^2\right]\right)^2\\
&=\underbrace{\frac{1}{n}\sum_{\substack{i,i',j,j'\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_{i'}\pi_{ij}^2\pi_{i'j'}^2\varepsilon_j^2\varepsilon_{j'}^2\right]-\left(\frac{1}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\pi_{ij}^2\varepsilon_j^2\right]\right)^2}_{\text{(II.1)}}\\
&\quad+\underbrace{\frac{1}{n}\sum_{\substack{i,j,j'\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left|\ddot{\mathbf{m}}_i\right|^2\pi_{ij}^2\pi_{ij'}^2\varepsilon_j^2\varepsilon_{j'}^2\right]}_{\text{(II.2)}}+\underbrace{\frac{2}{n}\sum_{\substack{i,i',j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_{i'}\pi_{ij}^2\pi_{ii'}^2\varepsilon_i^2\varepsilon_j^2\right]}_{\text{(II.3)}}\\
&\quad+\underbrace{\frac{1}{n}\sum_{\substack{i,i',j\\ \text{distinct}}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_{i'}\pi_{ij}^2\pi_{i'j}^2\varepsilon_j^4\right]}_{\text{(II.4)}}\\
&\quad+\underbrace{\frac{1}{n}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left|\ddot{\mathbf{m}}_i\right|^2\pi_{ij}^4\varepsilon_j^4\right]}_{\text{(II.5)}}+\underbrace{\frac{1}{n}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_j\pi_{ij}^4\varepsilon_i^2\varepsilon_j^2\right]}_{\text{(II.6)}}.
\end{aligned}$$

With (II.18) it is easy to see (together with the uniform bounded moments assumption) that (II.2)–(II.6) are of order $O_{\mathrm{p}}(n^{-1}\sum_i\pi_{ii}^2)=O_{\mathrm{p}}(k/n)$, hence asymptotically negligible. As for (II.1), note that

$$\begin{aligned}
\text{(II.1)}=&-\frac{1}{n}\sum_{\substack{i,j,j'\\ \text{distinct}}}\pi_{ij}^2\pi_{ij'}^2\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\varepsilon_j^2\right]^{\mathrm{T}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\varepsilon_{j'}^2\right]-\frac{2}{n}\sum_{\substack{i,i',j\\ \text{distinct}}}\pi_{ij}^2\pi_{ii'}^2\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\varepsilon_j^2\right]^{\mathrm{T}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_{i'}\varepsilon_i^2\right]\\
&-\frac{1}{n}\sum_{\substack{i,i',j\\ \text{distinct}}}\pi_{ij}^2\pi_{ij'}^2\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\varepsilon_j^2\right]^{\mathrm{T}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_{i'}\varepsilon_j^2\right]-\frac{1}{n}\sum_{i,j,i\neq j}\pi_{ij}^4\left(\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\varepsilon_j^2\right]\right)^2\\
&-\frac{1}{n}\sum_{i,j,i\neq j}\pi_{ij}^4\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\varepsilon_j^2\right]^{\mathrm{T}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_j\varepsilon_i^2\right].
\end{aligned}$$

Therefore we have (II.1) is of order $O_{\mathrm{p}}(n^{-1}\sum_i\pi_{ii}^2)=O_{\mathrm{p}}(k/n)$.

Variance, Term (III)

Next we consider (III), and still (II.19) implies

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\ddot{\mathbf{m}}_i\pi_{ij}\pi_{ii}\varepsilon_j\varepsilon_i\right)^2\right]=\frac{4}{n}\sum_{i,j,\text{distinct}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[|\ddot{\mathbf{m}}_i|^2\,\pi_{ij}^2\pi_{ii}^2\varepsilon_i^2\varepsilon_j^2\right]$$
$$+\frac{8}{n}\sum_{i,j,\text{distinct}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_j\pi_{ij}^2\pi_{ii}\pi_{jj}\varepsilon_i^2\varepsilon_j^2\right],$$

where the two terms are denoted by (III.1) and (III.2), respectively. For (III.1) it is bounded by

$$|(\mathrm{III.1})|\precsim\frac{1}{n}\sum_i\pi_{ii}^2\sum_{j\neq i}\pi_{ij}^2=\frac{1}{n}\sum_i\pi_{ii}^3,$$

which is bounded by k/n due to (II.17). Similarly

$$|(\mathrm{III.2})|\precsim\frac{1}{n}\sum_{i,j}\pi_{ii}\pi_{jj}\pi_{ij}^2\leq\frac{1}{n}\sum_{i,j}\pi_{jj}\pi_{ij}^2=O(k/n),$$

due to (II.17) and $\pi_{ii}\leq 1$.

Variance, Term (IV)

Finally we consider (IV), and the variance is

$$\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(\frac{1}{\sqrt{n}}\sum_i\ddot{\mathbf{m}}_i\pi_{ii}^2\varepsilon_i^2\right)^2\right]-\left(\frac{1}{\sqrt{n}}\sum_i\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\pi_{ii}^2\varepsilon_i^2\right]\right)^2$$
$$=\frac{1}{n}\sum_{i,j,i\neq j}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i^{\mathrm{T}}\ddot{\mathbf{m}}_j\pi_{ii}^2\pi_{jj}^2\varepsilon_i^2\varepsilon_j^2\right]-\left(\frac{1}{\sqrt{n}}\sum_i\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}_i\pi_{ii}^2\varepsilon_i^2\right]\right)^2+\frac{1}{n}\sum_i\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[|\ddot{\mathbf{m}}_i|^2\,\pi_{ii}^4\varepsilon_i^4\right].$$

And both terms are bounded by $O(k/n)$.

The last step is to show that one can essentially replace $\tilde{\mu}_i$ by μ_i in (II.11). This is trivial due to Assumption II.1, and the consistency assumption II.2. ■

II.9.4 Proof of Proposition II.1

By the condition $k=O(\sqrt{n})$, all terms of order $O_{\mathrm{p}}(\sqrt{k/n})$ can be ignored asymptotically. Also the bias term has order $\boldsymbol{\mathcal{B}}=O_{\mathrm{p}}(k/\sqrt{n})=O_{\mathrm{p}}(1)$. In particular, both (II.10) and (II.11) are of order $O_{\mathrm{p}}(1)$. By Assumption II.2, the remainder term in the quadratic expansion (after (II.11)) has the order $o_{\mathrm{p}}(|(\mathrm{II}.11)|)$, which is negligible. ■

II.9.5 Proof of Theorem II.1

We first make the following decomposition:

$$\tilde{\boldsymbol{\Psi}}_1 = \mathbb{E}[\bar{\boldsymbol{\Psi}}_1|\mathbf{Z}], \qquad \tilde{\boldsymbol{\Psi}}_2 = \bar{\boldsymbol{\Psi}}_1 - \mathbb{E}[\bar{\boldsymbol{\Psi}}_1|\mathbf{Z}] + \bar{\boldsymbol{\Psi}}_2.$$

Then note that $\tilde{\boldsymbol{\Psi}}_1$ is mean zero, and $\tilde{\boldsymbol{\Psi}}_2$ is conditionally mean zero (on $\mathbf{Z}$). One special case is that $\tilde{\boldsymbol{\Psi}}_1 = 0$ almost surely, which will happen if the moment condition for the second step is actually a conditional moment restriction. In what follows, we assume $\tilde{\boldsymbol{\Psi}}_1$ is nondegenerate.

By the usual central limit theorem, one has

$$\left(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_1]\right)^{-1/2}\tilde{\boldsymbol{\Psi}}_1 \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0},\mathbf{I}).$$

Next we consider the large sample distribution of $\tilde{\boldsymbol{\Psi}}_2$, which requires triangular array type argument. Let $\boldsymbol{\alpha}$ be a generic vector, and consider

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(a_i+b_i)^2\mathbb{1}\left[|a_i+b_i| > 2\varepsilon\sqrt{n}\right]\right],$$

where

$$a_i = \boldsymbol{\alpha}^{\mathrm{T}}\Big(\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) - \mathbb{E}[\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|\mathbf{z}_i]\Big), \qquad b_i = \boldsymbol{\alpha}^{\mathrm{T}}\Big(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j,\mu_j,\boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij}\Big)\varepsilon_i.$$

Note that

$$\begin{aligned}&\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[(a_i+b_i)^2\mathbb{1}\left[|a_i+b_i| > 2\varepsilon\sqrt{n}\right]\right]\\ &\precsim \frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\left(a_i^2+b_i^2\right)\left(\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right] + \mathbb{1}\left[|b_i|>\varepsilon\sqrt{n}\right]\right)\right],\end{aligned}$$

which is a sum of four terms.

The first case is the easiest:

$$\mathbb{E}\left|\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right]\right| = \frac{1}{n}\sum_i \mathbb{E}\left[a_i^2\mathbb{1}\left[|a_i|>\varepsilon\sqrt{n}\right]\right] \to 0,$$

where the first equality is true since the summands are nonnegative, and the last line comes from the i.i.d.ness of a_i. Therefore

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|a_i| > \varepsilon\sqrt{n}\right]\right] = o_{\mathrm{p}}(1).$$

For future reference, define $\tilde{b}_i = \boldsymbol{\alpha}^{\mathrm{T}}\Big(\sum_j \mathbb{E}[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)|\mathbf{z}_j]\pi_{ij}\Big)$. Then the second case becomes (where we used the union bound)

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|b_i| > \varepsilon\sqrt{n}\right]\right] \leq \frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|\tilde{b}_i| > \varepsilon\sqrt{n}/\log(n)\right]\right]$$
$$+\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|\varepsilon_i| > \log(n)\right]\right].$$

the last term in the above display is $o_{\mathrm{p}}(1)$ since it has expectation (note that it is nonnegative)

$$\lim_n \mathbb{E}\left[\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|\varepsilon_i| > \log(n)\right]\right]\right] = \lim_n \mathbb{E}\left[a_i^2\mathbb{1}\left[|\varepsilon_i| > \log(n)\right]\right]$$
$$= \mathbb{E}\left[a_i^2 \lim_n \mathbb{1}\left[|\varepsilon_i| > \log(n)\right]\right] = 0,$$

and interchanging limit and expectation is justified by dominated convergence, and the fact that $\mathbb{E}[a_i^2] < \infty$. The other terms is handled by the following:

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[a_i^2\mathbb{1}\left[|\tilde{b}_i| > \varepsilon\sqrt{n}/\log(n)\right]\right] = \frac{1}{n}\sum_i \mathbb{1}\left[|\tilde{b}_i| > \varepsilon\sqrt{n}/\log(n)\right]\mathbb{E}_{[\cdot|\mathbf{Z}]}[a_i^2]$$
$$\precsim \frac{1}{n}\sum_i \mathbb{1}\left[|\tilde{b}_i| > \varepsilon\sqrt{n}/\log(n)\right].$$

The first line comes from the fact that $\tilde{b}_i$ is constant after conditioning on $\mathbf{Z}$, and the second line is true since $\mathbb{E}_{[\cdot|\mathbf{Z}]}[a_i^2]$ is bounded. We show it is $o_{\mathrm{p}}(1)$ again by taking expectation, and the fact that $\tilde{b}_i$ is the projection of random variable with finite expectation.

The next case is again very simple:

$$\begin{aligned}
&\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[b_i^2 \mathbb{1}\left[|a_i| > \varepsilon\sqrt{n}\right]\right] \precsim \frac{1}{n}\sum_i \tilde{b}_i^2 \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\varepsilon_i^2 \mathbb{1}\left[|a_i| > \varepsilon\sqrt{n}\right]\right] \\
&\le \left(\max_{1\le i\le n} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\varepsilon_i^2 \mathbb{1}\left[|a_i| > \varepsilon\sqrt{n}\right]\right]\right) \frac{1}{n}\sum_i \tilde{b}_i^2 \precsim \left(\max_{1\le i\le n} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\varepsilon_i^2 \mathbb{1}\left[|a_i| > \varepsilon\sqrt{n}\right]\right]\right) \to 0.
\end{aligned}$$

The first inequality comes from the definition of $\tilde{b}_i$; the second is Hölder's inequality; the third inequality uses the fact $\sum_i \tilde{b}_i^2 = O(n)$; and the final inequality is true since we assumed bounded conditional moment.

Finally, the last case is

$$\frac{1}{n}\sum_i \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[b_i^2 \mathbb{1}\left[|b_i| > \varepsilon\sqrt{n}\right]\right] \precsim \frac{1}{n}\sum_i \tilde{b}_i^2 \mathbb{1}\left[|\tilde{b}_i| > \varepsilon\sqrt{n}/\log(n)\right] + o_{\mathrm{p}}(1) = o_{\mathrm{p}}(1),$$

since $\tilde{b}_i$ comes from projecting a bounded sequence.

To summarize, we have the following two convergence results: (1) $\tilde{\boldsymbol{\Psi}}_1$ converges unconditionally to a multivariate normal distribution; and (2) conditional on $\mathbf{Z}$, $\tilde{\boldsymbol{\Psi}}_2$ converges to a multivariate normal distribution (more precisely, conditional on $\mathbf{Z}$ the distribution function of $\tilde{\boldsymbol{\Psi}}_2$ converges to that of a multivariate normal in probability). The following remark shows how joint convergence can be established (not that it is not true in general that one can conclude joint convergence from marginal convergence)

Remark II.4 (From marginal convergence to joint convergence) Here we consider one special case where it is possible to deduce joint convergence from marginal convergence. Assume $X_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ and $Y_n|Z_n \xrightarrow{\mathrm{d}}_{\mathrm{p}} \mathcal{N}(0,1)$, and $X_n \in \sigma(Z_n)$, where $Y_n|Z_n \xrightarrow{\mathrm{d}}_{\mathrm{p}} \mathcal{N}(0,1)$. Then, $[X_n, Y_n]^{\mathrm{T}} \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{I})$.

This follows because

$$\begin{aligned}
\mathbb{P}\Big[X_n \le x, Y_n \le y\Big] &= \mathbb{E}\Big[\mathbb{1}[X_n \le x]\mathbb{P}[Y_n \le y|Z_n]\Big] \\
&= \mathbb{E}\Big[\mathbb{1}[X_n \le x]\Big(\mathbb{P}[Y_n \le y|Z_n] - \Phi(y)\Big)\Big] + \mathbb{P}\Big[X_n \le x\Big]\Phi(y) \\
&\to \Phi(x)\Phi(y),
\end{aligned}$$

using the dominated convergence theorem and the assumption that $\mathbb{P}[Y_n \le y|Z_n] \to_{\mathbb{P}} \Phi(y)$.

$\|$

Hence we are able to show

$$\begin{bmatrix} \left(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_1]\right)^{-1/2}\tilde{\boldsymbol{\Psi}}_1 \\ \left(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_2|\mathbf{Z}]\right)^{-1/2}\tilde{\boldsymbol{\Psi}}_2 \end{bmatrix} \xrightarrow{\mathrm{d}} \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\ \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\right),$$

and the desired result follows by considering the linear combination

$$\left(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_1] + \mathbb{V}[\tilde{\boldsymbol{\Psi}}_2|\mathbf{Z}]\right)^{-1/2}\left[\left(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_1]\right)^{1/2},\ \ \left(\mathbb{V}[\tilde{\boldsymbol{\Psi}}_2|\mathbf{Z}]\right)^{1/2}\right].$$

■

II.9.6 Proof of Theorem II.2

Part 1

For the ease of exposition we ignore (asymptotic negligible) remainder terms in the proof. Then $\hat{\boldsymbol{\theta}}$ has the expansion

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \frac{1}{\sqrt{n}}\sum_i \mathbf{a}_i + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_i\left(\hat{\mu}_i - \mu_i\right) + \frac{1}{\sqrt{n}}\sum_i \mathbf{c}_i\left(\hat{\mu}_i - \mu_i\right)^2,$$

where to save notations we used

$$\begin{aligned}
\mathbf{a}_i &= -\left(\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{M}_0\right)^{-1}\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{m}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\\
\mathbf{b}_i &= -\left(\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{M}_0\right)^{-1}\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)\\
\mathbf{c}_i &= -\frac{1}{2}\left(\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{M}_0\right)^{-1}\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0).
\end{aligned}$$

Denote the leave-j-out estimator by $\hat{\boldsymbol{\theta}}^{(j)}$, it is easy to see that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(j)} - \boldsymbol{\theta}_0\right) = \frac{\sqrt{n}}{n-1}\sum_{i,i\neq j} \mathbf{a}_i + \frac{\sqrt{n}}{n-1}\sum_{i,i\neq j} \mathbf{b}_i\left(\hat{\mu}_i^{(j)} - \mu_i\right) + \frac{\sqrt{n}}{n-1}\sum_{i,i\neq j} \mathbf{c}_i\left(\hat{\mu}_i^{(j)} - \mu_i\right)^2.$$

Recall that the jackknife estimator is defined as

$$\hat{\boldsymbol{\theta}}^{(\cdot)} = \frac{1}{n}\sum_j \hat{\boldsymbol{\theta}}^{(j)},$$

and with some algebraic manipulation,

$$
\begin{aligned}
(n-1)\cdot\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(\cdot)}-\hat{\boldsymbol{\theta}}\right) &= \frac{1}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat{\mu}_j - r_j\Big) && \text{(I)}\\
&\quad + \frac{1}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\Big(\hat{\mu}_j - r_j\Big)^2 && \text{(II)}\\
&\quad + \frac{2}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\mathbf{c}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i-\mu_i\right)\Big(\hat{\mu}_j - r_j\Big). && \text{(III)}
\end{aligned}
$$

By Assumption II.2, we could ignore the approximation error. And (I) becomes

$$
\begin{aligned}
\text{(I)} &= \frac{1}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat{\mu}_j-\mu_j+\mu_j - r_j\Big)\\
&= \underbrace{\frac{1}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\sum_{\ell}\pi_{j\ell}\varepsilon_\ell\Big)}_{\text{(I.1)}}\underbrace{-\frac{1}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\varepsilon_j}_{\text{(I.2)}}+o_{\mathrm{p}}(1).
\end{aligned}
$$

Then we have the following conditional expectations:

$$
\begin{aligned}
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{(I.1)}\right] &= \frac{1}{\sqrt{n}}\sum_{j}\sum_{i,i\neq j}\frac{\pi_{ij}^2}{1-\pi_{jj}}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i\varepsilon_i\right]\\
&= -\frac{1}{\sqrt{n}}\left(\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\mathbf{M}_0\right)^{-1}\mathbf{M}_0^{\mathrm{T}}\boldsymbol{\Omega}_0\left[\sum_{i}\mathbf{b}_{1,i}\pi_{ii}\right]\\
&\qquad + \frac{1}{\sqrt{n}}\sum_{i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i\varepsilon_i\right]\left(\sum_{j,j\neq i}\frac{\pi_{ij}^2}{1-\pi_{jj}}-\pi_{ii}\right)\\
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{(I.2)}\right] &= 0.
\end{aligned}
$$

To further simplify, note that

$$
\begin{aligned}
\left|\frac{1}{\sqrt{n}}\sum_{i}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{b}_i\varepsilon_i\right]\left(\sum_{j,j\neq i}\frac{\pi_{ij}^2}{1-\pi_{jj}}-\pi_{ii}\right)\right| &\precsim \frac{1}{\sqrt{n}}\sum_{i}\left|\sum_{j,j\neq i}\frac{\pi_{ij}^2}{1-\pi_{jj}}-\pi_{ii}\right|\\
&\precsim \frac{1}{\sqrt{n}}\sum_{i}\pi_{ii}^2 = o_{\mathrm{p}}(1).
\end{aligned}
$$

One could conduct variance calculation, which is tedious yet straightforward. Now we consider (II), which has the following expansion:

$$
\begin{aligned}
\text{(II)} &= \frac{1}{\sqrt{n}} \sum_j \sum_{i, i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \left(\hat{\mu}_j - \mu_j + \mu_j - r_j \right)^2 \\
&= \underbrace{\frac{1}{\sqrt{n}} \sum_j \sum_{i, i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \left(\sum_{\ell, m} \pi_{j\ell}\pi_{jm}\varepsilon_\ell \varepsilon_m \right)}_{\text{(II.1)}} \\
&\qquad + \underbrace{\frac{1}{\sqrt{n}} \sum_j \sum_{i, i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \varepsilon_j^2}_{\text{(II.2)}} \underbrace{- \frac{2}{\sqrt{n}} \sum_j \sum_{i, i\neq j} \mathbf{c}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \left(\sum_\ell \pi_{j\ell}\varepsilon_\ell \varepsilon_j \right)}_{\text{(II.3)}} + o_{\mathrm{p}}(1).
\end{aligned}
$$

Therefore

$$
\begin{aligned}
\left| \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\text{(II.1)} \right] \right| &= \left| \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \sum_\ell \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\mathbf{c}_i \varepsilon_\ell^2 \right] \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \pi_{j\ell}^2 \right| \\
&\precsim_{\mathrm{p}} \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \sum_\ell \pi_{ij}^2 \pi_{j\ell}^2 \leq \frac{1}{\sqrt{n}} \sum_{j,\ell} \pi_{j\ell}^2 \pi_{jj} = o_{\mathrm{p}}(1),
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\text{(II.2)} \right] &= \frac{1}{\sqrt{n}} \sum_j \sum_{i, i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\mathbf{c}_i \varepsilon_j^2 \right] \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\mathbf{c}_i \varepsilon_j^2 \right] \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 + o_{\mathrm{p}}(1) \\
&= -\frac{1}{\sqrt{n}} \left(\mathbf{M}_0^{\mathrm{T}} \mathbf{\Omega}_0 \mathbf{M}_0 \right)^{-1} \mathbf{M}_0^{\mathrm{T}} \mathbf{\Omega}_0 \left[\sum_{i,j} \mathbf{b}_{2,ij} \pi_{ij}^2 \right] \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i,j} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\mathbf{c}_i \varepsilon_j^2 \right] \frac{\pi_{ij}^2 \pi_{jj}}{(1-\pi_{jj})^2} + o_{\mathrm{p}}(1) \\
&= -\frac{1}{\sqrt{n}} \left(\mathbf{M}_0^{\mathrm{T}} \mathbf{\Omega}_0 \mathbf{M}_0 \right)^{-1} \mathbf{M}_0^{\mathrm{T}} \mathbf{\Omega}_0 \left[\sum_{i,j} \mathbf{b}_{2,ij} \pi_{ij}^2 \right] + o_{\mathrm{p}}(1), \qquad \text{(II.17)}
\end{aligned}
$$

and using (II.17) again,

$$
\left| \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\text{(II.3)} \right] \right| = \left| \frac{2}{\sqrt{n}} \sum_j \sum_{i, i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]} \left[\mathbf{c}_i \varepsilon_j^2 \right] \left(\frac{\pi_{ij}}{1-\pi_{jj}} \right)^2 \pi_{jj} \right| \precsim_{\mathrm{p}} \frac{1}{\sqrt{n}} \sum_{i,j} \pi_{ij}^2 \pi_{jj} = o_{\mathrm{p}}(1).
$$

Finally (III) has the expansion:

$$
\begin{aligned}
\text{(III)} = {} & \underbrace{\frac{2}{\sqrt{n}} \sum_{j} \sum_{i, i \neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\sum_{\ell, m} \pi_{i\ell}\pi_{jm}\varepsilon_\ell \varepsilon_m \right)}_{\text{III.1}} \\
& \underbrace{- \frac{2}{\sqrt{n}} \sum_{j} \sum_{i, i \neq j} \mathbf{c}_i \frac{\pi_{ij}}{1-\pi_{jj}} \left(\sum_{\ell} \pi_{i\ell}\varepsilon_\ell \varepsilon_j \right)}_{\text{III.2}} + o_{\mathrm{p}}(1).
\end{aligned}
$$

Again we consider the conditional expectations:

$$
\begin{aligned}
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.1}\right] &= \frac{2}{\sqrt{n}} \sum_{j} \sum_{i, i\neq j} \sum_{\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}, \\
\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.2}\right] &= -\frac{2}{\sqrt{n}} \sum_{j} \sum_{i, i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_j^2\right] \frac{\pi_{ij}^2}{1-\pi_{jj}}.
\end{aligned}
$$

Therefore using (II.17) and $\pi_{j'j'} \leq 1$

$$
\begin{aligned}
& \left| \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.1}\right] + \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\text{III.2}\right] \right| \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{i\ell}^2}{1-\pi_{\ell\ell}} \right| + o_{\mathrm{p}}(1) \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \pi_{ij}\pi_{i\ell}\pi_{j\ell} \right| + o_{\mathrm{p}}(1) \\
&= \left| \frac{2}{\sqrt{n}} \sum_{i,j,\ell} \mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\mathbf{c}_i \varepsilon_\ell^2\right] \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}} \right| + o_{\mathrm{p}}(1) \\
&\lesssim_{\mathrm{p}} \frac{1}{\sqrt{n}} \sqrt{\sum_{i,\ell} \pi_{i\ell}^2} \sqrt{\sum_{i,\ell} \left(\sum_{j} \frac{\pi_{ij}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}} \right)^2} \\
&= \frac{\sqrt{k}}{\sqrt{n}} \sqrt{\sum_{i,\ell} \sum_{jj'} \frac{\pi_{ij}\pi_{j\ell}\pi_{jj}\pi_{ij'}\pi_{j'\ell}\pi_{j'j'}}{(1-\pi_{jj})(1-\pi_{j'j'})}} \lesssim_{\mathrm{p}} \frac{\sqrt{k}}{\sqrt{n}} \sqrt{\sum_{jj'} \pi_{jj}\pi_{j'j'}\pi_{jj'}^2} = \frac{\sqrt{k}}{\sqrt{n}} \cdot o_{\mathrm{p}}(\sqrt{k}) = o_{\mathrm{p}}(1).
\end{aligned}
$$

Therefore we showed the desired result.

Part 2

First note that the jackknife variance estimator takes the form:

$$(n-1)\sum_j \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}^{(\cdot)}\right)^2,$$

where for a (column) vector $\mathbf{v}$, we use $\mathbf{v}^2$ to denote $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ to save space. Then the variance estimator could be rewritten as

$$\hat{\boldsymbol{\mathcal{V}}} = (n-1)\sum_j \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}\right)^2 - \frac{1}{n-1}\left(\hat{\boldsymbol{\mathcal{B}}}\right)^2 = (n-1)\sum_j \left(\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}}\right)^2 + O_{\mathrm{p}}\left(\frac{1}{n}\right).$$

Next recall that

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{(j)} - \hat{\boldsymbol{\theta}} = &\underbrace{\frac{1}{n-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)}_{\text{(I)}} \underbrace{-\frac{1}{n-1}\mathbf{a}_j}_{\text{(II)}} \underbrace{-\frac{1}{n-1}\mathbf{b}_j\left(\hat{\mu}_j - \mu_j\right)}_{\text{(III)}} \underbrace{-\frac{1}{n-1}\mathbf{c}_j\left(\hat{\mu}_j - \mu_j\right)^2}_{\text{(IV)}} \\
&+ \underbrace{\frac{1}{n-1}\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_j - r_j\right)}_{\text{(V)}} \\
&+ \underbrace{\frac{1}{n-1}\sum_{i,i\neq j}\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\hat{\mu}_j - r_j\right)^2}_{\text{(VI)}} + \underbrace{\frac{2}{n-1}\sum_{i,i\neq j}\mathbf{c}_i\frac{\pi_{ij}}{1-\pi_{jj}}\left(\hat{\mu}_i - \mu_i\right)\left(\hat{\mu}_j - r_j\right)}_{\text{(VII)}}.
\end{aligned}$$

Therefore we have to consider the square of each term, as well as their interactions. As the proof is quite tedious, we list the main steps here. First we would like to recover the variance terms in Theorem II.1 with

$$\begin{aligned}
(n-1)\sum_j \text{(II)}^2 = \mathbb{V}[\bar{\boldsymbol{\Psi}}_1] + o_{\mathrm{p}}(1),\ (n-1)\sum_j \text{(II)}\text{(V)}^{\mathrm{T}} \\
= \mathbb{C}\mathsf{ov}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_1, \bar{\boldsymbol{\Psi}}_2] + o_{\mathrm{p}}(1),\ (n-1)\sum_j \text{(V)}^2 = \mathbb{V}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_2] + o_{\mathrm{p}}(1).
\end{aligned}$$

Furthermore, all the other square terms and interactions are asymptotically negligible. We use the following fact repeatedly: For two sequences $\{u_i\}$ and $\{v_j\}$,

$$\left|\sum_{i,j} u_i\pi_{ij}v_j\right| \le \sqrt{\sum_i u_i^2}\sqrt{\sum_i\left(\sum_j \pi_{ij}v_j\right)^2} \le \sqrt{\sum_i u_i^2}\sqrt{\sum_i v_i^2}.$$

Term (I):

$$(n-1)\sum_j (\mathrm{I})^2 = \frac{1}{n-1}\sum_j \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^2 \asymp \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)^2 = o_{\mathrm{p}}(1),$$

by consistency. Then it is also easy to show that for $\dagger = \mathrm{II}, \cdots, \mathrm{VII}$

$$(n-1)\sum_j (\mathrm{I})(\dagger)^{\mathrm{T}} = (\mathrm{I})\frac{1}{n-1}\sum_j (\dagger)^{\mathrm{T}} = o_{\mathrm{p}}(1)\cdot\frac{1}{n-1}\sum_j (\dagger)^{\mathrm{T}} = o_{\mathrm{p}}(1),$$

since the summands are bounded in probability uniformly in j.

Next term (II):

$$(n-1)\sum_j (\mathrm{II})^2 = \frac{1}{n-1}\sum_j \mathbf{a}_j^2,$$

which is asymptotically equivalent to $\mathbb{V}[\bar{\boldsymbol{\Psi}}_1]$ in Theorem II.1. Now we consider the interactions:

$$\left|(n-1)\sum_j (\mathrm{II})(\mathrm{III})^{\mathrm{T}}\right| = \left|\frac{1}{n-1}\sum_j \mathbf{a}_j\mathbf{b}_j^{\mathrm{T}}\left(\hat{\mu}_j - \mu_j\right)\right| \leq o_{\mathrm{p}}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{a}_j\mathbf{b}_j^{\mathrm{T}}| = o_{\mathrm{p}}(1).$$

Similar techniques can be used to establish the following

$$(n-1)\sum_j (\mathrm{II})(\mathrm{IV})^{\mathrm{T}} = o_{\mathrm{p}}(1).$$

The interactions between (II) and (V), (VI) and (VII) are more involved. We first consider the interaction between (II) and (V):

$$\begin{aligned}(n-1)\sum_j (\mathrm{II})(\mathrm{V})^{\mathrm{T}} &= -\frac{1}{n}\sum_j \mathbf{a}_j\varepsilon_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}} + o_{\mathrm{p}}(1) \qquad \text{(Assumption II.2)}\\ &= \frac{1}{n}\sum_j \mathbf{a}_j\varepsilon_j\sum_{i,i\neq j}\mathbf{b}_i\pi_{ij} - \frac{1}{n}\sum_j \mathbf{a}_j\varepsilon_j\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} + o_{\mathrm{p}}(1)\\ &= \frac{1}{n}\sum_j \mathbf{a}_j\varepsilon_j\sum_{i,i\neq j}\mathbf{b}_i\pi_{ij} + o_{\mathrm{p}}(1),\end{aligned}$$

which is asymptotically equivalent to $\mathbb{C}\mathsf{ov}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_1, \bar{\boldsymbol{\Psi}}_2]$. And by symmetry, $(n-1)\sum_j (\mathrm{V})(\mathrm{II})^{\mathrm{T}}$ is equivalent to $\mathbb{C}\mathsf{ov}_{[\cdot|\mathbf{Z}]}[\bar{\boldsymbol{\Psi}}_2, \bar{\boldsymbol{\Psi}}_1]$. And as a short digression,

$$
\begin{aligned}
&(n-1)\sum_j (\mathrm{V})^2 \\
&= \frac{1}{n-1}\sum_j \varepsilon_j^2 \left(\sum_{i,i\neq j} \mathbf{b}_i \frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 + o_{\mathrm{p}}(1) \\
&= \frac{1}{n-1}\sum_j \varepsilon_j^2 \left(\sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i] \frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 + \frac{1}{n-1}\sum_j \varepsilon_j^2 \left(\sum_{i,i\neq j} \Big(\mathbf{b}_i - \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\Big) \frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \\
&\quad + \frac{1}{n-1}\sum_j \varepsilon_j^2 \left(\sum_{i,i\neq j} \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i] \frac{\pi_{ij}}{1-\pi_{jj}}\right)\left(\sum_{i,i\neq j} \Big(\mathbf{b}_i - \mathbb{E}_{[\cdot|\mathbf{Z}]}[\mathbf{b}_i]\Big) \frac{\pi_{ij}}{1-\pi_{jj}}\right)^{\mathrm{T}} + o_{\mathrm{p}}(1),
\end{aligned}
$$

where the first term in the above display recovers $\mathbb{V}_{[\cdot|\mathbf{Z}]}[\bar{\mathbf{\Psi}}_2]$, while the rest two are negligible by conditional expectation calculation. Therefore we recovered the asymptotic variance.

