The London School of Economics and Political Science

Essays on inference in Econometric models

Karun Adusumilli

A thesis submitted to the Department of Economics of the London School of Economics and Political Science for the degree of Doctor in Philosophy

May 2018

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it - see Statement of conjoint work below).

A version of Chapter 2 was published in the *Journal of the American Statistical Assosication*, Volume 112, Nunber 519, April 2017, p. 1064-1075.

I declare that my thesis consists of approximately 25,000 words.

Statement of conjoint work

I can confirm that Chapter 2 was jointly coauthored with Professor Taisuke Otsu.

In addition, I also declare that Chapter 3 was jointly coauthored with Professor Taisuke Otsu and Professor Yoon-Jae Whang (Seoul National University).

I have contributed to around 50% of the work in each case.

Abstract

This thesis contains three essays on inference in econometric models.

Chapter 1 considers the question of bootstrap inference for Propensity Score Matching. Propensity Score Matching, where the propensity scores are estimated in a first step, is widely used for estimating treatment effects. In this context, the naive bootstrap is invalid (Abadie and Imbens, 2008). This chapter proposes a novel bootstrap procedure for this context, and demonstrates its consistency. Simulations and real data examples demonstrate the superior performance of the proposed method relative to using the asymptotic distribution for inference, especially when the degree of overlap in propensity scores is poor. General versions of the procedure can also be applied to other causal effect estimators such as inverse probability weighting and propensity score subclassification, potentially leading to higher order refinements for inference in such contexts.

Chapter 2 tackles the question of inference in incomplete econometric models. In many economic and statistical applications, the observed data take the form of sets rather than points. Examples include bracket data in survey analysis, tumor growth and rock grain images in morphology analysis, and noisy measurements on the support function of a convex set in medical imaging and robotic vision. Additionally, nonparametric bounds on treatment effects under imperfect compliance can be expressed by means of random sets. This chapter develops a concept of nonparametric likelihood for random sets and its mean, known as the Aumann expectation, and proposes general inference methods by adapting the theory of empirical likelihood.

Chapter 3 considers inference on the cumulative distribution function (CDF) in the classical measurement error model. It proposes both asymptotic and bootstrap based uniform confidence bands for the estimator of the CDF under measurement error. The proposed techniques can also be used to obtain confidence bands for quantiles, and perform various CDF-based tests such as goodness-of-fit tests for parametric models of densities, two sample homogeneity tests, and tests for stochastic dominance; all for the first time under measurement error.

Acknowledgements

My deepest gratitude goes to my supervisor, Taisuke Otsu, for his invaluable support and guidance. This thesis wouldn't have been possible without his encouragement and insights.

Over the years, I have benefitted greatly from the discussions with, and comments from a number of the faculty at LSE and beyond. I would like to thank Kirill Evdokimov, Javier Hidalgo, Tatiana Komarova, Sokbae (Simon) Lee, Jörn-Steffen Pischke, Peter M. Robinson, Marcia Schafgans, Sorawoot Srisuma, Takuya Ura, and seminar participants at the Bristol Econometrics Study Group, Boston University, Georgetown University, Iowa State University, LSE, UC Berkeley, UC San Diego, University of Illinois - Urbana Champaign and University of Pennsylvania for helpful comments. I have also gained a lot from discussions with my fellow PhD 'supervisees', Hao Dong and Luke Taylor.

Special mention must also be made of my friends and fellow PhD students in 4.06: Dita Eckardt, Andres Barrios Fernandez, Friedrich Geiecke, Jay Lee, Shan Aman Rana, Claudio Schilter, Weihan Ding and Tianle Zhang. And outside the Economics department, I had the great privilege of knowing an equally incredible set of friends including Benjamin Arold, Amanda Diez-Fernandez, Sandra Gemayel, Dalia Gomez, Simran Kalra, Christian Lippl, Nelson A. Ruiz, Lucia Sanchez, Christina Alexandra Siomos, Damjan Temelkovski and Filippo Temporin. Apart from the great discussions and advice over the years, my PhD life would have been much more dull without the nights at George, MCR and beyond. You were the best friends and office mates anyone could ask for.

A great deal of thanks also go to my brother, Susheel, for patiently hearing out my research ideas and providing various suggesting for coding.

I would also like to acknowledge the financial support from the Economics Department at LSE.

I owe a lot to my parents, who are sadly not around to see this thesis finished. This work is dedicated to their memory.

Contents

List of Tables

List of Figures

Chapter 1

Bootstrap inference for Propensity Score Matching

1.1 Introduction

Inference on average treatment effects in the presence of confounding is a primary goal of many observational studies. Propensity Score Matching (PSM) is one of the most widely used methods for estimating treatment effects in such a setting. The propensity score is defined as the probability of obtaining treatment conditional on covariates. Under the assumption of selection on observables (i.e the treatment is as good as randomly assigned conditional on the covariates), Rosenbaum and Rubin (1983) show that matching on the propensity score is sufficient to remove confounding. Using the propensity score for matching reduces the dimensionality of the procedure by summarizing the information contained in the covariates in a single variable. Additionally, PSM can be flexibly combined with other strategies such as regression adjustment to further reduce the bias from the match (Abadie and Imbens, 2011; Imbens and Rubin, 2015). Such favourable properties have led to PSM becoming one of most commonly used methods for causal analysis of observational data. See for example Deheija and Wahba (1999), Heckman, Ichimura, Smith and Todd (1998), Lechner (2002) and Smith and Todd (2001) for some important applications and issues arising from its use in economics.

In practice, the propensity scores are usually estimated through a parametric first stage

based on a probit or logit model. Furthermore, to reduce the bias from the match, the number of matches is usually held fixed at small values, for example one. This introduces complications for inference since the matching function - defined as the number of times each unit is used as a match - is a highly non-linear function of the data. Abadie and Imbens (2016) show that the matching estimator under the estimated propensity score is consistent and asymptotically normal. Thus inference for the treatment effect can proceed based on a large sample approximation to the normal distribution, using the variance estimate suggested by the authors. At the same time, Abadie and Imbens (2008) show that the standard non-parametric bootstrap based on resampling fails to be consistent in this context. This is because the usual bootstrap procedure fails to reproduce the distribution of the matching function in the true sample.

In this chapter, I propose and demonstrate consistency of a bootstrap procedure for matching on the estimated propensity score. Both matching with and without replacement is considered. The proposed bootstrap is built around the concept of 'potential errors', introduced in this chapter as a general tool for causal inference. Potential errors formalize the idea that each observation can be associated with two possible error terms, corresponding to each of the potential states - treated or control - only one of which is actually realized. Thus, the variability of the estimator stems not only from the randomness of the potential errors themselves, but also from the probabilistic nature of treatment assignment, which randomly realizes one of the potential error terms. The proposed bootstrap takes both sources of randomness into account by resampling the potential errors as a pair, while also re-assigning new values for the treatments using the estimated propensity score. Implementing the procedure requires the construction of estimates of the error terms under both states. Since I only observe the errors under one of the potential states for any data point, I provide ways to impute these quantities for the other state.

The notion of potential errors is very general, and can be applied to causal effect estimators beyond propensity score matching. The exact form of the potential errors depends on both the estimator and the quantity being estimated (ATE, ATET, etc.), but a unifying

theme is that it is possible to obtain the 'error representation'[1]

$$\text{Estimator} - \text{Expected Value} = \text{Average(Realized errors)}.$$

Here, the terminology 'realized errors' refers to the observed values of the potential errors given the treatment status. For many estimators, directly resampling the realized errors suffices for valid inference, see e.g. Otsu and Rai (2017). However, such a strategy doesn't work for propensity score matching since the potential errors are functions of the estimated propensity score, which is itself a random quantity (see, Section 1.3.4). Taking the estimation of the propensity scores into account requires recreating the randomness of treatment assignment closely, since this determines the variability of the propensity scores. Doing so naturally leads to the proposed bootstrap statistic. Indeed, my bootstrap statistic is simply the average of the new realized errors - obtained after resampling the potential errors and reassigning treatments - and evaluated at propensity scores estimated from the bootstrap sample.

The proposed bootstrap can be easily extended to other causal effect estimators satisfying the error representation, for example inverse probability weighting or propensity score sub-classification (see, Section 1.6.3). Since it recreates all the sources of randomness more faithfully, it generally provides more precise inference compared to asymptotic methods or methods that only resample the realized errors. The gain in accuracy is especially pronounced when there is poor overlap between the propensity scores of the treated and control groups. Poor overlap usually occurs when there is heavy imbalance between the covariate distributions for the treated and control groups. In such situations, some observations gain disproportionate importance, for instance the few control units close to the treated units, and vice versa. The resulting causal estimate is then highly sensitive to possible switches to the treatment status of these observations. Failure to take this into account leads to severe under-estimation of the actual variance, as shown in simulations. By contrast, the proposed bootstrap is more accurate, and constitutes an attractive choice for inference when the overlap is poor.

[1]For matching estimators, this is equivalent to the martingale representation of Abadie and Imbens (2012).

I demonstrate consistency of this bootstrap procedure using Le Cam's framework of local limit experiments, applied on the bootstrap data generating process. To this end, I extend the techniques of local limit experiments previously employed by Abadie and Imbens (2016), and Andreou and Werker (2011) to obtain limiting distributions of non-smooth statistics to the setup of bootstrap inference. Thus, the techniques may be of independent theoretical interest.

The finite sample performance of the bootstrap is assessed through a number of simulations and real data examples. In almost all cases the bootstrap provides better size control than inference based on the asymptotic distribution. The results also confirm that the proposed bootstrap is particularly effective when the balance of covariates across treated and control samples is poor. Arguably, poor covariate balance is pervasive in observational studies.

The theoretical results in this chapter build on the properties of matching estimators with finite number of matches, established in an important series of papers by Abadie and Imbens (2006, 2008, 2011, 2012, 2016). When the number of matches is allowed to increase with sample size, as in the kernel matching method of Heckman, Smith and Todd (1997), the resulting estimator is asymptotically linear, and the usual non-parametric bootstrap can be employed. In the context of a fixed number of matches, Otsu and Rai (2016) propose a consistent bootstrap method for the version of nearest neighbor matching based on a distance measure (Euclidean, Mahalanobis etc.) over the full vector of covariates. The proposal of Otsu and Rai (2016) is equivalent to conditioning on both treatments and covariates, and resampling the realized errors in the error representation. However, their consistency result doesn't extend to propensity score matching because conditioning on both treatments and covariates precludes taking into account the effect of the estimation of propensity scores. Alternatives to the bootstrap that do provide consistent inference in this context include subsampling (Politis and Romano, 1994) and m-out-of-n bootstrap (Bickel, Götze and van Zwet, 2012).

1.2 Setup

The starting point of my analysis is the standard treatment effect model under selection on observables. I follow the same setup as Abadie and Imbens (2016). The aim is to estimate the effect of a binary treatment, denoted by W, on some outcome Y. A value of $W = 1$ implies the subject is treated, while $W = 0$ implies the subject hasn't received any treatment. The causal effect of the treatment is represented in the terminology of potential outcomes (Rubin, 1974). In particular, I introduce the random variables $(Y(0), Y(1))$, where $Y(0)$ denotes the potential outcome under no treatment, and $Y(1)$ denotes the potential outcome under treatment. I also have access to a set of covariates X, where $\dim(X) = k$. The goal is to estimate the average treatment effect

$$\tau = E[Y(1) - Y(0)].$$

In general, estimation of τ suffers from a missing data problem since only one of the potential outcomes is observable as the actual outcome variable, $Y = Y(W)$. To circumvent this, practitioners commonly impose the following identifying assumptions for τ:

Assumption 1. $(Y(1), Y(0))$ *is independent of* W *conditional on* X *almost surely, denoted as* $(Y(1), Y(0)) \perp\!\!\!\perp W \mid X$.

Assumption 2. (Y_i, W_i, X_i) *are i.i.d draws from the distribution of* (Y, W, X).

The first assumption is that of unconfoundedness, which implies that the treatment is as good as randomly assigned conditional on the covariates X. The second assumption implies that the potential outcome for individual i is independent of the treatment status and covariates of the other individuals. This rules out peer effects, for instance.

Define the propensity score, $p(X) = \Pr(W = 1|X)$, as the probability of being treated conditional on the covariates. Let $\bar{\mu}(w, X)$ and $\mu(w, p(X))$ denote the conditional means $E[Y|W = w, X]$ and $E[Y|W = w, p(X)]$ respectively. Additionally, let $\bar{\sigma}^2(w, X) = E[Y^2|W = w, X]$ and $\sigma^2(w, p(X)) = E[Y^2|W = w, p(X)]$ denote the conditional variances of Y given $W = w$ and X ; and that of Y given $W = w$ and $p(X)$ respectively. In a seminal paper, Rosenbaum and Rubin (1983) show that under Assumption 1, the potential

outcomes are also independent of the treatment conditional on the propensity scores, i.e $(Y(1), Y(0)) \perp\!\!\!\perp W \mid p(X)$. Thus, τ can be alternatively identified as

$$\tau = E[\mu(1, p(X)) - \mu(0, p(X))].$$

In the literature a number of propensity score matching techniques have been proposed that exploit the above characterization of τ, see e.g. Rosenbaum (2009) for a detailed survey. In this section, and for much of this chapter, I focus on matching with replacement, with a fixed number of matches for each unit, denoted by M. This is arguably the most commonly used matching procedure in economic applications. The case of matching without replacement is discussed in Section 1.6.2.

Suppose that I have a sample of N observations. The propensity score matching estimator for the average treatment effect, when matching with replacement, is defined as

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i; p(X))} Y_j \right),$$

where M is the number of matches for each unit, and $\mathcal{J}_M(i; p(X))$ is the set of matches for the individual i. In particular $\mathcal{J}_M(i; p(X))$ represents the set of M individuals from the opposite treatment arm whose propensity scores are closest to i's own, i.e,

$$\mathcal{J}_M(i; p(X)) = \{ j = 1, \ldots, N : \ W_j = 1 - W_i, \text{ and } \left(\sum_{l: W_l = 1 - W_i} \mathbb{I}_{[|p(X_i) - p(X_l)| \leq |p(X_i) - p(X_j)|]} \right) \leq M \Bigg\}.$$

Typically the value of M is taken to be quite small, for example $M = 1$, so as to reduce the bias.

The propensity scores are generally not known but have to be estimated. In this chapter, I consider parametric estimates for the propensity scores based on a generalized linear model $p(X) = F(X'\theta)$, where θ is a finite dimensional vector parameter, and $F(.)$ is a (known) link function, for instance a logistic or probit function. Let $(\mathbf{W}, \mathbf{X})$ denote the vector of treatments and covariates $(W_1, \ldots, W_N, X_1, \ldots, X_N)$. I denote the true value of θ by θ_0.

The latter is estimated through maximum likelihood as

$$\hat{\theta} = \arg\max_{\theta} L(\theta|\mathbf{W}, \mathbf{X}),$$

where

$$L(\theta|\mathbf{W}, \mathbf{X}) = \sum_{i=1}^{N} \left\{ W_i \ln F(X_i'\theta) + (1 - W_i) \ln(1 - F(X_i'\theta)) \right\},$$

denotes the log-likelihood function evaluated at θ.

Let $\mathcal{J}_M(i;\theta)$ denote the set of M closest matches to observation i for the match based on $F(X'\theta)$ as if it were the true propensity score, i.e

$$\mathcal{J}_M(i;\theta) = \{j = 1, \ldots, N : \ W_j = 1 - W_i, \text{ and } \left(\sum_{l:W_l = 1 - W_i} \mathbb{I}_{\left[\left| F(X_i'\theta) - F(X_l'\theta) \right| \leq \left| F(X_i'\theta) - F(X_j'\theta) \right| \right]} \right) \leq M \Bigg\}.$$

The matching estimator, for the match based on $F(X'\theta)$, is defined as

$$\hat{\tau}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i;\theta)} Y_j \right).$$

Let $K_M(i;\theta)$ denote the number of times observation i is used as a match based on $F(X'\theta)$, i.e

$$K_M(i;\theta) = \sum_{j=1}^{N} \mathbb{I}_{i \in \mathcal{J}_M(j;\theta)}.$$

Then an alternative way to represent $\hat{\tau}(\theta)$ is provided by the error representation

$$\hat{\tau}(\theta) - \tau - B(\theta) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i(W_i;\theta), \tag{1.1}$$

where

$$B(\theta) = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1) \cdot \left(\mu(1 - W_i, F(X_i'\theta)) - \frac{1}{M} \sum_{i \in \mathcal{J}_M(i;\theta)} \mu(1 - W_i, F(X_i'\theta)) \right)$$

denotes the bias from the match based on $F(X'\theta)$, and

$$\begin{aligned}\varepsilon_i(W_i;\theta) = &\left(\mu\left(1, F(X_i'\theta)\right) - \mu\left(0, F(X_i'\theta)\right) - \tau\right) \\ &+ (2W_i - 1)\left(1 + \frac{K_M(i;\theta)}{M}\right)\left(Y_i - \mu\left(W_i, F(X_i'\theta)\right)\right) \end{aligned} \tag{1.2}$$

denotes the effective error term for each observation. The variance is thus determined by the right hand side of equation (1.1). Consequently, this expression is of primary interest in approximating the distribution of $\hat{\tau}(\theta)$.