Back to the interaction terms,

$$
\left|(n-1)\sum_j (\mathrm{II})(\mathrm{VI})^{\mathrm{T}}\right| = \left|\frac{1}{n-1}\sum_j \mathbf{a}_j \sum_{i,i\neq j} \mathbf{c}_i^{\mathrm{T}} \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \Big(\hat{\mu}_j - r_j\Big)^2\right| \precsim_{\mathrm{p}} \frac{1}{n-1}\sum_{i,j} \pi_{ij}^2 = o_{\mathrm{p}}(1),
$$

and

$$
\begin{aligned}
&\left|(n-1)\sum_j (\mathrm{II})(\mathrm{VII})^{\mathrm{T}}\right| = \left|\frac{2}{n-1}\sum_j \mathbf{a}_j \Big(\hat{\mu}_j - r_j\Big) \sum_{i,i\neq j} \mathbf{c}_i^{\mathrm{T}} \frac{\pi_{ij}}{1-\pi_{jj}} \Big(\hat{\mu}_i - \mu_i\Big)\right| \\
&\asymp_{\mathrm{p}} \left|\frac{2}{n-1}\sum_{i,j} \mathbf{a}_j \Big(\hat{\mu}_j - r_j\Big) \mathbf{c}_i^{\mathrm{T}} \pi_{ij} \Big(\hat{\mu}_i - \mu_i\Big)\right| \qquad \text{(Assumption II.4)} \\
&\leq \frac{2}{n-1} \cdot \sqrt{\sum_j |\mathbf{a}_j|^2 \Big(\hat{\mu}_j - r_j\Big)^2} \sqrt{\sum_j |\mathbf{c}_j|^2 \Big(\hat{\mu}_j - \mu_j\Big)^2} \\
&\leq o_{\mathrm{p}}(1) \cdot \frac{2}{n-1} \cdot \sqrt{\sum_j |\mathbf{a}_j|^2 \Big(\hat{\mu}_j - r_j\Big)^2} \sqrt{\sum_j |\mathbf{c}_j|^2} = o_{\mathrm{p}}(1),
\end{aligned}
$$

With a quick inspection, the above method also applies to the following interactions

$$
\begin{aligned}
&(n-1)\sum_j (\mathrm{III})(\mathrm{V})^{\mathrm{T}} = o_{\mathrm{p}}(1), (n-1)\sum_j (\mathrm{III})(\mathrm{VI})^{\mathrm{T}} = o_{\mathrm{p}}(1), (n-1)\sum_j (\mathrm{III})(\mathrm{VII})^{\mathrm{T}} = o_{\mathrm{p}}(1), \\
&(n-1)\sum_j (\mathrm{IV})(\mathrm{V})^{\mathrm{T}} = o_{\mathrm{p}}(1), (n-1)\sum_j (\mathrm{IV})(\mathrm{VI})^{\mathrm{T}} = o_{\mathrm{p}}(1), (n-1)\sum_j (\mathrm{IV})(\mathrm{VII})^{\mathrm{T}} = o_{\mathrm{p}}(1).
\end{aligned}
$$

Next we consider the squared terms involving (III) and (IV):

$$
\begin{aligned}
(n-1)\sum_j (\mathrm{III})^2 &= \frac{1}{n-1}\sum_j (\mathbf{b}_j)^2\Big(\hat{\mu}_j-\mu_j\Big)^2 \le o_{\mathrm{p}}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{b}_j|^2 = o_{\mathrm{p}}(1),\\
(n-1)\sum_j (\mathrm{IV})^2 &= \frac{1}{n-1}\sum_j (\mathbf{c}_j)^2\Big(\hat{\mu}_j-\mu_j\Big)^4 \le o_{\mathrm{p}}(1)\cdot\frac{1}{n-1}\sum_j |\mathbf{c}_j|^2 = o_{\mathrm{p}}(1).
\end{aligned}
$$

What remains are $(\mathrm{V})(\mathrm{VI})^{\mathrm{T}}$, $(\mathrm{V})(\mathrm{VII})^{\mathrm{T}}$, $(\mathrm{VI})^2$, $(\mathrm{VI})(\mathrm{VII})^{\mathrm{T}}$ and $(\mathrm{VII})^2$.

$$
\begin{aligned}
\Big|(n-1)\sum_j (\mathrm{V})(\mathrm{VI})^{\mathrm{T}}\Big| &= \Big|\frac{1}{n-1}\sum_{i,j}\mathbf{b}_i\pi_{ij}\Big(\hat{\mu}_j-r_j\Big)^3\left(\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\left(\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\right)^2\right)^{\mathrm{T}}\Big| + o_{\mathrm{p}}(1)\\
&\precsim_{\mathrm{p}} \sqrt{\frac{1}{n}\sum_{j,i,\ell}\pi_{ij}^2\pi_{\ell j}^2} = o_{\mathrm{p}}(1).
\end{aligned}
$$

And

$$
\begin{aligned}
&\Big|(n-1)\sum_j (\mathrm{V})(\mathrm{VII})^{\mathrm{T}}\Big|\\
&= \Big|\frac{2}{n-1}\sum_j\left(\sum_{i,i\neq j}\mathbf{b}_i\frac{\pi_{ij}}{1-\pi_{jj}}\Big(\hat{\mu}_j-r_j\Big)\right)\left(\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\Big(\hat{\mu}_\ell-\mu_\ell\Big)\Big(\hat{\mu}_j-r_j\Big)\right)^{\mathrm{T}}\Big|\\
&\precsim_{\mathrm{p}} \sqrt{\frac{1}{n-1}\sum_j\Big(\hat{\mu}_j-r_j\Big)^4\left|\sum_{\ell,\ell\neq j}\mathbf{c}_\ell\frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\Big(\hat{\mu}_\ell-\mu_\ell\Big)\right|^2} = o_{\mathrm{p}}(1),
\end{aligned}
$$

where the last line uses Assumption II.2. Using techniques in the above results, we can show

$$
(n-1)\sum_j (\mathrm{VI})^2 = o_{\mathrm{p}}(1),\qquad (n-1)\sum_j (\mathrm{VII})^2 = o_{\mathrm{p}}(1),\qquad (n-1)\sum_j (\mathrm{VI})(\mathrm{VII})^{\mathrm{T}} = o_{\mathrm{p}}(1),
$$

which closes the proof. ■

II.9.7 Proof of Proposition II.2

Given consistency, we are able to linearize the bootstrap estimating equation with respect to $\hat{\boldsymbol{\theta}}^\star$, around $\hat{\boldsymbol{\theta}}$:

$$
\sqrt{n}\Big(\hat{\boldsymbol{\theta}}^\star-\hat{\boldsymbol{\theta}}\Big) = \boldsymbol{\Sigma}_0\left[\frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i,\hat{\mu}_i^\star,\hat{\boldsymbol{\theta}})\right]\Big(1+o_{\mathrm{p}}(1)\Big),
$$

where for notational simplicity, we define $\mathbf{m}^\star(\mathbf{w}_i,\cdot,\cdot) := (1+e_i^\star)\cdot\mathbf{m}(\mathbf{w}_i,\cdot,\cdot)$. We further expand the above with respect to the bootstrapped first step:

$$\frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i,\hat{\mu}_i^\star,\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) \tag{II.23}$$

$$+\frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star-\hat{\mu}_i\Big) \tag{II.24}$$

$$+\frac{1}{\sqrt{n}}\sum_i \frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i,\tilde{\mu}_i^\star,\hat{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star-\hat{\mu}_i\Big)^2 + o_{\mathrm{p}}(1). \tag{II.25}$$

Analyses of the above terms are similar to those of Lemma II.6 and II.7, with more delicate arguments.

Lemma II.8 (Term (II.23))
Assume Assumption II.1, II.2, II.3 and II.5 hold, and $k = O(\sqrt{n})$. Then

$$\textit{(II.23)} = \frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) + O_{\mathrm{p}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathrm{p}}(1). \qquad \|$$

Note that

$$\begin{aligned}\text{(II.23)} &= \frac{1}{\sqrt{n}}\sum_i \mathbf{m}^\star(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) + o_{\mathrm{p}}(1)\\ &= \frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0) + o_{\mathrm{p}}(1).\end{aligned}$$

the last equality comes from the argument that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}})\Big(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\Big) &\precsim_{\mathbb{P}} \frac{1}{n}\sum_i e_i^\star\cdot\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\tilde{\boldsymbol{\theta}})\\ &\xrightarrow{\mathrm{p}} \mathbb{E}\left[e_i^\star\cdot\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\right],\end{aligned}$$

given Assumption II.1. To further understand the last term, we still need to expand it with respect to $\hat{\mu}_i$, yielding

$$\frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) \tag{I}$$

$$+\frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_i-\mu_i\Big) \tag{II}$$

$$+\frac{1}{\sqrt{n}}\sum_i e_i^\star\cdot\frac{1}{2}\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\Big(\hat{\mu}_i-\mu_i\Big)^2\cdot(1+o_{\mathrm{p}}(1)). \tag{III}$$

(I) apparently contributes to the first order. For (II), note that it can be simplified using exactly the same argument used in Lemma II.6 and II.6. Equivalently, assuming Assumption II.1 and II.2, then

$$\text{(II)} = O_{\mathrm{p}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathrm{p}}(1).$$

By the same argument, (III) can be simplified with Lemma II.7 and II.7:

$$\text{(III)} = O_{\mathrm{p}}\left(\sqrt{\frac{k}{n}}\right) + o_{\mathrm{p}}(1).$$

$\blacksquare$

Lemma II.9 (Term (II.24))

Assume Assumption II.1, II.2, II.3 and II.5 hold, and $k = O(\sqrt{n})$. Then

$$\textit{(II.24)} = \frac{1}{\sqrt{n}}\sum_i \left(\sum_j \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0)|\mathbf{z}_j\right]\pi_{ij}\right)\varepsilon_i e_i^\star + \frac{1}{\sqrt{n}}\sum_i \mathbf{b}_{1,i}\cdot\pi_{ii} + o_{\mathrm{p}}(1),$$

where $\mathbf{b}_{1,i}$ is given in Lemma II.6. $\|$

For (II.24), we first show that it is possible to replace $\hat{\boldsymbol{\theta}}$ by $\boldsymbol{\theta}_0$, provided $\partial\dot{\mathbf{m}}/\partial\boldsymbol{\theta}$ is Hölder continuous in μ_i and $\boldsymbol{\theta}$:

$$\text{(II.24)} = \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big) + \frac{1}{n}\sum_i \frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big)\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big),$$

where the second term is bounded by the following

$$\begin{aligned}&\left|\frac{1}{n}\sum_i \frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big)\sqrt{n}\Big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\Big)\right| \\ &\precsim_{\mathrm{p}} \frac{1}{n}\sum_i \left|\frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big)\right| = o_{\mathrm{p}}(1)\cdot\frac{1}{n}\sum_i \left|\frac{\partial}{\partial\boldsymbol{\theta}}\dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \tilde{\boldsymbol{\theta}})\right| = o_{\mathrm{p}}(1),\end{aligned}$$

where the last one uses the uniform consistency of $\hat{\mu}_i^\star$ and $\hat{\mu}_i$. Hence

$$\text{(II.24)} = \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big(\hat{\mu}_i^\star - \hat{\mu}_i\Big) + o_\mathrm{p}(1) = \frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big(\sum_j \pi_{ij}\varepsilon_j e_j^\star\Big)$$

$$\underbrace{-\frac{1}{\sqrt{n}}\sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\Big(\sum_j \pi_{ij}(\hat{\mu}_j - \mu_j)e_j^\star\Big)}_{\text{(I)}} + o_\mathrm{p}(1).$$

For (I),

$$\mathbb{E}^\star\Big[\text{(I)(I)}^\mathrm{T}\Big] = \frac{1}{n}\mathbb{E}^\star\left[\sum_{i,i',j,j'} \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\dot{\mathbf{m}}^\star(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)^\mathrm{T}(\hat{\mu}_j - \mu_j)(\hat{\mu}_{j'} - \mu_{j'})e_j^\star e_{j'}^\star \pi_{ij}\pi_{i'j'}\right]$$

$$= \frac{1}{n}\sum_{\substack{i,i',j\\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)^\mathrm{T}(\hat{\mu}_j - \mu_j)^2\pi_{ij}\pi_{i'j} \tag{II}$$

$$+ \frac{2}{n}\sum_{\substack{i,i'\\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)^\mathrm{T}(\hat{\mu}_i - \mu_i)(\hat{\mu}_{i'} - \mu_{i'})\pi_{ii}\pi_{i'i'} \tag{III}$$

$$+ \frac{2}{n}\sum_{\substack{i,j\\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)^\mathrm{T}(\hat{\mu}_j - \mu_j)^2\pi_{ij}^2 \tag{IV}$$

$$+ \frac{C_1}{n}\sum_{\substack{i,j\\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_j, \hat{\mu}_j, \boldsymbol{\theta}_0)^\mathrm{T}(\hat{\mu}_j - \mu_j)^2\pi_{ij}\pi_{jj} \tag{V}$$

$$+ \frac{C_2}{n}\sum_{\substack{i\\ \text{distinct}}} \dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)^\mathrm{T}(\hat{\mu}_i - \mu_i)^2\pi_{ii}^2, \tag{VI}$$

where C_1 and C_2 are related to the third and fourth moments of $e_i^\star$. Then for each term,

$$\begin{aligned}|\text{(II)}| &\le \left(\max_{1\le i\le n}|\hat{\mu}_i - \mu_i|^2\right)\cdot\frac{1}{n}\sum_{\substack{i,i'\\ \text{distinct}}} |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|\,|\dot{\mathbf{m}}(\mathbf{w}_{i'}, \hat{\mu}_{i'}, \boldsymbol{\theta}_0)|\,\pi_{ii'}\\ &\le o_\mathrm{p}(1)\cdot\frac{1}{n}\sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|^2 = o_\mathrm{p}(1), \qquad \text{(projection and Assumption II.2)}\end{aligned}$$

provided $\dot{\mathbf{m}}$ is Hölder continuous in μ_i. (III) can be handled by observing that

$$
\begin{aligned}
|(\mathrm{III})| &\leq \left(\frac{1}{\sqrt{n}} \sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)| \, \pi_{ii} \, |\hat{\mu}_i - \mu_i| \right)^2 \\
&\leq o_{\mathrm{p}}(1) \cdot \left(\frac{1}{\sqrt{n}} \sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)| \, \pi_{ii} \right)^2 = o_{\mathrm{p}}\left(\frac{k^2}{n} \right).
\end{aligned}
$$

Similarly

$$
|(\mathrm{IV})| \leq o_{\mathrm{p}}(1) \cdot \frac{2}{n} \sum_{\substack{i,j \\ \text{distinct}}} |\dot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|^2 \, \pi_{ij}^2 = o_{\mathrm{p}}\left(\frac{k}{n} \right),
$$

and

$$
\begin{aligned}
|(\mathrm{V})| &\leq \frac{C_1}{n} \left(\sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|^2 \right)^{1/2} \left(\sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|^2 \, |\hat{\mu}_i - \mu_i|^4 \pi_{jj}^2 \right)^{1/2} \\
&\precsim_{\mathbb{P}} n^{-1} \cdot \sqrt{n} \cdot \sqrt{k} \cdot o_{\mathrm{p}}(1) = o_{\mathrm{p}}\left(\sqrt{\frac{k}{n}} \right).
\end{aligned}
$$

Finally,

$$
|(\mathrm{VI})| \leq \frac{C_2}{n} \sum_i |\dot{\mathbf{m}}(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0)|^2 \, |\hat{\mu}_i - \mu_i|^2 \pi_{ii}^2 = o_{\mathrm{p}}\left(\frac{k}{n} \right).
$$

To summarize, we have the following

$$
\begin{aligned}
(\mathrm{II.24}) &= \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \hat{\mu}_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \varepsilon_j e_j^\star \Big) + o_{\mathrm{p}}\left(\frac{k}{\sqrt{n}} \vee 1 \right) \\
&= \frac{1}{\sqrt{n}} \sum_i \dot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \varepsilon_j e_j^\star \Big) + o_{\mathrm{p}}\left(\frac{k}{\sqrt{n}} \vee 1 \right),
\end{aligned}
$$

where the second line relies on almost the same argument. Finally, we can apply the same techniques used to prove Lemma II.6 and II.6, yielding

$$
(\mathrm{II.24}) = \frac{1}{\sqrt{n}} \sum_i \left(\sum_j \mathbb{E}\left[\dot{\mathbf{m}}(\mathbf{w}_j, \mu_j, \boldsymbol{\theta}_0) | \mathbf{z}_j \right] \pi_{ij} \right) \varepsilon_i e_i^\star + \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,i} \cdot \pi_{ii} + o_{\mathrm{p}}\left(\frac{k}{\sqrt{n}} \vee 1 \right).
$$

■

Lemma II.10 (Term (II.25))

Assume Assumption II.1, II.2, II.3 and II.5 hold, and $k = O(\sqrt{n})$. Then

$$\text{(II.25)} = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_{2,ij} \cdot \pi_{ij}^2 + \frac{1}{\sqrt{n}} \sum_{i} \mathbf{b}_{2,ii} \cdot \pi_{ii}^2 \cdot \mathbb{E}[e_i^{\star 3}] + o_{\mathrm{p}}(1),$$

where $\mathbf{b}_{2,ij}$ is given in Lemma II.7. ‖

First note that

$$\begin{aligned}
\text{(II.25)} &= \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) \Big(\hat{\mu}_i^\star - \hat{\mu}_i \Big)^2 \\
&= \underbrace{\frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big(\hat{\mu}_i^\star - \hat{\mu}_i \Big)^2}_{\text{(I)}} \\
&\quad + \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \Big[\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) - \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big] \Big(\hat{\mu}_i^\star - \hat{\mu}_i \Big)^2,
\end{aligned}$$

where the second term is easily bounded by

$$\begin{aligned}
&\left| \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \Big[\ddot{\mathbf{m}}^\star(\mathbf{w}_i, \tilde{\mu}_i^\star, \hat{\boldsymbol{\theta}}) - \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big] \Big(\hat{\mu}_i^\star - \hat{\mu}_i \Big)^2 \right| \\
&\leq \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} (1 + e_i) \cdot \mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}}) \cdot (|\tilde{\mu}_i^\star - \mu_i| + |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|)^\alpha \cdot |\hat{\mu}_i^\star - \hat{\mu}_i|^2 \\
&\leq o_{\mathrm{p}}(1) \cdot \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} (1 + e_i) \cdot \mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}}) \cdot |\hat{\mu}_i^\star - \hat{\mu}_i|^2. \qquad \text{(II)}
\end{aligned}$$

Compare (I) and (II) and note that Assumption II.1 imposes the same restrictions on $\ddot{\mathbf{m}}$ and $\mathcal{H}_i^{\alpha,\delta}(\ddot{\mathbf{m}})$. Hence generically, (II) has the order

$$\text{(II)} = o_{\mathrm{p}}\left(|\text{(I)}| \right).$$

Next we consider (I), which can be written as

$$\text{(I)} = \frac{1}{\sqrt{n}} \sum_i \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \Big(\sum_j \pi_{ij} \hat{\varepsilon}_j e_j^\star \Big)^2 = \frac{1}{\sqrt{n}} \sum_{i,j,\ell} \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \hat{\varepsilon}_j \hat{\varepsilon}_\ell e_j^\star e_\ell^\star \pi_{ij} \pi_{i\ell}.$$

The key step, as before, is to replace $\hat{\varepsilon}$ by ε. Note that

$$\begin{aligned}(\mathrm{I}) &= \frac{1}{\sqrt{n}}\sum_{i,j,\ell}\frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\hat{\varepsilon}_j\varepsilon_\ell e_j^\star e_\ell^\star\pi_{ij}\pi_{i\ell}\\ &\quad-\frac{1}{\sqrt{n}}\sum_{i,\ell}\frac{1}{2}\ddot{\mathbf{m}}^\star(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)(\hat{\mu}_i^\star-\hat{\mu}_i)(\hat{\mu}_\ell-\mu_\ell)e_\ell^\star\pi_{i\ell}, && (\mathrm{III})\end{aligned}$$

and (for simplicity let $\mathbf{a}_i^\star = \ddot{\mathbf{m}}^\star(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)(\hat{\mu}_i^\star-\hat{\mu}_i)$)

$$\begin{aligned}\mathbb{E}^\star\left[(\mathrm{III})(\mathrm{III})^{\mathrm{T}}\right] &= \frac{1}{4n}\sum_{i,i',j,j'}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_{i'}^{\star\mathrm{T}}(\hat{\mu}_j-\mu_j)(\hat{\mu}_{j'}-\mu_{j'})e_j^\star e_{j'}^\star\pi_{ij}\pi_{i'j'}\right]\\ &= \frac{1}{4n}\sum_{\substack{i,i',j\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_{i'}^{\star\mathrm{T}}\right](\hat{\mu}_j-\mu_j)^2\pi_{ij}\pi_{i'j} && (\mathrm{IV})\\ &\quad+\frac{1}{4n}\sum_{\substack{i,j\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_i^{\star\mathrm{T}}\right](\hat{\mu}_j-\mu_j)^2\pi_{ij}^2 && (\mathrm{V})\\ &\quad+\frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star]\mathbb{E}^\star[\mathbf{a}_{i'}^\star e_{i'}^\star]^{\mathrm{T}}(\hat{\mu}_i-\mu_i)(\hat{\mu}_{i'}-\mu_{i'})\pi_{ii}\pi_{i'i'} && (\mathrm{VI})\\ &\quad+\frac{1}{2n}\sum_{\substack{i,i'\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\right]\mathbb{E}^\star\left[e_{i'}^{\star 2}\mathbf{a}_{i'}^{\star\mathrm{T}}\right](\hat{\mu}_{i'}-\mu_{i'})^2\pi_{ii'}\pi_{i'i'} && (\mathrm{VII})\\ &\quad+\frac{1}{4n}\sum_{i}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_i^{\star\mathrm{T}}e_i^{\star 2}\right](\hat{\mu}_i-\mu_i)^2\pi_{ii}^2. && (\mathrm{VIII})\end{aligned}$$

Then

$$\begin{aligned}&|(\mathrm{IV})| = \left|\frac{1}{4n}\sum_{\substack{i,i',j\\ \text{distinct}}}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_{i'}^{\star\mathrm{T}}\right](\hat{\mu}_j-\mu_j)^2\pi_{ij}\pi_{i'j}\right|\\ &\precsim o_{\mathrm{p}}(1)\cdot\frac{1}{n}\sum_{i,i'}\mathbb{E}^\star\left[\mathbf{a}_i^\star\mathbf{a}_{i'}^{\star\mathrm{T}}\right]\pi_{ii'} \le o_{\mathrm{p}}(1)\cdot\frac{1}{n}\sum_{i,i'}\mathbb{E}^\star\left[|\mathbf{a}_i^\star|\right]\mathbb{E}^\star\left[|\mathbf{a}_{i'}^\star|\right]\pi_{ii'}\\ &\le o_{\mathrm{p}}(1)\cdot\frac{1}{n}\sum_{i,i'}|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)||\ddot{\mathbf{m}}(\mathbf{w}_{i'},\mu_{i'},\boldsymbol{\theta}_0)|\pi_{ii'} \le o_{\mathrm{p}}(1)\cdot\frac{1}{n}\sum_{i}|\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)|^2 = o_{\mathrm{p}}(1),\end{aligned}$$

where the second line uses Assumption II.2, the fourth line uses Assumption II.5, and the last line uses projection property and Assumption II.1. Similarly, we have, for (V),

$$|(\mathrm{V})| = \left| \frac{1}{4n} \sum_{\substack{i,j \\ \text{distinct}}} \mathbb{E}^\star \left[\mathbf{a}_i^\star \mathbf{a}_i^{\star\mathrm{T}} \right] (\hat{\mu}_j - \mu_j)^2 \pi_{ij}^2 \right| \precsim o_{\mathrm{p}}(1) \cdot \frac{1}{n} \sum_{i,j} |\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|^2 \pi_{ij}^2 = o_{\mathrm{p}}\left(\frac{k}{n}\right),$$

and the last equality is a simple consequence of Assumption II.1. (VI) is the most difficult, which can be rewritten as

$$\begin{aligned} |(\mathrm{VI})| &= \frac{1}{2n} \sum_{\substack{i,i' \\ \text{distinct}}} \mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star] \mathbb{E}^\star[\mathbf{a}_{i'}^\star e_{i'}^\star]^{\mathrm{T}} (\hat{\mu}_i - \mu_i)(\hat{\mu}_{i'} - \mu_{i'}) \pi_{ii} \pi_{i'i'} \\ &\asymp \left(\frac{1}{\sqrt{n}} \sum_i \mathbb{E}^\star[\mathbf{a}_i^\star e_i^\star] (\hat{\mu}_i - \mu_i) \pi_{ii} \right)^2 \precsim o_{\mathrm{p}}(1) \cdot \left(\frac{1}{\sqrt{n}} \sum_i |\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)| \pi_{ii} \right)^2 = o_{\mathrm{p}}\left(\frac{k^2}{n}\right). \end{aligned}$$

And

$$\begin{aligned} |(\mathrm{VII})| &= \left| \frac{1}{2n} \sum_{\substack{i,i' \\ \text{distinct}}} \mathbb{E}^\star\left[\mathbf{a}_i^\star\right] \mathbb{E}^\star\left[e_{i'}^{\star 2} \mathbf{a}_{i'}^{\star\mathrm{T}} \right] (\hat{\mu}_{i'} - \mu_{i'})^2 \pi_{ii'} \pi_{i'i'} \right| \\ &\precsim \frac{1}{n} \left(\sum_i |\mathbb{E}^\star\left[\mathbf{a}_i^\star\right]|^2 \right)^{1/2} \left(\sum_i |\mathbb{E}^\star\left[e_i^{\star 2} \mathbf{a}_i^\star \right]|^2 |\hat{\mu}_i - \mu_i|^2 \pi_{ii}^2 \right)^{1/2} \qquad \text{(projection)} \\ &\leq o_{\mathrm{p}}(1) \cdot \frac{1}{n} \left(\sum_i |\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|^2 \right)^{1/2} \left(\sum_i |\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|^2 \pi_{ii}^2 \right)^{1/2} \\ &= o_{\mathrm{p}}(1) \cdot n^{-1} \cdot n^{1/2} \cdot k^{1/2} = o_{\mathrm{p}}\left(\sqrt{\frac{k}{n}} \right). \end{aligned}$$

Finally

$$|(\mathrm{VII})| = \frac{1}{4n} \sum_i \mathbb{E}^\star \left[\mathbf{a}_i^\star \mathbf{a}_i^{\star\mathrm{T}} e_i^{\star 2} \right] (\hat{\mu}_i - \mu_i)^2 \pi_{ii}^2 \precsim o_{\mathrm{p}}(1) \cdot \frac{1}{n} \sum_i |\ddot{\mathbf{m}}(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0)|^2 \pi_{ii}^2 = o_{\mathrm{p}}\left(\frac{k}{n}\right).$$

Hence we have shown that

$$(\mathrm{I}) = \frac{1}{\sqrt{n}} \sum_{i,j,\ell} \frac{1}{2} \ddot{\mathbf{m}}^\star(\mathbf{w}_i, \mu_i, \boldsymbol{\theta}_0) \hat{\varepsilon}_j \varepsilon_\ell e_j^\star e_\ell^\star \pi_{ij} \pi_{i\ell} + o_{\mathrm{p}}\left(\frac{k}{\sqrt{n}} \vee 1 \right).$$

Not surprisingly, we can replicate the above argument, and replace $\hat{\varepsilon}_j$ by ε_j in the above display, yielding

$$(\mathrm{I}) = \frac{1}{\sqrt{n}}\sum_{i,j,\ell}\frac{1}{2}\ddot{\mathbf{m}}^{\star}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_j\varepsilon_\ell e_j^{\star}e_\ell^{\star}\pi_{ij}\pi_{i\ell} + o_{\mathrm{p}}\left(\frac{k}{\sqrt{n}}\vee 1\right).$$

The next step is to apply Lemma II.7 to conclude that

$$\begin{aligned}(\mathrm{I}) &= \frac{1}{\sqrt{n}}\sum_{i,j}\frac{1}{2}\mathbb{E}_{[\cdot|\mathbf{Z}]}\left[\ddot{\mathbf{m}}^{\star}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)\varepsilon_j^2 e_j^{\star 2}\right]\pi_{ij}^2 + o_{\mathrm{p}}\left(\frac{k}{\sqrt{n}}\vee 1\right)\\ &= \frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{b}_{2,ij}\cdot\pi_{ij}^2 + \frac{1}{\sqrt{n}}\sum_{i}\mathbf{b}_{2,ii}\cdot\pi_{ii}^2\cdot\mathbb{E}[e_i^{\star 3}] + o_{\mathrm{p}}\left(\frac{k}{\sqrt{n}}\vee 1\right).\end{aligned}$$

■

Asymptotic Representation

This is a simple consequence of linearization, Lemma II.8, II.9 and II.10. ■

II.9.8 Proof of Theorem II.3

Part 1

For the ease of exposition we ignore (asymptotic negligible) remainder terms in the proof. Then $\hat{\boldsymbol{\theta}}^{\star}$ has the expansion

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star} - \hat{\boldsymbol{\theta}}\right) = \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^{\star}\hat{\mathbf{a}}_i + \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^{\star}\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star} - \hat{\mu}_i\right) + \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^{\star}\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star} - \hat{\mu}_i\right)^2,$$

where to save notations we used $\omega_i^{\star} = 1 + e_i^{\star}$, $n_\omega = \sum_i \omega_i^{\star}$, and

$$\hat{\mathbf{a}}_i = \boldsymbol{\Sigma}_0\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) \qquad \hat{\mathbf{b}}_i = \boldsymbol{\Sigma}_0\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) \qquad \hat{\mathbf{c}}_i = \boldsymbol{\Sigma}_0\frac{\ddot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}})}{2}.$$

For future reference, let

$$\mathbf{a}_i = \boldsymbol{\Sigma}_0\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) \qquad \mathbf{b}_i = \boldsymbol{\Sigma}_0\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) \qquad \mathbf{c}_i = \boldsymbol{\Sigma}_0\frac{\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)}{2}.$$

Denote the leave-j-out estimator by $\hat{\boldsymbol{\theta}}^{\star,(j)}$, it is easy to see that

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(j)}-\hat{\boldsymbol{\theta}}\right) &= \frac{\sqrt{n}}{n_\omega-1}\sum_i(\omega_i^\star-\delta_{ij})\hat{\mathbf{a}}_i+\frac{\sqrt{n}}{n_\omega-1}\sum_i(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)\\ &\quad+\frac{\sqrt{n}}{n_\omega-1}\sum_i(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)^2,\end{aligned}$$

where $\delta_{ij}=\mathbb{1}[i=j]$. Recall that the jackknife estimator is defined as

$$\hat{\boldsymbol{\theta}}^{\star,(\cdot)}=\frac{1}{n_\omega}\sum_j\omega_j^\star\hat{\boldsymbol{\theta}}^{\star,(j)},$$

hence

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)}-\hat{\boldsymbol{\theta}}\right) &= \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{a}}_i+\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)\\ &\quad+\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i\right)^2.\end{aligned}$$

To simplify, we further expand the leave-j-out propensity score, which satisfies

$$\hat{\mu}_i^{\star,(j)}-\hat{\mu}_i=\hat{\mu}_i^\star-\hat{\mu}_i+\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^\star-r_j^\star),$$

hence

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)}-\hat{\boldsymbol{\theta}}\right) &= \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{a}}_i\\ &\quad+\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)\\ &\quad+\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)^2\\ &\quad+\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^\star-r_j^\star)\\ &\quad+\frac{2\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_i^\star-\hat{\mu}_i)(\hat{\mu}_j^\star-r_j^\star)\\ &\quad+\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\hat{\mu}_j^\star-r_j^\star)^2.\end{aligned}$$

Note that

$$\begin{aligned}\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{a}}_i &= \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_i \hat{\mathbf{a}}_i \sum_j \left(\omega_j^\star(\omega_i^\star-\delta_{ij})\right)\\ &= \frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_i \hat{\mathbf{a}}_i \left((n_\omega-\omega_i^\star)\omega_i^\star+\omega_i^\star(\omega_i^\star-1)\right)\\ &= \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^\star\hat{\mathbf{a}}_i.\end{aligned}$$

Similarly, we have

$$\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right) = \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^\star\hat{\mathbf{b}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right),$$

and

$$\frac{\sqrt{n}}{n_\omega(n_\omega-1)}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)^2 = \frac{\sqrt{n}}{n_\omega}\sum_i \omega_i^\star\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star-\hat{\mu}_i\right)^2.$$

As a consequence,

$$\begin{aligned}(n_\omega-1)\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)}-\hat{\boldsymbol{\theta}}^\star\right) &= \frac{\sqrt{n}}{n_\omega}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^\star-r_j^\star)\\ &+\frac{2\sqrt{n}}{n_\omega}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_i^\star-\hat{\mu}_i)(\hat{\mu}_j^\star-r_j^\star)\\ &+\frac{\sqrt{n}}{n_\omega}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\hat{\mu}_j^\star-r_j^\star)^2\\ &= \frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{b}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_j^\star-r_j^\star) \qquad \text{(I)}\\ &+\frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_i^\star-\hat{\mu}_i)(\hat{\mu}_j^\star-r_j^\star) \qquad \text{(II)}\\ &+\frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\hat{\mu}_j^\star-r_j^\star)^2 \qquad \text{(III)}\\ &+o_{\mathrm{p}}(1).\end{aligned}$$

Next we analyze each term. For term (I), it is

$$
\begin{aligned}
(\mathrm{I}) &= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} (\hat{\mu}_j^\star - r_j^\star) \\
&= \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \Big(\sum_{\ell} \pi_{j\ell} e_\ell^\star \hat{\varepsilon}_\ell - e_j^\star \hat{\varepsilon}_j \Big) \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,\ell} \omega_j^\star(\omega_i^\star - \delta_{ij}) e_\ell^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{j\ell} \hat{\varepsilon}_\ell && (\mathrm{I}.1) \\
&- \frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star e_j^\star (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j . && (\mathrm{I}.2)
\end{aligned}
$$

Again we consider conditional expectation:

$$
\begin{aligned}
\mathbb{E}^\star[(\mathrm{I}.1)] &= \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,\ell} \omega_j^\star(\omega_i^\star - \delta_{ij}) e_\ell^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \pi_{j\ell} \hat{\varepsilon}_\ell \right] \\
&= \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star \omega_i^\star e_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i \right] + \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star \omega_i^\star e_j^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] \\
&+ \mathbb{E}^\star \left[\frac{1}{\sqrt{n}} \sum_{i} \omega_i^\star(\omega_i^\star - 1) e_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ii}^2}{1-\pi_{ii}} \hat{\varepsilon}_i \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i + \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}} \hat{\varepsilon}_j + \frac{1}{\sqrt{n}} \sum_{i} (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}^2}{1-\pi_{ii}} \hat{\varepsilon}_i .
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathbb{E}^\star[(\mathrm{I}.2)] &= \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_{i,j} \omega_j^\star e_j^\star (\omega_i^\star - \delta_{ij}) \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] \\
&= \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \omega_j^\star e_j^\star \omega_i^\star \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j \right] + \mathbb{E}^\star \left[-\frac{1}{\sqrt{n}} \sum_{i} \omega_i^\star e_i^\star (\omega_i^\star - 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}}{1-\pi_{ii}} \hat{\varepsilon}_i \right] \\
&= -\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} \hat{\varepsilon}_j - \frac{1}{\sqrt{n}} \sum_{i} (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \frac{\pi_{ii}}{1-\pi_{ii}} \hat{\varepsilon}_i .
\end{aligned}
$$

Therefore

$$\mathbb{E}^\star[(\mathrm{I})] = \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i \tag{I.3}$$

$$- \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \pi_{ij} \hat{\varepsilon}_j \tag{I.4}$$

$$- \frac{1}{\sqrt{n}} \sum_i (\mathbb{E}^\star[e_i^{\star 3}] + 1) \hat{\mathbf{b}}_i \pi_{ii} \hat{\varepsilon}_i. \tag{I.5}$$

Furthermore,

$$\begin{aligned}
(\mathrm{I.3}) &= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \hat{\varepsilon}_i \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \frac{\pi_{ij}^2}{1-\pi_{jj}} \varepsilon_i + o_{\mathrm{p}}(1) = \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \left(\pi_{ij}^2 + \frac{\pi_{ij}^2 \pi_{jj}}{1-\pi_{jj}} \right) \varepsilon_i + o_{\mathrm{p}}(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \mathbf{b}_i \pi_{ij}^2 \varepsilon_i + o_{\mathrm{p}}(1) = \frac{1}{\sqrt{n}} \sum_{i,j} \mathbf{b}_i \pi_{ij}^2 \varepsilon_i + o_{\mathrm{p}}(1) \\
&= \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_i \pi_{ii} \varepsilon_i + o_{\mathrm{p}}(1) = \frac{1}{\sqrt{n}} \sum_i \mathbb{E}[\mathbf{b}_i \varepsilon_i | \mathbf{z}_i] \pi_{ii} + o_{\mathrm{p}}(1) \\
&= \boldsymbol{\Sigma}_0 \frac{1}{\sqrt{n}} \sum_i \mathbf{b}_{1,i} \pi_{ii} + o_{\mathrm{p}}(1).
\end{aligned}$$

The second line follows from consistency and (II.16); the third line follows from Assumption II.4 and (II.17); the fourth line is a simple fact of Lemma II.6. Similar argument applies to (I.5), which implies

$$(\mathrm{I.5}) = -\frac{1}{\sqrt{n}} \sum_i \boldsymbol{\Sigma}_0 (\mathbb{E}^\star[e_i^{\star 3}] + 1) \mathbf{b}_{1,i} \pi_{ii} + o_{\mathrm{p}}(1).$$

Finally,

$$\begin{aligned}
(\mathrm{I.4}) &= -\frac{1}{\sqrt{n}} \sum_{i,j,i\neq j} \hat{\mathbf{b}}_i \pi_{ij} \hat{\varepsilon}_j = -\frac{1}{\sqrt{n}} \sum_{i,j} \hat{\mathbf{b}}_i \pi_{ij} \hat{\varepsilon}_j + \frac{1}{\sqrt{n}} \sum_i \hat{\mathbf{b}}_i \pi_{ii} \hat{\varepsilon}_i \\
&= \frac{1}{\sqrt{n}} \sum_i \hat{\mathbf{b}}_i \pi_{ii} \hat{\varepsilon}_i = \frac{1}{\sqrt{n}} \sum_i \boldsymbol{\Sigma}_0 \mathbf{b}_{1,i} \pi_{ii} + o_{\mathrm{p}}(1),
\end{aligned}$$

where, in the second line, we used the fact that $\sum_{ij} \pi_{ij} \hat{\varepsilon}_j = 0$ for all i. Therefore

$$\text{(I)} = (1 - \mathbb{E}^\star[e_i^{\star 3}])\frac{1}{\sqrt{n}}\sum_i \mathbf{\Sigma}_0 \mathbf{b}_{1,i}\pi_{ii} + o_{\mathrm{p}}(1).$$

Next we consider (II). Note that it has the expansion:

$$\begin{aligned}
\text{(II)} &= \frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\hat{\mu}_i^\star - \hat{\mu}_i)(\hat{\mu}_j^\star - r_j^\star) \\
&= \frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}(\sum_\ell \pi_{i\ell}e_\ell^\star\hat{\varepsilon}_\ell)(\sum_\ell \pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell - e_j^\star\hat{\varepsilon}_j) \\
&= \frac{2}{\sqrt{n}}\sum_{i,j,\ell,\ell'}\omega_j^\star(\omega_i^\star - \delta_{ij})e_\ell^\star e_{\ell'}^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{i\ell}\pi_{j\ell'}\hat{\varepsilon}_\ell\hat{\varepsilon}_{\ell'} && \text{(II.1)} \\
&- \frac{2}{\sqrt{n}}\sum_{i,j,\ell}\omega_j^\star e_j^\star(\omega_i^\star - \delta_{ij})e_\ell^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{i\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_j. && \text{(II.2)}
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{E}^\star[\text{(II.1)}] \\
&= \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_i \omega_i^\star(\omega_i^\star - 1)e_i^\star e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ii}}{1-\pi_{ii}}\pi_{ii}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_i\right] \\
&+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_i^\star e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ii}\pi_{ji}\hat{\varepsilon}_i\hat{\varepsilon}_i\right] + \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_j^\star e_j^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ij}\pi_{jj}\hat{\varepsilon}_j\hat{\varepsilon}_j\right] \\
&+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,\ell,i\neq \ell}\omega_i^\star(\omega_i^\star - 1)e_\ell^\star e_\ell^\star\hat{\mathbf{c}}_i\frac{\pi_{ii}}{1-\pi_{ii}}\pi_{i\ell}\pi_{i\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_\ell\right] \\
&+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_i^\star e_j^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ii}\pi_{jj}\hat{\varepsilon}_i\hat{\varepsilon}_j\right] \\
&+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\omega_j^\star\omega_i^\star e_j^\star e_i^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ij}\pi_{ji}\hat{\varepsilon}_j\hat{\varepsilon}_i\right] \\
&+ \mathbb{E}^\star\left[\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell \\ \text{distinct}}}\omega_j^\star(\omega_i^\star - \delta_{ij})e_\ell^\star e_\ell^\star\hat{\mathbf{c}}_i\frac{\pi_{ij}}{1-\pi_{jj}}\pi_{i\ell}\pi_{j\ell}\hat{\varepsilon}_\ell\hat{\varepsilon}_\ell\right] \\
&= \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}\pi_{jj}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j + \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j + \frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell \\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2 + o_{\mathrm{p}}(1),
\end{aligned}$$

where the $o_{\mathrm{p}}(1)$ terms follows from (II.17) and Assumption II.4. Similarly,

$$\begin{aligned}
\mathbb{E}^\star[\text{(II.2)}] &= \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_i \omega_i^\star e_i^\star(\omega_i^\star - 1)e_i^\star \hat{\mathbf{c}}_i \frac{\pi_{ii}}{1-\pi_{ii}}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_i\right] \\
&+ \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j} \omega_j^\star e_j^\star \omega_i^\star e_i^\star \hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_j\right] \\
&+ \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j} \omega_j^\star e_j^\star \omega_i^\star e_j^\star \hat{\mathbf{c}}_i \frac{\pi_{ij}}{1-\pi_{jj}}\pi_{ij}\hat{\varepsilon}_j\hat{\varepsilon}_j\right] \\
&= \frac{1}{\sqrt{n}}O_{\mathrm{p}}(\sum_i \pi_{ii}^2) - \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j - \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2 \\
&= -\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j - \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2 + o_{\mathrm{p}}(1).
\end{aligned}$$

Hence

$$\mathbb{E}^\star[\text{(II)}] = \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}\pi_{jj}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j \tag{II.3}$$

$$+ \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j \tag{II.4}$$

$$+ \frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell \\ \text{distinct}}}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2 \tag{II.5}$$

$$- \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}\pi_{ii}}{1-\pi_{jj}}\hat{\varepsilon}_i\hat{\varepsilon}_j \tag{II.6}$$

$$- \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}(\mathbb{E}^\star[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2 + o_{\mathrm{p}}(1). \tag{II.7}$$

First note that

$$\text{(II.4)} = \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbf{c}_i\frac{\pi_{ij}^3}{1-\pi_{jj}}\varepsilon_i\varepsilon_j + o_{\mathrm{p}}(1) = \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}[\mathbf{c}_i\varepsilon_i\varepsilon_j|\mathbf{z}_i,\mathbf{z}_j]\frac{\pi_{ij}^3}{1-\pi_{jj}} + o_{\mathrm{p}}(1) = o_{\mathrm{p}}(1).$$

Next

$$\begin{aligned}
\text{(II.3)+(II.6)} &= -\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\pi_{ij}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_j = -\frac{2}{\sqrt{n}}\sum_{i,j}\hat{\mathbf{c}}_i\pi_{ij}\pi_{ii}\hat{\varepsilon}_i\hat{\varepsilon}_j + \frac{2}{\sqrt{n}}\sum_i\hat{\mathbf{c}}_i\pi_{ii}^2\hat{\varepsilon}_i^2 \\
&= \frac{2}{\sqrt{n}}\sum_i\hat{\mathbf{c}}_i\pi_{ii}^2\hat{\varepsilon}_i^2 = o_{\mathrm{p}}(1),
\end{aligned}$$

where for the third line we used the fact $\sum_{i,j}\pi_{ij}\hat{\varepsilon}_j = 0$, and the last line follows from Assumption II.4. Hence

$$\mathbb{E}^\star[(\mathrm{II})] = \frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}} \hat{\mathbf{c}}_i \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}}\hat{\varepsilon}_\ell^2 - \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2 \tag{II.8}$$

$$-\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}^\star[e_i^{\star 3}]\hat{\mathbf{c}}_i \frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2 + o_{\mathrm{p}}(1). \tag{II.9}$$

For the first line, we have the following result:

$$\begin{aligned}
(\text{II.8}) &= \left|\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}} \hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\hat{\mathbf{c}}_i\hat{\varepsilon}_j^2\frac{\pi_{ij}^2}{1-\pi_{jj}}\right| \\
&= \left|\frac{2}{\sqrt{n}}\sum_{\substack{i,j,\ell\\ \text{distinct}}} \hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}}\sum_{i,\ell,i\neq \ell}\hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2\frac{\pi_{i\ell}^2}{1-\pi_{\ell\ell}}\right| && (\text{change } j \ \to\ \ell) \\
&= \left|\frac{2}{\sqrt{n}}\sum_{i,j,\ell} \hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}}\sum_{i,\ell,i\neq \ell}\hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2\frac{\pi_{i\ell}^2}{1-\pi_{\ell\ell}}\right| + o_{\mathrm{p}}(1) \\
& && ((\text{II.17}) \text{ and Assumption II.4}) \\
&= \left|\frac{2}{\sqrt{n}}\sum_{i,j,\ell} \hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}}\sum_{i,\ell,i\neq \ell}\hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2\pi_{i\ell}^2\right| + o_{\mathrm{p}}(1) && ((\text{II.17}) \text{ and Assumption II.4}) \\
&= \left|\frac{2}{\sqrt{n}}\sum_{i,j,\ell} \hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}}\sum_{i,\ell}\hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2\pi_{i\ell}^2\right| + o_{\mathrm{p}}(1) && (\text{Assumption II.4}) \\
&= \left|\frac{2}{\sqrt{n}}\sum_{i,j,\ell} \hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}}{1-\pi_{jj}} - \frac{2}{\sqrt{n}}\sum_{i,j,\ell}\hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2\pi_{ij}\pi_{i\ell}\pi_{j\ell}\right| + o_{\mathrm{p}}(1) \\
&= \left|\frac{2}{\sqrt{n}}\sum_{i,j,\ell} \hat{\mathbf{c}}_i\hat{\varepsilon}_\ell^2 \frac{\pi_{ij}\pi_{i\ell}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}}\right| + o_{\mathrm{p}}(1) \\
&\lesssim_{\mathrm{p}} \frac{1}{\sqrt{n}}\sqrt{\sum_{i,\ell}\pi_{i\ell}^2}\sqrt{\sum_{i,\ell}\left(\sum_j \frac{\pi_{ij}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}}\right)^2} = \frac{\sqrt{k}}{\sqrt{n}}\sqrt{\sum_{i,\ell}\left(\sum_j \frac{\pi_{ij}\pi_{j\ell}\pi_{jj}}{1-\pi_{jj}}\right)^2} \\
&= \frac{\sqrt{k}}{\sqrt{n}}\sqrt{\sum_{i,\ell}\sum_{jj'}\frac{\pi_{ij}\pi_{j\ell}\pi_{jj}\pi_{ij'}\pi_{j'\ell}\pi_{j'j'}}{(1-\pi_{jj})(1-\pi_{j'j'})}} = \frac{\sqrt{k}}{\sqrt{n}}\sqrt{\sum_{jj'}\frac{\pi_{jj}\pi_{j'j'}\pi_{jj'}^2}{(1-\pi_{jj})(1-\pi_{j'j'})}} \\
&\lesssim_{\mathrm{p}} \frac{\sqrt{k}}{\sqrt{n}}\sqrt{\sum_{jj'}\pi_{jj}\pi_{j'j'}\pi_{jj'}^2} = \frac{\sqrt{k}}{\sqrt{n}}\cdot o_{\mathrm{p}}(\sqrt{k}) = o_{\mathrm{p}}(1).
\end{aligned}$$

Hence we have:

$$\begin{aligned}
\text{(II)} &= -\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}^\star[e_i^{\star 3}]\hat{\mathbf{c}}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\hat{\varepsilon}_j^2 + o_{\mathrm{p}}(1) = -\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}^\star[e_i^{\star 3}]\mathbf{c}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\varepsilon_j^2 + o_{\mathrm{p}}(1) \\
&= -\frac{2}{\sqrt{n}}\sum_{i,j}\mathbb{E}^\star[e_i^{\star 3}]\mathbf{c}_i\frac{\pi_{ij}^2}{1-\pi_{jj}}\varepsilon_j^2 + o_{\mathrm{p}}(1) = -\mathbb{E}^\star[e_i^{\star 3}]\boldsymbol{\Sigma}_0\frac{2}{\sqrt{n}}\sum_{i,j}\mathbf{b}_{2,ij}\pi_{ij}^2 + o_{\mathrm{p}}(1),
\end{aligned}$$

and the last line follows essentially from Lemma II.7.

(III) has the following expansion:

$$\begin{aligned}
\text{(III)} &= \frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\hat{\mu}_j^\star - r_j^\star)^2 \\
&= \frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\sum_{\ell}\pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell - e_j^\star\hat{\varepsilon}_j)^2 \\
&= \frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\sum_{\ell}\pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell)^2 && \text{(III.1)} \\
&\quad - \frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(\sum_{\ell}\pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell)e_j^\star\hat{\varepsilon}_j && \text{(III.2)} \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2(e_j^\star\hat{\varepsilon}_j)^2. && \text{(III.3)}
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}^\star[\text{(III.1)}] &= \mathbb{E}^\star\left[\frac{1}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\sum_{\ell}\pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell\right)^2\right] \\
&= \frac{1}{\sqrt{n}}O_{\mathrm{p}}\left(\sum_i\pi_{ii}^4 + \sum_{i,j}\pi_{ij}^4 + \sum_{i,j}\pi_{ij}^2\pi_{jj}^2 + \sum_{i,\ell}\pi_{i\ell}^2\pi_{ii}^2 + \sum_{i,j}\pi_{ij}^3\pi_{jj} + \sum_{i,j,\ell}\pi_{ij}^2\pi_{j\ell}^2\right) \\
&= o_{\mathrm{p}}(1),
\end{aligned}$$

by (II.17), (II.18) and Assumption II.4. Next

$$\begin{aligned}
\mathbb{E}^\star[\text{(III.2)}] &= \mathbb{E}^\star\left[-\frac{2}{\sqrt{n}}\sum_{i,j}\omega_j^\star(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(\sum_{\ell}\pi_{j\ell}e_\ell^\star\hat{\varepsilon}_\ell\right)e_j^\star\hat{\varepsilon}_j\right] \\
&= -\frac{2}{\sqrt{n}}\sum_{i,j,i\neq j}\mathbb{E}[\mathbf{c}_i\varepsilon_i\varepsilon_j|\mathbf{z}_i,\mathbf{z}_j]\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\pi_{ij} + o_{\mathrm{p}}(1) = o_{\mathrm{p}}(1).
\end{aligned}$$

Finally

$$\begin{aligned}
\mathbb{E}^{\star}[(\text{III}.3)] &= \frac{1}{\sqrt{n}}\sum_{i,j}(\mathbb{E}^{\star}[e_i^{\star 3}]+1)\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\hat{\varepsilon}_j^2 + o_{\mathrm{p}}(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i,j}(\mathbb{E}^{\star}[e_i^{\star 3}]+1)\mathbf{c}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\varepsilon_j^2 + o_{\mathrm{p}}(1) \\
&= (\mathbb{E}^{\star}[e_i^{\star 3}]+1)\boldsymbol{\Sigma}_0\frac{1}{\sqrt{n}}\sum_{i,j}\mathbf{b}_{2,ij}\pi_{ij}^2 + o_{\mathrm{p}}(1).
\end{aligned}$$

Given the previous results,

$$(n_\omega - 1)\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}}^{\star}\right) = (1-\mathbb{E}^{\star}[e_i^{\star 3}])\boldsymbol{\Sigma}_0\frac{1}{\sqrt{n}}\left(\sum_i \mathbf{b}_{1,i}\pi_{ii} + \sum_{i,j}\mathbf{b}_{2,ij}\pi_{ij}^2\right) + o_{\mathrm{p}}(1)$$

$$= (1-\mathbb{E}^{\star}[e_i^{\star 3}])\boldsymbol{\mathcal{B}} + o_{\mathrm{p}}(1).$$

Part 2

We follow the notational convention used in the previous part:

$$\hat{\mathbf{a}}_i = \boldsymbol{\Sigma}_0\mathbf{m}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) \qquad \hat{\mathbf{b}}_i = \boldsymbol{\Sigma}_0\dot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}}) \qquad \hat{\mathbf{c}}_i = \boldsymbol{\Sigma}_0\frac{\ddot{\mathbf{m}}(\mathbf{w}_i,\hat{\mu}_i,\hat{\boldsymbol{\theta}})}{2}.$$

Similarly,

$$\mathbf{a}_i = \boldsymbol{\Sigma}_0\mathbf{m}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) \qquad \mathbf{b}_i = \boldsymbol{\Sigma}_0\dot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0) \qquad \mathbf{c}_i = \boldsymbol{\Sigma}_0\frac{\ddot{\mathbf{m}}(\mathbf{w}_i,\mu_i,\boldsymbol{\theta}_0)}{2}.$$

First note that the jackknife variance estimator for the bootstrap data takes the form:

$$(n-1)\sum_j\left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)}\right)^2,$$

where for a (column) vector $\mathbf{v}$, we use $\mathbf{v}^2$ to denote $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ to save space. Then the variance estimator could be rewritten as

$$\begin{aligned}
\hat{\boldsymbol{\mathcal{V}}}^{\star} &= (n-1)\sum_j\left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star}\right)^2 - \frac{1}{n-1}\left(\hat{\boldsymbol{\mathcal{B}}}^{\star}\right)^2 \\
&= (n-1)\sum_j\left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star}\right)^2 + O_{\mathrm{p}}\left(\frac{1}{n}\right).
\end{aligned}$$

Next recall that

$$\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}} = \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{a}}_i + \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right) + \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right)^2.$$

Then we make the following decomposition:

$$\frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{a}}_i = \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{a}}_i - \frac{1}{n_\omega - 1}\hat{\mathbf{a}}_j,$$

and

$$\begin{aligned}
&\frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right) = \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i - \frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)\\
&= \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) - \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)\\
&= \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) - \frac{1}{n_\omega - 1}\hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) - \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right),
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^{\star,(j)} - \hat{\mu}_i\right)^2 = \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i - \frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)^2\\
&= \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2\\
&+ \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2\\
&- \frac{2}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)\left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)\\
&= \frac{1}{n_\omega - 1}\sum_i \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2\\
&- \frac{1}{n_\omega - 1}\hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2\\
&+ \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2\\
&- \frac{2}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)\left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right).
\end{aligned}$$

Therefore

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}} &= \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{a}}_i - \frac{1}{n_\omega - 1}\hat{\mathbf{a}}_j \\
&+ \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) - \frac{1}{n_\omega - 1}\hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) \\
&- \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&+ \frac{1}{n_\omega - 1}\sum_i \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 - \frac{1}{n_\omega - 1}\hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&+ \frac{1}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega - 1}\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(\cdot)} - \hat{\boldsymbol{\theta}} &= \frac{1}{n_\omega}\sum_j \omega_j^\star \left(\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}\right) \\
&= \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{a}}_i - \frac{1}{n_\omega(n_\omega - 1)}\sum_j \omega_j^\star \hat{\mathbf{a}}_j \\
&+ \frac{1}{n_\omega - 1}\sum_i \omega_i^\star \hat{\mathbf{b}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) - \frac{1}{n_\omega(n_\omega - 1)}\sum_j \omega_i^\star \hat{\mathbf{b}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right) \\
&- \frac{1}{n_\omega(n_\omega - 1)}\sum_{i,j} (\omega_i^\star - \delta_{ij})\omega_j^\star \hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&+ \frac{1}{n_\omega - 1}\sum_i \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right)^2 - \frac{1}{n_\omega(n_\omega - 1)}\sum_j \omega_j^\star \hat{\mathbf{c}}_j \left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&+ \frac{1}{n_\omega(n_\omega - 1)}\sum_{i,j} (\omega_i^\star - \delta_{ij})\omega_j^\star \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2 \\
&- \frac{2}{n_\omega(n_\omega - 1)}\sum_{i,j} (\omega_i^\star - \delta_{ij})\omega_j^\star \hat{\mathbf{c}}_i \left(\hat{\mu}_i^\star - \hat{\mu}_i\right) \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right),
\end{aligned}$$

which means

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}^{\star,(j)} - \hat{\boldsymbol{\theta}}^{\star,(\cdot)} &= \frac{1}{n_\omega - 1}\left(\hat{\boldsymbol{\theta}}^\star - \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\mathcal{B}}}^\star/\sqrt{n_\omega}\right) \\
&\quad - \frac{1}{n_\omega - 1}\hat{\mathbf{a}}_j - \frac{1}{n_\omega - 1}\hat{\mathbf{b}}_j\left(\hat{\mu}_j^\star - \hat{\mu}_j\right) - \frac{1}{n_\omega - 1}\hat{\mathbf{c}}_j\left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 \\
&\quad - \frac{1}{n_\omega - 1}\sum_i(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}e_j^\star\hat{\varepsilon}_j\right) \\
&\quad + \frac{1}{n_\omega - 1}\sum_i(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(e_j^\star\hat{\varepsilon}_j\right)^2 \\
&\quad - \frac{2}{n_\omega - 1}\sum_i(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star - \hat{\mu}_i\right)\left(\frac{\pi_{ij}}{1-\pi_{jj}}e_j^\star\hat{\varepsilon}_j\right) \\
&= \frac{1}{n_\omega - 1}\left(\hat{\boldsymbol{\theta}}_{\mathsf{bc}}^\star - \hat{\boldsymbol{\theta}}\right) && \text{(I)} \\
&\quad - \frac{1}{n_\omega - 1}\hat{\mathbf{a}}_j && \text{(II)} \\
&\quad - \frac{1}{n_\omega - 1}\hat{\mathbf{b}}_j\left(\hat{\mu}_j^\star - \hat{\mu}_j\right) && \text{(III)} \\
&\quad - \frac{1}{n_\omega - 1}\hat{\mathbf{c}}_j\left(\hat{\mu}_j^\star - \hat{\mu}_j\right)^2 && \text{(IV)} \\
&\quad - \frac{1}{n_\omega - 1}\sum_i(\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}e_j^\star\hat{\varepsilon}_j\right) && \text{(V)} \\
&\quad + \frac{1}{n_\omega - 1}\sum_i(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2\left(e_j^\star\hat{\varepsilon}_j\right)^2 && \text{(VI)} \\
&\quad - \frac{2}{n_\omega - 1}\sum_i(\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\hat{\mu}_i^\star - \hat{\mu}_i\right)\left(\frac{\pi_{ij}}{1-\pi_{jj}}e_j^\star\hat{\varepsilon}_j\right). && \text{(VII)}
\end{aligned}
$$

Term (I) is the easiest:

$$
(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{I})^2 \asymp \left(\hat{\boldsymbol{\theta}}_{\mathsf{bc}}^\star - \hat{\boldsymbol{\theta}}\right)^2 = o_{\mathrm{p}}(1),
$$

by consistency. Similarly

$$
(n_\omega - 1)\sum_j \omega_j^\star(\mathrm{I})\Big((\mathrm{II}) + \cdots(\mathrm{VII})\Big)^{\mathrm{T}} = \left(\hat{\boldsymbol{\theta}}_{\mathsf{bc}}^\star - \hat{\boldsymbol{\theta}}\right)\sum_j \omega_j^\star\Big((\mathrm{II}) + \cdots(\mathrm{VII})\Big)^{\mathrm{T}} = o_{\mathrm{p}}(1).
$$

Next

$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{II})^2 = \frac{1}{n_\omega - 1}\sum_j \omega_j^\star \left(\hat{\mathbf{a}}_j\right)^2 \xrightarrow{\mathrm{p}} \mathbb{V}[\bar{\boldsymbol{\Psi}}_1].$$

By the uniform consistency of $\hat{\mu}_j^\star$, it is very easy to show that

$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{II})(\mathrm{III})^{\mathrm{T}} = o_{\mathrm{p}}(1), \qquad (n_\omega - 1)\sum_j \omega_j^\star (\mathrm{II})(\mathrm{IV})^{\mathrm{T}} = o_{\mathrm{p}}(1).$$

Then

$$\begin{aligned}
(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{II})(\mathrm{V})^{\mathrm{T}} &= \frac{1}{n_\omega - 1}\sum_{i,j} \hat{\mathbf{a}}_j \hat{\mathbf{b}}_i^{\mathrm{T}} \omega_j^\star (\omega_i^\star - \delta_{ij}) \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right) \\
&= \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \sum_i \left[\hat{\mathbf{b}}_i^{\mathrm{T}} (\omega_i^\star - \delta_{ij}) \frac{\pi_{ij}}{1-\pi_{jj}}\right] \\
&= \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \sum_i \left[\hat{\mathbf{b}}_i^{\mathrm{T}} \pi_{ij}\right] && \text{(i)} \\
&+ \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \sum_i \left[\hat{\mathbf{b}}_i^{\mathrm{T}} \frac{\pi_{ij}\pi_{jj}}{1-\pi_{jj}}\right] && \text{(ii)} \\
&+ \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \sum_{i, i\neq j} \left[\hat{\mathbf{b}}_i^{\mathrm{T}} e_i^\star \frac{\pi_{ij}}{1-\pi_{jj}}\right] && \text{(iii)} \\
&+ \frac{1}{n_\omega - 1}\sum_j \hat{\mathbf{a}}_j \omega_j^\star e_j^\star \hat{\varepsilon}_j \left[\hat{\mathbf{b}}_j^{\mathrm{T}} (e_j^\star - 1) \frac{\pi_{jj}}{1-\pi_{jj}}\right]. && \text{(iv)}
\end{aligned}$$