The matching estimator for τ based on the estimated propensity score is then given by

$$\hat{\tau} \equiv \hat{\tau}(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(2W_i - 1)\left(Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}_M(i;\hat{\theta})} Y_j\right).$$

Abadie and Imbens (2016) derive the large sample properties of the above estimator. Under some regularity conditions, they find that the bias term $B(\hat{\theta})$ converges in probability to zero at a rate faster than $\sqrt{N}$, and that $\hat{\tau}$ has an asymptotic normal distribution

$$\sqrt{N}(\hat{\tau} - \tau) \xrightarrow{d} N\left(0, \sigma^2 - c' I_{\theta_0}^{-1} c\right),$$

where σ^2 is the asymptotic variance for matching on the known propensity score,

$$c = E\left[\left\{\frac{\text{cov}[X, \mu(1,X)|F(X'\theta_0)]}{F(X'\theta_0)} + \frac{\text{cov}[X, \mu(0,X)|F(X'\theta_0)]}{1 - F(X'\theta_0)}\right\} f(X'\theta_0)\right]$$

with $f(.) \equiv F'(.)$; and for any value of θ, $I(\theta)$ denotes the information matrix evaluated at θ

$$I_\theta \equiv I(\theta) = E\left[\frac{f^2(X'\theta)}{F(X'\theta)(1 - F(X'\theta))} XX'\right].$$

The above result illustrates the well known 'Propensity Score Paradox': Matching on the estimated, as opposed to the true propensity scores, in fact reduces the asymptotic variance.

1.3 Bootstrap procedure

In this section I propose a bootstrap procedure for inference on the propensity score matching estimator. I fix the following notation: For each $w = 0, 1$, define $\mu(w, p; \theta) = E[Y(w)|F(X_i'\theta) = p]$. In what follows, I abuse notation a bit by dropping the index of $\mu(., .; \theta)$ with respect to θ when the context is clear. For $w = 0, 1$, denote[2]

$$e_{1i}(\theta) = \mu(1, F(X_i'\theta)) - \mu(0, F(X_i'\theta)) - \tau;$$
$$e_{2i}(w; \theta) = Y_i - \mu(w, F(X_i'\theta)).$$

Note that the above are distinct in general from the 'true' errors which are defined similarly but evaluated at θ_0.

I present here an informal description of the bootstrap procedure, relegating many of the formal details to the upcoming sub-sections. Given any value of θ, the pair of potential error terms for each observation i are given by

$$\varepsilon_i(w; \theta) \equiv e_{1i}(\theta) + (2w - 1)\left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M}\right) e_{2i}(w; \theta);\ w = 0, 1,$$

where $\tilde{K}_M(i; w, \theta)$ is a potential matching function, denoting the number of times observation i would have been used as a match depending on whether it is in the treated ($w = 1$) or control group ($w = 0$); see Section (1.3.2) for the formal definition of $\tilde{K}_M(i; w, \theta)$. Clearly, only one of the quantities $\varepsilon_i(w; \theta) : w = 0, 1$ is directly estimable; the other has to be imputed. Let $\hat{\varepsilon}_i(w; \theta)$ denote the estimated or imputed values of $\varepsilon_i(w; \theta)$. I then sample a set of N covariates denoted by X_j^* for $j = 1, \ldots, N$, along with the associated pair of (estimated) potential error terms $\left(\hat{\varepsilon}_{S_j^*}(0; \theta), \hat{\varepsilon}_{S_j^*}(1; \theta)\right)$, where S_j^* denotes the bootstrap index corresponding to the j-th observation in the draw. Subsequently, new bootstrap treatment values are generated using the estimated propensity scores as

$$W_j^* \sim \text{Bernoulli}(F(X_j^{*\prime}\hat{\theta})).$$

[2] I do not index τ with θ since the average treatment effect is independent of the propensity score.

Through this procedure I have sampled a new set of realized error terms given by $\varepsilon_j^*(\theta) \equiv \hat{\varepsilon}_{S_j^*}(W_j^*;\theta)$ for $j = 1,\ldots,N$. The bootstrap statistic, $T_N^*\left(\hat{\theta}^*\right)$, is the sample average of these errors, after some appropriate recentering using the function $\Xi^*\left(\hat{\theta}^*\right)$ [3], i.e

$$T_N^*\left(\hat{\theta}^*\right) \equiv \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\left\{\varepsilon_j^*\left(\hat{\theta}^*\right) - \Xi^*\left(\hat{\theta}^*\right)\right\}.$$

The errors above are being evaluated at $\hat{\theta}^*$- the bootstrap counterpart of $\hat{\theta}$ - obtained as

$$\hat{\theta}^* = \arg\max_{\theta} L(\theta|\mathbf{W}^*,\mathbf{X}^*).$$

Note that except for a negligible bias term $B(\hat{\theta})$, the construction of the bootstrap statistic closely mirrors the error representation for $\hat{\tau}(\hat{\theta}) - \tau$ given by

$$\hat{\tau}(\hat{\theta}) - \tau - B(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i(W_i;\hat{\theta}).$$

To formalize the above, I require techniques for: (i) constructing estimates of the error terms, $e_{1i}(\theta), e_{2i}(w;\theta)$, for each observation under both treated and control states; and (ii) constructing the potential matching function, $\tilde{K}_M(i;w,\theta)$, for each observation, also under both states. I now consider these in turn.

1.3.1 Constructing estimates of error terms

Denote by $\hat{\mu}(w, F(X_i'\theta))$ the estimates of the conditional expectation function $\mu(w, F(X_i'\theta))$ evaluated at $F(X_i'\theta)$. These can be obtained through non-parametric methods, for example series regression or smoothing splines. I then obtain the residuals

$$\begin{aligned}\hat{e}_{1i}(\theta) &= \hat{\mu}(1, F(X_i'\theta)) - \hat{\mu}(0, F(X_i'\theta)) - \hat{\tau}(\theta);\\ \hat{e}_{2i}(W_i;\theta) &= Y_i - \hat{\mu}(W_i, F(X_i'\theta)).\end{aligned}$$

These residuals serve as proxies for the unobserved terms $e_{1i}(\theta), e_{2i}(W_i;\theta)$, approximating the values of $e_{2i}(w;\theta)$ when $w = W_i$. For the bootstrap procedure, I also need estimates of

[3] The prcise expression for the re-centering term $\Xi^*(.)$ is provided in Section 1.3.3.

$\hat{e}_i(w;\theta)$ when $w \neq W_i$. I obtain these through a secondary matching: Define the secondary matching function as

$$\mathcal{J}_w(i) = \begin{cases} i & \text{if } W_i = w \\ \mathcal{J}_{\text{NN}}(i) & \text{if } W_i \neq w, \end{cases}$$

where $\mathcal{J}_{\text{NN}}(.)$ denotes the closest match (or nearest neighbor) to observation i from the opposite treatment arm, with the closeness measured in terms of a distance metric (Euclidean, Mahalanobis etc.) based on the full set of covariates. I then obtain:

$$\hat{e}_{2i}(w;\theta) = \hat{e}_{2\mathcal{J}_w(i)}(w;\theta).$$

The definition of $e_{2i}(w;\theta)$ proceeds in an analogous fashion.

Note that the secondary matching procedure matches on the full set of covariates, as opposed to matching on the propensity scores. This is done to preserve the conditional correlation between X and the error terms e_{1i}, e_{2i}, given the propensity scores. Indeed it is this correlation that helps drive down the asymptotic variance when using the estimated propensity score.

1.3.2 Constructing the matching function

As with the error terms, the bootstrap procedure requires values of the matching function under both treatment and non-treatment, even as only one of them is actually observed. To obtain the value of $\tilde{K}_M(i;w,\theta)$ in the opposite treatment arm (i.e when $w \neq W_i$), I employ another imputation procedure:

Let $\{\pi_1, \ldots, \pi_{q_N-1}\}$ denote the sample q_N-quantiles of $F(X'\hat{\theta})$. I let $q_N \to \infty$ as $N \to \infty$. Set $\pi_0 = 0$ and $\pi_{q_N} = 1$. Denote by $S_w(l)$, the set of all observations with $W_i = w$ in the l-th block, i.e

$$S_w(l) = \{i : \ \pi_{l-1} \leq F(X_i'\hat{\theta}) < \pi_l \cap W_i = w\},$$

and let $S(l) = S_1(l) \cup S_0(l)$. The number of untreated, treated and combined observations in the block l is given by

$$N_0(l) = \#S_0(l); \quad N_1(l) = \#S_1(l); \quad N(l) = N_0(l) + N_1(l),$$

respectively, where for any set A, $\#A$ denotes its cardinality. Suppose now that observation i falls in the block l. If $w = W_i$, I set $\tilde{K}_M(i; w, \theta) = K_M(i; \theta)$. If however $w \neq W_i$, I set $\tilde{K}_M(i; w, \theta)$ to the value $K_M(j; \theta)$, where j is drawn at random from the $S_w(l)$. Formally, denoting by $l(i)$ the block in which observation i resides, I obtain

$$\tilde{K}_M(i; w, \theta) = \begin{cases} K_M(i; \theta) & \text{if } w = W_i \\ \sum_{j \in S_w(l(i))} \{M_j(i) K_M(j; \theta)\} & \text{if } w \neq W_i, \end{cases}$$

where for each i, $\{M_j(i) : j \in S_w(l(i))\} \equiv \mathbf{M}(i)$ is a multinomial random vector with a single draw on $N_w(l(i))$ equal probability cells. These multinomial random variables are drawn independently for each observation i.

Based on these constructions I can define a combined error term excluding the effect of heterogeneity (i.e excluding $e_{1i}(\theta)$) as

$$\hat{\nu}_i(w; \theta) = \left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M}\right) \hat{e}_{2\mathcal{J}_w(i)}(w; \theta).$$

Thus the estimated potential errors are obtained as

$$\hat{\varepsilon}_i(w; \theta) = \hat{e}_{1i}(\theta) + (2w - 1)\hat{\nu}_i(w; \theta).$$

Remark 1. Unlike the error terms, the values of $K_M(i; \theta)$ cannot be imputed through nearest neighbor matching. Doing so renders the bootstrap inconsistent since $K_M(i; \theta)$ and $K_{\mathrm{NN}}(i)$ are correlated (here, $K_{\mathrm{NN}}(i)$ denotes the number of times observation i is used as a match, when closeness is measured in terms of a distance metric on the full set of covariates). Intuitively, a nearest neighbor based imputation over-selects observations that are already matched often, and hence fails to recreate the actual distribution of the matching function. A similar comment also applies to imputing the values through propensity score matching.

Remark 2. Let $\mathcal{F}_K^{(0)}(.)$ and $\mathcal{F}_K^{(1)}(.)$ denote the conditional distribution functions of $K_M(i; \theta)$ for the control and treated groups, given the own-propensity score $F(X_i'\theta)$. Consider the estimator, $\hat{\mathcal{F}}_K^{(w)}$, of $\mathcal{F}_K^{(w)}$ obtained by coarsening/blocking the propensity scores, and using the empirical distribution of $K_M(i; \theta)$ for $w = 0, 1$ within each block. The procedure described

in this section is equivalent to drawing a value from the distribution $\hat{\mathcal{F}}_K^{(w)}(F(X_i'\theta))$, independently for each i, and using it to impute the value of $\tilde{K}_M(i; w, \theta)$ when $w \neq W_i$. Coarsening is motivated by the fact $K_M(i; \theta)$ takes discrete values, which precludes smoothing. Clearly $\hat{\mathcal{F}}_K^{(w)} \equiv \mathcal{F}_K^{(w)}$ if the propensity scores are constant within the blocks. More generally, $\hat{\mathcal{F}}_K^{(w)}$ approaches $\mathcal{F}_K^{(w)}$ as $N \to \infty$ since I let $q_N \to \infty$. The optimal choice of q_N would minimize the variability in propensity scores within blocks while ensuring enough observations in each, thereby estimating $\mathcal{F}_K^{(w)}$ more accurately.

Sampling from $\hat{\mathcal{F}}_K^{(w)}$ also ensures each $K_M(i; \theta)$, for $i = 1, \ldots, N$, is used almost exactly once, on average, in the bootstrap: the term may drop out because $W_i^* \neq W_i$, but this probability is balanced by the number of times it may be used for imputations (for details, see Appendix A.2). Thus, the original set of matching functions is well reproduced in the bootstrap.

Remark 3. The variables $\mathbf{M} \equiv \{\mathbf{M}(i) : 1 \leq i \leq N\}$ do not enter the bootstrap distribution as the particular realization of $\mathbf{M}$ is fixed throughout the bootstrap procedure. This is equivalent to fixing an observation j that imputes for i in all the bootstrap draws. Thus the bootstrap distribution should be understood as conditional on both $\mathbf{M}$ and the observed data. This necessarily injects some randomness into the critical values obtained from the bootstrap (though the critical values do converge to the true ones almost surely for each sequence $\mathbf{M}$). To address this, I suggest repeating the bootstrap procedure for a number of different realizations of $\mathbf{M}$, and then taking an average (wrt $\mathbf{M}$) of the bootstrap distribution functions; see below.

1.3.3 The bootstrap algorithm

The bootstrap algorithm proceeds as follows.

Step 0: First obtain a set of multinomial probabilities $\mathbf{M}$ based on independent draws for each individual i as described in Section 1.3.2. Additionally calculate the nearest neighbor matching function $\mathcal{J}_w(i)$ for each i as defined in Section 1.3.1. Both these values are kept fixed throughout the bootstrap.

Step 1: Obtain new values of covariates $\mathbf{X}^* = (X_i^*, \ldots, X_N^*)$ through a non-parametric bootstrap draw. This involves drawing N independent categorical random variables $\mathbf{S}^* =$

$(S_1^*, \ldots, S_N^*)$.

Step 2: Based on the estimated propensity score, derive new treatment values $\mathbf{W}^* = (W_1^*, \ldots, W_N^*)$ through the random draws

$$W_i^* \sim \text{Bernoulli}(F(X_i^{*\prime}\hat{\theta})).$$

Step 3: Discard bootstrap samples for which $N_0^* \leq M+1$ or $N_1^* \leq M+1$, where N_0^* and N_1^* denote the number of control and treated observations in the bootstrap sample. For all the other samples, estimate the bootstrap statistic $\hat{\theta}^*$ using the MLE procedure on $(\mathbf{W}^*, \mathbf{X}^*)$

$$\hat{\theta}^* = \arg\max_{\theta} L(\theta|\mathbf{W}^*, \mathbf{X}^*).$$

Step 4: Based on $\hat{\theta}^*$, obtain the values of matching function $K_M(i;\hat{\theta}^*)$ for each i using the original sample of observations $\mathbf{W}, \mathbf{X}$. Additionally, derive the residuals $\left(\hat{e}_{1i}(\hat{\theta}^*), \hat{e}_{2i}(W_i;\hat{\theta}^*)\right)$, evaluated at $\hat{\theta}^*$, for each i through series regression (or any other nonparametric method) applied on the original sample of observations. From these, along with the values of $\mathbf{M}$ and $\mathcal{J}_w(i)$ from Step 0, determine the values of $\tilde{K}_M(i;w,\hat{\theta}^*)$ and $\hat{\nu}_i(w;\hat{\theta}^*)$ for $i = 1, \ldots, N$ by following the procedures laid down in Sections 1.3.1. and 1.3.2.

For the remaining steps, define the new 'bootstrap' realized errors $\varepsilon_i^*(\theta)$ as

$$\begin{aligned}\varepsilon_i^*(\theta) &\equiv \hat{\varepsilon}_{S_j^*}(W_j^*;\theta) \\ &= \hat{e}_{1S_i^*}(\theta) + W_i^*\hat{\nu}_{S_i^*}(1;\theta) - (1-W_i^*)\hat{\nu}_{S_i^*}(0;\theta).\end{aligned}$$

The bootstrap errors $\varepsilon_i^*(\theta)$ need to be re-centered; the expression for this is given by

$$\Xi^*(\theta) = \frac{1}{N}\sum_{k=1}^{N}\left\{\hat{e}_{1k}(\theta) + F(X_k'\theta)\hat{\nu}_k(1;\theta) - \left(1-F(X_k'\theta)\right)\hat{\nu}_k(0;\theta)\right\}.$$

Note that $\Xi^*(\theta) \equiv E_\theta^*[\varepsilon_i^*(\theta)]$, where $E_\theta^*[.]$ denotes the expectation over the probability distribution implied by $\mathbf{S}^*, \mathbf{W}^* \sim \text{Bernoulli}(F(\mathbf{X}^*\theta))$, conditional on the original data (see also Section 1.3.4 for a detailed explanation). Finally, for each value of θ, define the bootstrap

statistic

$$T_N^*(\theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \varepsilon_i^* (\theta) - \Xi^* (\theta) \right\}.$$

Step 5: Evaluate $T_N^*(\theta)$ at the parameter value $\hat{\theta}^*$ to obtain the bootstrap statistic $T_N^* \left(\hat{\theta}^* \right)$. This step utilizes the values of $\tilde{K}_M(i; w, \hat{\theta}^*)$ and $\hat{\nu}_i(w; \hat{\theta}^*)$ obtained in Step 4.

Step 6: Estimate the critical value by $c_{n,\alpha}^* = \inf\{t : F_n^*(t) \geq 1 - \alpha\}$, where $F_n^*(.)$ is the empirical distribution of $T_N^* \left(\hat{\theta}^* \right)$. This can be obtained by repeating Steps 1-5 for a set of B bootstrap repetitions.

Step 7: The critical value, $c_{n,\alpha}^*$, in Step 6 is based on a particular realization of $\mathbf{M}$. To reduce the dependence on the latter, repeat Steps 1-6 for L different values of $\mathbf{M}$ and average the resulting empirical distribution functions $F_n^*(.)$ to obtain $\bar{F}_n^*(.)$. The final estimated critical value is then given by $\bar{c}_{n,\alpha}^* = \inf\{t : \bar{F}_n^*(t) \geq 1 - \alpha\}$.

1.3.4 Discussion

This section elaborates further on key aspects of the bootstrap procedure.