Then we have (i) $\xrightarrow{\mathrm{p}} \mathbb{C}\mathrm{ov}[\bar{\boldsymbol{\Psi}}_1, \bar{\boldsymbol{\Psi}}_2 | \mathbf{Z}]$, and the other terms are asymptotically negligible. This essentially uses the same technique (conditional mean and variance calculation) used for Lemma II.6 and II.7, and we do not repeat here. By taking transpose, we have $(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{V})(\mathrm{II})^{\mathrm{T}} \xrightarrow{\mathrm{p}} \mathbb{C}\mathrm{ov}[\bar{\boldsymbol{\Psi}}_2, \bar{\boldsymbol{\Psi}}_1 | \mathbf{Z}]$. Further,

$$\begin{aligned}
\left|(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{II})(\mathrm{VI})^{\mathrm{T}}\right| &= \left|\frac{1}{n_\omega - 1}\sum_j \omega_j^\star \hat{\mathbf{a}}_j \sum_i (\omega_i^\star - \delta_{ij}) \hat{\mathbf{c}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)^2 \left(e_j^\star \hat{\varepsilon}_j\right)^2\right| \\
&\precsim_{\mathrm{p}} \frac{1}{n}\sum_{i,j} \pi_{ij}^2 = o_{\mathrm{p}}(1),
\end{aligned}$$

and

$$\left|(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{II})(\mathrm{VII})^{\mathrm{T}}\right| = \Big|\frac{2}{n_\omega-1}\sum_j \omega_j^\star e_j^\star \hat\varepsilon_j \hat{\mathbf{a}}_j \sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{c}}_i\left(\hat\mu_i^\star - \hat\mu_i\right)\left(\frac{\pi_{ij}}{1-\pi_{jj}}\right)\Big|$$

$$\lesssim_{\mathrm{p}} \frac{1}{n}\cdot \sqrt{\sum_j |\omega_j^\star e_j^\star\hat\varepsilon_j\hat{\mathbf{a}}_j|^2}\sqrt{\sum_j |(\omega_i^\star-\delta_{ij})\hat{\mathbf{c}}_i\left(\hat\mu_i^\star-\hat\mu_i\right)|^2}$$

$$= o_{\mathrm{p}}(1).$$

Due to uniform consistency of $\hat\mu_j^\star$, the following are easy to establish:

$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{III})^2 = o_{\mathrm{p}}(1),\ (n_\omega - 1)\sum_j \omega_j^\star (\mathrm{III})(\mathrm{IV})^{\mathrm{T}} = o_{\mathrm{p}}(1),$$
$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{III})(\mathrm{V})^{\mathrm{T}} = o_{\mathrm{p}}(1)$$
$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{III})(\mathrm{VI})^{\mathrm{T}} = o_{\mathrm{p}}(1),\ (n_\omega - 1)\sum_j \omega_j^\star (\mathrm{III})(\mathrm{VII})^{\mathrm{T}} = o_{\mathrm{p}}(1),$$

as well as

$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{IV})^2 = o_{\mathrm{p}}(1),\ (n_\omega - 1)\sum_j \omega_j^\star (\mathrm{IV})(\mathrm{V})^{\mathrm{T}} = o_{\mathrm{p}}(1),$$
$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{IV})(\mathrm{VI})^{\mathrm{T}} = o_{\mathrm{p}}(1),\ (n_\omega - 1)\sum_j \omega_j^\star (\mathrm{IV})(\mathrm{VII})^{\mathrm{T}} = o_{\mathrm{p}}(1).$$

Next it is easy to show that

$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{V})^2 \xrightarrow{\mathrm{p}} (1 + \mathbb{E}^\star[e_i^{\star 3}])\mathbb{V}[\bar{\mathbf{\Psi}}_2|\mathbf{Z}].$$

What remains are terms involving $(\mathrm{V})(\mathrm{VI})^{\mathrm{T}}$, $(\mathrm{V})(\mathrm{VII})^{\mathrm{T}}$, $(\mathrm{VI})^2$, $(\mathrm{VI})(\mathrm{VII})^{\mathrm{T}}$ and $(\mathrm{VII})^2$.

$$\left|(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{V})(\mathrm{VI})^{\mathrm{T}}\right|$$

$$= \Big|\frac{1}{n_\omega - 1}\sum_j \omega_j^\star \left(\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i\left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star\hat\varepsilon_j\right)\right)\left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j})\hat{\mathbf{c}}_\ell\left(\frac{\pi_{\ell j}}{1-\pi_{jj}}\right)^2\left(e_j^\star\hat\varepsilon_j\right)^2\right)^{\mathrm{T}}\Big|$$

$$\lesssim_{\mathrm{p}} \left(\frac{1}{n}\sum_j\left|\sum_\ell (\omega_\ell^\star - \delta_{\ell j})\hat{\mathbf{c}}_\ell\left(\frac{\pi_{\ell j}}{1-\pi_{jj}}\right)^2\right|^2\right)^{1/2} \asymp_p \sqrt{\frac{1}{n}\sum_{j,i,\ell}\pi_{ij}^2\pi_{\ell j}^2} = o_{\mathrm{p}}(1).$$

And

$$\begin{aligned}
&\Big|(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{V})(\mathrm{VII})^{\mathrm{T}}\Big| \\
&= \Big|\frac{2}{n_\omega - 1}\sum_j \left(\sum_i (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \left(\frac{\pi_{ij}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)\right)\left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j})\hat{\mathbf{c}}_\ell\left(\hat{\mu}_\ell^\star - \hat{\mu}_\ell\right)\left(\frac{\pi_{\ell j}}{1-\pi_{jj}} e_j^\star \hat{\varepsilon}_j\right)\right)^{\mathrm{T}}\Big| \\
&= \Big|\frac{2}{n_\omega - 1}\sum_{i,j} (\omega_i^\star - \delta_{ij})\hat{\mathbf{b}}_i \frac{\pi_{ij}}{1-\pi_{jj}} (e_j^\star \hat{\varepsilon}_j)^2 \left(\sum_\ell (\omega_\ell^\star - \delta_{\ell j})\hat{\mathbf{c}}_\ell\left(\hat{\mu}_\ell^\star - \hat{\mu}_\ell\right)\left(\frac{\pi_{\ell j}}{1-\pi_{jj}}\right)\right)^{\mathrm{T}}\Big| \\
&\lesssim_{\mathrm{p}} \sqrt{\frac{1}{n}\sum_j \left|\sum_\ell \frac{\pi_{\ell j}}{1-\pi_{\ell\ell}}\Big(\hat{\mu}_\ell^\star - \hat{\mu}_\ell\Big)\right|^2} = o_{\mathrm{p}}(1),
\end{aligned}$$

Using techniques in the above results, we can show

$$(n_\omega - 1)\sum_j \omega_j^\star (\mathrm{VI})^2 = o_{\mathrm{p}}(1),\ (n_\omega - 1)\sum_j \omega_j^\star (\mathrm{VII})^2 = o_{\mathrm{p}}(1),\ (n_\omega - 1)\sum_j \omega_j^\star (\mathrm{VI})(\mathrm{VII})^{\mathrm{T}} = o_{\mathrm{p}}(1),$$

which closes the proof. ■

CHAPTER III

Simple Local Polynomial Density Estimators

Abstract. *This chapter introduces an intuitive and easy-to-implement nonparametric density estimator based on local polynomial techniques. The estimator is fully boundary adaptive and automatic, but does not require pre-binning or any other transformation of the data. We study the main asymptotic properties of the estimator, and use these results to provide principled estimation, inference, and bandwidth selection methods. As a substantive application of our results, we develop a novel discontinuity in density testing procedure, an important problem in regression discontinuity designs and other program evaluation settings. An illustrative empirical application is provided. Two companion* `Stata` *and* `R` *software packages are provided.*

III.1 Introduction

Flexible (nonparametric) estimation of a probability density function features prominently in empirical work in statistics, economics, and many other disciplines. Sometimes the density function is the main object of interest, while in other cases it is a useful ingredient in forming other nonparametric or semiparametric procedures. In program evaluation and causal inference settings, for example, nonparametric density estimators are used for manipulation testing, distributional treatment effect and counterfactual analysis, instrumental variables treatment effect specification and heterogeneity analysis, and common support/overlap testing. See Imbens and Rubin (2015) and Abadie and Cattaneo (2018) for recent reviews and further references.

A common problem faced when implementing density estimators in empirical work is the presence of boundary evaluation points on the support of the variable of interest: whenever the density estimator is constructed at or near boundary points, which may or may not be known by the researcher, the finite- and large-sample properties of the estimator are

This chapter is based on the working paper "Simple Local Polynomial Density Estimators" (Cattaneo, Jansson and Ma, 2019b) affected. Standard kernel density estimators are invalid at or near boundary points, while other methods may remain valid but usually require choosing additional tuning parameters, transforming the data, a priori knowledge of the boundary point location, or some other boundary-related specific information or modification. Furthermore, it is usually the case that one type of density estimator is used for evaluation points at or near the boundary, while a different type is used for interior evaluation points.

We introduce a novel nonparametric estimator of a density function constructed using local polynomial techniques (Fan and Gijbels, 1996). The estimator is intuitive, easy to implement, does not require pre-binning of the data or a priori knowledge of the boundary location, and enjoys all the desirable features associated with local polynomial regression estimation. In particular, the estimator automatically adapts to the (possibly unknown) boundaries of the support of the density without requiring specific data modification or additional tuning parameter choices, a feature that is unavailable for most other density estimators in the literature: see Karunamuni and Albert (2005) for a review on this topic. The most closely related approaches currently available in the literature are the local polynomial density estimators of Cheng, Fan and Marron (1997) and Zhang and Karunamuni (1998), which require knowledge of the boundary location and pre-binning of the data (or, more generally, pre-estimation of the density near the boundary), and hence introduce additional tuning parameters that need to be chosen for implementation.

The heuristic idea underlying our estimator and differentiating the estimator from existing one is simple to explain: whereas other nonparametric density estimators are constructed by smoothing out a histogram-type estimator of the density, our estimator is constructed by smoothing out the empirical distribution function using local polynomial techniques. Accordingly, our density estimator is constructed using a preliminary tuning-parameter-free and $\sqrt{n}$-consistent distribution function estimator (where n denotes the sample size), implying in particular that the only tuning parameter required by our approach is the bandwidth associated with the local polynomial fit at each evaluation point. For the resulting density estimator, we establish (i) asymptotic expansions of the leading bias and variance, (ii) asymptotic Gaussian distributional approximation and valid statistical inference, (iii) consistent standard error estimates, and (iv) consistent data-driven bandwidth selection based on an asymptotic mean squared error (MSE) expansion. All these results apply to both interior and boundary points in a fully automatic and data-driven way, without requiring a prior knowledge of the boundary location, transforming the estimator or the data in specific ways, or employing additional tuning parameters (beyond the main bandwidth present in any kernel-based nonparametric method).

As a substantive methodological application of our proposed density estimator, we develop a novel discontinuity in density testing procedure. In a seminal paper, McCrary (2008) proposed the idea of manipulation testing via discontinuity in density testing for regression discontinuity (RD) designs, and developed an implementation thereof using the density estimator of Cheng, Fan and Marron (1997), which requires pre-binning of the data and choosing two tuning parameters. On the other hand, the new proposed discontinuity in density test employing our density estimator requires choosing only one tuning parameter, and enjoys other features associated with local polynomials methods. We also illustrate its performance with an empirical application employing the canonical Head Start data in the context of RD designs (Ludwig and Miller, 2007; Cattaneo, Titiunik and Vazquez-Bare, 2017). For recent practical introductions to RD methodology, and further references, see also Calonico, Cattaneo and Titiunik (2015), Cattaneo and Escanciano (2017), and Cattaneo, Idrobo and Titiunik (2018a,b).

Two general purpose software packages, for `Stata` and `R`, have been developed based on the main results discussed in the paper. Cattaneo, Jansson and Ma (2018c) discusses the package `rddensity`, which is specifically tailored to manipulation testing (i.e., two-sample discontinuity in density testing), while Cattaneo, Jansson and Ma (2019c) discusses the package `lpdensity`, which provides generic density estimation over the support of the data.

Section III.2 introduces the estimator and Section III.3 gives the main technical results. Bandwidth selection is discussed in Section III.4. Section III.5 applies these results to nonparametric testing of a discontinuity in a density at a boundary point (i.e., manipulation testing), while Section III.6 illustrates the new method with an empirical application. Section III.7 discusses extensions and concludes. Additional results, preliminary lemmas and proofs are collected in Section III.8 and III.9.

III.2 Boundary Adaptive Density Estimation

Suppose $\{x_1, x_2, \cdots, x_n\}$ is a random sample, where x_i is a continuous random variable with a smooth cumulative distribution function over its possibly unknown support $\mathcal{X} \subseteq \mathbb{R}$. The probability density function is $f(x) = F^{(1)}(x) = \frac{\partial}{\partial x}\mathbb{P}[x_i \leq x]$, where the derivative is interpreted as a one-sided derivative at a boundary point of $\mathcal{X}$, and F is the cumulative distribution function of x_i. Our results apply to known and unknown, as well as bounded or unbounded support $\mathcal{X}$, which is an important feature in most empirical applications employing density estimators. For example, in the context of manipulation testing (Section III.5), the random variable x_i is a running variable, score or index, and the parameter of interest is the potential discontinuity of the density function at an induced boundary point determined by the treatment eligibility cutoff.

Let $\tilde{F}(x) = n^{-1}\sum_{i=1}^{n} \mathbb{1}[x_i \le x]$ denote the classical empirical distribution estimator. Given $p \in \mathbb{N}$, our local polynomial distribution estimator is defined as

$$\hat{\boldsymbol{\beta}}_p(x) = \arg\min_{\mathbf{b}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \Big(\tilde{F}(x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\mathbf{b}\Big)^2 K\left(\frac{x_i - x}{h}\right),$$

where $\mathbf{r}_p(u) = [1, u, u^2, \cdots, u^p]$ is a (one-dimensional) polynomial expansion; K is a kernel function whose properties are to be specified later; $h = h_n$ is a bandwidth sequence. The estimator, $\hat{\boldsymbol{\beta}}_p(x)$, is motivated as a local Taylor series expansion, hence the target parameter is (i.e. the population counterpart, assuming exists)

$$\boldsymbol{\beta}_p(x) = \left[\frac{1}{0!}F(x),\ \frac{1}{1!}F^{(1)}(x),\ \cdots,\ \frac{1}{p!}F^{(p)}(x)\right]^{\mathrm{T}}.$$

Therefore, we also write[1]

$$\hat{\boldsymbol{\beta}}_p(x) = \left[\frac{1}{0!}\hat{F}_p(x),\ \frac{1}{1!}\hat{F}_p^{(1)}(x),\ \cdots,\ \frac{1}{p!}\hat{F}_p^{(p)}(x)\right]^{\mathrm{T}},$$

or equivalently, $\hat{F}_p^{(v)} = v!\mathbf{e}_v^{\mathrm{T}}\hat{\boldsymbol{\beta}}_p(x)$, provided that $v \le p$, and $\mathbf{e}_v$ is the $(v+1)$-th unit vector of $\mathbb{R}^{p+1}$. We also use $f = F^{(1)}$ to denote the corresponding probability density function for convenience. In other words, we take the empirical distribution function $\tilde{F}$ as the starting point, then construct a smooth local approximation to the distribution function using a polynomial expansion, and finally obtain the density estimator $\hat{f}_p$ as the slope coefficient in the local polynomial regression.

The idea behind the density estimator $\hat{f}_p(x)$ is explained graphically in Figure III.1. In this figure, we consider three distinct evaluation points on $\mathcal{X} = [-1, 1]$: a is near the lower boundary, b is an interior point, and $c = 1$ is the upper boundary. The conventional kernel

[1] The estimator has the following matrix form, which we will utilize:

$$\hat{\boldsymbol{\beta}}_p(x) = \mathbf{H}^{-1}\left(\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h\right)^{-1}\left(\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{Y}\right),$$

where

$$\mathbf{X}_h = \Big[\left(\frac{x_i - x}{h}\right)^j\Big]_{1\le i\le n,\ 0\le j\le p},$$

$\mathbf{K}_h$ is a diagonal matrix collecting $\{h^{-1}K((x_i - x)/h)\}_{1\le i\le n}$, and $\mathbf{Y}$ is a column vector collecting $\{\tilde{F}(x_i)\}_{1\le i\le n}$. We also use the convention $K_h(u) = h^{-1}K(u/h)$.

density estimator,

$$\hat{f}_{\mathtt{KD}}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right),$$

is valid for interior points, but otherwise inconsistent. See, e.g., Wand and Jones (1995) for a classical reference. On the other hand, our density estimator $\hat{f}_p(x)$ is valid for all evaluation points $x \in \mathcal{X}$ and can be used directly, without any modifications to approximate the unknown density. Figure III.1 is constructed using $n = 500$ observations. The top panel plots one realization of the empirical distribution function $\tilde{F}(x)$ in dark gray, and the local polynomial fits for the three evaluation points $x \in \{a, b, c\}$ in red, the latter implemented with $p = 2$ (quadratic approximation) and bandwidth h (different value for each evaluation point considered). The vertical light gray areas highlight the localization region controlled by the bandwidth choice, that is, only observations falling in these regions are used to smooth out the empirical distribution function via local polynomial approximation, depending on the evaluation point. The estimator $\hat{f}_p(x)$ is the slope coefficient accompanying the first-order term in the local polynomial approximation, which is depicted in the bottom panel of Figure III.1 as the solid line in red. The bottom panel also plots three other curves: dashed blue line corresponding to the population density function, dashed-dotted green line corresponding to the average of our density estimate over simulations, and dashed black line corresponding to average of the standard kernel density estimates obtained using $\hat{f}_{\mathtt{KD}}(x)$.

Figure III.1 illustrates how our proposed density estimator adapts to (near) boundary points automatically, showing graphically its good performance in repeated samples. Evaluation point b is an interior point and, consequently, a symmetric smoothing around that point is employed, just like the standard estimator $\hat{f}_{\mathtt{KD}}(x)$ does. On the other hand, evaluation points a and c both exhibit boundary bias if the standard kernel density estimator is used: point a is near the boundary and hence employs asymmetric smoothing, while point c is at the upper boundary and hence employs one-sided smoothing. In contrast, our proposed density estimator $\hat{f}_p(x)$ automatically adapts to the (possibly unknown) boundary point, as the bottom panel in Figure III.1 illustrates. This feature makes $\hat{f}_p(x)$ particularly well-suited for empirical applications where there is known or unknown finite boundaries on the support of the data.

III.3 Main Technical Results

We summarize three main large sample results concerning the proposed estimator: (i) an asymptotic distributional approximation with precise leading bias and variance characterizations, (ii) a consistent standard error estimator which is also data-driven and fully automatic, and (iii) bandwidth selection. We report additional theoretical results, preliminary lemmas and detailed proofs in Section III.8 and III.9 to conserve space.

We first give detailed assumptions supporting results, including preliminary lemmas and our main results. Other specific assumptions will be given in corresponding sections. Let $\mathcal{O}$ be a connected subset of $\mathbb{R}$ with nonempty interior, $\mathcal{C}^s(\mathcal{O})$ denotes functions that are at least s-times continuously differentiable in the interior of $\mathcal{O}$, and that the derivatives can be continuously extended to the boundary of $\mathcal{O}$.

Assumption III.1 (DGP)
$\{x_i\}_{1\leq i\leq n}$ is a random sample from distribution F, supported on $\mathcal{X} = [x_{\mathtt{L}}, x_{\mathtt{U}}]$. Further, $F \in \mathcal{C}^{\alpha_x}(\mathcal{X})$ for some $\alpha_x \geq 1$, and $f(x) = F^{(1)}(x) > 0$ for all $x \in \mathcal{X}$. ‖

This assumption imposes basic regularity conditions on the data generating process, ensuring that $f(x)$ is well-defined and possesses enough smoothness.

Assumption III.2 (Kernel)
The kernel function $K(\cdot)$ is nonnegative, symmetric, and belongs to $\mathcal{C}^0([-1,1])$. Further, it integrates to one: $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$. ‖

This assumption is standard in nonparametric estimation, and is satisfied for common kernel functions. We exclude kernels with unbounded support (e.g., Gaussian kernel) for simplicity, since such kernels will always hit boundaries. Our results, however, can be extended to accommodate unbounded support kernels, albeit more cumbersome notation would be needed.

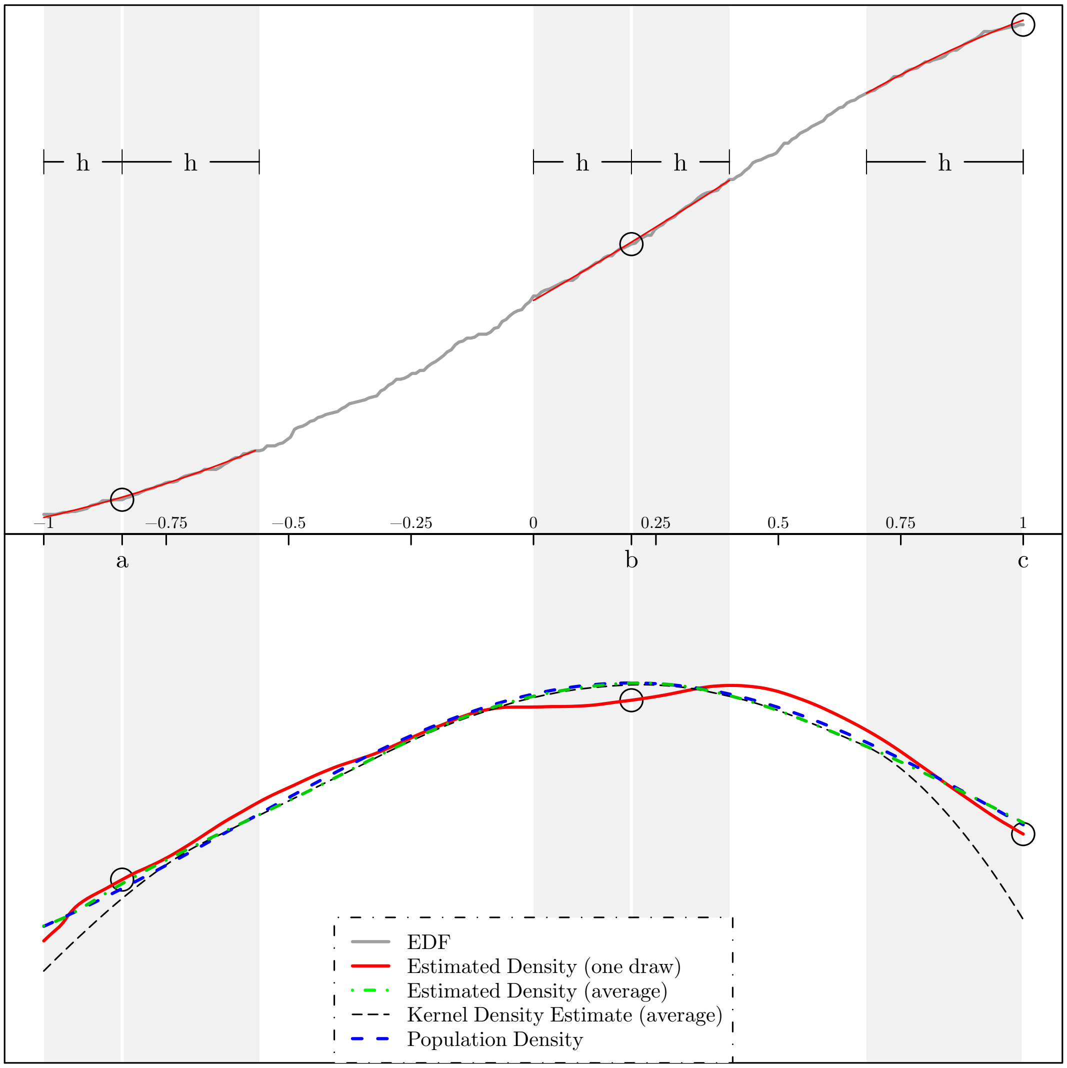
We also collect some matrices which will be used throughout this chapter. They show up in asymptotic results as components of bias and variance. Note that x can be either a fixed point, or it can be a drifting sequence to capture the issue of estimation and inference in boundary regions. For the latter, x takes the form $x = x_{\mathtt{L}} + ch$ or $x = x_{\mathtt{U}} - ch$ for some $c \in [0, 1)$.

$$\mathbf{S}_{p,x} = \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\mathbf{r}_p(u)^{\mathrm{T}} K(u)\mathrm{d}u,$$

$$\mathbf{c}_{p,x} = \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)u^{p+1} K(u)\mathrm{d}u, \quad \tilde{\mathbf{c}}_{p,x} = \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)u^{p+2} K(u)\mathrm{d}u,$$

$$\mathbf{\Gamma}_{p,x} = \iint_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} (u \wedge v)\mathbf{r}_p(u)\mathbf{r}_p(v) K(u)K(v)\mathrm{d}u\mathrm{d}v, \quad \mathbf{T}_{p,x} = \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\mathbf{r}_p(u)^{\mathrm{T}} K(u)^2\mathrm{d}u.$$

Figure III.1. Graphical illustration of density estimator.



Note. (i) Constructed using companion `R` (and `Stata`) package described in Cattaneo *et al.* (2019c) with simulated data.

Later we will assume the kernel function K being supported on $[-1, 1]$, hence with bandwidth $h \downarrow 0$, the region of integration in the above display can be replaced by

x	$(x_{\mathtt{L}} - x)/h$	$(x_{\mathtt{U}} - x)/h$
x interior	-1	$+1$
$x = x_{\mathtt{L}} + ch$ in lower boundary	$-c$	$+1$
$x = x_{\mathtt{U}} - ch$ in upper boundary	-1	$+c$

Since we do not allow $x_{\mathtt{L}} = x_{\mathtt{U}}$, no drifting sequence x can be in both boundary regions, at least asymptotically.

The following theorem gives a characterization of the asymptotic bias and variance of our estimator, as well as a valid distributional approximation.

Theorem III.1 (Asymptotic Normality)
Assume Assumptions III.1 and III.2 hold with $\alpha_x \geq p+1$ for some integer $p \geq 0$. Further $h \to 0$, $nh^2 \to \infty$ and $nh^{2p+1} = O(1)$. Then

$$\sqrt{nh^{2v-1}}\Big(\hat{F}_p^{(v)}(x) - F^{(v)}(x) - h^{p+1-v}\mathcal{B}_{p,v}(x)\Big) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0,\ \mathcal{V}_{p,v}(x)\Big), \qquad 1 \leq v \leq p,$$
$$\sqrt{\frac{n}{\mathcal{V}_{p,0}(x)}}\Big(\hat{F}_p(x) - F(x) - h^{p+1}\mathcal{B}_{p,0}(x)\Big) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0,\ 1\Big).$$

The constants are

$$\mathcal{B}_{p,v}(x) = v!\frac{F^{(p+1)}(x)}{(p+1)!}\mathbf{e}_v^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x},$$

and

$$\mathcal{V}_{p,v}(x) = \begin{cases} (v!)^2 f(x)\mathbf{e}_v^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v & 1 \leq v \leq p \\ F(x)(1-F(x)) & v = 0,\ x \text{ interior} \\ hf(x)\left(\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_0 + c\right) & v = 0,\ x = x_{\mathtt{L}} + ch \text{ or } x_{\mathtt{U}} - ch. \end{cases} \qquad \|$$

In this theorem, the integration region reflects the effect of boundaries. Because $K(\cdot)$ is compactly supported, if x is an interior point, we have $h^{-1}(\mathcal{X} - x) \supset [-1, 1]$ for h small enough, thus ensuring the kernel function is not truncated and the local approximation is symmetric around x. On the other hand, for x near or at a boundary of $\mathcal{X}$ (i.e., for h not small enough relative to the distance of x to the boundary), we have $h^{-1}(\mathcal{X} - x) \not\supset [-1, 1]$, and the local approximation is asymmetric (or one-sided). It follows that the density estimator $\hat{f}(x)$ is boundary adaptive and design adaptive, as in the case of local polynomial regression (Fan and Gijbels, 1996).

Remark III.1 (On $nh^{2p+1} = O(1)$) This condition ensures that higher order bias, after scaling, is asymptotically negligible. $\|$

Remark III.2 (On $nh^2 \to \infty$) This condition ensures that a second-order remainder term, which turns out to take a U-statistic form, has smaller order compared to the leading term. This second order remainder term arises as our estimator involves a double summation: one is used to construct the empirical distribution function $\tilde{F}$, and the other comes from the local polynomial smoothing step. Note that this condition can be dropped for boundary x or when the parameter of interest is the distribution function $\hat{F}_p$. $\|$

Remark III.3 (On $\mathcal{V}_{p,0}(x)$) It may seem that the variance formula has a discontinuity in x for the smoothed empirical distribution function (i.e. $v = 0$), when x switches from interior to boundary. This phenomenon, however, is purely an artifact of employing different asymptotic frameworks. To see this, assume $x_{\mathtt{L}} = 0$ and $x_{\mathtt{U}} = 1$, and for some sample the bandwidth $h = 0.2$ is used. Given our convention, the point $x = 0.3$ is not a boundary point, hence we should consider $\sqrt{n}$ as the correct scaling for $\hat{F}_p(0.3)$. On the other hand, one can also consider 0.3 as part of the asymptotic sequence $x = 1.5h$, in which case one promises to move the evaluation point closer to the lower boundary as sample size increases. Then despite the fact that such x is not a boundary point, $\hat{F}_p(x)$ is still an estimator of zero, which means it is super consistent and the correct scaling is $\sqrt{n/h}$.

This discussion also applies to the usual empirical distribution function $\tilde{F}(x)$. Such phenomenon, however, does not occur for other components of $\hat{\boldsymbol{\beta}}_p(x)$, for which the evaluation point only affects the exact form of multiplicative constants, but not the rate of convergence. ‖

Now we consider the problem of variance estimation. Given the formula in Theorem III.1, it is possible to estimate the asymptotic variance by "plug-in" unknown quantities regarding the data generating process. For example consider $\mathcal{V}_{p,1}(x)$ for the estimated density. Assume the researcher knows the location of the boundary $x_{\mathtt{L}}$ and $x_{\mathtt{U}}$, the matrices $\mathbf{S}_{p,x}$ and $\mathbf{\Gamma}_{p,x}$ can be constructed with numerical integration, since they are related to features of the kernel function, not the data generating process. The unknown density $f(x)$ can also be replaced by its estimate, as long as $p \geq 1$.

Another approach is to estimate the unknown quantities in an "automatic" way. To introduce our variance estimator, we make the following definitions.

$$\hat{\mathbf{S}}_{p,x} = \frac{1}{n}\mathbf{X}_h\mathbf{K}_h\mathbf{X}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_p\left(\frac{x_i - x}{h}\right)\mathbf{r}_p\left(\frac{x_i - x}{h}\right)^{\mathrm{T}} K_h(x_i - x)$$

$$\hat{\mathbf{\Gamma}}_{p,x} = \frac{1}{n^3}\sum_{i,j,k=1}^{n}\mathbf{r}_p\left(\frac{x_j - x}{h}\right)\mathbf{r}_p\left(\frac{x_k - x}{h}\right)^{\mathrm{T}} K_h(x_j - x)K_h(x_k - x)$$
$$\Big(\mathbb{1}[x_i \leq x_j] - \tilde{F}(x_j)\Big)\Big(\mathbb{1}[x_i \leq x_k] - \tilde{F}(x_k)\Big).$$

Following is the main result regarding variance estimation. It is automatic and fully-adaptive, in the sense that no knowledge about the boundary location is needed.