1.3.4.1 Asymptotic Linearity

Efron and Stein (1981) have shown that an estimator typically needs to be asymptotically linear in the observations for the standard (nonparametric) bootstrap to be valid. However, the matching estimator fails to satisfy asymptotic linearity under the regime of fixed number of matches. Indeed, fixing the number of matches is qualitatively similar to choosing 'small bandwidth asymptotics' for semiparametric estimators, wherein it is known that asymptotic linearity fails (see, e.g Cattaneo, Jansson & Newey, 2016). The same reasoning also implies the standard bootstrap is invalid for the kernel matching estimator of Heckman, Smith and Todd (1997) under small bandwidth asymptotics. Nevertheless, while the matching estimators are not generally asymptotically linear in the observations $(\mathbf{X}, \mathbf{W}, \mathbf{Y})$, they *are* linear in the potential errors, by construction. Thus, by changing the unit of resampling to potential errors (rather than the observations), we can regain bootstrap consistency.

1.3.4.2 Randomization of treatments

A distinctive feature of the bootstrap procedure is the randomization of the treatments, $\mathbf{W}^*$. For many causal effect estimators, such as nearest neighbor matching using the vector of covariates, it suffices to resample the realized errors (see, e.g. Otsu and Rai, 2017). However, such a strategy doesn't work for propensity score matching because the potential and realized errors are functions of the random quantity $F(\mathbf{X}'\hat{\theta})$. The variability of $\mathbf{W}$ conditional on $\mathbf{X}$ has a first order effect on inference through the estimation of $\hat{\theta}$, necessitating the re-drawing of $\mathbf{W}^*$ in the bootstrap. The precise mechanism is as follows: Suppose that one of the covariates is heavily imbalanced between the treatment and control groups. Then the magnitude of $\hat{\theta}$ corresponding to the covariate increases, and the procedure places greater emphasis on balancing that covariate. This reduces the conditional (on $\mathbf{X}, \mathbf{W}$) bias, eventually showing up as (unconditional) asymptotic variance reduction, see Section 3.2. But for a fixed $\mathbf{X}$, the level of imbalance depends on the assignment of $\mathbf{W}$; hence the conditional distribution of $\mathbf{W}$ given $\mathbf{X}$ has a large effect on the variability of the estimate.

1.3.4.3 Bootstrap Recentering

An interesting feature of the recentering term, $\Xi^*(\theta)$, is that it is based on taking the bootstrap expectation over $T_N^*(\theta)$ as if $\mathbf{W}^* \sim \text{Bernoulli}(F(\mathbf{X}^*\theta))$, even though in fact $\mathbf{W}^* \sim \text{Bernoulli}(F(\mathbf{X}^*\hat{\theta}))$. If θ were an exogenous parameter, this would mean the bootstrap expectation of $T_N^*(\theta)$ is exactly 0 only when $\theta = \hat{\theta}$. However $T_N^*()$ is evaluated at $\hat{\theta}^*$, itself a function of the bootstrap random variables. In this case the precise form of the recentering ensures $T_N^*\left(\hat{\theta}^*\right)$ converges in distribution to a mean zero random variable. The reasoning is broadly as follows (see proof of Theorem 1 for details): Suppose for the sake of argument that $\mathbf{W}^* \sim \text{Bernoulli}(F(\mathbf{X}^*\theta))$. Together with $\mathbf{S}^*$, this parametrizes the bootstrap probability distribution, denoted by P_θ^*. Under P_θ^* the test statistic $T_N^*(\theta)$ is exactly mean 0, and therefore converges to a mean 0 random variable in distribution. However for values of θ that are sufficiently close to $\hat{\theta}$, such as $\hat{\theta}^*$, the probability distributions P_θ^* and $P_{\hat{\theta}}^*$ largely coincide. Hence $T_N^*(\theta)$ also converges to a mean 0 random variable under $P_{\hat{\theta}}^*$ - the actual bootstrap distribution.

1.3.4.4 Imputation

The imputation step, Step 0, is critical to the bootstrap. Here, prior to the bootstrap draws, each observation is linked with two others from the opposite treatment arm: the first for imputing the errors (cf Section 1.3.1), and the second for imputing the matching functions (cf Section 1.3.2). In general these observations do not coincide. However conditional on the propensity score, the variables $K_M(i;\theta), e_{1i}, e_{2i}$ are independent of each other even in the true DGP. Thus the approximation properties of the bootstrap are not adversely affected.

Alternatively, one may choose to sort the observations into blocks based on the full set of covariates rather than the propensity scores as in Section 1.3.2. Then a single observation, drawn at random from the block, can be used to impute both the errors and the matching functions. However, even with a binary categorization of the covariates, the number of blocks increases as 2^k with the dimension k. Hence even for moderate k (e.g. $k \geq 5$), it is highly likely that many of the blocks only contain observations from a single treatment arm.

1.4 Asymptotic properties

In this section, I derive the asymptotic properties of the bootstrap procedure outlined in Section 3.1, and demonstrate its consistency. Let P_θ denote the joint distribution of $\{\mathbf{Y}, \mathbf{W}, \mathbf{X}\}$ implied by $W \sim \text{Bernoulli}(F(X_i'\theta))$, the marginal distribution of X, and the conditional distribution of Y given W, X. The corresponding expectation over P_θ is denoted by $E_\theta[.]$. Also, denote by $\tilde{P}_\theta$ the joint probability distribution over both $\{\mathbf{Y}, \mathbf{W}, \mathbf{X}\}$ and $\mathbf{M}$; with $\tilde{E}_\theta[.]$ as the corresponding expectation. For convenience, I set $P_0 \equiv P_{\theta_0}$ $E_0[.] \equiv E_{\theta_0}[.]$, $\tilde{P}_0 \equiv \tilde{P}_{\theta_0}$ and $\tilde{E}_0[.] \equiv \tilde{E}_{\theta_0}[.]$.

Because the matching function $K_M(i;\theta)$ is highly non-linear in θ, it is not possible to use linearization to derive the asymptotic distribution of $T_N^*(\hat{\theta}^*)$. I therefore obtain the limiting distribution by employing a version of Le Cam's skeleton argument, analogous to the proof technique of Abadie and Imbens (2016). Let $\mathcal{N} \equiv \{\theta : \|\theta - \theta_0\| < \epsilon\}$ denote a neighborhood of θ_0 for some $\epsilon > 0$ arbitrarily small. The following regularity conditions are similar to Abadie and Imbens (2016):

Assumption 3. *(i)* $\theta_0 \in int(\Theta)$ *with* Θ *compact,* X *has bounded support and* $E[XX']$ *is non-singular; (ii)* $F(.)$ *is twice continuously differentiable on* $\mathbb{R}$ *with derivatives* $f(.), f'(.)$ *strictly bounded and* $f(.)$ *strictly positive; (iii) for each* $\theta \in \mathcal{N}$ *the random variable* $F(X'\theta)$ *is continuously distributed with interval support; and its pdf* $g_\theta(.)$ *is such that the collection* $\{g_\theta : \theta \in \mathcal{N}\}$ *is uniformly Lipschitz continuous; (iv) at least one component of* X *is continuously distributed, has non-zero coefficient in* θ_0*, and has a continuous density function conditional on the rest of* X*; (v) for each* $\theta \in \mathcal{N}$ *and* $w = 0, 1$*, the functions* $\mu(w, p; \theta)$*,* $Var[\bar{\mu}(w, X)|F(X'\theta) = p]$*,* $Cov[X, \bar{\mu}(w, X)|F(X'\theta) = p]$ *and* $E[\bar{\sigma}^2(w, X)|F(X'\theta) = p]$ *are Lipschitz continuous in* p *with the Lipschitz constants independent of* θ*; furthermore there exists some* $\delta > 0$ *such that* $E[Y^{4+\delta}|W = w, X = x]$ *is uniformly bounded.*

Assumption 4. *There exists some* $\epsilon > 0$ *such that for all* θ *satisfying* $\|\theta - \theta_0\| < \epsilon$*, and for any sequence* $\theta_N \to \theta$*,* $E_{\theta_N}[r(Y, W, X)|W, F(X'\theta_N)]$ *converges to* $E_\theta[r(Y, W, X)|W, F(X'\theta)]$ *almost surely, for any* $\mathbb{R}^{k+2}$*-to-*$\mathbb{R}$ *bounded and measurable function* $r(y, w, x)$ *that is continuous in* x*.*

The above assumptions rule out the case where all the regressors are discrete. In this case the matching estimator reduces to the propensity score sub-classification estimator, inference for which is easily obtained using standard methods. Assumptions 3(i),(ii) ensure that the propensity scores for all the observations are bounded away from zero and one. Khan and Tamer (2010) show that under full support, the usual parametric rate is not attainable, and the rate of convergence depends on the tail behavior of the regressors and error terms. Hence inference in this context would necessarily be at a non-standard rate, and is beyond the scope of this chapter.

Assumption 3 is taken almost directly from Abadie and Imbens (2016). The only substantive difference is in Assumptions 3(iii) and 3(v) which demand uniform extensions of related assumptions in Abadie and Imbens (2016) - in the sense of holding uniformly in a neighborhood $\mathcal{N}$ of θ_0. Assumption 4 is similarly stronger than the corresponding one in Abadie and Imbens (2016). However sufficient conditions for the latter (Theorem S.12 in Abadie and Imbens, 2016) also imply the former.

I shall also require assumptions to ensure the residuals $\{\hat{e}_{1i}(\theta), \hat{e}_{2i}(W_i; \theta)\}$ are 'close' to the unobserved errors $\{e_{1i}(\theta), e_{2i}(W_i; \theta)\}$. I impose the following high level condition:

Assumption 5. *Uniformly over all $\theta \in \mathcal{N}$, it holds under P_0,*

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{e}_{1i}(\theta) - e_{1i}(\theta))^2 = o_p(N^{-\xi}), \text{ and}$$
$$\frac{1}{N}\sum_{i=1}^{N}(\hat{e}_{2i}(W_i;\theta) - e_{2i}(W_i;\theta))^2 = o_p(N^{-\xi}).$$

for some $\xi > 0$.

The assumption posits that the vector of residuals is close to the vector of true errors in terms of the Euclidean metric. For many of the commonly used non-parametric methods such as series or kernel regression, Assumption 5 can be verified under fairly weak continuity conditions, for instance when $\sup_{\theta\in\mathcal{N}} |\partial\mu(w,x;\theta)/\partial x| < \infty$ under $w = 0, 1$. It is usually straightforward to select the tuning parameters for estimation, such as the number of series terms, either visually or through cross-validation. In simulations, low order polynomial series, such as first or second order polynomials, appear to work reasonably well, and constitute an attractive choice in practice.

The final assumption concerns the number of quantile partitions q_N.

Assumption 6. *The number of quantile partitions satisfies $q_N \to \infty$ and $q_N^{2+\eta}/N \to 0$ as $N \to \infty$ for some $\eta > 0$.*

Assumption 6 is fairly weak in that a wide range of choices for q_N are allowed. Here, the choice of q_N determines how close the bootstrap variance estimate $\hat{V}^*$ is to the true variance (due to re-centering, the bootstrap mean is asymptotically 0). Higher values of q_N increase the balance in the propensity scores within the blocks (thus lowering the bias of $\hat{V}^*$), but reduce the number of observations in the treatment and control groups in each block (thus increasing the variance of $\hat{V}^*$), see Remark 2. In fact, this is the same trade-off faced by sub-classification estimators for average treatment effects. In this case, there exists extensive theoretical and empirical literature suggesting that small values of q_N are sufficient to reduce most of the bias due to the stratification of the propensity score (see e.g. Rosenbaum and Rubin, 1984; Imbens and Rubin, 2015). Indeed, under some reasonable conditions, Rosenbaum and Rubin (1984), drawing on previous work by Cochran (1968), find that 4 blocks/sub-classes are sufficient to reduce the bias by over 85%, while having 5

blocks reduces it by more than 90%. These values are independent of sample size since the bias depends solely on q_N. Consequently, following the recommendation of Rosenbaum and Rubin (1984), I suggest a default choice of $q_N = 5$.

Based on the above assumptions, I can derive the asymptotic properties of the bootstrap estimator. Following the techniques of Abadie and Imbens (2016) and Andreou and Werker (2012), I employ the Le Cam skeleton or discretization device for formalizing the theorem. In particular, I discretize both the bootstrap and sample estimators , $\hat{\theta}^*, \hat{\theta}$ along a grid of cubes of length $d/\sqrt{N}$. For instance, if the j-th component of $\hat{\theta}^*$, $\hat{\theta}^*_j$, falls in the q-th cube where $q = \left\lfloor \sqrt{N}\hat{\theta}_j/d \right\rfloor$ with $\lfloor . \rfloor$ being the nearest integer function, then the corresponding component of the discretized estimator is given by $\tilde{\theta}^*_j = dq/\sqrt{N}$. Analogously, I also discretize $\hat{\theta}$ as $\bar{\hat{\theta}} = d\left\lfloor \sqrt{N}\hat{\theta}/d \right\rfloor/\sqrt{N}$. The theoretical results are thus based on using $\bar{\hat{\theta}}$ rather than $\hat{\theta}$ to construct the bootstrap samples. The discretization is only a theoretical device for applying the skeleton arguments and not necessary in practice; indeed, the theory doesn't specify any minimum grid size d.

Let P^* denote the bootstrap probability distribution conditional on both the observations, $(\mathbf{Y}, \mathbf{W}, \mathbf{X})$, and $\mathbf{M}$. In other words, P^* represents joint probability distribution of $W^* \sim \text{Bernoulli}(F(X_i^{*\prime}\bar{\hat{\theta}}))$ and $\mathbf{S}^*$ conditional on $(\mathbf{Y}, \mathbf{W}, \mathbf{X}, \mathbf{M})$. The asymptotic properties of the bootstrap procedure are summarized in the following theorem:

Theorem 1. *Suppose that Assumptions 1-6 hold. Then for d sufficiently small,*

$$P^*\left(T_N^*\left(\tilde{\theta}^*\right) \leq z\right) \xrightarrow{p} Pr(Z \leq z) + O(d)$$

under $\tilde{P}_0$, where Z is a normal random variable with mean 0 and variance $V = \sigma^2 - c' I_{\theta_0}^{-1} c$.

I refer to Appendix A.1 for the formal proof Theorem 1. The derivation parallels that of Abadie and Imbens (2016) in using Le Cam's skeleton argument to obtain the limiting distribution. Let P^*_θ denote the joint distribution of $W^* \sim \text{Bernoulli}(F(X_i^{*\prime}\theta))$ and $\mathbf{S}^*$, conditional on both the observed data and $\mathbf{M}$. Note that $P^* \equiv P^*_{\bar{\hat{\theta}}}$. I consider the bootstrap distribution of the estimator under a local sequence of bootstrap probability distributions $P^*_{\theta_N}$, indexed by $\theta_N = \hat{\theta} + h/\sqrt{N}$. Here θ_N can be thought of as local 'shift' of the estimated propensity score parameter. More precisely, I aim to characterize the limiting distribution

- under the bootstrap sequence of probabilities, $P^*_{\theta_N}$ - of the vector

$$\begin{pmatrix} T^*_N(\theta_N) \\ \sqrt{N}(\hat{\theta}^*_N - \theta_N) \\ \Lambda^*_N\left(\bar{\hat{\theta}}|\theta_N\right) \end{pmatrix},$$

where $\hat{\theta}^*_N$ is the bootstrap estimator of θ under $P^*_{\theta_N}$, and $\Lambda^*_N(\theta|\theta') \equiv \log(dP^*_\theta/dP^*_{\theta'})$ denotes the difference in log-likelihood of the bootstrap probability distributions evaluated at θ and θ' . The limiting distribution of $T^*_N(\hat{\theta}^*)$ under P^* can then be obtained by invoking Le Cam's third lemma (to switch from $P^*_{\theta_N}$ to the actual bootstrap probability P^*), and using the discretization device. A technical difficulty is that $\hat{\theta}$ is also random under $\tilde{P}_0$. To this end, I extend the proof techniques of Abadie and Imbens (2016).

Theorem 1 assures that the bootstrap statistic $T^*_N(\hat{\theta}^*)$ has the same limiting distribution as the true sample. A practical consequence of this theorem is $c^*_{n,\alpha} \xrightarrow{p} c_\alpha$ under $\tilde{P}_0$, where c_α is the critical value from the asymptotic distribution of $\sqrt{N}\left(\hat{\tau}(\hat{\theta}) - \tau(\theta_0)\right)$. Thus, the bootstrap procedure is consistent.

As noted earlier, a drawback of the above result is that in finite samples the value of $c^*_{n,\alpha}$ depends on the particular realization of $\mathbf{M}$. To reduce this dependence, it is possible to proceed as in Step 7 of the bootstrap procedure (cf Section 1.3.3) and average the bootstrap empirical distribution over different values of $\mathbf{M}$. The resulting bootstrap critical value is denoted by $\bar{c}_{n,\alpha}$ (see Section 1.3.3). The following corollary, proved in Appendix A.1, assures that $\bar{c}_{n,\alpha}$ is consistent with respect to P_0 - the probability distribution of the original data.

Corollary 1. *Suppose that Assumptions 1-6 hold. Then* $\bar{c}_{n,\alpha} \xrightarrow{p} c_\alpha + O(d)$ *under* P_0.

1.5 On higher order refinements

In this section I argue that the proposed bootstrap provides a closer approximation to the true distribution of the propensity score matching estimator, as compared to the asymptotic normal limit. I focus in particular on the role played by the randomization of the treatment values and matching functions, and their effect on variance estimation. Previous remarks have already emphasized the importance of redrawing $\mathbf{W}^*$ for inference with propensity

score matching. Here, I show by examples that the bootstrap can generate second order refinements even with other causal effect estimators, especially when the overlap in propensity scores is poor.

As the first example, consider the estimation of the variance for the unadjusted treatment effect estimator $\hat{\tau}_a = \bar{Y}_t - \bar{Y}_c$, where $\bar{Y}_t$,$\bar{Y}_c$ denote the sample averages of the outcomes for the treated and control groups. The estimator is consistent when the data is obtained from a Bernoulli trial RCT, for example. Neglecting the heterogeneity term $E[Y(1)|X] - E[Y(0)|X] - \tau_0$ for simplicity, the potential errors in this example are given by $e(1;X) = Y(1) - E[Y(1)|X]$ and $e(0;X) = E[Y(0)|X] - Y(0)$. Suppose that both the propensity scores, $p(X_i)$, and the potential errors, $\{e(1;X_i), e(0;X_i)\}$, are known. The asymptotic variance estimate is

$$\hat{V} = \frac{1}{N}\sum_{i=1}^{N} e^2(W_i;X_i).$$

A straightforward extension of the bootstrap procedure can also be used to provide inference for $\hat{\tau}_a$. The resulting bootstrap variance estimate is

$$\hat{V}_{\text{boot}} = \frac{1}{N}\sum_{i=1}^{N}\Big\{p(X_i)e^2(1;X_i) + (1-p(X_i))e^2(0;X_i)\Big\} - \Xi_a^2,$$

where Ξ_a is the re-centering term. Since $\Xi_a^2 = O(N^{-1})$, I neglect this in further analysis. Let $\Delta_1 = \hat{V} - V$, $\Delta_2 = \hat{V}_{\text{boot}} - V$, and $\Delta_3 = \hat{V} - \hat{V}_{\text{boot}}$, where V denotes the true variance of the estimate. It is possible to decompose $\Delta_1 = \Delta_2 + \Delta_3$, where Δ_2 and Δ_3 are asymptotically independent, since $\hat{V}_{\text{boot}} \approx E[\hat{V}|\mathbf{X}]$. This immediately implies $\hat{V}_{\text{boot}}$ is a more accurate estimator of V than $\hat{V}$. The extent of the gain in accuracy can be characterized using anti-concentration inequalities: with high probability, $\Delta_3 \geq cN^{-1/2}$ for some $c > 0$. Also, the superior performance of the bootstrap holds even if the potential errors have to be estimated. Let $\tilde{V}_{\text{boot}}$ denote the bootstrap estimator based on estimates, $\hat{e}(w;X_i)$, of the potential errors. If, for instance, X is univariate, and the conditional means of $Y(1)$ and $Y(0)$ are linear in X, the values of $\{\hat{e}(1;X_i), \hat{e}(0;X_i)\}$ can be obtained from linear regressions, and it follows $\tilde{V}_{\text{boot}} - \hat{V}_{\text{boot}} = O_p(N^{-1})$. More generally, as long as the dimension of X is not high (in particular $k \leq 5$), it can be shown that $\tilde{V}_{\text{boot}} - \hat{V}_{\text{boot}} = o_p(N^{-1/2})$ and the bootstrap variance estimate is preferable.

The above example demonstrates that for any given realization of the observations, the bootstrap variance estimate is typically closer to the truth. This can translate to large gains when the degree of overlap in propensity scores is poor. The following example is based on propensity score matching for concreteness, but the intuition applies to causal effect estimators more broadly (for example, simply replacing the matching function with inverse propensity scores gives the Horvitz-Thomson estimator):

Consider a dataset where the range of propensity scores falls within an arbitrarily narrow interval centered around a (known) value p_0 that is close to 0. This implies the number of treated observations is very low, but they have a disproportionately high influence, being used as matches very often. Suppose now that the conditional variances (i.e $\sigma(w;X) = \text{Var}(Y(w)|X)$) are independent of w, and determined by a single binary covariate X_1 with $\sigma(x_1) \equiv \sigma(w;X)$ taking the values H (high) and L (low) when $x_1 = 0, 1$ respectively. I also suppose that X_1 takes the values $0, 1$ with equal probability. For simplicity I focus on the within sample variance, by neglecting the first term (corresponding to e_{1i}) in equation (1.2). In this example, the Abadie-Imbens variance estimate is

$$\hat{V}_{\text{AI}} = \frac{1}{N}\sum_{W_i=1}\left(1+\frac{K_M(i)}{M}\right)^2\sigma^2(X_{1i}) \;+\; \frac{1}{N}\sum_{W_i=0}\left(1+\frac{K_M(i)}{M}\right)^2\sigma^2(X_{1i}), \tag{1.3}$$

with $\sigma(X_{1i}) = L + X_{1i}(H-L)$. The bootstrap (within-sample) variance estimate is[4]

$$\hat{V}_{\text{boot}} = \frac{1}{N}\sum_{i=1}^{N} p_0\left(1+\frac{\tilde{K}_M(i;1)}{M}\right)^2\sigma^2(X_{1i}) \;+\; \frac{1}{N}\sum_{i=1}^{N}(1-p_0)\left(1+\frac{\tilde{K}_M(i;0)}{M}\right)^2\sigma^2(X_{1i}). \tag{1.4}$$

The Abadie-Imbens variance estimator - particularly the first term in (1.3) - is highly sensitive to the relative proportion of observations with $X_{1i} = 0$ or 1 in the treated group. Thus when the value of p_0 is low and $H \gg L$, the estimator is highly variable, and therefore inaccurate. On the other hand $\hat{V}_{\text{boot}}$ is more stable. This is because of the re-randomization of treatment values for all the observations, due to which $\hat{V}_{\text{boot}}$ only depends on the observed density of X_{1i} for the entire sample - a much less variable quantity.

[4]This is based on neglecting the recentering term which is of the order N^{-1}. Also I have modified the bootstrap to take into account the known values of the variances and propensity scores. Even if these modifications were not made, the error from approximating the resulting bootstrap variance estimator with $\hat{V}_{\text{boot}}$ can be made arbitrarily small compared to the effect of moving the value of p_0 closer to 0.

A second robustness property of the bootstrap stems from the random imputation of the matching functions (c.f Section 1.3.2). In the previous example, a low value of p_0 implies greater variability in the matching functions for the treatments. Indeed, it can be shown that

$$\mathrm{Var}[K_M(i)|W_i = 1] \approx \frac{M}{2}\left(\frac{1-p_0}{p_0}\right)^2 + M\frac{1-p_0}{p_0}.$$

Suppose that the variances $\sigma(w;X)$ were not exactly known, then both $\hat{V}_{\mathrm{AI}}, \hat{V}_{\mathrm{boot}}$ would be modified by replacing $\sigma^2(X_{1i})$ with estimated (or imputed, in the case of bootstrap) residuals $\hat{e}^2(w;X_i)$. Consequently $\hat{V}_{\mathrm{AI}}$ is heavily influenced by the error terms of those treated observations that are used as a match most often. Since $\max_{W_i=1} K_M(i) \to \infty$ as $p_0 \to 0$, this again implies greater variability and slow rates of convergence for $\hat{V}_{\mathrm{AI}}$. By contrast, the bootstrap also imputes $\tilde{K}_M(i;1)$ for all the control observations from the conditional distribution of $K_M(i)$ given $W_i = 1$. Thus the high values of $K_M(i)$ are paired with a greater range of the error terms from $\{\hat{e}^2(1;X_i) : i = 1,\ldots N\}$, reducing the influence of a few particular observations.

The above arguments demonstrate as much the benefits of the imputation procedures as those of the bootstrap. However there are other advantages specific to the bootstrap as well. For instance, the bootstrap employs the exact values of the matching functions. By contrast, in the setup of estimated propensity scores, the asymptotic distribution relies on large sample approximations to the same. When the degree of overlap is poor, or when $p_0 \to 0$ in the above example, the rate of convergence of the matching function to its asymptotic approximation can be very slow, as evidenced by the large variances for $K_M(i;\theta)$. As a result the bootstrap would have better approximation properties.

1.6 Extensions

1.6.1 Average treatment effect on the treated

Thus far this chapter has focused on inference for the average treatment effect. An alternative quantity of interest could be the average treatment effect on the treated (ATET), defined as

$$\tau_t(\theta) = E\left[Y_i(1) - Y_i(0)|W_i = 1\right],$$

when the propensity score is given by $F(X'\theta)$. The estimator is indexed with θ since it can now be a function of the propensity score. The parameter of interest is the quantity $\tau_t(\theta_0)$. In this section we show how the bootstrap procedure can be extended to provide inference for $\tau_t(\theta_0)$.

The matching estimator for the ATET for a match based on $F(X'\theta)$ is defined as

$$\hat{\tau}_t(\theta) = \frac{1}{N}\sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i;\theta)} Y_j \right).$$

This has an error representation given by

$$\begin{aligned}\hat{\tau}_t(\theta) - \tau_t(\theta) - B_t(\theta) = &\frac{1}{N_1}\sum_{i=1}^{N} W_i e_{t,1i}(\theta) + \frac{1}{N_1}\sum_{i=1}^{N} e_{2i}(W_i;\theta) \\ &- \frac{1}{N_1}\sum_{i=1}^{N}(1-W_i)\left(1 + \frac{K_M(i;\theta)}{M}\right) e_{2i}(W_i;\theta)\end{aligned}$$

where $B_t(\theta)$ denotes the bias term and

$$e_{t,1i}(\theta) \equiv \mu(1, F(X_i'\theta)) - \mu(0, F(X_i'\theta)) - \tau_t(\theta).$$

The large sample properties of this estimator under the estimated propensity score have been derived by Abadie and Imbens (2016). In particular they show that the bias is asymptotically negligible (i.e $\sqrt{N}B_t(\hat{\theta}) \xrightarrow{p} 0$) and that

$$\sqrt{N}\left(\hat{\tau}_t(\hat{\theta}) - \tau_t(\theta_0)\right) \xrightarrow{d} N\left(0, \sigma_t^2 - c_t' I_{\theta_0}^{-1} c_t + \frac{\partial \tau(\theta_0)}{\partial \theta}' I_{\theta_0}^{-1} \frac{\partial \tau(\theta_0)}{\partial \theta}\right),$$

subject to discretization. We refer to Abadie and Imbens (2016) for the values of σ_t and c_t.

As with the ATE, the error representation motivates our bootstrap procedure. For each θ, denote

$$\hat{e}_{t,1i}(\theta) \equiv \hat{\mu}(1, F(X_i'\theta)) - \hat{\mu}(0, F(X_i'\theta)) - \hat{\tau}_t(\theta).$$

Then our proposed bootstrap statistic for the ATET is

$$T_{t,N}^*\left(\hat{\theta}^*\right) = \frac{\sqrt{N}}{N_1^*}\sum_{i=1}^{N}\left\{\varepsilon_{t,i}^*\left(\hat{\theta}^*\right) - \Xi_t^*\left(\hat{\theta}^*\right)\right\}.$$

where for each θ,

$$\varepsilon^*_{t,i}(\theta) = W^*_i \left\{ \hat{e}_{t,1S^*_i}(\theta) + \hat{e}_{2S^*_i}(1;\theta) \right\} + (1 - W^*_i) \left\{ \hat{e}_{2S^*_i}(0;\theta) - \hat{\nu}_{S^*_i}(0;\theta) \right\}$$

and $\Xi^*_t(\theta) \equiv E^*_\theta[\varepsilon^*_{t,i}(\theta)]$ is the centering term for the bootstrap. In particular the latter can be expanded as

$$\Xi^*_t(\theta) = \frac{1}{N} \sum_{k=1}^{N} \left\{ F(X'_k\theta) \left(\hat{e}_{t,1k}(\theta) + \hat{e}_{2k}(1;\theta) \right) + \left(1 - F(X'_k\theta)\right) \left(\hat{e}_{2k}(0;\theta) - \hat{\nu}_k(0;\theta) \right) \right\}.$$

The empirical distribution, $F^*_{t,n}(.)$, of $T^*_{t,N}\left(\hat{\theta}^*\right)$ can be obtained by a similar algorithm as in Section 1.3.3. Using $F^*_{t,n}(.)$, and a particular realization of $\mathbf{M}$, the critical value is obtained as $c^*_{t,n,\alpha} = \inf\{u : F^*_{t,n}(u) \geq 1 - \alpha\}$. Alternatively, averaging the empirical distribution $F^*_{t,n}(.)$ over L different values of $\mathbf{M}$ gives $\bar{F}^*_{t,n}(.)$. The resulting critical values are given by $\bar{c}^*_{t,n,\alpha} = \inf\{u : \bar{F}^*_{t,n}(u) \geq 1 - \alpha\}$.

Let $c_{t,\alpha}$ denote the critical value from the asymptotic distribution of $\sqrt{N}\left(\hat{\tau}_t(\hat{\theta}) - \tau_t(\theta_0)\right)$. The following theorem assures that the bootstrap procedure for the ATET is consistent. As with Theorem 1, the formal statement relies on discretization.

Theorem 2. *Suppose that Assumptions 1-6 hold. Then for d sufficiently small,*

$$P^*\left(T^*_{t,N}\left(\tilde{\theta}^*\right) \leq z\right) \xrightarrow{p} Pr(Z_t \leq z) + O(d)$$

under $\tilde{P}_0$*, where* Z_t *is a normal random variable with mean* 0 *and variance* $V_t = \sigma^2_t - c'_t I^{-1}_{\theta_0} c_t + \frac{\partial \tau(\theta_0)}{\partial \theta}' I^{-1}_{\theta_0} \frac{\partial \tau(\theta_0)}{\partial \theta}$*. Furthermore,* $\bar{c}_{t,n,\alpha} \xrightarrow{p} c_{t,\alpha} + O(d)$ *under* P_0*.*

The proof of the theorem is similar to that of Theorem 1, and therefore omitted. Similar results also hold for related estimators like the average treatment effect on the controls.

Remark 4. In empirical examples pertaining to the ATET, it is frequently the case that $N_1 \ll N_0$. In such cases, the bootstrap resamples would be predominantly dominated by observations from the control arm. However the error terms and matching functions are imputed from the treated variables. Hence the information from the treated sample is still incorporated in each bootstrap draw.

1.6.2 Matching without replacement

In this section I consider matching without replacement as an alternative for estimating the ATET. This has the advantage of having a lower variance, compared to matching with replacement. At the same time, if the pool of controls is sufficiently large, the increase in bias is not substantial. Here I focus on so called optimal-matching (Rosenbaum, 1989), which is one procedure for matching without replacement. However, the proposed bootstrap is applicable more generally, for instance to greedy or sequential matching.

Suppose that the propensity scores are given by $F(\mathbf{X}'\theta)$. The matching indices, $\mathcal{J}_M^{\text{opt}}(i;\theta)$, for optimal-matching are obtained as the ones that minimize the sum of matching discrepancies, i.e

$$\mathcal{J}_M^{\text{opt}}(\cdot;\theta) \in \text{argmin}_{\{J(i):i=1,\ldots,N\}} \sum_{i=1}^{N} W_i \sum_{j\in J(i)} \left\| F(X_i'\theta) - F(X_j'\theta) \right\|,$$

where $J(.) : \{i : W_i = 1\} \mapsto \{i : W_i = 0\}$ is any one-one mapping from the indices of the treated observations to that of the controls. The corresponding matching function is denoted by

$$K_M^{\text{opt}}(i;\theta) = \sum_{j=1}^{N} \mathbb{I}_{i\in\mathcal{J}_M^{\text{opt}}(j;\theta)}.$$

By definition $K_M^{\text{opt}}(i;\theta) \in \{0,1\}$ for every unit i in the treatment group. For matching based on $F(X'\theta)$, the optimal-matching estimator for the ATET is then

$$\hat{\tau}_t^{\text{opt}}(\theta) = \frac{1}{N}\sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j\in\mathcal{J}_M^{\text{opt}}(i;\theta)} Y_j \right).$$

The estimators $\hat{\tau}_t^{\text{opt}}(\theta)$ and $\hat{\tau}_t(\theta)$ only differ in employing $K_M^{\text{opt}}(i;\theta)$ instead of $K_M(i;\theta)$ as the matching function. With the estimated propensity score, the quantity of interest is $\hat{\tau}_t^{\text{opt}}(\hat{\theta})$. To obtain its large sample properties, I impose the following condition, based on Abadie and Imbens (2012): (Let $\mathcal{N}$ denote some neighborhood of θ_0)

Assumption 7. *Uniformly over all $\theta \in \mathcal{N}$, it holds under P_0 that as $N_1 \to \infty$,*

$$\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N} W_i \sum_{j \in \mathcal{J}^{opt}(i;\theta)} \left\| F(X_i'\theta) - F(X_j'\theta) \right\| \xrightarrow{p} 0.$$

Assumption 7 is a high level condition ensuring the bias from the optimal matching decays fast enough to 0. Suppose that $g_{1,\theta}$ and $g_{0,\theta}$ denote the conditional pdfs of $F(X'\theta)$ conditional on $W_i = 1$ and $W_i = 0$ respectively. Following the arguments of Abadie and Imbens (2012, Proposition 1), sufficient conditions for Assumption 7 can be provided as: (i) $\sup_{\theta \in \mathcal{N}} g_{1,\theta}, g_{0,\theta} \leq C < \infty$ and $\inf_{\theta \in \mathcal{N}} g_{0,\theta} \geq c > 0$; and (ii) $N_1^r \leq cN_0$ for some $c > 0$ and $r > 1$. Here, the requirement of $N_1 \ll N_0$ is crucial for driving down the bias. Using Assumptions 1-7, it is possible to derive the limiting distribution of $\hat{\tau}_t^{\text{opt}}(\hat{\theta})$,

$$\sqrt{N}\left(\hat{\tau}_t^{\text{opt}}(\hat{\theta}) - \tau_t(\theta_0)\right) \xrightarrow{d} N\left(0, \sigma_w^2 - c_w' I_{\theta_0}^{-1} c_w + \frac{\partial \tau(\theta_0)}{\partial \theta}' I_{\theta_0}^{-1} \frac{\partial \tau(\theta_0)}{\partial \theta}\right),$$

subject to discretization. The proof of the above, together with the expressions for $\sigma_w^2,\ c_w$, can be obtained by adapting the arguments of Abadie and Imbens (2012, 2016). In general $\sigma_w^2,\ c_w$ are distinct from the corresponding quantities, $\sigma_t^2,\ c_t$, for matching with replacement.