Theorem III.2 (Variance Estimation)
Assume Assumptions III.1 and III.2 hold with $\alpha_x \geq p + 1$ for some integer $p \geq 0$. Further

$h \to 0$, $nh^2 \to \infty$ and $nh^{2p+1} = O(1)$. Then

$$\hat{\mathcal{V}}_{p,v}(x) \equiv (v!)^2 \mathbf{e}_v^{\mathrm{T}} \mathbf{N}_x \hat{\mathbf{S}}_{p,x}^{-1} \hat{\mathbf{\Gamma}}_{p,x} \hat{\mathbf{S}}_{p,x}^{-1} \mathbf{N}_x \mathbf{e}_v \ \overset{\mathrm{p}}{\to}\ \mathcal{V}_{p,v}(x).$$

Define the standard error as

$$\hat{\sigma}_{p,v}(x) \equiv (v!)\sqrt{\frac{1}{nh^{2v}} \mathbf{e}_v^{\mathrm{T}} \hat{\mathbf{S}}_{p,x}^{-1} \hat{\mathbf{\Gamma}}_{p,x} \hat{\mathbf{S}}_{p,x}^{-1} \mathbf{e}_v},$$

then

$$\hat{\sigma}_{p,v}(x)^{-1}\Big(\hat{F}_p^{(v)}(x) - F^{(v)}(x) - h^{p+1-v}\mathcal{B}_{p,v}(x)\Big) \overset{\mathrm{d}}{\to} \mathcal{N}\Big(0,\ 1\Big). \qquad \|$$

Although constructing $\hat{\mathcal{V}}_{p,v}(x)$ requires the knowledge of the location of boundaries, it is not needed for inference. This is why the standard error, $\hat{\sigma}_{p,v}(x)$, is automatic and fully-adaptive. In addition, although we have to split the definition of $\mathcal{V}_{p,v}(x)$ by different v and x, $\hat{\sigma}_{p,v}(x)$ automatically adapts to these different scenarios, and hence it provides a unified approach for variance estimation/inference.

Finally, we recommend implementing the density estimator $\hat{f}_p(x)$ with $p = 2$ (and generally implementing $F_p^{(v)}(x)$ with $p = v+1$), which corresponds to the minimal odd polynomial order choice (i.e., equivalent to local-linear local polynomial regression). Higher-order local polynomials could be used, but they typically exhibit erratic behavior near boundary points (usually known as the Runge's Phenomenon, see Calonico, Cattaneo and Titiunik, 2015, pp. 1756-1757), and lead to counter-intuitive weighting schemes (Gelman and Imbens, 2018). See Fan and Gijbels (1996) for further discussion and automatic polynomial order selection methods that can be applied to our estimator as well.

III.4 Bandwidth Selection

In this section we consider the problem of constructing MSE-optimal bandwidth for our local polynomial regression-based distribution estimators. We focus exclusively on the case $v \geq 1$, hence the object of interest will be either the density function or derivatives thereof. Valid bandwidth choice for the distribution function $\hat{F}_p(x)$ is also an interesting topic, but difficulty arises since it is estimated with (at least) parametric rate. We will briefly mention MSE expansion of the estimated distribution function at the end.

III.4.1 For Density and Derivatives Estimates ($v \geq 1$)

Consider some $1 \leq v \leq p$, the following lemma gives finer characterization of the bias.

Lemma III.1 (Bias)
Assume Assumptions III.1 and III.2 hold with $\alpha_x \geq p+2$, $h \to 0$ and $nh^3 \to \infty$. Then the leading bias of $\hat{F}_p^{(v)}(x)$ is

$$h^{p+1-v}\mathcal{B}_{p,v}(x) = h^{p+1-v}\Big\{\frac{F^{(p+1)}(x)}{(p+1)!}v!\mathbf{e}_v^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x} + h\left(\frac{F^{(p+2)}(x)}{(p+2)!} + \frac{F^{(p+1)}(x)}{(p+1)!}\frac{F^{(2)}(x)}{f(x)}\right)v!\mathbf{e}_v^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}\Big\}. \qquad \|$$

The above lemma characterizes the higher-order bias. To see its necessity, we note that when $p-v$ is even and x is an interior evaluation point, the leading bias is zero. This is because $\mathbf{e}_v^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x}$ is zero, which is explained in Fan and Gijbels (1996). Except for rare cases such as $F^{(p+1)}(x) = 0$ or $F^{(p+2)}(x) = 0$, we have

Order of bias: $h^{p+1-v}\mathcal{B}_{p,v}(x) \asymp$

	$p-v$ odd	even
x interior	h^{p+1-v}	h^{p+2-v}
boundary	h^{p+1-v}	h^{p+1-v}

Note that for boundary evaluation points, the leading bias never vanishes.

The leading variance is also characterized by Theorem III.1, and we reproduce it here:

$$\frac{1}{nh^{2v-1}}\mathcal{V}_{p,v}(x) = \frac{1}{nh^{2v-1}}(v!)^2 f(x)\mathbf{e}_v^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\boldsymbol{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v.$$

The MSE-optimal bandwidth is defined as a minimizer of the following

$$h_{p,v}(x) = \arg\min_{h>0}\left[\frac{1}{nh^{2v-1}}\mathcal{V}_{p,v}(x) + h^{2p+2-2v}\mathcal{B}_{p,v}(x)^2\right].$$

Given the discussion we had earlier on the bias, it is easy to see that the MSE-optimal bandwidth has the following asymptotic order:

Order of MSE-optimal bandwidth: $h_{p,v}(x) \asymp$

	$p-v$ odd	even
x interior	$n^{-\frac{1}{2p+1}}$	$n^{-\frac{1}{2p+3}}$
boundary	$n^{-\frac{1}{2p+1}}$	$n^{-\frac{1}{2p+1}}$

Again only the case where $p-v$ is even and x is interior needs special attention.

There are two notions of bandwidth consistency. Let h be some non-stochastic bandwidth sequence, and $\hat{h}$ be an estimated bandwidth. Then $\hat{h}$ is consistent *in rate* if $\hat{h} \asymp_{\mathrm{p}} h$ (in most cases it is even true that $\hat{h}/h \overset{\mathrm{p}}{\to} C \in (0,\infty)$). And $\hat{h}$ is consistent *in rate and constant* if $\hat{h}/h \overset{\mathrm{p}}{\to} 1$.

To construct consistent bandwidth, either rate consistent or consistent in both rate and constant, we need estimates of both the bias and variance. The variance part is relatively easy, as we have already demonstrated in Theorem III.2:

$$n\ell^{2v-1}\frac{\hat{\sigma}_{p,v}(x)^2}{\mathcal{V}_{p,v}(x)} \overset{\mathrm{p}}{\to} 1,$$

where ℓ is some preliminary bandwidth used to construct $\hat{\sigma}_{p,v}(x)$.

For the bias, there are two approaches. The first one is more common in the literature, where one distinguishes between boundary and interior cases, and propose consistent bias estimators separately. This method is appealing in the sense that the bandwidth constructed will be consistent both in rate and constant. The drawback, however, is that it requires the precise knowledge of the location of x relative to the boundaries, which is not always obvious.

We will follow the second approach, where we replace the unknown bias by an estimate which is consistent in rate (but not necessarily in constant). To be precise, our bias estimator will be consistent in rate and constant if either x is boundary or $p-v$ is odd, and will be consistent in rate otherwise. This bias estimator has an appealing feature: it is purely data-driven and no precise knowledge about the positioning of x relative to the boundaries is needed, with the price that it (and the bandwidth constructed thereof) is not consistent in constant when x is interior and $p-v$ is even.

To introduce this approach, first assume there are consistent estimators for $F^{(p+1)}(x)$ and $F^{(p+2)}(x)$, denoted by $\hat{F}^{(p+1)}(x)$ and $\hat{F}^{(p+2)}(x)$. They can be obtained, for example, using our local polynomial regression-based approach, or can be constructed with some reference model (such as the normal distribution). The critical step is to obtain consistent estimators of the matrices, which are given in the following lemma.

Lemma III.2

Assume Assumptions III.1 and III.2 hold, $\ell \to 0$ and $n\ell \to \infty$. Then

$$\widehat{\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x}} = \left(\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{\ell}\right)\mathbf{r}_p\left(\frac{x_i - x}{\ell}\right)^{\mathrm{T}} K_\ell(x_i - x)\right)^{-1}$$
$$\left(\frac{1}{n}\sum_i \left(\frac{x_i - x}{\ell}\right)^{p+1}\mathbf{r}_p\left(\frac{x_i - x}{\ell}\right) K_\ell(x_i - x)\right) \xrightarrow{\mathrm{p}} \mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x},$$

and

$$\widehat{\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}} = \left(\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{\ell}\right)\mathbf{r}_p\left(\frac{x_i - x}{\ell}\right)^{\mathrm{T}} K_\ell(x_i - x)\right)^{-1}$$
$$\left(\frac{1}{n}\sum_i \left(\frac{x_i - x}{\ell}\right)^{p+2}\mathbf{r}_p\left(\frac{x_i - x}{\ell}\right) K_\ell(x_i - x)\right) \xrightarrow{\mathrm{p}} \mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}. \qquad \|$$

Note that we used different notation, ℓ, for bandwidth.

Now we have enough ingredients for bandwidth selection. Define:

$$h^{p+1-v}\hat{\mathcal{B}}_{p,v}(x) = h^{p+1-v}\left\{\frac{\hat{F}^{(p+1)}(x)}{(p+1)!}v!\mathbf{e}_v^{\mathrm{T}}\widehat{\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x}} + h\frac{\hat{F}^{(p+2)}(x)}{(p+2)!}v!\mathbf{e}_v^{\mathrm{T}}\widehat{\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}}\right\},$$

and assume that $\hat{\sigma}_{p,v}(x)$ is constructed using the preliminary bandwidth ℓ. Then

$$\hat{h}_{p,v}(x) = \arg\min_{h>0}\left[\frac{\ell^{2v-1}}{h^{2v-1}}\hat{\sigma}_{p,v}(x)^2 + h^{2p+2-2v}\hat{\mathcal{B}}_{p,v}(x)^2\right].$$

We make three remarks here.

Remark III.4 (Preliminary bandwidth ℓ) The optimization argument h enters the RHS of the previous display in three places. First it is part of the variance component, by $1/h^{2v-1}$. Second it shows as a multiplicative factor of the bias component, $h^{2p-2v+2}$. Finally within the definition of $\hat{\mathcal{B}}_{p,v}(x)$, there is another multiplicative h, in front of the higher order bias.

The preliminary bandwidth ℓ, serves a different role. It is used to estimate the variance and bias components. Of course one can use different preliminary bandwidths for $\hat{\sigma}_{p,v}(x)$, $\widehat{\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x}}$ and $\widehat{\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}}$. $\|$

Remark III.5 (Known boundaries) If boundary locations are known, either from *a priori* knowledge or suggested by the data, then it is possible to simplify the problem, and closed-form solution for $\hat{h}_{p,v}(x)$ is feasible. To be precise, if it is known that x is a boundary point *or* $p - v$ is odd, one can simply ignore the second component in $\hat{\mathcal{B}}_{p,v}(x)$. Similarly, if it is the case that x is interior and $p-v$ is even, then the first component in $\hat{\mathcal{B}}_{p,v}(x)$ can be dropped.

The option we opt-for is more flexible in the sense that it adapts to any $p-v$ (odd or even) and any x (interior or boundary). ‖

Remark III.6 (Consistent bias estimator) The bias estimator we proposed, $h^{p-v+1}\hat{\mathcal{B}}_{p,v}(x)$, is consistent in rate for the true leading bias, but not necessarily in constant. Compare $\hat{\mathcal{B}}_{p,v}(x)$ and $\mathcal{B}_{p,v}(x)$, it is easily seen that the term involving $F^{(p+1)}(x)F^{(2)}(x)/f(x)$ is not captured. To capture this term, we need one additional nonparametric estimator for $F^{(2)}(x)$. This is indeed feasible, and one can employ our local polynomial regression-based estimator for this purpose. ‖

Theorem III.3 (Consistent bandwidth)
Let $1 \leq v \leq p$. Assume the preliminary bandwidth ℓ is chosen such that $nh^{2v-1}\hat{\sigma}_{p,v}(x)^2/\mathcal{V}_{p,v}(x)$ $\xrightarrow{\text{p}} 1$, $\widehat{\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x}} \xrightarrow{\text{p}} \mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x}$, and $\widehat{\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}} \xrightarrow{\text{p}} \mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}$. Under the conditions of Lemma III.3 (in Section III.8) and Theorem III.2:

- *If either x is in boundary regions or $p-v$ is odd, let $\hat{F}^{(p+1)}(x)$ be consistent for $F^{(p+1)} \neq 0$. Then*

$$\frac{\hat{h}_{p,v}(x)}{h_{p,v}(x)} \xrightarrow{\text{p}} 1.$$

- *If x is in interior and $p-v$ is even, let $\hat{F}^{(p+2)}(x)$ be consistent for $F^{(p+2)} \neq 0$. Further assume $nh^3 \to 0$ and $h_{p,v}(x)$ is well-defined. Then*

$$\frac{\hat{h}_{p,v}(x)}{h_{p,v}(x)} \xrightarrow{\text{p}} C \in (0,\infty).$$

‖

III.4.2 For Distribution Function Estimate ($v = 0$)

In this subsection we mention briefly how to choose bandwidth for the distribution function estimate, $\hat{F}_p^{(0)}(x) \equiv \hat{F}_p(x)$. We assume x is in interior. Previous discussions on bias remains to apply:

$$h^{p+1}\mathcal{B}_{p,0}(x) = h^{p+1}\left\{\frac{F^{(p+1)}(x)}{(p+1)!}\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x} + h\left(\frac{F^{(p+2)}(x)}{(p+2)!} + \frac{F^{(p+1)}(x)}{(p+1)!}\frac{F^{(2)}(x)}{f(x)}\right)\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{c}}_{p,x}\right\},$$

which means the bias of $\hat{F}_p(x)$ has order h^{p+1} if either x is boundary or p is odd, and h^{p+2} otherwise. Difficulty arises since the distribution function estimator has leading variance of order

$$\mathcal{V}_{p,0}(x) \asymp \frac{\mathbb{1}[x \text{ interior}] + h}{n},$$

which cannot be used for bandwidth selection, because the above is proportional to the bandwidth (i.e., there is no bias-variance trade-off).

The trick is to use a higher order variance term. In Section III.8 we show that the local polynomial regression-based estimator is essentially a second order U-statistic, which is then decomposed into two terms, a linear term $\hat{\mathbf{L}}$ and a quadratic term $\hat{\mathbf{R}}$, where the latter is a degenerate second-order U-statistic. The variance of the quadratic term $\hat{\mathbf{R}}$ has been ignored so far, as it is negligible compared to the variance of the linear term. For the distribution function estimator, however, it is the variance of this quadratic term that leads to a bias-variance trade-off. The exact form of this variance is given in Lemma III.6 in Section III.8. With this additional variance term included, we have (with some abuse of notation)

$$\mathcal{V}_{p,0}(x) \asymp \frac{\mathbb{1}[x \text{ interior}] + h}{n} + \frac{\mathbb{1}[x \text{ interior}] + h}{n^2 h}.$$

Provided x is an interior point, the additional variance term increases as the bandwidth shrinks. As a result, a MSE-optimal bandwidth for $\hat{F}_p(x)$ is well-defined, and estimating this bandwidth is also straightforward.

Order of MSE-optimal bandwidth: $h_{p,0}(x) \asymp$

	$p - v$ odd	even
x interior	$n^{-\frac{2}{2p+3}}$	$n^{-\frac{2}{2p+5}}$
boundary	undefined	undefined

What if x is in a boundary region? Then the MSE-optimal bandwidth for $\hat{F}_p(x)$ is not well defined. The leading variance now takes the form $h/n + 1/n^2$, which is proportional to the bandwidth. (This is not surprising, since for boundary x the distribution function is known, and a very small bandwidth gives a super-consistent estimator.). Although MSE-optimal bandwidth for $\hat{F}_p(x)$ is not well-defined for boundary x, it is still feasible to minimize the empirical MSE. To see how this works, one first estimate the bias term and variance term with some preliminary bandwidth ℓ, leading to $\hat{\mathcal{B}}_{p,0}(x)$ and $\hat{\mathcal{V}}_{p,0}(x)$. Then the MSE-optimal bandwidth can be constructed by minimizing the empirical MSE. Under regularity conditions, $\hat{\mathcal{B}}_{p,0}(x)$ will converge to some nonzero constant, while, if x is boundary, $\hat{\mathcal{V}}_{p,0}(x)$ has order ℓ, the same as the preliminary bandwidth. Then the MSE-optimal bandwidth constructed in this way will have the following order:

Order of estimated MSE-optimal bandwidth: $\hat{h}_{p,0}(x) \asymp$

	$p-v$ odd	even
x interior	$n^{-\frac{2}{2p+3}}$	$n^{-\frac{2}{2p+5}}$
boundary	$(n^2/\ell)^{-\frac{1}{2p+3}}$	$(n^2/\ell)^{-\frac{1}{2p+5}}$

Note that the preliminary bandwidth enters the rate of $\hat{h}_{p,0}(x)$ for boundary x, because it determines the rate at which the variance estimator $\hat{\mathcal{V}}_{p,0}(x)$ vanishes. Although this estimated bandwidth is not consistent for any well-defined object, it can be useful in practice, and it reflects the fact that for boundary x it is appropriate to use bandwidth shrinks fast when the object of interest is the distribution function

III.5 Application to Manipulation Testing

Testing for manipulation is useful when units are assigned to two (or more) distinct groups using a hard-thresholding rule based on an observable variable, as it provides an intuitive and simple method to check empirically whether units are able to alter (i.e., manipulate) their assignment. Manipulation tests are used in empirical work both as falsification tests of RD designs and as empirical tests with substantive implications in other program evaluation settings.

Available implementations require choosing multiple tuning parameters (McCrary, 2008), or employ empirical likelihood methods together with boundary-corrected kernels (Otsu, Xu and Matsushita, 2014). In contrast, our proposed method requires choosing only one tuning parameter, avoids pre-binning the data, and permits the use of simple well-known weighting schemes (e.g., uniform or triangular kernel), thereby avoiding the need of choosing the length and positions of bins or of employing more complicated boundary kernels. In addition, our method is intuitive, easy-to-implement, and fully data-driven and principled: bandwidth selection methods are formally developed and implemented, along with valid inference methods based on robust bias correction.

To describe the manipulation testing setup, suppose units are assigned to one group ("control") if $x_i < \bar{x}$ and to another group ("treatment") if $x_i \geq \bar{x}$. For example, in the application discussed below we employ the Head Start data, where x_i is a poverty index at the county level, $\bar{x} = 59.1984$ is a fixed cutoff determining eligibility to the program (see panel (a) in Figure III.2 below). The goal is to test formally whether the density $f(x)$ is continuous at $\bar{x}$, using the two subsamples $\{x_i : x_i < \bar{x}\}$ and $\{x_i : x_i \geq \bar{x}\}$, and thus the null and alternative hypotheses are:

$$\mathsf{H}_0 : \lim_{x\uparrow\bar{x}} f(x) = \lim_{x\downarrow\bar{x}} f(x) \qquad \text{vs} \qquad \mathsf{H}_1 : \lim_{x\uparrow\bar{x}} f(x) \neq \lim_{x\downarrow\bar{x}} f(x).$$

This hypothesis testing problem induces a nonparametric boundary point at $x = \bar{x}$ because two distinct densities need to be estimated, one from the left and the other from the right. Our proposed density estimator $\hat{f}_p(x)$ is readily applicable because it is boundary adaptive and fully automatic, and it can also be used to plot the density near the cutoff in an automatic way. See panel (b) of Figure III.2 below for an example using the Head Start data.

To start, consider the following polynomial basis $\mathbf{r}_p$

$$\mathbf{r}_p(u) = \begin{bmatrix} \mathbf{1}_{\{u<0\}} & u\mathbf{1}_{\{u<0\}} & \cdots & u^p\mathbf{1}_{\{u<0\}} & \Big| & \mathbf{1}_{\{u\geq 0\}} & u\mathbf{1}_{\{u\geq 0\}} & \cdots & u^p\mathbf{1}_{\{u\geq 0\}} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{2p+2}.$$

The following two vectors will arise later, which we give the definition here:

$$\mathbf{r}_{-,p}(u) = \begin{bmatrix} 1 & u & \cdots & u^p & 0 & \cdots & 0 \end{bmatrix}^{\mathrm{T}}, \quad \mathbf{r}_{+,p}(u) = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & \cdots & u^p \end{bmatrix}^{\mathrm{T}}.$$

Also we define the vectors to extract the corresponding derivatives

$$\mathbf{I}_{2p+2} = \begin{bmatrix} \mathbf{e}_{0,-} & \mathbf{e}_{1,-} & \cdots & \mathbf{e}_{p,-} & \mathbf{e}_{0,+} & \mathbf{e}_{1,+} & \cdots & \mathbf{e}_{p,+} \end{bmatrix}.$$

With the above definition, the estimator at the cutoff is[2]

$$\hat{\boldsymbol{\beta}}_p(\bar{x}) = \arg\min_{\mathbf{b}\in\mathbb{R}^{2p+2}} \sum_i \Big(\tilde{F}(x_i) - \mathbf{r}_p(x_i - \bar{x})^{\mathrm{T}}\mathbf{b} \Big)^2 K_h(x_i - \bar{x}).$$

We assume the same bandwidth is used below and above the cutoff to avoid cumbersome notation. Generalizing to using different bandwidths is straightforward. Other notations (for example $\mathbf{X}$ and $\mathbf{X}_h$) are redefined similarly, with the scaling matrix $\mathbf{H}$ adjusted so that $\mathbf{H}^{-1}\mathbf{r}_p(u) = \mathbf{r}_p(h^{-1}u)$ is always true. we denote the estimates by

$$\hat{F}^{(v)}_p(\bar{x}-) = v!\mathbf{e}_{v,-}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_p(\bar{x}), \qquad \hat{F}^{(v)}_p(\bar{x}+) = v!\mathbf{e}_{v,+}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_p(\bar{x}).$$

Now we state the main result concerning the manipulation testing. Let $\hat{\mathbf{S}}_{p,\bar{x}}$ and $\hat{\mathbf{\Gamma}}_{p,\bar{x}}$ be constructed as in Section III.3, and

$$\hat{\mathcal{V}}_{p,1}(\bar{x}) = \frac{1}{h}(\mathbf{e}_{1,+} - \mathbf{e}_{1,-})^{\mathrm{T}}\hat{\mathbf{S}}_{p,\bar{x}}\hat{\mathbf{\Gamma}}_{p,\bar{x}}\hat{\mathbf{S}}_{p,\bar{x}}(\mathbf{e}_{1,+} - \mathbf{e}_{1,-}).$$

Corollary III.1 (Manipulation testing)
Assume Assumptions III.1 and III.2 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$ with $\alpha_x \geq p+1$ for some

[2] The empirical distribution function is defined with the whole sample as before: $\tilde{F}(u) = n^{-1}\sum_i \mathbb{1}[x_i \leq u]$.

integer $p \geq 1$. Further, $n \cdot h^2 \to \infty$ and $n \cdot h^{2p+1} \to 0$. Then under the null hypothesis $\mathsf{H}_0 : f(\bar{x}+) = f(\bar{x}-)$,

$$T_p(h) = \frac{\hat{f}_p(\bar{x}+) - \hat{f}_p(\bar{x}-)}{\sqrt{\frac{1}{nh}\hat{\mathcal{V}}_{p,1}(\bar{x})}} \rightsquigarrow \mathcal{N}(0,1).$$

As a result, under the alternative hypothesis $\mathsf{H}_1 : f(\bar{x}+) \neq f(\bar{x}-)$,

$$\lim_{n\to\infty} \mathbb{P}[|T_p(h)| \geq \Phi_{1-\alpha/2}] = 1.$$

Here $\Phi_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. ‖

Remark III.7 (Separate estimation) An alternative implementation is to apply our local polynomial-based estimator separately to the two samples, one with observations below the cutoff, and the other with observations above the cutoff. To be precise, let $\tilde{F}_-(\cdot)$ and $\tilde{F}_+(\cdot)$ be the empirical distribution functions constructed by the two samples. That is,

$$\tilde{F}_-(x) = \frac{1}{n_-}\sum_{i:\ x_i<\bar{x}} \mathbb{1}[x_i \leq x], \qquad \tilde{F}_+(x) = \frac{1}{n_+}\sum_{i:\ x_i\geq\bar{x}} \mathbb{1}[x_i \leq x],$$

where n_- and n_+ denote the size of the two samples, respectively. The the local polynomial approach, applied to $\tilde{F}_-(\cdot)$ and $\tilde{F}_+(\cdot)$ separately, will yield two sets of estimates, which we denote by $\hat{F}^{(v)}_{p,-}(\bar{x})$ and $\hat{F}^{(v)}_{p,+}(\bar{x})$. To see the relation between joint and separate estimations, we note the following (which can be easily seen using least squares algebra)

$$\begin{array}{lll} v=0 & \hat{F}_{p,-}(\bar{x}) = \dfrac{n}{n_-}\hat{F}_p(\bar{x}-), & \hat{F}_{p,+}(\bar{x}) = \dfrac{n}{n_+}\hat{F}_p(\bar{x}+) - \dfrac{n_-}{n_+} \\ v\geq 1 & \hat{F}^{(v)}_{p,-}(\bar{x}) = \dfrac{n}{n_-}\hat{F}^{(v)}_p(\bar{x}-), & \hat{F}^{(v)}_{p,+}(\bar{x}) = \dfrac{n}{n_+}\hat{F}^{(v)}_p(\bar{x}+). \end{array}$$

The difference comes from the fact that by separate estimation, one obtains estimates of the conditional distribution function and the derivatives.

For manipulation testing, let $\hat{f}_{p,-}(\bar{x})$ and $\hat{f}_{p,+}(\bar{x})$ be the two density estimates, and $\hat{\mathcal{V}}_{p,1,-}(\bar{x})$ and $\hat{\mathcal{V}}_{p,1,+}(\bar{x})$ be the associated variance estimates. Then the test statistic is equivalently:

$$T_p(h) = \frac{\frac{n_+}{n}\hat{f}_{p,+}(\bar{x}) - \frac{n_-}{n}\hat{f}_{p,-}(\bar{x})}{\sqrt{\frac{1}{nh}\left(\frac{n_+}{n}\hat{\mathcal{V}}_{p,1,+}(\bar{x}) + \frac{n_-}{n}\hat{\mathcal{V}}_{p,1,-}(\bar{x})\right)}}.$$ ‖

A key implementation issue of our manipulation test is the choice of bandwidth h, a problem common to all nonparametric manipulation tests available in the literature. To select h in an automatic and data-driven way, we obtain an approximate MSE-optimal bandwidth choice for the point estimator $\hat{f}_p(\bar{x}+) - \hat{f}_p(\bar{x}-)$, and then propose a consistent implementation thereof, which is denoted by $\hat{h}_p$. Given the data-driven bandwidth choice $\hat{h}_p$, or its theoretical (infeasible) counterpart h_p, we propose a simple robust bias-corrected test statistic implementation following ideas in Calonico, Cattaneo and Titiunik (2014) and Calonico, Cattaneo and Farrell (2018); see the later reference for theoretical results on higher-order refinements and the important role of pre-asymptotic variance estimation. Specifically, our proposed data-driven robust bias-corrected test statistic is $T_{p+1}(\hat{h}_p)$, which rejects H_0 iff $|T_{p+1}(\hat{h}_p)| \geq \Phi_{1-\alpha/2}$ for a nominal α-level test. This approach corresponds to a special case of manual bias-correction together with the corresponding adjustment of Studentization. In practice, most common choices are $p = 2$, and this is the default in the `Stata` and `R` software implementations (Cattaneo, Jansson and Ma, 2018c, 2019c).

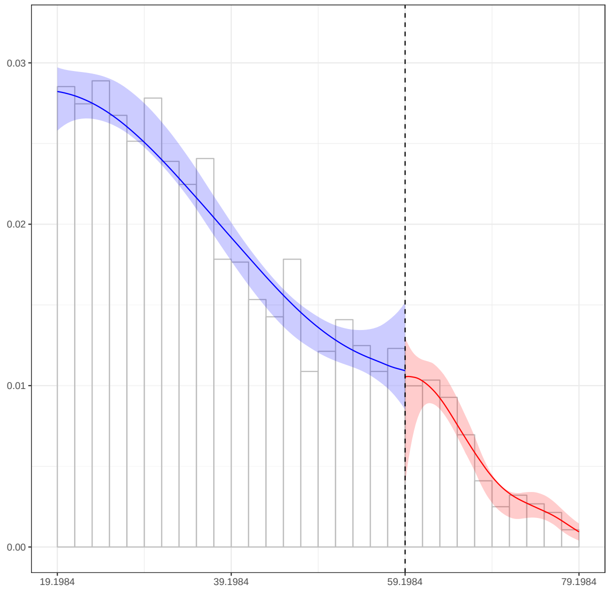
Finally, we point out that it is possible to impose additional assumptions to improve the power of the manipulation testing. Recall that our construction of the test statistic T_p essentially applies the local polynomial density estimator twice, and separately on the two sides of the cutoff point using the corresponding subsamples. However, one may argue that, under the null hypothesis that there is no manipulation, the distribution function may exhibit additional smoothness properties. We explore this possibility in Section III.8, and demonstrate how to estimate the density function on the two sides of a cutoff point, while at the same time imposing the assumption that higher order derivatives of the density function remain continuous across this cutoff point.

III.6 Empirical Illustration

We apply our proposed manipulation test to the data of Ludwig and Miller (2007) on the original Head Start implementation in the U.S. In this empirical application, a discontinuity on access to program funds at the county level occurred in 1965 when the program was first implemented: the federal government provided assistance to the 300 poorest counties, thus creating a discontinuity in program participation. Using our notation, x_i denotes the poverty index for county i, which was computed in 1965 using 1960 Census variables, and $\bar{x} = 59.1984$ is the cutoff point and poverty index of the 300-th poorest municipality.

A manipulation test in this context amounts to testing whether there is a disproportional number of counties are situated above $\bar{x}$ relative to those present below the cutoff. Figure III.2(a) presents the histogram of counties below and above the cutoff, while Figure III.2(b) presents our local polynomial density estimate along with pointwise robust bias-corrected

Figure III.2. Manipulation testing: Head Start data.



Note. (i) panel (a) reports histogram estimate of the running variable (poverty index) computed with default values in R, and panel (b) reports local polynomial density using companion R (and Stata) package described in Cattaneo, Jansson and Ma (2018c); and (ii) $n_- = 2,504$, $n_+ = 300$, and $\bar{x} = 59.1984$.

confidence intervals over a grid of points near the cutoff $\bar{x}$, implemented using $p = 2$ and the corresponding MSE-optimal data-driven bandwidth estimate. Table III.1 presents the empirical results from our manipulation test. We consider two main approaches, both covered by our theoretical work and available in our software implementation: (i) using two distinct bandwidths on each side of the cutoff ($h_- \neq h_+$), and (ii) using a common bandwidth for each side of the cutoff ($h_- = h_+$), with h_- and h_+ denoting the bandwidth on the left and on the right, respectively. For each case, we consider three distinct implementations of our manipulation test, which varies the degree of polynomial approximation used to smooth out the empirical distribution function: $T_q(h_p)$ denotes the test statistic constructed using a q-th order local polynomial density estimator, with bandwidth choice that is MSE-optimal for p-th order local polynomial density estimator. For example, our recommended choice is $T_3(h_2)$, with either common bandwidth or two different bandwidths, which amounts to first choose MSE-optimal bandwidth(s) for a local quadratic fit, and then conduct inference using a cubic approximation. This approach is the simplest implementation of the robust bias correction inference: $T_p(h_p)$ does not lead to a valid inference approach because a first-order bias will make the test over-reject the null hypothesis.