Given the close analogy with $\hat{\tau}_t(\theta)$, it is straightforward to modify the bootstrap procedure of Section 1.6.1 to obtain valid inference for $\hat{\tau}_t^{\text{opt}}(\hat{\theta})$. The primary difference is that the matching functions are obtained as $K_M^{\text{opt}}(i;\hat{\theta}^*)$ rather than $K_M(i;\hat{\theta}^*)$ in Step 4. Also, only the values of the potential matching function $\tilde{K}_M^{\text{opt}}(i;w,\theta)$ for $w = 0$ need to be known, since the optimal-matching function is defined solely for control variables. The proposed bootstrap test statistic for $\hat{\tau}_t^{\text{opt}}(\hat{\theta})$, denoted by $T_{t,N}^{(\text{opt})*}\left(\hat{\theta}^*\right)$, thus has the same form as $T_{t,N}^*\left(\hat{\theta}^*\right)$, with the sole change being the matches are now given by $K_M^{\text{opt}}(i;\theta)$. Consistency of the bootstrap procedure can be demonstrated by analogous arguments to Theorem 1, using results from Abadie and Imbens (2012).

Theorem 3. *Suppose that Assumptions 1-7 hold. Then for d sufficiently small,*

$$P^*\left(T_{t,N}^{(opt)*}\left(\tilde{\theta}^*\right) \leq z\right) \xrightarrow{p} Pr(Z_t^{(opt)} \leq z) + O(d)$$

under $\tilde{P}_0$, where $Z_t^{(opt)}$ is a normal random variable with mean 0 and variance V_t^{opt} =

$\sigma_w^2 - c_w' I_{\theta_0}^{-1} c_w + \frac{\partial \tau(\theta_0)}{\partial \theta}' I_{\theta_0}^{-1} \frac{\partial \tau(\theta_0)}{\partial \theta}$.

1.6.3 Other causal effect estimators

The bootstrap procedure can easily be extended to other causal effect estimators. Indeed, estimators of the ATE that are linear in the outcome variables, for instance propensity score sub-classification or Horvitz-Thompson estimators, have a common structure in terms of an error representation of the form

$$\hat{\tau}^{(c)} - E[\hat{\tau}^{(c)}] = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i^{(c)}(W_i; \theta),$$

where the potential errors are given by[5]

$$\varepsilon_i^{(c)}(w; \theta) = e_{1i}(\theta) + (2w - 1)\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w) e_{2i}(w; \theta).$$

Here, $\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w)$ may interpreted as quantifying the importance of each observation in terms of estimating the ATE, depending on whether it is in the treated ($w = 1$), or control group ($w = 0$). The estimators differ only in the choice of $\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w)$. The propensity score matching estimator sets $\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w) = 1 + \tilde{K}_M(i; w, \theta)$, while setting

$$\Lambda_i^{-1}(\mathbf{X}'\theta, \mathbf{W}_{-i}, w) = wF(X_i'\theta) + (1 - w)(1 - F(X_i'\theta))$$

gives the Horvitz-Thompson estimator. In a similar vein, the propensity score sub-classification estimator sets

$$\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w) = w\frac{N_1(b_i(\theta)) + N_0(b_i(\theta))}{N_1(b_i(\theta))} + (1 - w)\frac{N_1(b_i(\theta)) + N_0(b_i(\theta))}{N_0(b_i(\theta))},$$

where $b_i(\theta)$ denotes the block in which observation i resides when the blocks are obtained by partitioning $F(\mathbf{X}'\theta)$; and $N_1(b)$, $N_0(b)$ denote the number of treated and control observations in block b. A common theme across all choices is that control (treated) units with high (low) propensity scores gain greater importance, to compensate for them being fewer

[5]The term $\mathbf{W}_{-i}$ denotes the vector of treatments $\mathbf{W}$ excluding W_i.

in number.

The techniques in Section 1.3 provide a template for estimating and imputing the potential error terms, $\varepsilon_i^{(c)}(w;\theta)$. In particular, the values of $e_{2i}(w;\theta)$ for $w = 0, 1$ can be obtained through secondary matching as in Section 1.3.1. Additionally, the unobserved values of the importance function $\Lambda_i(\mathbf{X}'\theta, \mathbf{W}_{-i}, w)$ can be imputed either through a blocking scheme as in Section 1.3.2, or directly, if the functional form is known, as in the case of Horvitz-Thompson and propensity score sub-classification estimators. Consequently, given the potential errors, a bootstrap algorithm can be constructed by analogy with Section 1.3.3; indeed, the bootstrap drawing and re-centering schemes continue to apply.

The consistency of the bootstrap procedure for this more general class of estimators follows by the same reasoning as in Theorem 1.

1.7 Simulation

In this section I investigate the finite sample performance of the bootstrap procedure outlined in Section 1.3.3 using simulation exercises. These confirm my theoretical results and demonstrate the accuracy of the bootstrap procedure.

1.7.1 Simulation designs

I consider different four data generating processes. The first DGP (DGP1) is taken from Abadie and Imbens (2016, Supplementary material). I generate a two dimensional vector (X_1, X_2) of covariates by drawing both variables from a uniform $[-1/2, 1/2]$ distribution independently of each other. The potential outcomes are generated as $Y(0) = 3X_1 - 3X_2 + U_0$ and $Y(1) = 5 + 5X_1 + X_2 + U_1$, where U_1 and U_0 are mutually independent standard normal random variables. The propensity score is given by the logistic function

$$p(X) \equiv P(W = 1|X) = \frac{\exp(X_1 + 2X_2)}{1 + \exp(X_1 + 2X_2)},$$

and the treatments are generated as $W \sim \text{Bernoulli}(p(X))$. Finally, the outcome variables are generated as $Y = WY(1) + (1 - W)Y(0)$.

The second DGP (DGP2) is similar to the first except that the potential outcomes are

generated as $Y(0) = -3X_1 + 3X_2 + U_0$ and $Y(1) = 5 + 7X_1 + 12X_2^2 + U_1$. In this DGP the treatment effect varies more widely with X. Additionally, it also incorporates some non-linearity through the quadratic term in $Y(1)$.

The third DGP (DGP3) is also similar to the first except that the propensity scores are given by

$$P(W = 1|X) = \frac{\exp(X_1 + 7X_2)}{1 + \exp(X_1 + 7X_2)}.$$

The effect of this is to greatly reduce to amount of overlap in the propensity scores between the treated and control samples, as compared to DGP1. For instance, out of a set of 1000 observations, less than 5% of the first 200 observations as ordered by the propensity score are from the treated sample .

The final DGP (DGP4) is adapted from Kang and Schafer (2007). This is chosen for its resemblance with a real data study.[6] For each observation I draw covariates X_1, X_2, X_3, X_4 independently of each other from a standard normal distribution. The potential outcomes are given by $Y(1) = 210 + 27.4X_1 + 13.7X_2 + 13.7X_3 + 13.7X_4 + U_1$ and $Y(0) = U_0$ where U_1 and U_0 are independent standard normal random variables. The propensity scores are given by

$$P(W = 1|X) = \frac{\exp(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)}{1 + \exp(-X_1 + 0.5X_2 - 0.25X_3 - 0.1X_4)}.$$

In all DGPs I consider the case of a single match, i.e $M = 1$. I consider four different sample sizes: $N = 100,\ 200,\ 500,\ 1000$. In all cases, the number bootstrap repetitions is $B = 399$, and the number of Monte-Carlo repetitions is 2500. To ease the computational burden, I only present results for the bootstrap procedure based on a single realization of the multinomial random vector $\mathbf{M}$ (i.e I only follow steps 1-6 of the algorithm in Section 1.3.3).

1.7.2 Choice of tuning parameters

The procedure requires choosing the type of series regression and the number of quantile partitions q_N. Based on visual inspection, I used third order polynomials for the series regression for all DGPs. In practice the number of series terms can also be chosen through

[6]The original simulation study of Kang and Schafer (2007) uses a version of this DGP to evaluate the performance of different procedures under missing data. Here I adapt it to study average treatment effects.

		Sample size			
		$N = 100$	$N = 200$	$N = 500$	$N = 1000$
DGP1	*Bootstrap*	0.057	0.044	0.050	0.050
	Asymptotic	0.070	0.059	0.054	0.049
DGP2	*Bootstrap*	0.058	0.052	0.054	0.058
	Asymptotic	0.063	0.056	0.058	0.057
DGP3	*Bootstrap*	0.090	0.069	0.052	0.047
	Asymptotic	0.141	0.100	0.092	0.069
DGP4	*Bootstrap*	0.079	0.066	0.057	0.058
	Asymptotic	0.118	0.077	0.061	0.057

Table 1.1: Rejection probabilities under the null for various DGPs

cross-validation. I also experimented with different types of non-parametric estimators and show the procedure is not sensitive to the particular choices employed. The choice of q_N was discussed in Section 1.4; there I recommended setting a value of $q_N = 5$. Correspondingly, for the baseline results I set $q_N = 5$ throughout. In a separate table I also report results for different choices of q_N.

For the secondary matching (c.f Section 1.3.1), I employ nearest neighbor matching based on the Euclidean metric. Since in all the DGPs the covariates are standard normal and independent of each other, this is practically equivalent to matching on the Mahalanobis metric.

1.7.3 Simulation results

Table 1.1 reports the performance of the bootstrap inferential procedure for all the DGPs, along with inference based on the asymptotic distribution. The nominal coverage probability is 0.95. The tuning parameters of the number of series terms and q_N haven been deliberately kept unchanged with sample size to emphasize that the values reported are not due to the particular selection of these parameters. In all cases, the bootstrap critical values are very close to nominal even for relatively small sample sizes, for example $N = 100$.

The bootstrap outperforms inference based on the asymptotic distribution in almost all cases. The performance of the bootstrap is particularly advantageous when the sample size is small, see e.g. the results for $N = 100$; and when the extent of imbalance in propensity

	Non-parametric estimators			
	Linear	Poly-3	Poly-4	Spline
DGP1	0.056	0.050	0.055	0.051
DGP2	0.052	0.054	0.047	0.048
DGP3	0.030	0.052	0.048	0.049
DGP4	0.064	0.057	0.062	0.062

Table 1.2: Rejection probabilities under the null for different non-parametric estimators when $N = 500$

scores is high, e.g. DGP3. At the same time, when there is sufficient overlap between the propensity scores, and the effect of estimation of propensity scores is negligible, as in DGP2, there is very little difference between the inferential procedures.

To assess the sensitivity of the bootstrap, I repeated the Monte-Carlo simulations for different choices of tuning parameters. In Table 1.2, I experiment with different non-parametric specifications to estimate the residuals, namely: linear, third and fourth order polynomials, and cubic smoothing splines (with smoothing parameter 0.99). The bootstrap is quite insensitive to the choice of the specification. I found similar results for the other sample sizes; for brevity I do not report these results.

In Table 1.3, I repeat the procedure for different values of q_N under the sample sizes $N = 200$ and $N = 500$ for all the DGPs. I find that the bootstrap procedure is largely robust to the actual choice of q_N, except for the value of $q_N = 1$, which corresponds to no partitioning. This is consistent with the observation, made in Section 1.4, that small values of q_N are sufficient to reduce most of the bias. At the same time, even for larger sample sizes, the reduction in bias is marginal as q_N increases beyond a certain amount. For example, there is not much variability in the results between $q_N = 5$ and $q_N = 8$.

1.7.4 Robustness to Mis-specification

To check the robustness of the inference to mis-specification, I modify the DGPs by using a Probit link function for the true propensity scores, even as the estimation and inferential procedures themselves employ the Logistic regression. Table 1.4 reports the results of the simulation under various DGPs when $N = 200$ and 500. While performance of both

		Number of quantile partitions			
		$q_N = 1$	$q_N = 2$	$q_N = 5$	$q_N = 8$
DGP1	$N = 200$	0.067	0.057	0.046	0.049
	$N = 500$	0.057	0.054	0.050	0.050
DGP2	$N = 200$	0.072	0.058	0.060	0.048
	$N = 500$	0.070	0.047	0.047	0.060
DGP3	$N = 200$	0.202	0.082	0.069	0.076
	$N = 500$	0.190	0.078	0.052	0.045
DGP4	$N = 200$	0.103	0.079	0.066	0.064
	$N = 500$	0.093	0.078	0.057	0.054

Table 1.3: Rejection probabilities under the null for different values of q_N

		DGP1	DGP2	DGP3	DGP4
N = 200	*Bootstrap*	0.050	0.046	0.107	0.087
	Asymptotic	0.063	0.062	0.141	0.133
N = 500	*Bootstrap*	0.052	0.052	0.091	0.069
	Asymptotic	0.052	0.062	0.142	0.101

Table 1.4: Rejection probabilities for the null under mis-specification

inferential procedure degrades somewhat, the bootstrap remains much more robust. A reason for this could be that the residuals $\hat{e}_1(.), \hat{e}_2(.,.)$ - obtained under the mis-specified propensity score - still approximate the actual errors under mis-specification.

1.8 Case study - The LaLonde datasets

The National Supported Work (NSW) demonstration was a randomized evaluation of a job training program, first analyzed by LaLonde (1986), and later the focus of papers by Heckman and Hotz (1989), Deheija and Wahba (1999), Smith and Todd (2005) among others. The original dataset is based on a randomized study. LaLonde (1986) set aside the experimental control group and replaced it with two other sets of observations from the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS). In this section I simulate observations resembling the LaLonde experimental and observational datasets, and use them as test cases for analyzing the relative performance of the bootstrap

and asymptotic inferential procedures[7].

1.8.1 Description of the data and the data generating process

The datasets comprise of the following pre-treatment variables: age (age), years of education (edu), indicator for high school dropout (nodeg), indicator for married (mar), real earnings (in thousands of dollars) in 1974 (re74), indicator for unemployed in 1974 (un74), real earnings (in thousands of dollars) in 1975 (re75), indicator for unemployed in 1975 (un75), and finally two indicators for race: (black) and (hispanic). The outcome variable is real earnings in 1978 (Y). For the results in this section I consider only the African American subsample, which comprises the bulk ($> 85\%$) of the original experimental data. This selects $N_0 = 215$ and $N_1 = 156$ control and treated observations respectively from the experimental dataset, for a total of $N = 371$ observations.

For the observational data, I follow LaLonde (1986) in replacing the experimental control group with the subgroup of all men from PSID and CPS samples who were not working when surveyed in the spring of 1976 (denoted as PSID-2 and CPS-2 respectively). I further extract the African-American subsample from these datasets. This selects $N_0 = 99$ and $N_0 = 286$ observations for the control groups based on the PSID and CPS samples, for a total of $N = 255$ and $N = 442$ observations respectively (given the $N_1 = 156$ treated observations).

I simulate observations mimicking the experimental and observational datasets by broadly following the algorithm described in Busso, DiNardo and McRary (2014). Denote by $\tilde{\mathbf{X}}$ the original set of covariates, and let $\mathbf{Z}$ denote the set of variables comprised of an intercept, $\tilde{\mathbf{X}}$, all the squared terms in $\tilde{\mathbf{X}}$, and the following interaction terms: un75×un74, edu×re75, re74×re75. For each simulation draw, I generate N observations using the following procedure: (1) Draw new covariates $\mathbf{X}$ using the population model specified in the next paragraph; (2) Estimate the propensity scores as $p(\mathbf{X}) = F(V'\theta_0)$ where $F(.)$ is the Logistic function, V is a vector of covariates described below, and θ_0 is the parameter vector obtained given by running a Logistic regression on the original datasets; (3) construct $Y_i(0) = Z_i'\delta_0 + \sigma_0\epsilon_i$,

[7]LaLonde (1986) replaced the experimental control group to analyze the accuracy of non-experimental statistical methods. Here I abstract away from this issue by explicitly imposing selection on observables in simulations.

where δ_0 is obtained by regressing the control observations of the original datasets with $\mathbf{Z}$, σ_0^2 is the root mean squared error of the regression, and ϵ_{0i} are iid standard normal errors; (4) construct $Y_i(1)$ analogously using the treated observations from the original datasets; (5) construct treatment values as $\mathbf{W} \sim \text{Bernoulli}(p(\mathbf{X}))$; (6) construct outcome values as $Y_i = W_i Y_i(1) + (1 - W_i)Y_i(0)$.

Following Busso, DiNardo and McRary (2014), I draw the new covariates $\mathbf{X}$ in the following way: (1) draw the indicator variables mar, un74, un75 by sapling with replacement from the original datasets; (2) fix the pair (mar, un74, un75) as a group and simulate the other variables, i.e (age, edu, re74, re75), from a group-specific multivariate normal distribution, where the distributional parameters are the group means and covariances estimated from the original data; (3) round the values of age, edu to the nearest integer values.

For the experimental data I use a linear specification for the propensity scores with $V =$ (age, edu, nodeg, mar, re74, re75, un74, un75). For the observational designs I employ a somewhat modified version of the propensity score specification used by Deheija and Wahba (1999): $V =$ (age, edu, mar, nodeg, re74, re75, age^2, edu^2, re74^2, re75^2, edu$\times$re74).

The simulations are designed to replicate the broad features of both the experimental and observational datasets. Of particular interest is the degree of overlap in the propensity scores between the treated and control groups. Figure 1.1 presents a representative plot for the simulated datasets. In the experimental design there is a high degree of overlap in the propensity scores which are also bounded away from 0 and 1. On the other hand, the degree of overlap is quite poor in the observational designs with many of the treated observations concentrated around the propensity score value of 0. This has a significant impact on the performance of inferential methods for matching.

1.8.2 Simulation results

I first describe the bootstrap procedure: For secondary matching (cf Section 1.3.1), I used nearest neighbor matching based on the Mahalanobis metric, applied over the unique set of covariates in the data, i.e overage, edu, mar, re74, re75.[8] Additionally, based on a visual

[8] Indeed the other covariates are defined as functions of these with nodeg=1(edu < 12), un74=1(re74 $= 0$) and un75=1(re75 $= 0$).

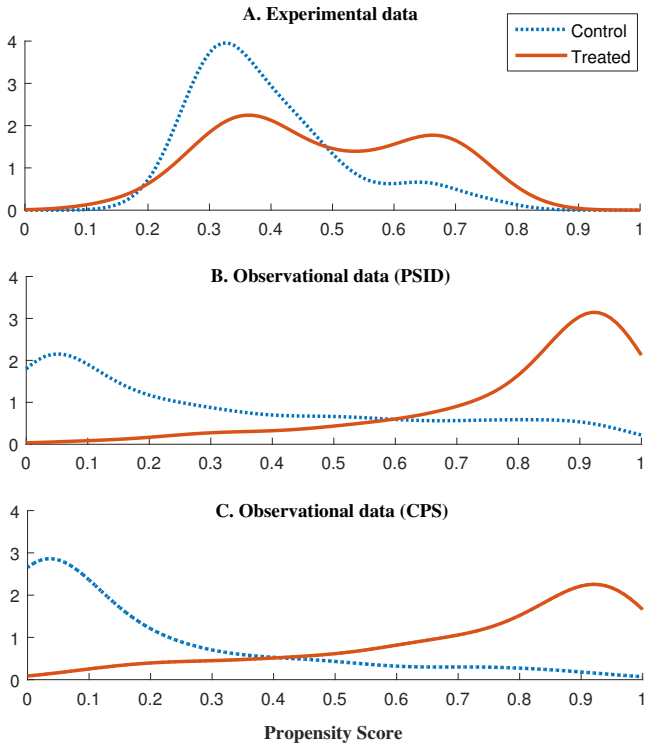

Figure 1.1: Representative overlap plots based on kernel density estimates of propensity scores for control (dotted line) and treated units (solid line)

inspection, I employed a linear specification for the series regression in all designs. For the number of quantile partitions, I employed $q_N = 5$ for the experimental and PSID designs, and $q_N = 4$ for the CPS design. The reason for the lower value of q_N in the latter case is due to the poor overlap in the propensity scores, which results in some cells having no treated observations when q_N is higher.[9] Table 1.5 reports the performance of the bootstrap and asymptotic inferential procedures for the matching estimator of the ATE. All values are based on 5000 Monte-Carlo repetitions with $B = 399$. The results are provided after bias correction, which in any case is an order of magnitude smaller than the standard deviation.

The first three rows of Table 1.5 present the simulation results with the same sample sizes as in the original datasets. For the experimental design, both the bootstrap and asymptotic methods provide very similar performance. This is an example in which estimation of the propensity scores hardly affects variance. The asymptotic method appears to be slightly preferable, even if the difference is not statistically significant. This is possibly due to the

[9] In the rather rare instance where one of the cells has no treated observations even with the lower value of q_N, I impute the matching function by drawing treated observations randomly from the neighboring cell.

	Rejection probability		Confidence Interval length		
	Bootstrap	*Asymptotic*	*Bootstrap*	*Asymptotic*	*True*
Experimental (N = 371)	0.061	0.054	3.474	3.528	3.666
Observational (PSID) (N = 255)	0.079	0.214	7.426	5.621	8.335
Observational (CPS) (N = 422)	0.076	0.233	8.669	5.985	9.288
Experimental (N = 150)	0.075	0.075	5.270	5.312	5.756
Observational (PSID) (N = 500)	0.063	0.148	5.984	4.848	6.147
Observational (CPS) (N = 1000)	0.062	0.169	7.242	5.363	7.194

Table 1.5: Rejection probabilities and average length of confidence intervals (in thousands of dollars) under experimental and observational designs

bias introduced by the nearest-neighbor-matching technique while imputing the error terms.

The performance of the inferential methods declines under both observational designs. Nevertheless, the asymptotic procedure performs considerably worse than the bootstrap, and underestimates the length of the confidence interval by close to 33% of the true length. (I also found that in about 4-5% of the cases, the asymptotic procedure actually reported a negative value for the variance!) By contrast, the bootstrap provides good size control, despite the fact the propensity scores are not bounded away from 0 and 1.

Figure 1.2 plots the estimates of the finite sample distribution (after centering by the true value) using bootstrap and asymptotic methods for representative simulation samples. For the observational data, the estimate from the asymptotic method is highly biased and heavily underestimates the true variance. The bootstrap distribution is much closer to the actual one.

In fact, not only is the asymptotic variance estimate heavily biased for the observational data, it is also highly variable. Figure 1.3 demonstrates this by plotting the finite sample distributions using bootstrap and asymptotic methods for 20 different simulation samples under experimental and PSID designs (the CPS dataset is omitted for brevity). For the

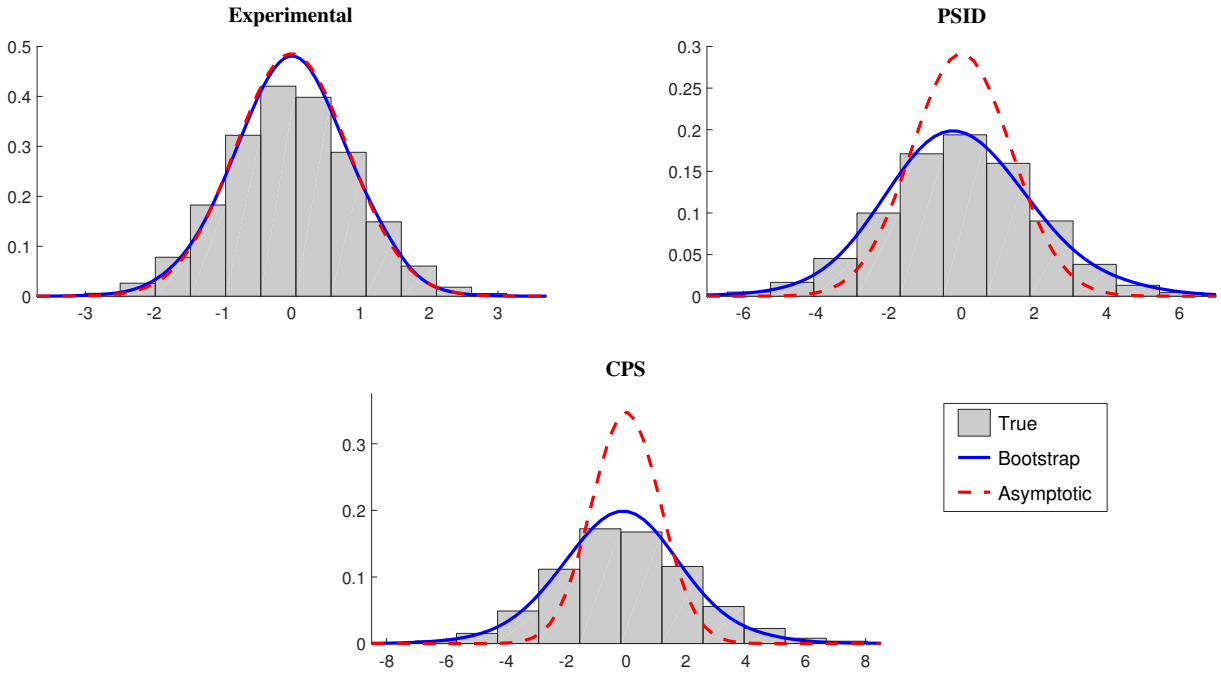

Figure 1.2: Estimates of the finite sample distribution using bootstrap (solid blue) and asymptotic methods (dashed red) for representative simulation samples. The bars represent the actual finite sample distribution.

PSID data, the bootstrap estimate of the finite sample distribution is much more stable over the different simulation samples. However both methods perform very similarly on the experimental dataset, suggesting that most of the differences in the PSID design are generated by poor overlap. This is consistent with the discussion in Section 1.5, where I showed that the asymptotic variance estimate is much more sensitive to a few influential observations, as compared to the bootstrap.

In the last three rows of Table 1.5, I redo the simulation with different sample sizes. Here, I employ $q_N = 5$ for all the designs. For the experimental design, both inferential methods perform well even on a sample size that is about half the original one. However, for the observational designs the bootstrap outperforms asymptotic inference by a considerable margin even after doubling the number of observations.

1.9 Conclusion

In this chapter, I propose a bootstrap procedure for propensity score matching estimators of the ATE and ATET, and demonstrate its consistency. The procedure can be easily

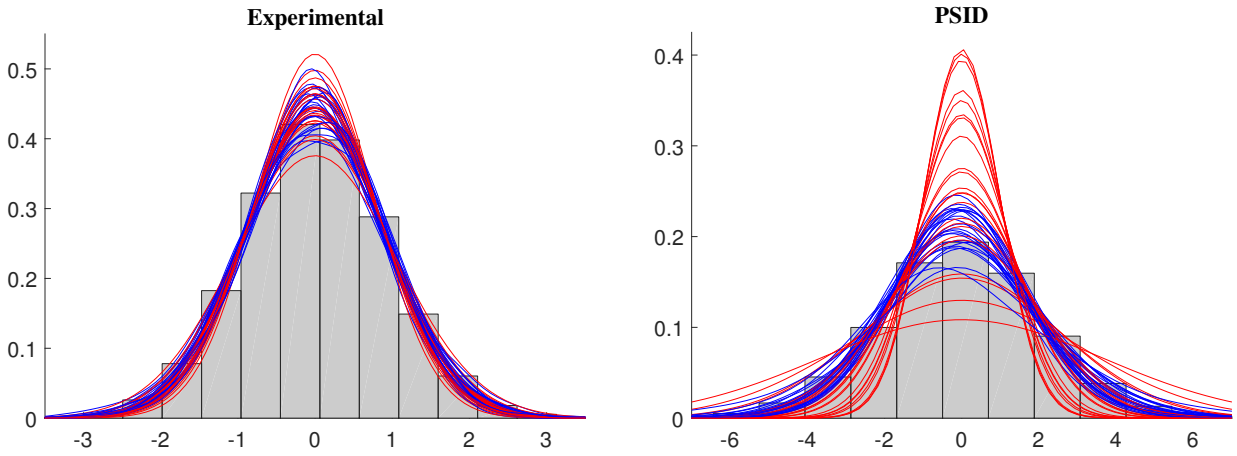

Figure 1.3: Estimates of finite sample distributions using bootstrap (blue) and asymptotic methods (red) for 20 different simulation samples. The bars represent the actual finite sample distribution. Note the difference in scaling of the axes.

extended to other estimators, including, but not limited to, inverse probability weighting (e.g. Horvitz-Thompson) and propensity score sub-classification. It is built around the concepts of potential errors and the error representation, introduced in this chapter. Both these concepts are also applicable very generally. Together, they constitute a powerful new formalism for describing causal effect estimators.

Simulations and theoretical examples suggest the proposed bootstrap achieves greater accuracy than asymptotic methods, particularly when the overlap in propensity scores is poor. They also highlight the key role played by the (re-)randomization of treatment values in obtaining more precise inference. While beyond the scope of this chapter, it would be interesting to formally investigate the higher order properties of the bootstrap procedure.

This chapter focuses on treatment effects. However, the techniques and results in this chapter may also be useful in other contexts, for instance where the outcome data is missing at random (i.e the propensity for missingness is only a function of observed covariates).

Chapter 2

Empirical Likelihood for random sets

2.1 Introduction

In many statistical applications, the observed data take the form of sets rather than points. For example, in survey analysis, we often observe bracket data instead of precise measurements. In mathematical morphology, geostatistics, and particle statistics, the observations often take the form of two or three dimensional sets reflecting models for tumor growth or sand rock grains (e.g., Cressie and Hulting, 1992, and Stoyan, 1998, for a review). Also, in the context of medical imaging and robotic vision, researchers sometimes need to infer a convex set from noisy measurements of its support function (Fisher *et al.*, 1997). Furthermore, in studies of treatment effects (e.g., Balke and Pearl, 1997, and Horowitz and Manski, 2000), researchers often wish to conduct statistical inference on nonparametric bounds for the average treatment effects which can be expressed by means of random sets, as shown in Beresteanu, Molchanov and Molinari (2012).

In this chapter, we develop a nonparametric likelihood concept for the Aumann expectation of a random sample of convex sets - this is a generalization of the conventional mathematical expectation to random sets - and propose general inference methods by adapting the theory of empirical likelihood (Owen, 2001). In particular, by relying upon the isomorphism between a convex set and its support function, we convert the testing problem on the

random set to one on its support function which implies a continuum of moment constraints indexed by the direction of the support function. Based on this conversion, we construct two nonparametric likelihood statistics for testing the moment constraints which we term the marked and sieve empirical likelihood statistics. We study the asymptotic properties of these statistics and describe how to compute critical values for testing. Moreover, to enhance the applicability of our methods, we also discuss testing directed hypotheses and projections, along with situations where the random set of interest is not directly observable due to nuisance parameters to be estimated and where inference is based on noisy measurements of the support function.

We demonstrate the usefulness of the proposed methods by four numerical examples. First, we consider the setup of best linear prediction with interval dependent variables. In this case, the set of all possible coefficients for the best linear predictor is characterized by an Aumann expectation involving the interval data. We illustrate our empirical likelihood methods via inference on the parameters for the best linear predictor of interval wages given years of education using the Current Population Survey (CPS) data. Second, we consider a Boolean model for tumor growth studied by Cressie and Hulting (1992) and numerically evaluate the marked and sieve empirical likelihood tests. Third, we employ the empirical example in Balke and Pearl (1997) on the treatment effect of Vitamin A supplementation under imperfect compliance to study the numerical performance of our empirical likelihood based inference on the bounds of the average treatment effect. Finally, based on Fisher *et al.* (1997), we study the problem of testing the shape of a convex set based on noisy measurements of its support function; the results are provided in Appendix B. Both parameter hypothesis and goodness-of-fit testing problems are investigated. In all of the examples, the proposed empirical likelihood tests perform well in terms of size and power.

After early developments in e.g., Kendall (1974) and Matheron (1975), the literature on the probabilistic and statistical theory of random sets is steadily growing (see, Molchanov, 2005, for a modern and comprehensive treatment of random set theory). Most of the statistical literature on random sets focuses on inference via capacity functionals (e.g., Cressie and Hulting, 1992) and support functions (e.g., Fisher *et al.*, 1997) which provide equivalent

characterizations of random sets. The population mean of random sets is typically characterized by the so-called Aumann expectation. Beresteanu and Molinari (2008) developed a Wald type test for the Aumann mean of random sets. This chapter introduces a nonparametric likelihood-based approach for inference on the Aumann expectation by modifying the empirical likelihood method. Thus, this chapter also contributes to the literature on empirical likelihood (see Owen, 2001, for a review) by extending its scope to random sets rather than points. To establish the asymptotic theory, we adapt the theoretical results developed in Hjort, McKeague and van Keilegom (2009) to our context.

Recently, applications of random set methods have been discussed in the context of partial identification and inference in econometrics; see Molchanov and Molinari (2014) for a review of such applications, Tamer (2010) for a review of partial identification in econometrics, and Manski (2003) for a thorough treatment of partial identification. Partial identification concerns the situation wherein a parameter of interest is not point identified but identified only as a set. This could be because of limitations in the data, e.g. interval or categorical data, or because the theoretical models do not provide enough restrictions to identify a unique value for the parameter, e.g. game theoretic models with multiple equilibria. In this context, Balke and Pearl (1997) and Horowitz and Manski (2000) made fundamental contributions to partial identification of treatment effects and probability distributions with missing data, respectively. However, these papers did not connect the inference problems on the identified sets to random set theory. Beresteanu and Molinari (2008) were the first to employ random set methods to conduct estimation and inference for partially identified models.

An important application of random set theory is in the context of inference for parameters characterized by moment inequalities. In this setup, the parameters are typically partially identified, and thus the aim is to propose a confidence region that covers the identified set. Examples of this strand of literature include Chernozhukov, Kocatulum and Menzel (2015), Kaido (2012), and Kaido and Santos (2014) among others. See also Andrews and Shi (2015) for an extension to conditional moment inequalities. On the other hand, Canay (2010) developed an empirical likelihood-based inference method for moment inequality models using "standard" probability theory. Our chapter is the first to to bring

together random set theory and empirical likelihood. Although sharing applications with the moment inequality setup, our approach, which is based on random sets as observations, is fundamentally different. Indeed, there are situations where the moment inequality setup is not directly applicable unlike ours (e.g., the Boolean model and image analysis via support function), and vice versa. In addition, the focus of our chapter is on testing, which may have other uses over and above the construction of confidence regions (cf. the Boolean model example). Closer to our setup, Beresteanu and Molinari (2008) were the first to consider tests for expectations of general random sets. Bontemps, Magnac and Maurin (2012) and Chandrasekhar *et al.* (2012) obtained related inferential results in the context of best linear predictors for set identified functions under a variety of extensions but did not consider other formulations of random sets.

This chapter is organized as follows. Section 2.2 introduces the basic setup and presents two inference approaches, the marked and sieve empirical likelihood methods. Section 2.3 discusses various extensions of these approaches for wider applicability. In Section 2.4, numerical examples are provided. Assumptions and some definitions are presented in Appendix B. The appendix also contains proofs and additional simulation results.