Our empirical results show no evidence of manipulation. In fact, this finding is con-

Table III.1. Manipulation testing: Head Start data.

	Pre-binning		**Bandwidths**		**Eff.** n		**Test**	
	left	right	left	right	left	right	T	p-val
$h_- \neq h_+$								
$T_2(\hat{h}_1)$			15.771	2.326	581	65	0.024	0.981
$T_3(\hat{h}_2)$			19.776	8.296	762	210	-1.146	0.252
$T_4(\hat{h}_3)$			32.487	10.808	1598	232	-1.083	0.279
$h_- = h_+$								
$T_2(\hat{h}_1)$			3.274	3.274	99	95	-1.355	0.175
$T_3(\hat{h}_2)$			9.213	9.213	316	221	-0.515	0.607
$T_4(\hat{h}_3)$			12.270	12.270	419	243	-0.712	0.477
McCrary	76	60	13.950	13.950	24	24	0.142	0.887

Note. (i) $T_p(h)$ denotes the manipulation test statistic using p-th order density estimators with bandwidth choice h (which could be common on both sides or different on either side of the cutoff), and $\hat{h}_p$ denotes the estimated MSE-optimal bandwidths for p-th order density estimator or difference of estimators (depending on the case considered); (ii) Columns under "Bandwidths" report estimated MSE-optimal bandwidths, Columns under "Eff. n" report effective sample size on either side of the cutoff, and Columns under "Test" report value of test statistic (T) and two-sided p-value (p-val); and (iii) first three rows allow for different bandwidths on each side of the cutoff, while last three rows employ a common bandwidth on both sides of the cutoff (chosen to be MSE-optimal for the difference of density estimates). All estimates are obtained using companion `R` (and `Stata`) package described in Cattaneo, Jansson and Ma (2018c).

sistent with the underlying institutional knowledge of the program: the poverty index was constructed in 1965 at the federal level using county level information from the 1960 Census, which implies it is indeed highly implausible that individual counties could have manipulated their assigned poverty index. Our findings are robust to different bandwidth and local polynomial order specifications.

III.7 Conclusion

We introduced a boundary adaptive kernel-based density estimator employing local polynomial methods, which requires choosing only one tuning parameter and does not require boundary-specific data transformations (such as pre-binning). We studied its main asymptotic properties, including bias, variance and distributional approximations, consistent variance estimation, and consistent bandwidth selection. We used these results to develop a new manipulation test via discontinuity in density testing at a boundary point. Several extensions and generalizations of our results are underway in ongoing work (Cattaneo, Jansson and Ma, 2019a), and two distinct general purpose software packages in `Stata` and `R` are readily available Cattaneo, Jansson and Ma (2018c, 2019c).

III.8 Additional Results and Preliminary Lemmas

III.8.1 Other Standard Error Estimators

The standard error $\hat{\sigma}_{p,v}(x)$ (see Theorem III.2) is fully automatic and adapts to both interior and boundary regions. In this section we consider two other ways to construct a standard error.

Plug-in Standard Error

Take $v \geq 1$. Then the asymptotic variance of $\hat{F}_p^{(v)}(x)$ takes the following form:

$$\mathcal{V}_{p,v}(x) = (v!)^2 f(x)\mathbf{e}_v^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v.$$

One way of constructing estimate of the above quantity is to plug-in a consistent estimator of $f(x)$, which is simply the estimated density. Hence we can use

$$\hat{\mathcal{V}}_{p,v}(x) = (v!)^2 \hat{f}_p(x)\mathbf{e}_v^{\mathrm{T}}\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v.$$

The next question is how $\mathbf{S}_{p,x}$ and $\mathbf{\Gamma}_{p,x}$ should be constructed. Note that they are related to the kernel, evaluation point x and the bandwidth h, but *not* the data generating process. Therefore the three matrices can be constructed by either analytical integration or numerical method.

Jackknife-based Standard Error

The standard error $\hat{\sigma}_{p,v}(x)$ is obtained by inspecting the asymptotic linear representation. It is fully automatic and adapts to both interior and boundaries. In this part, we present another standard error which resembles $\hat{\sigma}_{p,v}(x)$, albeit with a different motivation.

Recall that $\hat{\boldsymbol{\beta}}_p(x)$ is essentially a second order U-statistic, and the following expansion is justified:

$$\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_p(x)\right)$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\left(\tilde{F}(x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\right)K_h(x_i - x)\\
&= \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\left(\frac{1}{n-1}\sum_{j;j\neq i}\Big(\mathbb{1}(x_j \le x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\Big)\right)K_h(x_i - x) + O_{\mathrm{p}}\left(\frac{1}{n}\right)\\
&= \frac{1}{n(n-1)}\sum_{i,j;i\neq j}\mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}(x_j \le x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\Big)K_h(x_i - x) + O_{\mathrm{p}}\left(\frac{1}{n}\right),
\end{aligned}
$$

where the remainder represents leave-in bias. Note that the above could be written as a U-statistic, and to apply the Hoeffding decomposition, define

$$
\begin{aligned}
\mathbf{U}(x_i, x_j) = \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}(x_j \le x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\Big)K_h(x_i - x)\\
+ \mathbf{r}_p\left(\frac{x_j - x}{h}\right)\Big(\mathbb{1}(x_i \le x_j) - \mathbf{r}_p(x_j - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\Big)K_h(x_j - x),
\end{aligned}
$$

which is symmetric in its two arguments. Then

$$
\begin{aligned}
\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_p(x)\right) &= \mathbb{E}\left[\mathbf{U}(x_i, x_j)\right] + \frac{1}{n}\sum_i\Bigg(\mathbf{U}_1(x_i) - \mathbb{E}\left[\mathbf{U}(x_i, x_j)\right]\Bigg)\\
&\quad + \binom{n}{2}^{-1}\sum_{i,j;i<j}\Bigg(\mathbf{U}(x_i, x_j) - \mathbf{U}_1(x_i) - \mathbf{U}_1(x_j) + \mathbb{E}\left[\mathbf{U}(x_i, x_j)\right]\Bigg).
\end{aligned}
$$

Here $\mathbf{U}_1(x_i) = \mathbb{E}\left[\mathbf{U}(x_i, x_j)\middle| x_i\right]$. The second line in the above display is the analogue of $\hat{\mathbf{L}}$, which contributes to the leading variance, and the third line is negligible. The new standard error, we call the jackknife-based standard error, is given by the following:

$$
\hat{\sigma}^{(\mathtt{JK})}_{p,v}(x) \equiv (v!)\sqrt{\frac{1}{nh^{2v}}\mathbf{e}_v^{\mathrm{T}}\hat{\mathbf{S}}_{p,x}^{-1}\hat{\boldsymbol{\Gamma}}^{\mathtt{JK}}_{p,x}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{e}_v},
$$

with

$$
\begin{aligned}
\hat{\boldsymbol{\Gamma}}^{\mathtt{JK}}_{p,x} &= \frac{1}{n}\sum_i\left(\frac{1}{n-1}\sum_{j;j\neq i}\hat{\mathbf{U}}(x_i, x_j)\right)\left(\frac{1}{n-1}\sum_{j;j\neq i}\hat{\mathbf{U}}(x_i, x_j)\right)^{\mathrm{T}}\\
&\quad - \left(\binom{n}{2}^{-1}\sum_{i,j;i\neq j}\hat{\mathbf{U}}(x_i, x_j)\right)\left(\binom{n}{2}^{-1}\sum_{i,j;i\neq j}\hat{\mathbf{U}}(x_i, x_j)\right)^{\mathrm{T}},
\end{aligned}
$$

and

$$\begin{aligned}\hat{\mathbf{U}}(x_i, x_j) &= \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}(x_j \le x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_p(x)\Big)K_h(x_i - x)\\ &\quad + \mathbf{r}_p\left(\frac{x_j - x}{h}\right)\Big(\mathbb{1}(x_i \le x_j) - \mathbf{r}_p(x_j - x)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_p(x)\Big)K_h(x_j - x).\end{aligned}$$

The name jackknife comes from the fact that we use leave-one-out "estimator" for $\mathbf{U}_1(x_i)$: with x_i fixed,

$$\text{“}\frac{1}{n-1}\sum_{j;j\neq i}\hat{\mathbf{U}}(x_i, x_j) \overset{\mathrm{p}}{\to} \mathbf{U}_1(x_i)\text{”}.$$

Under the same conditions specified in Theorem III.2, one can show that the jackknife-based standard error is consistent.

III.8.2 Manipulation Testing: Restricted Estimation

In Section III.5, we provide a test procedure on the discontinuity of the density by estimating on the two sides of the cutoff separately. This procedure is flexible and requires minimum assumptions. There are ways, however, to improve the power of the test when the densities are estimated with additional assumptions on the smoothness of the distribution function

In a restricted model, the polynomial basis is re-defined as

$$\mathbf{r}_p(u) = \Big[1 \quad u\mathbf{1}(u<0) \quad u\mathbf{1}(u\ge 0) \quad u^2 \quad u^3 \quad \cdots \quad u^p\Big]^{\mathrm{T}} \in \mathbb{R}^{p+2},$$

and the estimator in the fully restricted model is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_p(\bar{x}) &= \Big[\hat{F}_p(\bar{x}) \quad \hat{f}_p(\bar{x}-) \quad \hat{f}_p(\bar{x}+) \quad \tfrac{1}{2}\hat{F}_p^{(2)}(\bar{x}) \quad \cdots \quad \tfrac{1}{p!}\hat{F}_p^{(p)}(\bar{x})\Big]^{\mathrm{T}}\\ &= \arg\max_{\mathbf{b}\in\mathbb{R}^{p+2}}\sum_i\Big(\tilde{F}(x_i) - \mathbf{r}_p(x_i - \bar{x})^{\mathrm{T}}\mathbf{b}\Big)^2 K_h(x_i - \bar{x}).\end{aligned}$$

Again the notations (for example $\mathbf{X}$ and $\mathbf{X}_h$) are redefined similarly, with the scaling matrix $\mathbf{H}$ adjusted to ensure $\mathbf{H}^{-1}\mathbf{r}_p(u) = \mathbf{r}_p(h^{-1}u)$. Here $\hat{F}_p(\bar{x})$ is the estimated distribution function and $\frac{1}{2}\hat{F}_p^{(2)}(\bar{x})$, $\cdots$, $\frac{1}{p!}\hat{F}_p^{(p)}(\bar{x})$ are the estimated higher order derivatives, which we assume are all continuous at $\bar{x}$, while $\hat{f}_p(\bar{x}-)$ and $\hat{f}_p(\bar{x}+)$ are the estimated densities on the two sides of $\bar{x}$. Therefore we call the above model restricted, since it only allows discontinuity of the first derivative of F (i.e. the density) but not the other derivatives.

With the modification of the polynomial basis, all other matrices in the previous subsection are redefined similarly, and

$$\mathbf{I}_{p+2} = \begin{bmatrix} \mathbf{e}_0 & \mathbf{e}_{1,-} & \mathbf{e}_{1,+} & \mathbf{e}_2 & \cdots & \mathbf{e}_p \end{bmatrix}_{(p+2)\times(p+2)}.$$

where the subscripts indicate the corresponding derivatives to extract. Moreover

$$\mathbf{r}_{-,p}(u) = \begin{bmatrix} 1 & u & 0 & u^2 & \cdots & u^p \end{bmatrix}, \qquad \mathbf{r}_{+,p}(u) = \begin{bmatrix} 1 & 0 & u & u^2 & \cdots & u^p \end{bmatrix}.$$

Now we state the main result concerning the manipulation testing. Let $\hat{\mathbf{S}}_{p,\bar{x}}$ and $\hat{\mathbf{\Gamma}}_{p,\bar{x}}$ be constructed as in Section III.3, and

$$\hat{\mathcal{V}}_{p,1}(\bar{x}) = \frac{1}{h}(\mathbf{e}_{1,+} - \mathbf{e}_{1,-})^{\mathrm{T}} \hat{\mathbf{S}}_{p,\bar{x}} \hat{\mathbf{\Gamma}}_{p,\bar{x}} \hat{\mathbf{S}}_{p,\bar{x}} (\mathbf{e}_{1,+} - \mathbf{e}_{1,-}).$$

Corollary III.2 (Manipulation testing: restricted estimation)
Assume Assumptions III.1 and III.2 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$ with $\alpha_x \geq p+1$ for some integer $p \geq 1$. Further, $n \cdot h^2 \to \infty$ and $n \cdot h^{2p+1} \to 0$. Then under the null hypothesis $\mathsf{H}_0 : f(\bar{x}+) = f(\bar{x}-)$,

$$T_p(h) = \frac{\hat{f}_p(\bar{x}+) - \hat{f}_p(\bar{x}-)}{\sqrt{\frac{1}{nh}\hat{\mathcal{V}}_{p,1}(\bar{x})}} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

As a result, under the alternative hypothesis $\mathsf{H}_1 : f(\bar{x}+) \neq f(\bar{x}-)$,

$$\lim_{n\to\infty} \mathbb{P}[|T_p(h)| \geq \Phi_{1-\alpha/2}] = 1.$$

Here $\Phi_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. ‖

III.8.3 Preliminary Lemmas for Section III.3

We first consider the object $\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h/n$

Lemma III.3
Assume Assumptions III.1 and III.2 hold, $h \to 0$ and $nh \to \infty$. Then

$$\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h = f(x)\mathbf{S}_{p,x} + o(1) + O_{\mathrm{p}}\left(1/\sqrt{nh}\right).$$ ‖

Lemma III.3 shows that the matrix $\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h/n$ is asymptotically invertible. Also note that this result covers both interior and boundary evaluation point x, and depending on the nature of x, the exact form of $\mathbf{S}_{p,x}$ differs.

With simple algebra, one has

$$\hat{\boldsymbol{\beta}}_p(x) - \boldsymbol{\beta}_p(x) = \mathbf{H}^{-1}\left(\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h\right)^{-1}\left(\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_p(x))\right),$$

and the following gives a further decomposition of the "numerator."

$$\begin{aligned}
\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_p(x)) &= \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\tilde{F}(x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\Big)K_h(x_i - x)\\
&= \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(F(x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\Big)K_h(x_i - x)\\
&\quad + \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\Big(\tilde{F}(x + hu) - F(x + hu)\Big)K(u)f(x + hu)\mathrm{d}u\\
&\quad + \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\tilde{F}(x_i) - F(x_i)\Big)K_h(x_i - x)\\
&\quad - \int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p(u)\Big(\tilde{F}(x + hu) - F(x + hu)\Big)K(u)f(x + hu)\mathrm{d}u.
\end{aligned}$$

The first part represents the smoothing bias, and the second part can be analyzed as a sample average. The real challenge comes from the third term, which can have a nonnegligible (first order) contribution. We further decompose it as

$$\begin{aligned}
&\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\tilde{F}(x_i) - F(x_i)\Big)K_h(x_i - x)\\
&= \frac{1}{n^2}\sum_{i,j} \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x)\\
&= \frac{1}{n^2}\sum_{i} \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(1 - F(x_i)\Big)K_h(x_i - x)\\
&\qquad + \frac{1}{n^2}\sum_{i,j;i\neq j} \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x).
\end{aligned}$$

As a result,

$$\begin{aligned}
&\frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\tilde{F}(x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\Big)K_h(x_i - x)\\
&\qquad = \frac{1}{n}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(F(x_i) - \mathbf{r}_p(x_i - x)^{\mathrm{T}}\boldsymbol{\beta}_p(x)\Big)K_h(x_i - x)
\end{aligned}$$

(smoothing bias $\hat{\mathbf{B}}_{\mathsf{S}}$)

$$+ \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p(u)\Big(\tilde{F}(x+hu) - F(x+hu)\Big)K(u)f(x+hu)\mathrm{d}u \qquad \text{(linear variance } \hat{\mathbf{L}}\text{)}$$

$$+ \frac{1}{n^2}\sum_i \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(1 - F(x_i)\Big)K_h(x_i - x) \qquad \text{(leave-in bias } \hat{\mathbf{B}}_{\mathtt{LI}}\text{)}$$

$$+ \frac{1}{n^2}\sum_{i,j;i\neq j}\Big\{\mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x)$$

$$- \mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x)\Big| x_j\right]\Big\}.$$

$$\text{(quadratic variance } \hat{\mathbf{R}}\text{)}$$

To provide intuition for the above decomposition, the smoothing bias is a typical feature of nonparametric estimators; leave-in bias arises since each observation is used twice, in constructing the empirical distribution function $\tilde{F}$ and as a design point (that is, $\tilde{F}$ has to be evaluated at x_i); and a second order U-statistic shows up because the "dependent variable," $\mathbf{Y}$, is estimated, which leads to a double sum.

We first analyze the bias terms.

Lemma III.4

Assume Assumptions III.1 and III.2 hold with $\alpha_x \ge p+1$, $h \to 0$ and $nh \to \infty$. Then

$$\hat{\mathbf{B}}_{\mathtt{S}} = h^{p+1}\frac{F^{(p+1)}(x)f(x)}{(p+1)!}\mathbf{c}_{p,x} + o_{\mathrm{p}}(h^{p+1}), \qquad \hat{\mathbf{B}}_{\mathtt{LI}} = O_{\mathrm{p}}\left(n^{-1}\right). \qquad \|$$

By imposing additional smoothness, it is also possible to characterize the next term in the smoothing bias, which has order h^{p+2}. We report the higher order bias in a later section as it is used for bandwidth selection.

Next we consider the "influence function" part, $\hat{\mathbf{L}}$. This term is crucial in the sense that (under suitable conditions such that $\hat{\mathbf{R}}$ is negligible) it determines the asymptotic variance of our estimator, and with correct scaling, it is asymptotically normally distributed.

Lemma III.5

Assume Assumptions III.1and III.2 hold with $\alpha_x \ge 2$, $h \to 0$ and $nh \to \infty$. Define the scaling matrix

$$\mathbf{N}_x = \begin{cases} \mathrm{diag}\Big\{1, \quad h^{-1/2},\ h^{-1/2},\ \cdots,\ h^{-1/2}\Big\} & x \textit{ interior},\\ \mathrm{diag}\Big\{h^{-1/2},\ h^{-1/2},\ h^{-1/2},\ \cdots,\ h^{-1/2}\Big\} & x \textit{ boundary}, \end{cases}$$

then

$$\sqrt{n}\mathbf{N}_x\Big[f(x)\mathbf{S}_{p,x}\Big]^{-1}\hat{\mathbf{L}} \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0},\ \mathsf{V}_{p,x}),$$

with

$$\mathsf{V}_{p,x} = \begin{cases} F(x)(1-F(x))\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}} + f(x)(\mathbf{I}-\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}})\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}(\mathbf{I}-\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}) & x \textit{ interior} \\ f(x)\left(\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1} + c\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\right) & x = x_{\mathtt{L}} + ch \\ f(x)\left(\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1} + c\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}} - (\mathbf{e}_1\mathbf{e}_0^{\mathrm{T}}+\mathbf{e}_0\mathbf{e}_1^{\mathrm{T}})\right) & x = x_{\mathtt{U}} - ch. \end{cases} \quad \|$$

The scaling matrix depends on whether the evaluation point is located in the interior or boundary, which is a unique feature of our estimator. To see the intuition, consider an interior point x, and recall that the first element of $\hat{\boldsymbol{\beta}}_p(x)$ is the smoothed empirical distribution function, which is $\sqrt{n}$-estimable. Therefore, the property of $\hat{F}_p(x)$ is very different from those of the estimated density and higher order derivatives.

When x is either in the lower or upper boundary region, $\hat{F}_p(x)$ essentially estimates 0 or 1, respectively, hence it is super-consistent in the sense that it converges even faster than $1/\sqrt{n}$. In this case, the leading $1/\sqrt{n}$-variance vanishes, and higher order residual noise dominates, which makes $\hat{F}_p(x)$ no longer independent of the estimated density and derivatives, justifying the formula of boundary evaluation points.

Finally we consider the second order U-statistic component.

Lemma III.6

Assume Assumptions III.1 and III.2 hold, $h \to 0$ and $nh \to \infty$. Then

$$\mathbb{V}[\hat{\mathbf{R}}] = \frac{2}{n^2 h} f(x)F(x)(1-F(x))\mathbf{T}_{p,x} + O(n^{-2}).$$

In particular, when x is in the boundary region, the above has order $O(n^{-2})$. ‖

III.8.4 Preliminary Lemmas for Section III.5

In the following lemmas, we will give asymptotic results for the estimation problem in Section III.5. Proofs are omitted.

Lemma III.7

Let Assumptions of Lemma III.3 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then

$$\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h = f(\bar{x}-)\mathbf{S}_{-,p} + f(\bar{x}+)\mathbf{S}_{+,p} + O\Big(h\Big) + O_{\mathrm{p}}\Big(1/\sqrt{nh}\Big),$$

where

$$\mathbf{S}_{-,p} = \int_{-1}^{0} \mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(u)^{\mathrm{T}} K(u)\mathrm{d}u, \qquad \mathbf{S}_{+,p} = \int_{0}^{1} \mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(u)^{\mathrm{T}} K(u)\mathrm{d}u. \qquad \|$$

Again we decompose the estimator into four terms, namely $\hat{\mathbf{B}}_{\mathtt{LI}}$, $\hat{\mathbf{B}}_{\mathtt{S}}$, $\hat{\mathbf{L}}$ and $\hat{\mathbf{R}}$.

Lemma III.8

Let Assumptions of Lemma III.4 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then

$$\hat{\mathbf{B}}_{\mathtt{S}} = h^{p+1}\left\{ \frac{F^{(p+1)}(\bar{x}-)f(\bar{x}-)}{(p+1)!}\mathbf{c}_{-,p} + \frac{F^{(p+1)}(\bar{x}+)f(x+)}{(p+1)!}\mathbf{c}_{+,p} \right\} + o_{\mathrm{p}}(h^{p+1}),\ \hat{\mathbf{B}}_{\mathtt{LI}} = O_{\mathrm{p}}\left(\frac{1}{n}\right),$$

where

$$\mathbf{c}_{-,p} = \int_{-1}^{0} u^{p+1}\mathbf{r}_{-,p}(u)K(u)\mathrm{d}u, \qquad \mathbf{c}_{+,p} = \int_{0}^{1} u^{p+1}\mathbf{r}_{+,p}(u)K(u)\mathrm{d}u. \qquad \|$$

Lemma III.9

Let Assumptions of Lemma III.5 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then

$$\begin{aligned}&\mathbb{V}\left[\sqrt{\frac{n}{h}}\Big(\mathbf{e}_{1,+} - \mathbf{e}_{1,-}\Big)^{\mathrm{T}}\Big(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p}\Big)^{-1}\hat{\mathbf{L}}\right] \\ &= f(\bar{x}-)\mathbf{e}_{1,-}^{\mathrm{T}}\mathbf{S}_{-,p}^{-1}\mathbf{\Gamma}_{-,p}\mathbf{S}_{-,p}^{-1}\mathbf{e}_{1,-} + f(\bar{x}+)\mathbf{e}_{1,+}^{\mathrm{T}}\mathbf{S}_{+,p}^{-1}\mathbf{\Gamma}_{+,p}\mathbf{S}_{+,p}^{-1}\mathbf{e}_{1,+} + O(h),\end{aligned}$$

where

$$\begin{aligned}\mathbf{\Gamma}_{-,p} &= \iint_{[-1,0]^2} (u \wedge v)\mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(v)^{\mathrm{T}}K(u)K(v)\ \mathrm{d}u\mathrm{d}v, \\ \mathbf{\Gamma}_{+,p} &= \iint_{[0,1]^2} (u \wedge v)\mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(v)^{\mathrm{T}}K(u)K(v)\ \mathrm{d}u\mathrm{d}v. \qquad \|\end{aligned}$$

Note that the above gives the asymptotic variance of the difference $\hat{f}(\bar{x}+) - \hat{f}(\bar{x}-)$, and the variance takes an additive form. This is not surprising, since the two density estimates, $\hat{f}(\bar{x}+)$ and $\hat{f}(\bar{x}-)$, rely on distinctive subsamples, meaning that they are asymptotically independent.

Finally the order of $\hat{\mathbf{R}}$ can also be established.

Lemma III.10

Let Assumptions of Lemma III.6 hold separately on $\mathcal{X}_-$ and $\mathcal{X}_+$, then

$$\hat{\mathbf{R}} = O_{\mathrm{p}}\left(\sqrt{\frac{1}{n^2h}}\right). \qquad \|$$

III.8.5 Preliminary Lemmas for Manipulation Testing with Restricted Estimation

Lemma III.11

Let Assumptions of Lemma III.3 hold with the exception that f may be discontinuous across $\bar{x}$, then

$$\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h = \{f(\bar{x}-)\mathbf{S}_{-,p} + f(\bar{x}+)\mathbf{S}_{+,p}\} + O(h) + O_{\mathrm{p}}(1/\sqrt{nh}),$$

where

$$\mathbf{S}_{-,p} = \int_{-1}^{0} \mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(u)^{\mathrm{T}}K(u)\mathrm{d}u, \qquad \mathbf{S}_{+,p} = \int_{0}^{1} \mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(u)^{\mathrm{T}}K(u)\mathrm{d}u. \qquad \|$$

Again we decompose the estimator into four terms, $\hat{\mathbf{B}}_{\mathtt{LI}}$, $\hat{\mathbf{B}}_{\mathtt{S}}$, $\hat{\mathbf{L}}$ and $\hat{\mathbf{R}}$, which correspond to leave-in bias, smoothing bias, linear variance and quadratic variance, respectively.

Lemma III.12

Let Assumptions of Lemma III.4 hold with the exception that f may be discontinuous across $\bar{x}$, then

$$\hat{\mathbf{B}}_{\mathtt{S}} = h^{p+1}\left\{\frac{F^{(p+1)}(\bar{x}-)f(\bar{x}-)}{(p+1)!}\mathbf{c}_{-,p} + \frac{F^{(p+1)}(\bar{x}+)f(\bar{x}+)}{(p+1)!}\mathbf{c}_{+,p}\right\} + o_{\mathrm{p}}(h^{p+1}),\ \hat{\mathbf{B}}_{\mathtt{LI}} = O_{\mathrm{p}}\left(\frac{1}{n}\right),$$

where

$$\mathbf{c}_{-,p} = \int_{-1}^{0} u^{p+1}\mathbf{r}_{-,p}(u)K(u)\mathrm{d}u, \qquad \mathbf{c}_{+,p} = \int_{0}^{1} u^{p+1}\mathbf{r}_{+,p}(u)K(u)\mathrm{d}u. \qquad \|$$

Lemma III.13

Let Assumptions of Lemma III.5 hold with the exception that f may be discontinuous across $\bar{x}$, then

$$\begin{aligned}
&\mathbb{V}\left[\sqrt{\frac{n}{h}}\Big(\mathbf{e}_{1,+} - \mathbf{e}_{1,-}\Big)^{\mathrm{T}}\Big(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p}\Big)^{-1}\hat{\mathbf{L}}\right] \\
&= (\mathbf{e}_{1,+} - \mathbf{e}_{1,-})^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})^{-1}(f(\bar{x}+)^3\mathbf{\Gamma}_{+,p} \\
&\qquad + f(\bar{x}-)^3\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi})(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})^{-1}(\mathbf{e}_{1,+} - \mathbf{e}_{1,-}) + O(h),
\end{aligned}$$

where

$$\begin{aligned}\mathbf{\Gamma}_{-,p} &= \iint_{[-1,0]^2} (u \wedge v)\mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(v)^{\mathrm{T}} K(u)K(v)\ \mathrm{d}u\mathrm{d}v,\\ \mathbf{\Gamma}_{+,p} &= \iint_{[0,1]^2} (u \wedge v)\mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(v)^{\mathrm{T}} K(u)K(v)\ \mathrm{d}u\mathrm{d}v.\end{aligned}$$

and

$$\mathbf{\Psi} = \begin{bmatrix} (-1)^0 & & & & & & \\ & & (-1)^1 & & & & \\ & (-1)^1 & & & & & \\ & & & (-1)^2 & & & \\ & & & & (-1)^3 & & \\ & & & & & \ddots & \\ & & & & & & (-1)^p \end{bmatrix}. \qquad \|$$

Again we can show that the quadratic part is negligible.

Lemma III.14
Let Assumptions of Lemma III.6 hold with the exception that f may not be continuous across $\bar{x}$, then

$$\hat{\mathbf{R}} = O_{\mathrm{p}}\left(\sqrt{\frac{1}{n^2 h}}\right). \qquad \|$$

III.9 Proof

III.9.1 Proof of Theorem III.1

This follows from the preliminary lemmas. ■

III.9.2 Proof of Theorem III.2

First we note that the second half of the theorem follows from the first half and the asymptotic normality result of Theorem III.1, hence it suffices to prove the first half, i.e. the consistency of $\hat{\mathcal{V}}_{p,v}(x)$.

The analysis of this estimator is quite involved, since it takes the form of a third order V-statistic. Moreover, since the empirical distribution function $\tilde{F}$ is involved in the formula, a full expansion leads to a fifth order V-statistic. However, some simple tricks will greatly simplify the problem.