2.2 Methodology

Suppose we observe a set-valued random variable (SVRV) $X : \Omega \mapsto \mathbb{K}^d$, where $\mathbb{K}^d$ is the collection of all non-empty compact and convex subsets of the Euclidean space $\mathbb{R}^d$. The collection $\mathbb{K}^d$ is endowed with the Hausdorff norm defined as $\|A\|_H = \sup\{\|a\| : a \in A\}$ for every set A, where $\|\cdot\|$ is the Euclidean norm. Let μ denote some underlying probability measure on Ω. The mean of the SVRV X is characterized by the Aumann expectation

$$\mathbb{E}[X] = \left\{ \int_\Omega x d\mu : x \in \{x(\omega) \in X(\omega) \text{ a.s. and } \int_\Omega \|x\| \, d\mu < \infty\} \right\},$$

(see, Molchanov, 2005, for details). We restrict our attention to compact and convex valued SVRVs; however, similar results hold for general compact sets since $\mathbb{E}[X] = \mathbb{E}[\text{co}(X)]$ for compact valued X if μ is non-atomic, with $\text{co}(X)$ denoting the convex hull operation on X (Molchanov, 2005, p. 154). A fundamental statistical question is to test hypotheses on the

Aumann expectation of the form:

$$\mathrm{H}_0 : \mathbb{E}[X] = \Theta_0(\nu) \text{ vs. } \mathrm{H}_1 : \mathbb{E}[X] \neq \Theta_0(\nu), \tag{2.1}$$

based on a random sample of SVRVs $\{X_1, \ldots, X_n\}$, where $\Theta_0(\nu)$ is a hypothetical set that may depend on real-valued nuisance parameters $\nu \in \mathbb{R}^r$. In general, there is no restriction on the relationship between the dimension d of X and r of ν.

To test the null hypothesis H_0, we focus on the dual representation of convex sets by their support functions. Let $\langle \cdot, \cdot \rangle$ denote the inner product and $\mathbb{S}^d$ the unit sphere in $\mathbb{R}^d$. The support function of a set $A \in \mathbb{K}^d$ is defined as $s(A, p) = \sup_{x \in A} \langle p, x \rangle$ for $p \in \mathbb{S}^d$. If X is integrably bounded, the testing problem in (2.1) is equivalent to (Molchanov, 2005, p. 157)

$$\mathrm{H}_0 : E[s(X, p)] = s(\Theta_0(\nu), p) \text{ for all } p \in \mathbb{S}^d \text{ vs. } \mathrm{H}_1 : E[s(X, p)] \neq s(\Theta_0(\nu), p) \text{ for some } p \in \mathbb{S}^d, \tag{2.2}$$

where $E[\cdot]$ is the ordinary mathematical expectation with respect to μ. Therefore, inference on the Aumann mean of the random set is equivalent to inference on the support function (or continuum of moment restrictions over $p \in \mathbb{S}^d$). Since this is a testing problem for infinite dimensional parameters without any parametric distributional assumptions on the population μ, it is of interest to develop a nonparametric likelihood inference method. In particular, we adopt the empirical likelihood approach (Owen, 2001) to our testing problem.

2.2.1 Marked empirical likelihood

We now introduce the first empirical likelihood approach to test the hypothesis in (2.1) for the Aumann expectation of random sets. We assume that a consistent estimator $\hat{\nu}$ for the nuisance parameters ν is available. Typically ν is a smooth function of population moments which can be estimated by the method of moments.

One method to construct a nonparametric likelihood function to test H_0 in (2.1) is to fix a direction $p \in \mathbb{S}^d$ for the support function defining the equivalent form of H_0 in (2.2) and employ the empirical likelihood approach. For given p, the marked empirical likelihood

function under the restriction $E[s(X,p)] = s(\Theta_0(\nu),p)$ is given by

$$\ell_n(p) = \max\left\{ \prod_{i=1}^{n} nw_i \middle| \sum_{i=1}^{n} w_i s(X_i,p) = s(\Theta_0(\hat{\nu}),p),\ w_i \geq 0,\ \sum_{i=1}^{n} w_i = 1 \right\}. \tag{2.3}$$

In practice, $\ell_n(p)$ can be computed from its dual form based on the Lagrange multiplier method, that is

$$\ell_n(p) = \prod_{i=1}^{n} \frac{1}{1+\lambda\{s(X_i,p) - s(\Theta_0(\hat{\nu}),p)\}}, \tag{2.4}$$

where λ solves the first-order condition $\sum_{i=1}^{n} \frac{s(X_i,p)-s(\Theta_0(\hat{\nu}),p)}{1+\lambda\{s(X_i,p)-s(\Theta_0(\hat{\nu}),p)\}} = 0$. Since the direction p is given, the object $\ell_n(p)$ imposes only a single restriction implied from the null H_0. In order to guarantee consistency against any departure from H_0, we need to assess the whole process $\{\ell_n(p) : p \in \mathbb{S}^d\}$ over the range of $\mathbb{S}^d$. Taking the supremum over p leads to the Kolmogorov-Smirnov type test statistic

$$K_n = \sup_{p\in\mathbb{S}^d} \{-2\log \ell_n(p)\}.$$

Suppose there exists a function $G(p;\nu)$ continuous in $p \in \mathbb{S}^d$ such that

$$\sup_{p\in\mathbb{S}^d} |s(\Theta_0(\hat{\nu}),p) - s(\Theta_0(\nu),p) - G(p;\nu)'(\hat{\nu}-\nu)| = o_p(n^{-1/2}). \tag{2.5}$$

In Section 2.4.1, we provide an example of $G(p;\nu)$ for the case of the best linear prediction with an interval valued dependent variable. The asymptotic properties of K_n are summarized in the following theorem.

Theorem 4. *Under Assumption M in Appendix B.1, it holds*

$$K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{\{Z(p) - G(p;\nu)'Z_1\}^2}{\mathrm{Var}(s(X,p))}, \quad \textit{under } \mathrm{H}_0, \tag{2.6}$$

where $(Z(p), Z_1')' \sim N(0, V(p))$ *and* $V(p)$ *is the limiting covariance matrix of* $(n^{-1/2}\sum_{i=1}^{n}\{s(X_i,p) - E[s(X,p)]\}, \sqrt{n}(\hat{\nu}-\nu)')'$. *In addition,* K_n *diverges to infinity under* H_1.

By a slight modification of the proof, we can also show that under the local alternative

$$\mathrm{H}_{1n} : E[s(X,p)] = s(\Theta_0(\nu),p) + n^{-1/2}\eta(p) \text{ over } p \in \mathbb{S}^d,$$

for some continuous function η, the marked empirical likelihood statistic satisfies $K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{\{Z(p)-G(p;\nu)'Z_1+\eta(p)\}^2}{\mathrm{Var}(s(X,p))}$. Therefore, the test statistic K_n has non-trivial local power against a local alternative at the parametric rate.

One major advantage of the conventional empirical likelihood approach is that it yields an asymptotically pivotal statistic even for nonparametric objects of interest under complicated data structures. However, the proposed statistic K_n (or other statistics constructed from the process $\{\ell_n(p) : p \in \mathbb{S}^d\}$) does not share such attractiveness, and its limiting distribution contains several unknowns to be estimated. To deal with this problem, Section 2.2.1.1 proposes a bootstrap procedure to approximate the null distribution of K_n. In Section 2.2.2, we develop an alternative test statistic which is asymptotically pivotal (but requires a choice of a tuning parameter). In the current setup, we are not aware of any test statistic which is both asymptotically pivotal and free from tuning parameters.

We note that lack of pivotalness of process-based tests emerges commonly in the context of goodness-of-fit testing (e.g., Stute, 1997). In the literature on empirical likelihood, Chan *et al.* (2009) propose an integral version of the empirical likelihood statistic to test hypotheses on Lévy processes via characteristic functions and derive a non-pivotal limiting distribution; this is approximated by a bootstrap procedure due to its complicated form. Li (2003) obtained similar results for an empirical likelihood test of survival data. Furthermore, Hjort, McKeague and van Keilegom (2009) provided various extensions of empirical likelihood to the cases of (infinite-dimensional) nuisance parameters and growing numbers of estimating equations. They argued that the empirical likelihood statistic is not necessarily pivotal but can be approximated by bootstrap methods.

Since the marked empirical likelihood statistic is not asymptotically pivotal, one may seek to employ alternative likelihood concepts. For instance, we can generate the likelihood

process from the Euclidean likelihood (Owen, 2001, Section 3.15):

$$L_n^E(p) = \max\left\{ -\frac{1}{2}\sum_{i=1}^n (nw_i - 1)^2 \middle| \sum_{i=1}^n w_i s(X_i, p) = s(\Theta_0(\hat{\nu}), p),\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\},$$

whose dual form is explicitly given by

$$-2L_n^E(p) = \frac{\left(\sum_{i=1}^n \{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}\right)^2}{\sum_{i=1}^n \{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}^2},$$

for each p. Inspection of the proof of Theorem 4 shows that $L_n^E(p)$ is asymptotically equivalent to $\log \ell_n(p)$ for each p and the test statistic $K_n^E = \sup_{p\in\mathbb{S}^d}\{-2L_n^E(p)\}$ obeys the same limiting distribution as K_n. One practical advantage of the Euclidean likelihood-based statistic K_n^E over K_n is that K_n^E does not require a numerical search for the Lagrange multiplier λ as in (2.4).

2.2.1.1 Bootstrap calibration

The limiting null distribution of the process $\{\ell_n(p) : p \in \mathbb{S}^d\}$ is generally difficult to approximate as it contains parameters to be estimated. Thus, we suggest approximating the distribution of K_n by a bootstrap procedure. Let $\{X_i^*\}_{i=1}^n$ denote the bootstrap draws of $\{X_i\}_{i=1}^n$ with replacement and $\hat{\nu}^*$ the bootstrap counterpart of $\hat{\nu}$.[1] Denote $\bar{s}(p) = n^{-1}\sum_{i=1}^n s(X_i, p)$ and $\hat{V}(p) = n^{-1}\sum_{i=1}^n \{s(X_i, p) - \bar{s}(p)\}^2$. For the bootstrap counterpart of the empirical likelihood function $\ell_n(p)$, we propose

$$\ell_n^*(p) = \max\left\{ \prod_{i=1}^n nw_i \middle| \sum_{i=1}^n w_i\{s(X_i^*, p) - s(\Theta_0(\hat{\nu}^*), p)\} = \{\bar{s}(p) - s(\Theta_0(\hat{\nu}), p)\},\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\}. \tag{2.7}$$

Note that $\ell_n^*(p)$ does not directly mimic the original statistic but rather evaluates the likelihood after recentering by $\bar{s}(p) - s(\Theta_0(\hat{\nu}), p)$. Such a recentering is necessary to account for the effect of the estimated nuisance parameters.[2] Indeed, by Giné and Zinn (1990), after

[1] If ν is a smooth function of means, then $\hat{\nu}^*$ is given by replacing the moments with the bootstrap counterparts. If $\hat{\nu}$ is an M-estimator, we obtain $\hat{\nu}^*$ through properly recentered estimating equations as in Shorack (1982) and Lahiri (1992).

[2] The idea of recentering estimating equations is developed in Shorack (1982) and Lahiri (1992). It is interesting to see whether such recentering induces a desirable higher-order property in our setup as in Lahiri (1992).

imposing bootstrap analogs of Assumption M (i)-(iii), a similar argument to the proof of Theorem 4 implies that $-2\log \ell_n^*(p)$ is approximated by $\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n \{s(X_i^*,p)-\bar{s}(p)\}-\{s(\Theta_0(\hat{\nu}^*),p)-s(\Theta_0(\hat{\nu}),p)\}\right]^2/\hat{V}(p)$. However, in the absence of recentering, the additional term $\bar{s}(p)-s(\Theta_0(\hat{\nu}),p)$ appears in the numerator which makes the bootstrap invalid. This is reminiscent of Stute, Gonzalez-Manteiga and Quindimil (1998) who showed inconsistency of the classical bootstrap in the context of model checks for regression. Using the quadratic expansion above, standard arguments based on Giné and Zinn (1990) enable us to prove the following consistency result for the proposed bootstrap statistic.

Proposition 1. *Under Assumptions M and M', the process* $\{\ell_n^*(p) : p \in \mathbb{S}^d\}$ *converges in distribution to the Gaussian process* $\{\{Z(p)-G(p;\nu)'Z_1\}^2/\mathrm{Var}(s(X,p)) : p \in \mathbb{S}^d\}$ *in* P^**-probability, where* P^* *denotes the probability computed under the bootstrap distribution conditional on the data.*

Therefore, the bootstrap critical values of K_n are given by the quantiles of $K_n^* = \sup_{p\in\mathbb{S}^d}\{-2\log \ell_n^*(p)\}$.

2.2.1.2 Case of no nuisance parameter

If there is no nuisance parameter to be estimated (i.e., $\Theta_0(\nu)=\Theta_0$), Assumption M is implied by the sole requirement that $E[\|X\|_H^\xi]<\infty$ for some $\xi>2$,[3] and the null distribution of K_n becomes

$$K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d}\frac{Z(p)^2}{E[Z(p)^2]},$$

where Z is a Gaussian process with zero mean and covariance kernel $\mathrm{Cov}(s(X,p),s(X,q))$.

For comparison, let us consider the Wald type statistic of Beresteanu and Molinari (2008) adapted to the case of no nuisance parameters. In this case the statistic is simply $W_n=\sqrt{n}d_H\left(\frac{1}{n}\oplus_{i=1}^n X_i,\Theta_0\right)$, i.e., the contrast provided by the Hausdorff distance between the Minkowski average $\frac{1}{n}\oplus_{i=1}^n X_i$ and the null hypothetical set Θ_0. For convex sets, the Wald type statistic W_n may be alternatively characterized using the support functions

[3] This follows from the Lipschitz property of the support function, $|s(X,p)-s(X,q)|\le \|X\|_H\|p-q\|$ a.s. for any $p,q\in\mathbb{S}^d$, which ensures that $\{s(X,p):p\in\mathbb{S}^d\}$ is μ-Donsker by a standard empirical process argument (e.g., van der Vaart, 1998, Example 19.7).

as $W_n = \sqrt{n}\sup_{p\in\mathbb{S}^d}\left|\frac{1}{n}\sum_{i=1}^n s(X_i,p) - s(\Theta_0,p)\right|$ (Beresteanu and Molinari, 2008, equation (A.1)). Based on the proof of Theorem 4, we can then see that

$$K_n^{1/2} = \sqrt{n}\sup_{p\in\mathbb{S}^d} E[Z(p)^2]^{-1/2}\left|\frac{1}{n}\sum_{i=1}^n s(X_i,p) - s(\Theta_0,p)\right| + o_p(1),$$

under H_0. Therefore, while the Wald type statistic W_n of Beresteanu and Molinari (2008) evaluates the contrast $\frac{1}{n}\sum_{i=1}^n s(X_i,p) - s(\Theta_0,p)$ over $p\in\mathbb{S}^d$, the empirical likelihood statistic K_n evaluates the same contrast but normalized by its standard deviation. This normalization ensures that our statistic K_n is invariant to scale transformations (i.e., multiplication of both $\{X_i\}_{i=1}^n$ and Θ_0 by some non-singular matrix independent of i), unlike the Wald type statistic W_n which is sensitive to such transforms.[4] In Section 2.4.1, we illustrate that the lack of invariance of the Wald type statistic can yield different size properties depending on what scaling is used.

When there is no nuisance parameter, it is possible to invert K_n to obtain an approximate confidence region within which the Aumann expectation $\mathbb{E}[X]$ lies with some desired probability. Indeed, using the quadratic approximation for the empirical likelihood process (cf. proof of Theorem 4), it follows that with probability α, the support function for the set $\mathbb{E}[X]$ asymptotically satisfies $s(\mathbb{E}[X],p) \le n^{-1}\sum_{i=1}^n s(X_i,p) + \sqrt{\frac{\hat{c}_\alpha}{n}}\hat{V}(p)^{1/2}$ for all $p\in\mathbb{S}^d$, where $\hat{c}_\alpha$ is the bootstrap estimate of the α-th quantile of the limiting distribution of K_n. Based on the right hand side of this inequality, we can thus recover the confidence region that covers $\mathbb{E}[X]$ with the desired probability level α.

2.2.2 Sieve empirical likelihood

Another way to construct an empirical likelihood for testing H_0 in (2.2) is to incorporate the continuum of moment conditions $E[s(X,p)] = s(\Theta_0(\nu),p)$ for all $p\in\mathbb{S}^d$ into a vector of moments with growing dimension. Let $k = k_n$ be a sequence of positive integers satisfying $k\to\infty$ as $n\to\infty$, and choose points (or sieve) $\{p_1,\ldots,p_k\}$ from $\mathbb{S}^d$ so that in the limit

[4] For the identified set $\Theta_0 = \{\theta : E[m(\theta)] \le 0\}$ defined by a finite number of moment inequalities, Chernozhukov, Kocatulum and Menzel (2015) proposed a confidence region that is invariant to arbitrary one-to-one mappings of the form $\tau : \Theta_0 \to \Psi$. However, their construction does not apply in general to our setup which is concerned with testing $\mathbb{E}[X_i] = \Theta_0$ implying the continuum of moment inequalities. In contrast, invariance of K_n is restricted to particular transformations (i.e., multiplication of both $\{X_i\}_{i=1}^n$ and Θ_0 by some non-singular matrix independent of i).

they form a dense subset of $\mathbb{S}^d$. By plugging in the nuisance parameter estimator $\hat{\nu}$, the sieve empirical likelihood function under the restrictions $E[s(X, p_j)] = s(\Theta_0(\nu), p_j)$ for $j = 1, \ldots, k$ is defined as

$$l_n = \max \left\{ \prod_{i=1}^{n} n w_i \middle| \sum_{i=1}^{n} w_i s(X_i, p_j) = s(\Theta_0(\hat{\nu}), p_j) \text{ for } j = 1, \ldots, k,\ w_i \geq 0,\ \sum_{i=1}^{n} w_i = 1 \right\}. \tag{2.8}$$

If there is no nuisance parameter (i.e., $\Theta_0(\nu) = \Theta_0$), we can simplify the proof of Theorem 5 below to show that $(-2 \log l_n - k)/\sqrt{2k} \xrightarrow{d} N(0, 1)$ under the null $\mathrm{H}_0 : \mathbb{E}[X] = \Theta_0$. When there are nuisance parameters, the statistic l_n containing $\hat{\nu}$ is not internally studentized (i.e., $(-2 \log l_n - k)/\sqrt{2k}$ does not converge to the standard normal) due to the variance of $\hat{\nu}$. To recover internal studentization, we penalize the dual form of l_n as

$$L_n = \sup_{\lambda \in \Lambda_n} 2 \sum_{i=1}^{n} \log(1 + \lambda' m_k(X_i)) - n \lambda' (\bar{V}_k - \hat{V}_k) \lambda, \tag{2.9}$$

where $m_k(X_i) = [s(X_i, p_1) - s(\Theta_0(\hat{\nu}), p_1), \ldots, s(X_i, p_k) - s(\Theta_0(\hat{\nu}), p_k)]'$ and Λ_n,$\bar{V}_k$, and $\hat{V}_k$ are defined in Appendix B.1. The limiting null distribution of the penalized statistic L_n is obtained as follows.