We first split $\hat{\mathbf{\Gamma}}_{p,x}$ into four terms, respectively

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_{p,x,1} &= \frac{1}{n^3}\sum_{i,j,k}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)^{\mathrm{T}}K_h(x_j-x)K_h(x_k-x)\\
&\qquad\Big(\mathbb{1}[x_i\le x_j]-F(x_j)\Big)\Big(\mathbb{1}[x_i\le x_k]-F(x_k)\Big)\\
\hat{\mathbf{\Sigma}}_{p,x,2} &= \frac{1}{n^3}\sum_{i,j,k}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)^{\mathrm{T}}K_h(x_j-x)K_h(x_k-x)\\
&\qquad\Big(F(x_j)-\tilde{F}(x_j)\Big)\Big(\mathbb{1}[x_i\le x_k]-\tilde{F}(x_k)\Big)\\
\hat{\mathbf{\Sigma}}_{p,x,3} &= \frac{1}{n^3}\sum_{i,j,k}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)^{\mathrm{T}}K_h(x_j-x)K_h(x_k-x)\\
&\qquad\Big(\mathbb{1}[x_i\le x_j]-\tilde{F}(x_j)\Big)\Big(F(x_k)-\tilde{F}(x_k)\Big)\\
\hat{\mathbf{\Sigma}}_{p,x,4} &= \frac{1}{n^3}\sum_{i,j,k}\mathbf{r}_p\left(\frac{x_j-x}{h}\right)\mathbf{r}_p\left(\frac{x_k-x}{h}\right)^{\mathrm{T}}K_h(x_j-x)K_h(x_k-x)\\
&\qquad\Big(F(x_j)-\tilde{F}(x_j)\Big)\Big(F(x_k)-\tilde{F}(x_k)\Big).
\end{aligned}$$

Leaving $\hat{\mathbf{\Sigma}}_{p,x,1}$ for a while, since it is the key component in this variance estimator. We first consider $\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\mathbf{\Sigma}}_{p,x,4}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x$. By the uniform consistency of the empirical distribution function, it can be shown easily that

$$\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\mathbf{\Sigma}}_{p,x,4}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x = O_{\mathrm{p}}\left((nh)^{-1}\right).$$

Note that the extra h^{-1} comes from the scaling matrix $\mathbf{N}_x$, but not the kernel function K_h. Next we consider $\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\mathbf{\Sigma}}_{p,x,2}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x$, which takes the following form (up to the negligible smoothing bias):

$$\begin{aligned}
&\mathbf{N}_x\hat{\mathbf{S}}_{p,x}^{-1}\hat{\mathbf{\Sigma}}_{p,x,2}\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x\\
=&\mathbf{N}_x\mathbf{H}(\boldsymbol{\beta}_p(x)-\hat{\boldsymbol{\beta}}_p(x))\left(\frac{1}{n^2}\sum_{i,k}\mathbf{r}_p\left(\frac{x_k-x}{h}\right)^{\mathrm{T}}K_h(x_k-x)\Big(\mathbb{1}[x_i\le x_k]-\tilde{F}(x_k)\Big)\right)\hat{\mathbf{S}}_{p,x}^{-1}\mathbf{N}_x\\
=&O_{\mathrm{p}}((nh)^{-1/2})=o_{\mathrm{p}}(1),
\end{aligned}$$

where the last line uses the asymptotic normality of $\hat{\boldsymbol{\beta}}_p(x)$. For $\hat{\mathbf{\Sigma}}_{p,x,1}$, we make the observation that it is possible to ignore all "diagonal" terms, meaning that

$$\hat{\mathbf{\Sigma}}_{p,x,1} = \frac{1}{n^3} \sum_{\substack{i,j,k \\ \text{distinct}}} \mathbf{r}_p\left(\frac{x_j - x}{h}\right)\mathbf{r}_p\left(\frac{x_k - x}{h}\right)^{\mathrm{T}} K_h(x_j - x)K_h(x_k - x)$$
$$\Big(\mathbb{1}[x_i \le x_j] - F(x_j)\Big)\Big(\mathbb{1}[x_i \le x_k] - F(x_k)\Big) + o_{\mathrm{p}}(h),$$

under the assumption that $nh^2 \to \infty$. As a surrogate, define

$$\mathbf{U}_{i,j,k} = \mathbf{r}_p\left(\frac{x_j - x}{h}\right)\mathbf{r}_p\left(\frac{x_k - x}{h}\right)^{\mathrm{T}} K_h(x_j - x)K_h(x_k - x)$$
$$\Big(\mathbb{1}[x_i \le x_j] - F(x_j)\Big)\Big(\mathbb{1}[x_i \le x_k] - F(x_k)\Big),$$

which means

$$\hat{\mathbf{\Sigma}}_{p,x,1} = \frac{1}{n^3} \sum_{\substack{i,j,k \\ \text{distinct}}} \mathbf{U}_{i,j,k}.$$

The critical step is to further decompose the above into

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_{p,x,1} &= \frac{1}{n^3} \sum_{\substack{i,j,k \\ \text{distinct}}} \mathbb{E}[\mathbf{U}_{i,j,k}|x_i] && \text{(I)} \\
&+ \frac{1}{n^3} \sum_{\substack{i,j,k \\ \text{distinct}}} \Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i,j,k}|x_i, x_j]\Big) && \text{(II)} \\
&+ \frac{1}{n^3} \sum_{\substack{i,j,k \\ \text{distinct}}} \Big(\mathbb{E}[\mathbf{U}_{i,j,k}|x_i, x_j] - \mathbb{E}[\mathbf{U}_{i,j,k}|x_i]\Big). && \text{(III)}
\end{aligned}$$

We already investigated the properties of term (I) in Lemma III.5, hence it remains to show that both (II) and (III) are $o(h)$, hence does not affect the estimation of asymptotic variance. We consider (II) as an example, and the analysis of (III) is similar. Since (II) has zero expectation, we consider its variance (for simplicity treat $\mathbf{U}$ as a scaler):

$$\mathbb{V}[\text{(II)}] = \mathbb{E}\left[\frac{1}{n^6} \sum_{\substack{i,j,k \\ \text{distinct}}} \sum_{\substack{i',j',k' \\ \text{distinct}}} \Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i,j,k}|x_i, x_j]\Big)\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i',j',k'}|x_{i'}, x_{j'}]\Big)\right].$$

The expectation will be zero if the six indices are all distinct. Similarly, when there are only two indices among the six are equal, the expectation will be zero *unless* $k = k'$, hence

$$
\begin{aligned}
\mathbb{V}[(\mathrm{II})] &= \mathbb{E}\left[\frac{1}{n^6}\sum_{\substack{i,j,k\\ \text{distinct}}}\sum_{\substack{i',j',k'\\ \text{distinct}}}\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i,j,k}|x_i, x_j]\Big)\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i',j',k'}|x_{i'}, x_{j'}]\Big)\right]\\
&= \mathbb{E}\left[\frac{1}{n^6}\sum_{\substack{i,j,k,i'j'\\ \text{distinct}}}\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i,j,k}|x_i, x_j]\Big)\Big(\mathbf{U}_{i,j,k} - \mathbb{E}[\mathbf{U}_{i',j',k}|x_{i'}, x_{j'}]\Big)\right]\\
&\quad + \cdots,
\end{aligned}
$$

where $\cdots$ represent cases where more than two indices among the six are equal. We can easily compute the order from the above as

$$
\mathbb{V}[(\mathrm{II})] = O(n^{-1}) + O((nh)^{-2}),
$$

which shows that

$$
(\mathrm{II}) = O_{\mathrm{p}}(n^{-1/2} + (nh)^{-1}) = o_{\mathrm{p}}(h),
$$

which closes the proof. ■

III.9.3 Proof of Lemma III.1

We rely on Lemma III.3 and III.4 (note that whether the weights are estimated is irrelevant here), hence will not repeat arguments already established there. Instead, extra care will be given to ensure the characterization of higher order bias.

Consider the case where with enough smoothness on F, then the bias is characterized by

$$
\begin{aligned}
& h^{-v}v!\mathbf{e}_v^{\mathrm{T}}\left[f(x)\mathbf{S}_{p,x} + hF^{(2)}(x)\tilde{\mathbf{S}}_{p,x} + o(h) + O_{\mathrm{p}}(1/\sqrt{nh})\right]^{-1}\\
&\qquad \left[h^{p+1}\frac{F^{(p+1)}(x)}{(p+1)!}f(x)\mathbf{c}_{p,x} + h^{p+2}\left[\frac{F^{(p+2)}(x)}{(p+2)!}f(x) + \frac{F^{(p+1)}(x)}{(p+1)!}F^{(2)}(x)\right]\tilde{\mathbf{c}}_{p,x} + o(h^{p+2})\right]\\
&= h^{-v}v!\mathbf{e}_v^{\mathrm{T}}\left[\frac{1}{f(x)}\mathbf{S}_{p,x}^{-1} - h\frac{F^{(2)}(x)}{[f(x)]^2}\mathbf{S}_{p,x}^{-1}\tilde{\mathbf{S}}_{p,x}\mathbf{S}_{p,x}^{-1} + O_{\mathrm{p}}\left(1/\sqrt{nh}\right)\right]\\
&\qquad \left[h^{p+1}\frac{F^{(p+1)}(x)}{(p+1)!}f(x)\mathbf{c}_{p,x} + h^{p+2}\left[\frac{F^{(p+2)}(x)}{(p+2)!}f(x) + \frac{F^{(p+1)}(x)}{(p+1)!}F^{(2)}(x)\right]\tilde{\mathbf{c}}_{p,x} + o(h^{p+2})\right]\\
&\qquad \{1 + o_{\mathrm{p}}(1)\},
\end{aligned}
$$

which gives the desired result. Here $\tilde{\mathbf{S}}_{p,x} = \int_{\frac{x_\mathrm{L}-x}{h}}^{\frac{x_\mathrm{U}-x}{h}} u\mathbf{r}_p(u)\mathbf{r}_p(u)^\mathrm{T}k(u)\mathrm{d}u$. And for the last line to hold, one needs the extra condition $nh^3 \to \infty$ so that $O_\mathrm{p}\left(1/\sqrt{nh}\right) = o_\mathrm{p}(h)$. See Fan and Gijbels (1996) (Theorem 3.1, pp. 62). ■

III.9.4 Proof of Lemma III.2

The proof resembles that of Lemma III.3, and is omitted here. ■

III.9.5 Proof of Theorem III.3

The proof splits into two cases. We sketch one of them. Assume either x is boundary or $p - v$ is odd, the MSE-optimal bandwidth is asymptotically equivalent to the following:

$$\frac{\tilde{h}_{p,v}(x)}{h_{p,v}(x)} \to 1, \qquad \tilde{h}_{p,v}(x) = \left(\frac{1}{n}\frac{(2v-1)f(x)\mathbf{e}_v^\mathrm{T}\mathbf{S}_{p,x}^{-1}\mathbf{\Gamma}_{p,x}\mathbf{S}_{p,x}^{-1}\mathbf{e}_v}{(2p-2v+2)(\frac{F^{(p+1)}(x)}{(p+1)!}\mathbf{e}_v^\mathrm{T}\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x})^2}\right)^{\frac{1}{2p+1}},$$

which is obtained by optimizing MSE ignoring the higher order bias term. With consistency of the preliminary estimates, it can be shown that

$$\hat{h}_{p,v}(x) = \left(\frac{1}{n}\frac{(2v-1)\hat{\sigma}_{p,v}(x)^2 n\ell^{2v-1}}{(2p-2v+2)(v!\frac{\hat{F}^{(p+1)}(x)}{(p+1)!}\mathbf{e}_v^\mathrm{T}\mathbf{S}_{p,x}^{-1}\mathbf{c}_{p,x})^2}\right)^{\frac{1}{2p+1}}\{1 + o_\mathrm{p}(1)\}.$$

Apply the consistency assumption of the preliminary estimates again, one can easily show that $\hat{h}_{p,v}(x)$ is consistent both in rate and constant.

A similar argument can be made for the other case, and is omitted here. ■

III.9.6 Proof of Corollary III.1

This follows from the previous lemmas and verifying the Lindeberg condition. See also the proof of Lemma III.5, Theorem III.1 and Theorem III.2. ■

III.9.7 Proof of Corollary III.2

This follows from the previous lemmas and verifying the Lindeberg condition. See also the proof of Lemma III.5, Theorem III.1 and Theorem III.2. ■

III.9.8 Proof of Lemma III.3

A generic element of the matrix $\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h$ takes the form:

$$\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right), \qquad 0 \le s \le 2p.$$

Then we compute the expectation:

$$\mathbb{E}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right]$$
$$= \int_{x_{\mathtt{L}}}^{x_{\mathtt{U}}} \frac{1}{h}\left(\frac{u - x}{h}\right)^s K\left(\frac{u - x}{h}\right) f(u)\mathrm{d}u = \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} v^s K\left(v\right) f(x + vh)\mathrm{d}v$$
$$= \int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} v^s K\left(v\right) f(x + vh)\mathrm{d}v,$$

hence for x in the interior,

$$\mathbb{E}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = f(x)\int_{\mathbb{R}} \mathbf{r}_p(v)\mathbf{r}_p(v)^{\mathrm{T}}K(v)\mathrm{d}v + o(1),$$

and for $x = x_{\mathtt{L}} + ch$ with $c \in [0, 1]$,

$$\mathbb{E}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = f(x_{\mathtt{L}})\int_{-c}^{\infty} \mathbf{r}_p(v)\mathbf{r}_p(v)^{\mathrm{T}}K(v)\mathrm{d}v + o(1),$$

and for $x = x_{\mathtt{U}} - ch$ with $c \in [0, 1]$,

$$\mathbb{E}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = f(x_{\mathtt{U}})\int_{-\infty}^{c} \mathbf{r}_p(v)\mathbf{r}_p(v)^{\mathrm{T}}K(v)\mathrm{d}v + o(1),$$

provided that $F \in \mathcal{C}^1$.

The variance satisfies

$$\mathbb{V}\left[\frac{1}{n}\sum_i \frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right] = \frac{1}{n}\mathbb{V}\left[\frac{1}{h}\left(\frac{x_i - x}{h}\right)^s K\left(\frac{x_i - x}{h}\right)\right]$$
$$\le \frac{1}{n}\mathbb{E}\left[\frac{1}{h^2}\left(\frac{x_i - x}{h}\right)^{2s} K\left(\frac{x_i - x}{h}\right)^2\right] = O\left(\frac{1}{nh}\right),$$

provided that $F \in \mathcal{C}^1$. ■

III.9.9 Proof of Lemma III.4

First consider the smoothing bias. The leading term can be easily obtain by taking expectation together with Taylor expansion of F to power $p+1$. The variance of this term has order $n^{-1}h^{-1}h^{2p+2}$, which gives the residual estimate $o_{\mathrm{p}}(h^{p+1})$ since it is assumed that $nh \to \infty$.

Next for the leave-in bias, note that it has expectation of order n^{-1}, and variance of order $n^{-3}h^{-1}$, hence overall this term of order $O_{\mathrm{p}}(n^{-1})$. ■

III.9.10 Proof of Lemma III.5

We first compute the variance. Note that

$$
\begin{aligned}
&\int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p\left(u\right)\Big(\tilde{F}(x+hu)-F(x+hu)\Big)K(u)f(x+hu)\mathrm{d}u\\
&=\frac{1}{n}\int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p\left(u\right)\Big(\mathbb{1}[x_i\leq x+hu]-F(x+hu)\Big)K(u)f(x+hu)\mathrm{d}u,
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{V}\left[\int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p\left(u\right)\Big(\mathbb{1}[x_i\leq x+hu]-F(x+hu)\Big)K(u)f(x+hu)\mathrm{d}u\right]\\
&=\iint_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}}K(u)K(v)f(x+hu)f(x+hv)\\
&\qquad\left[\int_{\mathbb{R}}\left(\mathbb{1}[t\leq x+hu]-F(x+hu)\right)\left(\mathbb{1}[t\leq x+hv]-F(x+hv)\right)f(t)\mathrm{d}t\right]\mathrm{d}u\mathrm{d}v\\
&=\iint_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}}K(u)K(v)f(x+hu)f(x+hv)\\
&\qquad\Big(F(x+h(u\wedge v))-F(x+hu)F(x+hv)\Big)\mathrm{d}u\mathrm{d}v. \qquad \text{(I)}
\end{aligned}
$$

We first consider the interior case, where the above reduces to:

$$
\begin{aligned}
&(\mathrm{I})_{\text{interior}}\\
&=\iint_{\mathbb{R}}\mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}}K(u)K(v)f(x)^2\Big(F(x)-F(x)^2\Big)\mathrm{d}u\mathrm{d}v\\
&\qquad+h\iint_{\mathbb{R}}(u\wedge v)\mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}}K(u)K(v)f(x)^3\mathrm{d}u\mathrm{d}v
\end{aligned}
$$

$$
\begin{aligned}
&\qquad - h \iint_{\mathbb{R}} (u+v)\mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v)f(x)^3F(x)\mathrm{d}u\mathrm{d}v \\
&\qquad + h \iint_{\mathbb{R}} (u+v)\mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v)f(x)F^{(2)}(x)\Big(F(x)-F(x)^2\Big)\mathrm{d}u\mathrm{d}v + o(h) \\
&= f(x)^2\Big(F(x)-F(x)^2\Big)\mathbf{S}_{p,x}\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{p,x} \\
&\qquad - hf(x)^3F(x)\mathbf{S}_{p,x}(\mathbf{e}_1\mathbf{e}_0^{\mathrm{T}}+\mathbf{e}_0\mathbf{e}_1^{\mathrm{T}})\mathbf{S}_{p,x} \\
&\qquad + hf(x)F^{(2)}(x)\Big(F(x)-F(x)^2\Big)\mathbf{S}_{p,x}(\mathbf{e}_1\mathbf{e}_0^{\mathrm{T}}+\mathbf{e}_0\mathbf{e}_1^{\mathrm{T}})\mathbf{S}_{p,x} \\
&\qquad + hf(x)^3\mathbf{\Gamma}_{p,x} + o(h).
\end{aligned}
$$

For $x = x_{\mathtt{L}} + hc$ with $c \in [0,1)$ in the lower boundary region,

$$
\begin{aligned}
&(\mathrm{I})_{\text{lower boundary}} \\
&= h \iint_{\mathbb{R}} (u \wedge v + c)\mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v)f(x_{\mathtt{L}})^3\mathrm{d}u\mathrm{d}v + o(h) \\
&= hf(x_{\mathtt{L}})^3\left(\mathbf{\Gamma}_{p,x} + c\mathbf{S}_{p,x}\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{p,x}\right) + o(h).
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&(\mathrm{I})_{\text{upper boundary}} \\
&= h \iint_{\mathbb{R}} (u \wedge v - c)\mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v)f(x_{\mathtt{U}})^3\mathrm{d}u\mathrm{d}v \\
&\qquad - h \iint_{\mathbb{R}} (u + v - 2c)\mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v)f(x_{\mathtt{U}})^3\mathrm{d}u\mathrm{d}v + o(h) \\
&= hf(x_{\mathtt{U}})^2f(x_{\mathtt{U}})\left(\mathbf{\Gamma}_{p,x} + c\mathbf{S}_{p,x}\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{p,x} - \mathbf{S}_{p,x}(\mathbf{e}_1\mathbf{e}_0^{\mathrm{T}}+\mathbf{e}_0\mathbf{e}_1^{\mathrm{T}})\mathbf{S}_{p,x}\right) + o(h).
\end{aligned}
$$

With the above results, it is easy to verify the variance formula, provided that we can show the asymptotic normality.

We first consider the interior case, and verify the Lindeberg condition on the fourth moment. Let $\boldsymbol{\alpha} \in \mathbb{R}^{p+1}$ be an arbitrary nonzero vector, then

$$
\begin{aligned}
&\sum_i \mathbb{E}\left(\frac{1}{\sqrt{n}}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{N}_x(f(x)\mathbf{S}_{p,x})^{-1}\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p\left(u\right)\Big(\mathbb{1}[x_i \le x+hu] - F(x+hu)\Big)K(u)f(x+hu)\mathrm{d}u\right)^4 \\
&= \frac{1}{n}\mathbb{E}\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{N}_x(f(x)\mathbf{S}_{p,x})^{-1}\int_{\frac{x_{\mathtt{L}}-x}{h}}^{\frac{x_{\mathtt{U}}-x}{h}} \mathbf{r}_p\left(u\right)\Big(\mathbb{1}[x_i \le x+hu] - F(x+hu)\Big)K(u)f(x+hu)\mathrm{d}u\right)^4 \\
&= \frac{1}{n}\iiiint_{\mathcal{A}} \prod_{j=1,2,3,4}\Big(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{N}_x(f(x)\mathbf{S}_{p,x})^{-1}\mathbf{r}_p\left(u_j\right)K(u_j)\Big)f(x+hu_j)
\end{aligned}
$$

$$\left[\int_{\mathbb{R}} \prod_{j=1,2,3,4} \Big(\mathbb{1}[t \le x + hu_j] - F(x+hu_j)\Big) f(t)\mathrm{d}t\right] \mathrm{d}u_1\mathrm{d}u_2\mathrm{d}u_3\mathrm{d}u_4$$

$$\le \frac{C}{n} \cdot \iiiint_{\mathcal{A}} \prod_{j=1,2,3,4} \Big(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{N}_x (f(x)\mathbf{S}_{p,x})^{-1}\mathbf{r}_p(u_j) K(u_j)\Big) f(x)\mathrm{d}u_1\mathrm{d}u_2\mathrm{d}u_3\mathrm{d}u_4 + O\left(\frac{1}{nh}\right),$$

where $\mathcal{A} = [\frac{x_{\mathtt{L}}-x}{h}, \frac{x_{\mathtt{U}}-x}{h}]^4 \subset \mathbb{R}^4$. The first term in the above display is asymptotically negligible, since it is takes the form $C \cdot (\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{N}_x\mathbf{e}_0)^4/n$ where the constant C depends on the DGP, and is finite. The order of the next term is $1/(nh)$, which comes from multiplying n^{-1}, h^{-2} (from the scaling matrix $\mathbf{N}_x$), and h (from linearization), hence is also negligible.

Under the assumption that $nh \to \infty$, the Lindeberg condition is verified for interior case. The same logic applies to the boundary case, whose proof is easier than the interior case, since the leading term in the calculation is identically zero for x in either the lower or upper boundary. ■

III.9.11 Proof of Lemma III.6

For $\hat{\mathbf{R}}$, we rewrite it as a second order degenerate U-statistic:

$$\hat{\mathbf{R}} = \frac{1}{n^2}\sum_{i,j;i<j} \hat{\mathbf{U}}_{ij},$$

where

$$\begin{aligned}
\hat{\mathbf{U}}_{ij} &= \mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x)\\
&\quad + \mathbf{r}_p\left(\frac{x_j - x}{h}\right)\Big(\mathbb{1}[x_i \le x_j] - F(x_j)\Big)K_h(x_j - x)\\
&\quad - \mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i - x}{h}\right)\Big(\mathbb{1}[x_j \le x_i] - F(x_i)\Big)K_h(x_i - x)\Big|x_j\right]\\
&\quad - \mathbb{E}\left[\mathbf{r}_p\left(\frac{x_j - x}{h}\right)\Big(\mathbb{1}[x_i \le x_j] - F(x_j)\Big)K_h(x_j - x)\Big|x_i\right].
\end{aligned}$$

To compute the leading term, it suffices to consider

$$\begin{aligned}
&2\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i - x}{h}\right)\mathbf{r}_p\left(\frac{x_i - x}{h}\right)^{\mathrm{T}}\left(\mathbb{1}[x_j \le x_i] - F(x_i)\right)^2 K_h(x_i - x)^2\right]\\
&= 2\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i - x}{h}\right)\mathbf{r}_p\left(\frac{x_i - x}{h}\right)^{\mathrm{T}}\Big(F(x_i) - F(x_i)^2\Big)K_h(x_i - x)^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{h}\int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p\left(v\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}}\Big(F(x+hv)-F(x+hv)^2\Big)K(v)^2 f(x+hv)\mathrm{d}v\\
&= \frac{2}{h}\int_{\frac{x_{\mathrm{L}}-x}{h}}^{\frac{x_{\mathrm{U}}-x}{h}} \mathbf{r}_p\left(v\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}}\Big(F(x)-F(x)^2\Big)K(v)^2 f(x)\mathrm{d}v + O(1)\\
&=_{\text{interior}} \frac{2}{h}f(x)\left[F(x)-F(x)^2\right]\mathbf{T}_{p,x} + O(1),\\
&=_{\text{boundary}} O(1),
\end{aligned}
$$

which closes the proof. ■

III.9.12 Proof of Lemma III.7

This resembles the proof of Lemma III.3, and we only perform the mean computation. To start,

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\mathbf{X}_h^{\mathrm{T}}\mathbf{K}_h\mathbf{X}_h\right] &= \mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)^{\mathrm{T}}\frac{1}{h}K\left(\frac{x_i-\bar{x}}{h}\right)\right]\\
= \mathbb{E}&\left[\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)^{\mathrm{T}}\frac{1}{h}K\left(\frac{x_i-\bar{x}}{h}\right)\middle| x_i<\bar{x}\right]F(\bar{x})\\
&+\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)^{\mathrm{T}}\frac{1}{h}K\left(\frac{x_i-\bar{x}}{h}\right)\middle| x_i\geq\bar{x}\right](1-F(\bar{x})).
\end{aligned}
$$

Then by Lemma III.3, the first term takes the form:

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)^{\mathrm{T}}\frac{1}{h}K\left(\frac{x_i-\bar{x}}{h}\right)\middle| x_i<\bar{x}\right]F(\bar{x})\\
&= f(\bar{x}-|x_i<\bar{x})F(\bar{x})\int_{-1}^{0}\mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(u)^{\mathrm{T}}K(u)\mathrm{d}u + O(h),
\end{aligned}
$$

where $f(\bar{x}-|x_i<\bar{x})$ is the one-sided density of x_i at the cutoff, conditional on $x_i<\bar{x}$. Alternatively, we can simplify by the fact that $f(\bar{x}|x_i<\bar{x})F(\bar{x}) = f(\bar{x}-)$. Similarly, one has

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)\mathbf{r}_p\left(\frac{x_i-\bar{x}}{h}\right)^{\mathrm{T}}\frac{1}{h}K\left(\frac{x_i-\bar{x}}{h}\right)\middle| x_i\geq\bar{x}\right](1-F(\bar{x}))\\
&= f(\bar{x}+|x_i\geq\bar{x})(1-F(\bar{x}))\int_{0}^{1}\mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(u)^{\mathrm{T}}K(u)\mathrm{d}u + O(h),
\end{aligned}
$$

and that $f(\bar{x}+|x_i \geq \bar{x})(1-F(\bar{x})) = f(\bar{x}+)$. The rest of the proof follows standard variance calculation, and is not repeated here. ■

III.9.13 Proof of Lemma III.8

This follows from Lemma III.4 by splitting the bias calculation for the two subsamples, below and above the cutoff $\bar{x}$. ■

III.9.14 Proof of Lemma III.9

To start,

$$\begin{aligned}&\int_{-1}^{1} \mathbf{r}_p\left(u\right)\Big(\tilde{F}(\bar{x}+hu) - F(\bar{x}+hu)\Big)K(u)f(\bar{x}+hu)\mathrm{d}u\\&\qquad = \frac{1}{n}\int_{-1}^{1} \mathbf{r}_p\left(u\right)\Big(\mathbb{1}[x_i \leq \bar{x}+hu] - F(\bar{x}+hu)\Big)K(u)f(\bar{x}+hu)\mathrm{d}u,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{V}\left[\int_{-1}^{1} \mathbf{r}_p\left(u\right)\Big(\mathbb{1}[x_i \leq \bar{x}+hu] - F(\bar{x}+hu)\Big)K(u)f(\bar{x}+hu)\mathrm{d}u\right]\\&= \iint_{-1}^{1} \mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv)\\&\qquad\left[\int_{\mathbb{R}} \left(\mathbb{1}[t \leq \bar{x}+hu] - F(\bar{x}+hu)\right)\left(\mathbb{1}[t \leq \bar{x}+hv] - F(\bar{x}+hv)\right)f(t)\mathrm{d}t\right]\mathrm{d}u\mathrm{d}v\\&= \iint_{-1}^{1} \mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv)\\&\qquad\Big(F(\bar{x}+h(u\wedge v)) - F(\bar{x}+hu)F(\bar{x}+hv)\Big)\mathrm{d}u\mathrm{d}v. \qquad \text{(I)}\end{aligned}$$

Now we split the integral of (I) into four regions.

$$\begin{aligned}&(u<0, v<0)\ \text{(I)}\\&= \iint_{-1}^{0} \mathbf{r}_{-,p}\left(u\right)\mathbf{r}_{-,p}\left(v\right)^{\mathrm{T}} K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv)\\&\qquad\Big(F(\bar{x}+h(u\wedge v)) - F(\bar{x}+hu)F(\bar{x}+hv)\Big)\mathrm{d}u\mathrm{d}v\end{aligned}$$

$$
\begin{aligned}
&= f(\bar{x}-)^2\Big(F(\bar{x}) - F(\bar{x})^2\Big)\mathbf{S}_{-,p}\mathbf{e}_{0,-}\mathbf{e}_{0,-}^{\mathrm{T}}\mathbf{S}_{-,p} \\
&\quad - hf(\bar{x}-)^3F(\bar{x})\mathbf{S}_{-,p}(\mathbf{e}_{1,-}\mathbf{e}_{0,-}^{\mathrm{T}} + \mathbf{e}_{0,-}\mathbf{e}_{1,-}^{\mathrm{T}})\mathbf{S}_{-,p} \\
&\quad + hf(\bar{x}-)F^{(2)}(\bar{x}-)\Big(F(\bar{x}) - F(\bar{x})^2\Big)\mathbf{S}_{-,p}(\mathbf{e}_{1,-}\mathbf{e}_{0,-}^{\mathrm{T}} + \mathbf{e}_{0,-}\mathbf{e}_{1,-}^{\mathrm{T}})\mathbf{S}_{-,p} \\
&\quad + hf(\bar{x}-)^3\mathbf{\Gamma}_{-,p} + O(h^2),
\end{aligned}
$$

and

$$
\begin{aligned}
&(u \geq 0, v \geq 0)\ (\mathrm{I}) \\
&= \iint_0^1 \mathbf{r}_{+,p}\left(u\right)\mathbf{r}_{+,p}\left(v\right)^{\mathrm{T}}K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv) \\
&\qquad \Big(F(\bar{x}+h(u\wedge v)) - F(\bar{x}+hu)F(\bar{x}+hv)\Big)\mathrm{d}u\mathrm{d}v \\
&= f(\bar{x}+)^2\Big(F(\bar{x}) - F(\bar{x})^2\Big)\mathbf{S}_{+,p}\mathbf{e}_{0,+}\mathbf{e}_{0,+}^{\mathrm{T}}\mathbf{S}_{+,p} \\
&\quad - hf(\bar{x}+)^3F(\bar{x})\mathbf{S}_{+,p}(\mathbf{e}_{1,+}\mathbf{e}_{0,+}^{\mathrm{T}} + \mathbf{e}_{0,+}\mathbf{e}_{1,+}^{\mathrm{T}})\mathbf{S}_{+,p} \\
&\quad + hf(\bar{x}+)F^{(2)}(\bar{x}+)\Big(F(\bar{x}) - F(\bar{x})^2\Big)\mathbf{S}_{+,p}(\mathbf{e}_{1,+}\mathbf{e}_{0,+}^{\mathrm{T}} + \mathbf{e}_{0,+}\mathbf{e}_{1,+}^{\mathrm{T}})\mathbf{S}_{+,p} \\
&\quad + hf(\bar{x}+)^3\mathbf{\Gamma}_{+,p} + O(h^2),
\end{aligned}
$$

and

$$
\begin{aligned}
&(u < 0, v \geq 0)\ (\mathrm{I}) \\
&= \iint_{[-1,0]\times[0,1]} \mathbf{r}_{-,p}\left(u\right)\mathbf{r}_{+,p}\left(v\right)^{\mathrm{T}}K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv) \\
&\qquad F(\bar{x}+hu)\Big(1 - F(\bar{x}+hv)\Big)\mathrm{d}u\mathrm{d}v \\
&= \left[\int_{-1}^0 \mathbf{r}_{-,p}\left(u\right)K(u)f(\bar{x}+hu)F(\bar{x}+hu)\mathrm{d}u\right] \\
&\qquad \left[\int_0^1 \mathbf{r}_{+,p}\left(v\right)^{\mathrm{T}}K(v)f(\bar{x}+hv)\Big(1 - F(\bar{x}+hv)\Big)\mathrm{d}v\right] \\
&= \Big[f(\bar{x}-)F(\bar{x})\mathbf{S}_{-,p}\mathbf{e}_{0,-} + h\Big(f(\bar{x}-)^2 + F^{(2)}(\bar{x}-)F(\bar{x})\Big)\mathbf{S}_{-,p}\mathbf{e}_{1,-} + O(h^2)\Big] \\
&\qquad \Big[f(\bar{x}+)(1-F(\bar{x}))\mathbf{S}_{+,p}\mathbf{e}_{0,+} + h\Big(-f(\bar{x}+)^2 + F^{(2)}(\bar{x}+)(1-F(\bar{x}))\Big)\mathbf{S}_{+,p}\mathbf{e}_{1,+} + O(h^2)\Big]^{\mathrm{T}},
\end{aligned}
$$

and

$$
\begin{aligned}
&(u \geq 0, v < 0) \text{ (I)} \\
&= \iint_{[0,1]\times[-1,0]} \mathbf{r}_{-,p}\,(u)\,\mathbf{r}_{+,p}\,(v)^{\mathrm{T}}\,K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv) \\
&\qquad F(\bar{x}+hv)\Big(1-F(\bar{x}+hu)\Big)\mathrm{d}u\mathrm{d}v \\
&= \left[\int_0^1 \mathbf{r}_{+,p}\,(u)\,K(u)f(\bar{x}+hu)(1-F(\bar{x}+hu))\mathrm{d}u\right] \\
&\qquad \left[\int_{-1}^0 \mathbf{r}_{-,p}\,(v)^{\mathrm{T}}\,K(v)f(\bar{x}+hv)F(\bar{x}+hv)\Big)\mathrm{d}v\right]
\end{aligned}
$$

$$
\begin{aligned}
= &\Big[f(\bar{x}+)(1-F(\bar{x}))\mathbf{S}_{+,p}\mathbf{e}_{0,+} + h\Big(-f(\bar{x}+)^2 + F^{(2)}(\bar{x}+)(1-F(\bar{x}))\Big)\mathbf{S}_{+,p}\mathbf{e}_{1,+} + O(h^2)\Big] \\
&\qquad \Big[f(\bar{x}-)F(\bar{x}-)\mathbf{S}_{-,p}\mathbf{e}_{0,-} + h\Big(f(\bar{x}-)^2 + F^{(2)}(\bar{x}-)F(\bar{x})\Big)\mathbf{S}_{-,p}\mathbf{e}_{1,-} + O(h^2)\Big]^{\mathrm{T}}.
\end{aligned}
$$

Let $\mathbf{S}_{-,p}^{-1}$ and $\mathbf{S}_{+,p}^{-1}$ be the Moore–Penrose inverse of $\mathbf{S}_{-,p}$ and $\mathbf{S}_{+,p}$, respectively. Then

$$
\begin{aligned}
&\mathbb{V}\left[(\mathbf{e}_{1,+}-\mathbf{e}_{1,-})^{\mathrm{T}}\sqrt{\frac{n}{h}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})^{-1}\hat{\mathbf{L}}\right] \\
&= f(\bar{x}-)\mathbf{e}_{1,-}^{\mathrm{T}}\mathbf{S}_{-,p}^{-1}\boldsymbol{\Gamma}_{-,p}\mathbf{S}_{-,p}^{-1}\mathbf{e}_{1,-} + f(\bar{x}+)\mathbf{e}_{1,+}^{\mathrm{T}}\mathbf{S}_{+,p}^{-1}\boldsymbol{\Gamma}_{+,p}\mathbf{S}_{+,p}^{-1}\mathbf{e}_{1,+} + O(h).
\end{aligned}
$$

■

III.9.15 Proof of Lemma III.10

This follows from Lemma III.6 by splitting the bias calculation for the two subsamples, below and above the cutoff $\bar{x}$. ■

III.9.16 Proof of Lemma III.11

This follows from Lemma III.3 by splitting the bias calculation for the two subsamples, below and above the cutoff $\bar{x}$. See also the proof of Lemma III.7. ■

III.9.17 Proof of Lemma III.12

This follows from Lemma III.4 by splitting the bias calculation for the two subsamples, below and above the cutoff $\bar{x}$. ■

III.9.18 Proof of Lemma III.13

To start,

$$\begin{aligned}
&\int_{-1}^{1} \mathbf{r}_p\left(u\right) \Big(\tilde{F}(\bar{x}+hu) - F(\bar{x}+hu)\Big) K(u) f(\bar{x}+hu)\mathrm{d}u \\
&= \frac{1}{n}\int_{-1}^{1} \mathbf{r}_p\left(u\right) \Big(\mathbb{1}[x_i \le \bar{x}+hu] - F(\bar{x}+hu)\Big) K(u) f(\bar{x}+hu)\mathrm{d}u,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{V}\left[\int_{-1}^{1} \mathbf{r}_p\left(u\right) \Big(\mathbb{1}[x_i \le \bar{x}+hu] - F(\bar{x}+hu)\Big) K(u) f(\bar{x}+hu)\mathrm{d}u\right] \\
&= \iint_{-1}^{1} \mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v) f(\bar{x}+hu) f(\bar{x}+hv) \\
&\qquad \left[\int_{\mathbb{R}} \left(\mathbb{1}[t \le \bar{x}+hu] - F(\bar{x}+hu)\right)\left(\mathbb{1}[t \le \bar{x}+hv] - F(\bar{x}+hv)\right) f(t)\mathrm{d}t\right] \mathrm{d}u\mathrm{d}v \\
&= \iint_{-1}^{1} \mathbf{r}_p\left(u\right)\mathbf{r}_p\left(v\right)^{\mathrm{T}} K(u)K(v) f(\bar{x}+hu) f(\bar{x}+hv) \qquad \text{(I)} \\
&\qquad \Big(F(\bar{x}+h(u\wedge v)) - F(\bar{x}+hu)F(\bar{x}+hv)\Big)\mathrm{d}u\mathrm{d}v.
\end{aligned}$$

Now we split the integral of (I) into four regions.

$$\begin{aligned}
&(u<0, v<0)\ \text{(I)} \\
&= \iint_{-1}^{0} \mathbf{r}_{-,p}\left(u\right)\mathbf{r}_{-,p}\left(v\right)^{\mathrm{T}} K(u)K(v) f(\bar{x}+hu) f(\bar{x}+hv) \\
&\qquad \Big(F(\bar{x}+h(u\wedge v)) - F(\bar{x}+hu)F(\bar{x}+hv)\Big)\mathrm{d}u\mathrm{d}v \\
&= f(\bar{x}-)^2\Big(F(\bar{x}) - F(\bar{x})^2\Big)\mathbf{S}_{-,p}\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{-,p} \\
&\qquad - hf(\bar{x}-)^3 F(\bar{x})\mathbf{S}_{-,p}(\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}} + \mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}})\mathbf{S}_{-,p} \\
&\qquad + hf(\bar{x}-)F^{(2)}(\bar{x})\Big(F(\bar{x}) - F(\bar{x})^2\Big)\mathbf{S}_{-,p}(\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}} + \mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}})\mathbf{S}_{-,p} \\
&\qquad + hf(\bar{x}-)^3\mathbf{\Gamma}_{-,p} + O(h^2),
\end{aligned}$$

and

$$
\begin{aligned}
&(u \geq 0, v \geq 0)\ (\mathrm{I}) \\
&= \iint_0^1 \mathbf{r}_{+,p}\left(u\right)\mathbf{r}_{+,p}\left(v\right)^{\mathrm{T}} K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv) \\
&\qquad \Big(F(\bar{x}+h(u\wedge v)) - F(\bar{x}+hu)F(\bar{x}+hv)\Big)\mathrm{d}u\mathrm{d}v \\
&= f(\bar{x}+)^2\Big(F(\bar{x})-F(\bar{x})^2\Big)\mathbf{S}_{+,p}\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{+,p} \\
&\qquad - hf(\bar{x}+)^3F(\bar{x})\mathbf{S}_{+,p}(\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}+\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}})\mathbf{S}_{+,p} \\
&\qquad + hf(\bar{x}+)F^{(2)}(\bar{x})\Big(F(\bar{x})-F(\bar{x})^2\Big)\mathbf{S}_{+,p}(\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}+\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}})\mathbf{S}_{+,p} \\
&\qquad + hf(\bar{x}+)^3\mathbf{\Gamma}_{+,p}+O(h^2),
\end{aligned}
$$

and

$$
\begin{aligned}
&(u < 0, v \geq 0)\ (\mathrm{I}) \\
&= \iint_{[-1,0]\times[0,1]} \mathbf{r}_{-,p}\left(u\right)\mathbf{r}_{+,p}\left(v\right)^{\mathrm{T}} K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv) \\
&\qquad F(\bar{x}+hu)\Big(1-F(\bar{x}+hv)\Big)\mathrm{d}u\mathrm{d}v \\
&= \left[\int_{-1}^0 \mathbf{r}_{-,p}\left(u\right)K(u)f(\bar{x}+hu)F(\bar{x}+hu)\mathrm{d}u\right] \\
&\qquad \left[\int_0^1 \mathbf{r}_{+,p}\left(v\right)^{\mathrm{T}} K(v)f(\bar{x}+hv)\Big(1-F(\bar{x}+hv)\Big)\mathrm{d}v\right] \\
&= \Big[f(\bar{x}-)F(\bar{x})\mathbf{S}_{-,p}\mathbf{e}_0 + h\Big(f(\bar{x}-)^2+F^{(2)}(\bar{x})F(\bar{x})\Big)\mathbf{S}_{-,p}\mathbf{e}_{1,-}+O(h^2)\Big] \\
&\qquad\qquad \Big[f(\bar{x}+)(1-F(\bar{x}))\mathbf{S}_{+,p}\mathbf{e}_0 + h\Big(-f(\bar{x}+)^2+F^{(2)}(\bar{x})(1-F(\bar{x}))\Big)\mathbf{S}_{+,p}\mathbf{e}_{1,+}+O(h^2)\Big]^{\mathrm{T}},
\end{aligned}
$$

and

$$
\begin{aligned}
&(u \geq 0, v < 0)\ (\mathrm{I}) \\
&= \iint_{[0,1]\times[-1,0]} \mathbf{r}_{-,p}\left(u\right)\mathbf{r}_{+,p}\left(v\right)^{\mathrm{T}} K(u)K(v)f(\bar{x}+hu)f(\bar{x}+hv) \\
&\qquad F(\bar{x}+hv)\Big(1-F(\bar{x}+hu)\Big)\mathrm{d}u\mathrm{d}v \\
&= \left[\int_0^1 \mathbf{r}_{+,p}\left(u\right)K(u)f(\bar{x}+hu)(1-F(\bar{x}+hu))\mathrm{d}u\right] \\
&\qquad \left[\int_{-1}^0 \mathbf{r}_{-,p}\left(v\right)^{\mathrm{T}} K(v)f(\bar{x}+hv)F(\bar{x}+hv)\Big)\mathrm{d}v\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \Big[f(\bar{x}+)(1-F(\bar{x}))\mathbf{S}_{+,p}\mathbf{e}_0 + h\Big(-f(\bar{x}+)^2 + F^{(2)}(\bar{x})(1-F(\bar{x}))\Big)\mathbf{S}_{+,p}\mathbf{e}_{1,+} + O(h^2)\Big] \\
&\qquad \Big[f(\bar{x}-)F(\bar{x})\mathbf{S}_{-,p}\mathbf{e}_0 + h\Big(f(\bar{x}-)^2 + F^{(2)}(\bar{x})F(\bar{x})\Big)\mathbf{S}_{-,p}\mathbf{e}_{1,-} + O(h^2)\Big]^{\mathrm{T}}.
\end{aligned}
$$

By collecting terms, one has

$$
\begin{aligned}
\text{(I)} = &\Big(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}+)\mathbf{S}_{-,p}\Big)\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\Big(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}+)\mathbf{S}_{-,p}\Big)^{\mathrm{T}} \\
&- hf(\bar{x}-)F(\bar{x})f(\bar{x}-)\mathbf{S}_{-,p}\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p}) \\
&+ h\frac{F^{(2)}(\bar{x})}{f(\bar{x}-)}F(\bar{x})(1-F(\bar{x}))f(\bar{x}-)\mathbf{S}_{-,p}\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p}) \\
\\
&- hf(\bar{x}-)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}f(\bar{x}-)\mathbf{S}_{-,p} \\
&+ h\frac{F^{(2)}(\bar{x})}{f(\bar{x}-)}F(\bar{x})(1-F(\bar{x}))(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}f(\bar{x}-)\mathbf{S}_{-,p} \\
&- hf(\bar{x}+)F(\bar{x})f(\bar{x}+)\mathbf{S}_{+,p}\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p}) \\
&+ h\frac{F^{(2)}(\bar{x})}{f(\bar{x}+)}(1-F(\bar{x}))F(\bar{x})f(\bar{x}+)\mathbf{S}_{+,p}\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p}) \\
&- hf(\bar{x}+)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}}f(\bar{x}+)\mathbf{S}_{+,p} \\
&+ h\frac{F^{(2)}(\bar{x})}{f(\bar{x}+)}F(\bar{x})(1-F(\bar{x}))(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}}f(\bar{x}+)\mathbf{S}_{+,p} \\
&+ hf(\bar{x}-)f(\bar{x}-)\mathbf{S}_{-,p}\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}f(\bar{x}+)\mathbf{S}_{+,p} \\
&+ hf(\bar{x}-)f(\bar{x}+)\mathbf{S}_{+,p}\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}f(\bar{x}-)\mathbf{S}_{-,p} \\
&+ h(f(\bar{x}+)^3\mathbf{\Gamma}_{+,p} + f(\bar{x}-)^3\mathbf{\Gamma}_{-,p}).
\end{aligned}
$$

Next, we note that

$$
\mathbf{S}_{+,p}\mathbf{e}_{1,-} = \mathbf{S}_{-,p}\mathbf{e}_{1,+} = \mathbf{0},
$$

which implies

$$
\begin{aligned}
\text{(I)} = &\Big(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}+)\mathbf{S}_{-,p}\Big)\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\Big(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}+)\mathbf{S}_{-,p}\Big)^{\mathrm{T}} \\
&- hf(\bar{x}-)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p}) \\
&+ h\frac{F^{(2)}(\bar{x})}{f(\bar{x}-)}F(\bar{x})(1-F(\bar{x}))(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p}) \\
&- hf(\bar{x}-)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p} + f(\bar{x}-)\mathbf{S}_{-,p})
\end{aligned}
$$

$$\begin{aligned}
&+h\frac{F^{(2)}(\bar{x})}{f(\bar{x}-)}F(\bar{x})(1-F(\bar{x}))(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&-hf(\bar{x}+)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+h\frac{F^{(2)}(\bar{x})}{f(\bar{x}+)}(1-F(\bar{x}))F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&-hf(\bar{x}+)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+h\frac{F^{(2)}(\bar{x})}{f(\bar{x}+)}F(\bar{x})(1-F(\bar{x}))(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+hf(\bar{x}-)(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}f(\bar{x}+)\mathbf{S}_{+,p}\\
&+hf(\bar{x}-)f(\bar{x}+)\mathbf{S}_{+,p}\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+h(f(\bar{x}+)^3\mathbf{\Gamma}_{+,p}+f(\bar{x}-)^3\mathbf{\Gamma}_{-,p}).
\end{aligned}$$

Next note that

$$\begin{aligned}
\mathbf{\Gamma}_{-,p}&=\iint_{[-1,0]^2}(u\wedge v)\mathbf{r}_{-,p}(u)\mathbf{r}_{-,p}(v)^{\mathrm{T}}K(u)K(v)dudv\\
&=\iint_{[0,1]^2}((-u)\wedge(-v))\mathbf{r}_{-,p}(-u)\mathbf{r}_{-,p}(-v)^{\mathrm{T}}K(u)K(v)dudv\\
&=\iint_{[0,1]^2}(u\wedge v-u-v)\mathbf{\Psi}\mathbf{r}_{+,p}(u)\mathbf{r}_{+,p}(v)^{\mathrm{T}}\mathbf{\Psi}K(u)K(v)dudv\\
&=\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi}-\mathbf{\Psi}\mathbf{S}_{+,p}\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{+,p}\mathbf{\Psi}-\mathbf{\Psi}\mathbf{S}_{+,p}\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}}\mathbf{S}_{+,p}\mathbf{\Psi}\\
&=\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi}+\mathbf{S}_{-,p}\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}\mathbf{S}_{-,p}+\mathbf{S}_{-,p}\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}\mathbf{S}_{-,p},
\end{aligned}$$

then

$$\begin{aligned}
\text{(I)}=&\Big(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}+)\mathbf{S}_{-,p}\Big)\mathbf{e}_0\mathbf{e}_0^{\mathrm{T}}\Big(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}+)\mathbf{S}_{-,p}\Big)^{\mathrm{T}}\\
&-hf(\bar{x}-)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+h\frac{F^{(2)}(\bar{x})}{f(\bar{x}-)}F(\bar{x})(1-F(\bar{x}))(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&-hf(\bar{x}-)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+h\frac{F^{(2)}(\bar{x})}{f(\bar{x}-)}F(\bar{x})(1-F(\bar{x}))(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&-hf(\bar{x}+)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})
\end{aligned}$$

$$
\begin{aligned}
&+ h\frac{F^{(2)}(\bar{x})}{f(\bar{x}+)}(1-F(\bar{x}))F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,+}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&- hf(\bar{x}+)F(\bar{x})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+ h\frac{F^{(2)}(\bar{x})}{f(\bar{x}+)}F(\bar{x})(1-F(\bar{x}))(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,+}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+ hf(\bar{x}-)(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_{1,-}\mathbf{e}_0^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+ hf(\bar{x}-)(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\mathbf{e}_0\mathbf{e}_{1,-}^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})\\
&+ h(f(\bar{x}+)^3\mathbf{\Gamma}_{+,p}+f(\bar{x}-)^3\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{V}\left[(\mathbf{e}_{1,+}-\mathbf{e}_{1,-})^{\mathrm{T}}\sqrt{\frac{n}{h}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})^{-1}\hat{\mathbf{L}}\right]\\
&=(\mathbf{e}_{1,+}-\mathbf{e}_{1,-})^{\mathrm{T}}(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})^{-1}(f(\bar{x}+)^3\mathbf{\Gamma}_{+,p}\\
&+f(\bar{x}-)^3\mathbf{\Psi}\mathbf{\Gamma}_{+,p}\mathbf{\Psi})(f(\bar{x}+)\mathbf{S}_{+,p}+f(\bar{x}-)\mathbf{S}_{-,p})^{-1}(\mathbf{e}_{1,+}-\mathbf{e}_{1,-})+O(h).
\end{aligned}
$$

■

III.9.19 Proof of Lemma III.14

This follows from Lemma III.6 by splitting the bias calculation for the two subsamples, below and above the cutoff $\bar{x}$. ■

BIBLIOGRAPHY

Abadie, Alberto, (2003). "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, *113*(2), pp. 231–263.

Abadie, Alberto, (2005). "Semiparametric difference-in-differences estimators," *Review of Economic Studies*, *72*(1), pp. 1–19.

Abadie, Alberto and Matias D. Cattaneo, (2018). "Econometric methods for program evaluation," *Annual Review of Economics*, *10*, pp. 465–503.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens, (2000). "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish," *Review of Economic Studies*, *67*(3), pp. 499–527.

Arcones, Miguel A. and Evarist Giné, (1991). "Additions and correction to 'The bootstrap of the mean with arbitrary bootstrap sample size'," *Annals of the Institute Henri Poincaré*, *27*(4), pp. 583–595.

Athey, Susan, Guido W. Imbens, and Stefan Wager, (2018). "Approximate residual balancing: Debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B*, *80*(4), pp. 597–623.

Athreya, K. B., (1987). "Bootstrap of the mean in the infinite variance case," *Annals of Statistics*, *15*(2), pp. 724–731.

Bang, Heejung and James M. Robins, (2005). "Doubly robust estimation in missing data and causal inference models," *Biometrics*, *61*(4), pp. 962–972.

Belloni, Alexandre, Victor Chernozhukov, Denis Chetverikov, Christian Hansen, and Kengo Kato, (2018). "High-dimensional econometrics and regularized GMM," arXiv:1806.01888.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernández-Val, and Chris Hansen, (2017). "Program evaluation and causal inference with high-dimensional data," *Econometrica*, *85*(1), pp. 233–298.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, (2014). "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies*, *81*(2), pp. 608–650.

Berkes, Istvan, Lajos Horváth, and Johannes Schauer, (2012). "Asymptotic behavior of trimmed sums," *Stochastics and Dynamics*, *12*(1), pp. 12–29.

Björklund, Anders and Robert Moffitt, (1987). "The estimation of wage gains and welfare gains in self-selection models," *Review of Economics and Statistics*, *69*(1), pp. 42–49.

Busso, Matias, John DiNardo, and Justin McCrary, (2014). "New evidence on the finite sample properties of propensity score matching and reweighting estimators," *Review of Economics and Statistics*, *96*(5), pp. 885–897.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell, (2018). “On the effect of bias estimation on coverage accuracy in nonparametric inference,” *Journal of the American Statistical Association*, *113*(522), pp. 767–779.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik, (2014). “Robust nonparametric confidence intervals for regression-discontinuity designs,” *Econometrica*, *82*(6), pp. 2295–2326.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik, (2015). “Optimal data-driven regression discontinuity plots,” *Journal of the American Statistical Association*, *110*(512), pp. 1753–1769.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil, (2011). “Estimating marginal returns to education,” *American Economic Review*, *101*(6), pp. 2754–2781.

Cattaneo, Matias D., (2010). “Efficient semiparametric estimation of multi-valued treatment effects under ignorability,” *Journal of Econometrics*, *155*(2), pp. 138–154.

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson, (2010). “Robust data-driven inference for density-weighted average derivatives,” *Journal of the American Statistical Association*, *105*(491), pp. 1070–1083.

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson, (2013). “Generalized jackknife estimators of weighted average derivatives (with comments and rejoinder),” *Journal of the American Statistical Association*, *108*(504), pp. 1243–1256.

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson, (2014a). “Bootstrapping density-weighted average derivatives,” *Econometric Theory*, *30*(6), pp. 1135–1164.

Cattaneo, Matias D., Richard K. Crump, and Michael Jansson, (2014b). “Small bandwidth asymptotics for density-weighted average derivatives,” *Econometric Theory*, *30*(1), pp. 176–200.

Cattaneo, Matias D. and Juan Carlos Escanciano, (2017). *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*: Emerald Group Publishing.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik, (2018a). *A Practical Introduction to Regression Discontinuity Designs: Volume I*: Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik, (2018b). *A Practical Introduction to Regression Discontinuity Designs: Volume II*: Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

Cattaneo, Matias D. and Michael Jansson, (2018). “Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency,” *Econometrica*, *86*(3), pp. 955–995.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, (2018a). “Manipulation testing based on density discontinuity,” *Stata Journal*, *18*(1), pp. 234–261.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, (2018b). “Two-step estimation and inference with possibly many included covariates,” *Review of Economic Studies*, forthcoming.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, (2019a). “Local regression distribution estimators,” Working paper.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, (2019b). “Simple local polynomial density estimators,” Working paper.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, (2019c). “`lpdensity`: Local polynomial density estimation and inference,” Working paper.

Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey, (2018a). “Alternative asymptotics and the partially linear model with many regressors,” *Econometric Theory*, *34*(2), pp. 277–301.

Cattaneo, Matias D., Michael Jansson, and Whitney K. Newey, (2018b). “Inference in linear regression models with many covariates and heteroskedasticity,” *Journal of the American Statistical Association*, forthcoming.

Cattaneo, Matias D., Rocío Titiunik, and Gonzalo Vazquez-Bare, (2017). “Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality,” *Journal of Policy Analysis and Management*, *36*(3), pp. 643–681.

Chatterjee, Samprit and Ali S. Hadi, (1988). *Sensitivity Analysis in Linear Regression*, New York: Wiley.

Chaudhuri, Saraswata and Jonathan B. Hill, (2016). “Heavy tail robust estimation and inference for average treatment effects,” Working paper.

Chen, Xiaohong, (2007). “Large sample sieve estimation of semi-nonparametric models,” In J. J. Heckman and E. Leamer (eds.) *Handbook of Econometrics, Volume VI*, New York: Elsevier Science B.V. pp. 5549–5632.

Cheng, Ming-Yen, Jianqing Fan, and J. S. Marron, (1997). “On automatic boundary corrections,” *Annals of Statistics*, *25*(4), pp. 1691–1708.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney K. Newey, and James M. Robins, (2018). “Double/debiased machine learning for treatment and structural parameters,” *Econometrics Journal*, *21*(1), pp. C1–C68.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins, (2018). “Locally robust semiparametric estimation,” arXiv:1608.00033.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik, (2009). “Dealing with limited overlap in estimation of average treatment effects,” *Biometrika*, *96*(1), pp. 187–199.

Csörgő, Sándor, Erich Haeusler, and David M. Mason, (1988). “The asymptotic distribution of trimmed sums,” *Annals of Probability*, *16*(2), pp. 672–699.

Dehejia, Rajeev H. and Sadek Wahba, (1999). “Causal effects in nonexperimental studies: Reevaluating the evaluations of training programs,” *Journal of the American Statistical Association*, *94*(448), pp. 1053–1062.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux, (1996). “Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach,” *Econometrica*, *64*(5), pp. 1001–1044.

El Karoui, Noureddine, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu, (2013). “On robust regression with high-dimensional predictors,” *Proceedings of the National Academy of Sciences*, *110*(36), pp. 14557–14562.

Fan, Jianqing and Irene Gijbels, (1996). *Local Polynomial Modelling and Its Applications*: Chapman & Hall/CRC.

Fan, Jianqing, Jinchi Lv, and Lei Qi, (2011). “Sparse high-dimensional models in economics,” *Annual Review of Economics*, *3*, pp. 291–317.

Farrell, Max H., (2015). “Robust inference on average treatment effects with possibly more covariates than observations,” *Journal of Econometrics*, *189*(1), pp. 1–23.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra, (2018). “Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands,” arXiv:1809.09953.

Feller, William, (1991). *An Introduction to Probability Theory and Its Applications (Volume II)*: John Wiley & Sons, 2nd edition.

Fernandez-Val, Ivan and Martin Weidner, , “Fixed effects estimation of large-T panel data models,” *Annual Review of Economics*, *10*.

Gelman, Andrew and Guido W. Imbens, (2018). “Why high-order polynomials should not be used in regression discontinuity designs,” *Journal of Business & Economic Statistics*, forthcoming.

Hahn, Jinyong and Geert Ridder, (2013). “Asymptotic variance of semiparametric estimators with generated regressors,” *Econometrica*, *81*(1), pp. 315–340.

Hahn, Marjorie G. and Daniel C. Weiner, (1992). “Asymptotic behavior of self-normalized trimmed sums: Nonnormal limits,” *Annals of Probability*, *20*(1), pp. 455–482.

Heckman, James J., Sergio Urzua, and Edward J. Vytlacil, (2006). "Understanding instrumental variables in models with essential heterogeneity," *Review of Economics and Statistics*, *88*(3), pp. 389–432.

Heckman, James J. and Edward J. Vytlacil, (2005). "Structural equations, treatment effects and econometric policy evaluation," *Econometrica*, *73*(3), pp. 669–738.

Heiler, Phillip and Ekaterina Kazak, (2018). "Valid inference for treatment effect parameters under irregular identification and extreme propensity scores," Working paper.

Hernán, Miguel A. and James M. Robins, (2018). *Causal Inference*: Chapman & Hall/CRC, forthcoming.

Hill, Jonathan B. and Eric Renault, (2012). "Variance targeting for heavy tailed time series," Working paper.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, (2003). "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, *71*(4), pp. 1161–1189.

Hong, Han, Michael Leung, and Jessie Li, (2018). "Inference on finite population treatment effects under limited overlap," ssrn.3128546.

Horowitz, Joel L., (2001). "The bootstrap," In J. J. Heckman and E. Leamer (eds.) *Handbook of Econometrics, Volume V*: Elsevier Science B.V. pp. 3159–3228.

Ichimura, Hidehiko and Petra E. Todd, (2007). "Implementing nonparametric and semiparametric estimators," In J. J. Heckman and E. Leamer (eds.) *Handbook of Econometrics, Volume VIB*: Elsevier Science B.V. pp. 5370–5468.

Imbens, Guido W. and Joshua D. Angrist, (1994). "Identification and estimation of local average treatment effects," *Econometrica*, *62*(2), pp. 467–475.

Imbens, Guido W. and Donald B. Rubin, (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*: Cambridge University Press.

Karunamuni, R.J. and T. Albert, (2005). "On boundary correction in kernel density estimation," *Statistical Methodology*, *2*, pp. 191–212.

Khan, Shakeeb and Elie Tamer, (2010). "Irregular identification, support conditions, and inverse weight estimation," *Econometrica*, *78*(6), pp. 2021–2042.

Kline, Patrick and Andres Santos, (2012). "Higher order properties of the wild bootstrap under misspecification," *Journal of Econometrics*, *171*(1), pp. 54–70.

Knight, Keith, (1989). "On the bootstrap of the sample mean in the infinite variance case," *Annals of Statistics*, *17*(3), pp. 1168–1175.

LaLonde, Robert J., (1986). "Evaluating the econometric evaluations of training programs with experimental data," *American Economic Review*, *76*(4), pp. 604–620.

Li, Chenchuan and Ulrich K. Müller, (2017). "Linear regression with many controls of limited explanatory power," Working paper.

Logan, B. F., C. L. Mallows, S. O. Rice, and L. A. Shepp, (1973). "Limit distributions of self-normalized sums," *Annals of Probability*, *1*(5), pp. 788–809.

Ludwig, Jens and Douglas L. Miller, (2007). "Does head start improve children's life chances? Evidence from a regression discontinuity design," *Quarterly Journal of Economics*, *122*(1), pp. 159–208.

Ma, Xinwei and Jingshen Wang, (2019). "Robust inference using inverse probability weighting," Working paper.

Mammen, Enno, (1989). "Asymptotics with increasing dimension for robust regression with applications to the bootstrap," *Annals of Statistics*, *17*(1), pp. 382–400.

Mammen, Enno, (1993). "Bootstrap and wild bootstrap for high dimensional linear models," *Annals of Statistics*, *21*, pp. 255–285.

McCrary, Justin, (2008). "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, *142*(2), pp. 698–714.

Newey, Whitney K., (1994). "The asymptotic variance of semiparametric estimators," *Econometrica*, *62*(6), pp. 1349–82.

Newey, Whitney K. and Daniel L. McFadden, (1994). "Large sample estimation and hypothesis testing," In R. F. Engle and D. L. McFadden (eds.) *Handbook of Econometrics, Volume IV*: Elsevier Science B.V. pp. 2111–2245.

Newey, Whitney K. and James M. Robins, (2018). "Cross-fitting and fast remainder rates for semiparametric estimation," arXiv:1801.09138.

Otsu, Taisuke, Ke-Li Xu, and Yukitoshi Matsushita, (2014). "Estimation and inference of discontinuity in density," *Journal of Business and Economic Statistics*, *31*(4), pp. 507–524.

Politis, Dimitris N. and Joseph P. Romano, (1994). "Large sample confidence regions based on subsamples under minimal assumptions," *Annals of Statistics*, *22*(4), pp. 2031–2050.

Politis, Dimitris N., Joseph P. Romano, and Michael Wolf, (1999). *Subsampling*: Springer.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao, (1994). "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, *89*(427), pp. 846–866.

Romano, Joseph P. and Michael Wolf, (1999). "Subsampling inference for the mean in the heavy-tailed case," *Metrika*, *50*(1), pp. 55–69.

Sasaki, Yuya and Takuya Ura, (2018). "Estimation and inference for moments of ratios with robustness against large trimming bias," arXiv:1709.00981.

Vaynman, Igor and Brendan K. Beare, (2014). "Stable limit theory for the variance targeting estimator," In Y. Chang, T. B. Fomby, and J. Y. Park (eds.) *Advances in Econometrics, Vol 33: Essays in Honor of Peter CB Phillips*: Emerald Group Publishing, pp. 639–672.

Vershynin, Roman, (2018). *High-Dimensional Probability*: Cambridge University Press.

Vytlacil, Edward J., (2002). "Independence, monotonicity, and latent index models: An equivalence result," *Econometrica*, *70*(1), pp. 331–341.

Wand, M. P. and M. C. Jones, (1995). *Kernel Smoothing*: Chapman & Hall/CRC.

Webb, Matthew D., (2014). "Reworking wild bootstrap based inference for clustered errors," Working paper.

Wooldridge, Jeffrey M., (1999). "Asymptotic properties of weighted M-estimators for variable probability samples," *Econometrica*, *67*(6), pp. 1385–1406.

Wooldridge, Jeffrey M., (2007). "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, *141*(2), pp. 1281–1301.

Yang, S. and P. Ding, (2018). "Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores," *Biometrika*, *105*(2), pp. 487–493.

Zhang, Shunpu and Rohana J. Karunamuni, (1998). "On kernel density estimation near endpoints," *Journal of Statistical Planning and Inference*, *70*(1), pp. 301–316.