Theorem 5. *Under Assumption S in Appendix B.1, it holds that* $(L_n - k)/\sqrt{2k} \xrightarrow{d} N(0, 1)$ *under* H_0. *In addition,* $(L_n - k)/\sqrt{2k}$ *diverges to infinity under* H_1.

By adapting the proof of Theorem 5, we can show that under the local alternative

$$\mathrm{H}_{1n} : E[s(X, p)] = s(\Theta_0(\nu), p) + a_n \eta(p) \text{ over } p \in \mathbb{S}^d,$$

for some continuous function η, where $a_n = k^{1/4}/\sqrt{n \eta_k' \dot{V}_k \eta_k}$ and $\eta_k = (\eta(p_1), \ldots, \eta(p_k))'$, the sieve empirical likelihood statistic satisfies $(L_n - k)/\sqrt{2k} \xrightarrow{d} N(2^{-1/2}, 1)$. Therefore, the test statistic $(L_n - k)/\sqrt{2k}$ has non-trivial local power against a local alternative at the a_n-rate. Also, we note that similar to the marked empirical likelihood statistic K_n, both l_n and L_n are invariant to scale transformations (i.e., multiplication of both $\{X_i\}_{i=1}^n$ and Θ_0 by some non-singular matrix independent of i).

Compared to the marked empirical likelihood statistic studied in Section 2.2.1, the sieve

empirical likelihood statistic L_n is asymptotically pivotal but requires choosing the sieve $\{p_1, \ldots, p_k\}$. A natural choice for locations of the sieve $\{p_1, \ldots, p_k\}$ is a grid of equidistant angle values in $\mathbb{S}^d$. The main remaining problem for practical implementation is choosing the tuning parameter k. In the literature on empirical likelihood, several statistics have been proposed possessing the same feature (i.e., asymptotically pivotal but depending on smoothing parameters), see for instance Fan, Zhang and Zhang (2001), Chen, Härdle and Li (2003), and Fan and Zhang (2004). Following the insight of Fan, Zhang and Zhang (2001) and Fan and Zhang (2004), one may choose k to be the maximizer $\arg\max_{k\in[n^c,n^{c'}]}(L_n - k)/\sqrt{2k}$ for some constants $c' \geq c > 0$. This results in a multi-scale test whose critical value can be obtained by bootstrap. For goodness-of-fit testing of parametric regression models, Fan and Huang (2001) showed adaptive minimaxity of such a test. A thorough analysis of mutli-scale testing in our setup is beyond the scope of this chapter.

2.3 Discussion and extensions

2.3.1 Test for directed hypotheses

It is possible to extend the methodology of marked empirical likelihood to test directed hypotheses of the form[5]

$$\mathrm{H}_0 : \Theta_0(\nu) \subseteq \mathbb{E}[X] \text{ vs. } \mathrm{H}_1 : \Theta_0(\nu) \nsubseteq \mathbb{E}[X]. \tag{2.10}$$

Beresteanu and Molinari (2008) were the first to develop a Wald type test for this problem. Here we propose empirical likelihood tests. By analogy with the testing problem in (2.1), the above is equivalent to testing the continuum of moment inequalities

$$\mathrm{H}_0 : s(\Theta_0(\nu), p) \leq E[s(X, p)] \text{ for all } p \in \mathbb{S}^d \text{ vs. } \mathrm{H}_1 : s(\Theta_0(\nu), p) > E[s(X, p)] \text{ for some } p \in \mathbb{S}^d.$$

[5]The null for the opposite direction $\mathrm{H}_0 : \mathbb{E}[X] \subseteq \Theta_0(\nu)$ can be treated analogously.

For a given direction p and preliminary estimator $\hat{\nu}$, the moment inequality restriction can be used to form the directed-marked empirical likelihood function

$$\vec{\ell}_n(p) = \max\left\{ \prod_{i=1}^n nw_i \middle| s(\Theta_0(\hat{\nu}), p) \leq \sum_{i=1}^n w_i s(X_i, p),\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\},$$

which can be equivalently written in the dual form as (see, Canay, 2010)

$$\vec{\ell}_n(p) = \min_{\lambda \leq 0} \prod_{i=1}^n \frac{1}{1 + \lambda\{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}}.$$

Therefore, the directed hypothesis in (2.10) can be tested by assessing the process $\{\vec{\ell}_n(p) : p \in \mathbb{S}^d\}$. In particular, we propose the directed Kolmogorov-Smirnov type statistic $\vec{K}_n = \sup_{p \in \mathbb{S}^d}\{-2\log \vec{\ell}_n(p)\}$. By similar arguments as in the proof of Theorem 4 (in particular, by modifying the proof of Hjort, McKeague and van Keilegom 2009, Theorem 2.1), we can show that $\vec{K}_n \xrightarrow{d} \sup_{p \in \mathbb{S}^d} \frac{\min\{Z(p) - G(p;\nu)' Z_1, 0\}^2}{\mathrm{Var}(s(X,p))}$ under H_0. The same also applies for testing the hypothesis $\mathrm{H}_0 : \theta_0 \in \mathbb{E}[X]$ for a singleton $\theta_0 \in \mathbb{R}^d$. In this case, we simply set $s(\Theta_0(\nu), p) = s(\Theta_0(\hat{\nu}), p) = p'\theta_0$.

It may be possible to extend the construction of the sieve empirical likelihood statistic to test the directed hypotheses in (2.10) by replacing the equality constraints $\sum_{i=1}^n w_i s(X_i, p_j) = s(\Theta_0(\hat{\nu}), p_j)$ in (2.8) with the inequalities $\sum_{i=1}^n w_i s(X_i, p_j) \geq s(\Theta_0(\hat{\nu}), p_j)$ for $j = 1, \ldots, k$. If k is fixed, we can apply the results of Canay (2010) to investigate its asymptotic properties. However, for the case of $k \to \infty$, the asymptotic analysis of the statistic is very different and is beyond the scope of this chapter.

2.3.2 Linear transform and projection

Our empirical likelihood approach can be easily modified to test hypotheses on a linear transform $R\mathbb{E}[X]$ of the Aumann mean, where R is an $l \times d$ constant matrix with $l < d$ and full row rank. The first test for such hypotheses was proposed by Beresteanu and Molinari (2008) who employed a Wald type statistic based on the Hausdorff metric. Here we provide empirical likelihood based alternatives. Since the null hypothesis $\mathrm{H}_0^R : R\mathbb{E}[X] = R\Theta_0(\nu)$ is equivalent to $\mathrm{H}_0^R : E[s(X, R'q)] = s(\Theta_0(\nu), R'q)$ for all $q \in \mathbb{S}^l$, this motivates the use of the

marked empirical likelihood function $\ell_n(R'q)$ for $q \in \mathbb{S}^l$, and the Kolmogorov-Smirnov type statistic $K_n^R = \sup_{q\in\mathbb{S}^l}\{-2\log \ell_n(R'q)\}$ for testing the null. By the invariance property, the latter is simply $K_n^R = \sup_{p\in\Delta}\{-2\log \ell_n(p)\}$, where $\Delta = \{R'q/\left\|R'q\right\| : q \in \mathbb{S}^l\}$ is a subset of $\mathbb{S}^d$. Thus, the test statistic K_n^R for the linear transform is given by taking the supremum over a particular subset $\Delta \subset \mathbb{S}^d$ rather than the whole set $\mathbb{S}^d$ as is the case with K_n. A modification of Theorem 4 then implies $K_n^R \xrightarrow{d} \sup_{p\in\Delta} \frac{\{Z(p)-G(p;\nu)'Z_1\}^2}{\mathrm{Var}(s(X,p))}$ under H_0^R. It is also possible to extend the sieve empirical likelihood approach to test H_0^R by choosing a sieve on Δ.

Now let us discuss one of the most important examples: testing for the projection of $\mathbb{E}[X]$ to one of its components. We argue that in this case the sieve empirical likelihood (with profiling out for ν) is particularly attractive. Suppose we are interested in the first component (i.e., $R = [1, 0, \ldots, 0]$). In this case, the null hypothesis $\mathrm{H}_0^R : R\mathbb{E}[X] = R\Theta_0(\nu)$ reduces to the two moment constraints $\mathrm{H}_0^R : E[s(X, R'q)] = s(\Theta_0(\nu), R'q)$ for $q = \pm 1$. Let ν be defined through the estimating equations $E[m(z_i, \nu)] = 0$ for observables z_i.[6] Then the sieve empirical likelihood reduces to the conventional empirical likelihood:

$$l_n(\nu) = \max\left\{ \prod_{i=1}^n nw_i \,\middle|\, \sum_{i=1}^n w_i \begin{pmatrix} s(X_i, R') - s(\Theta_0(\nu), R') \\ s(X_i, -R') - s(\Theta_0(\nu), -R') \\ m(z_i, \nu) \end{pmatrix} = 0,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\}.$$

By Qin and Lawless (1994), mild regularity conditions guarantee Wilks' theorem, that is $-2\max_\nu\{\log l_n(\nu)\} \xrightarrow{d} \chi^2_2$ under H_0^R. In this case, we recommend internalizing the nuisance parameters ν and profiling them out because the statistic $l_n(\hat{\nu})$ with a preliminary estimator $\hat{\nu}$ is not asymptotically pivotal in general. See Section 2.3.3 below for further discussion.

2.3.3 Profile likelihood

In Section 2.2, we considered empirical likelihood statistics where the nuisance parameters ν are replaced with a preliminary estimator $\hat{\nu}$. This approach is particularly practical when the dimension of ν is high. On the other hand, as explained in the last subsection, there are some situations where profiling out ν may be desirable to achieve asymptotic pivotalness.

[6] When ν is defined by a smooth function of means, it can be treated as in Owen (2001, Section 3.4).

Here we discuss some such extensions for profiling out ν. Again, suppose throughout that ν is defined by some estimating equations $E[m(z_i, \nu)] = 0$ for observables z_i.

The marked profile empirical likelihood can be defined as $\ell_n^P(p) = \max_\nu \ell_n(p, \nu)$, where

$$\ell_n(p,\nu) = \max\left\{ \prod_{i=1}^n nw_i \middle| \sum_{i=1}^n w_i \begin{pmatrix} s(X_i,p) - s(\Theta_0(\nu),p) \\ m(z_i,\nu) \end{pmatrix} = 0,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\}.$$

There is a computational drawback of this approach: it requires optimization with respect to ν for each p. Although the technical arguments would be more involved than the plug-in case, by extending the argument in Qin and Lawless (1994, Corollary 5) we can obtain the limiting distribution of the process $\ell_n^P(p)$. In particular, defining $g_i(p,\nu) = [s(X_i,p) - s(\Theta_0(\nu),p), m(z_i,\nu)]'$, we can show that $\sup_{p\in\mathbb{S}^d}\{-2\log \ell_n^P(p)\}$ will converge to $\sup_{p\in\mathbb{S}^d}\{\tilde{Z}(p)'\tilde{Z}(p)\}$, where $\tilde{Z}(p)' = [Z(p), Z_1'] \left(I - S(p)\left(S(p)'\Omega(p)^{-1}S(p)\right)^{-1} S(p)'\right)\Omega(p)^{-1/2}$, with Z_1 denoting the limiting distribution of $n^{-1/2}\sum_{i=1}^n m(z_i,\nu_0)$, $S(p) = \begin{bmatrix} G(p;\nu_0)' \\ E[\partial m(z_i,\nu_0)/\partial \nu'] \end{bmatrix}$ (here $G(p;\nu_0)'$ is as defined in (2.5) and the existence of $E[\partial m(z_i,\nu_0)/\partial\nu']$ is assumed), and $\Omega(p) = \mathrm{Var}(g_i(p,\nu_0))$. We note the limiting distribution is still not pivotal, and the critical value needs to be approximated by bootstrap.

Similarly, the sieve profile empirical likelihood can be defined as $l_n^P = \max_\nu l_n(\nu)$, where

$$l_n(\nu) = \max\left\{ \prod_{i=1}^n nw_i \middle| \sum_{i=1}^n w_i \begin{pmatrix} s(X_i,p_1) - s(\Theta_0(\nu),p_1) \\ \vdots \\ s(X_i,p_k) - s(\Theta_0(\nu),p_k) \\ m(z_i,\nu) \end{pmatrix} = 0,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\}.$$

Compared to the marked profile empirical likelihood $\ell_n^P(p)$, the sieve statistic l_n^P is more tractable because it requires optimization with respect to ν only once. Additionally, by arguing as in Donald, Imbens and Newey (2003, Theorems 6.3-6.4), it can be shown that the null distribution is standard normal, i.e. $(l_n^P - k)/\sqrt{2k} \xrightarrow{d} N(0,1)$ under certain conditions. Thus, the profile statistic l_n^P is asymptotically pivotal without the need for penalization as in (2.9).

2.3.4 Inference based on estimated random sets

In some applications, the random set of interest X is not directly observable because it contains some parameters to be estimated. For example, in the context of treatment effect analysis in experimental studies, Balke and Pearl (1997) proposed nonparametric bounds on the average treatment effect when the treatment assignment is random but subject compliance is imperfect. In a general form, Balke and Pearl's (1997) bound on the average treatment (ATE) can essentially be written as

$$\max_{1\leq j\leq J_L} \frac{E[g_{Li}^j]}{E[h_{Li}^j]} \leq ATE \leq \max_{1\leq j\leq J_U} \frac{E[g_{Ui}^j]}{E[h_{Ui}^j]}, \tag{2.11}$$

where g_{Li}^j $(j = 1, \ldots, J_L)$ and g_{Ui}^j $(j = 1, \ldots, J_U)$ are observable scalar random variables. By applying the "smooth-max" approximation (Chernozhukov, Kocatulum and Menzel, 2015), these bounds can be approximated by $\sum_{j=1}^{J_L} w_A^j E[g_{Ai}^j]/E[h_{Ai}^j]$ with

$$w_A^j = e^{\varrho E[g_{Ai}^j]/E[h_{Ai}^j]} \Big/ \left(\sum_{j=1}^{J_A} e^{\varrho E[g_{Ai}^j]/E[h_{Ai}^j]} \right)$$

for $A = L$ and U. Indeed, the approximation error satisfies

$$\left| \sum_{j=1}^{J_A} w_A^j E[g_{Ai}^j]/E[h_{Ai}^j] - \max_{1\leq j\leq J_A} E[g_{Ai}^j]/E[h_{Ai}^j] \right| = O(\varrho^{-1})$$

for $A = L$ and U. Thus by choosing ϱ large enough, the bounds on the ATE given above are well approximated by the Aumann expectation $\mathbb{E}[X_i(\gamma)]$ of the SVRV

$$X_i(\gamma) = \left[\sum_{j=1}^{J_L} w_L^j g_{Li}^j/E[h_{Li}^j], \sum_{j=1}^{J_U} w_U^j g_{Ui}^j/E[h_{Ui}^j] \right],$$

where $\gamma = (E[g_{Li}^1], \ldots, E[g_{Li}^{J_L}], E[h_{Li}^1], \ldots, E[h_{Li}^{J_L}], E[g_{Ui}^1], \ldots, E[g_{Ui}^{J_U}], E[h_{Ui}^1], \ldots, E[h_{Ui}^{J_U}])'$. In this case, the SVRV of interest $X_i(\gamma)$ is not observable because it contains unknown parameters γ.

In order to test null hypotheses of the form $\mathrm{H}_0 : \mathbb{E}[X(\gamma)] = \Theta_0(\nu)$, the marked empirical likelihood function $\ell_n(p)$ in (2.3) can be modified by replacing X_i with the estimated

counterpart $X_i(\hat{\gamma})$, where $\hat{\gamma}$ is an estimator of γ. By imposing assumptions analogous to Assumption M (i)-(iii) to deal with the estimation error of $X_i(\hat{\gamma}) - X_i(\gamma)$ along with the assumption $\sup_{p\in\mathbb{S}^d} E[|s(X_i(\gamma_m), p) - s(X_i(\gamma), p)|^2] \to 0$ for all $\gamma_m \to \gamma$, we can show that

$$K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{\{Z(p) - G(p;\nu)'Z_1 + \Gamma(p;\gamma)'Z_2\}^2}{\mathrm{Var}(s(X(\gamma), p))},$$

where $(Z(p), Z_1', Z_2')' \sim N(0, \tilde{V}(p))$, $\tilde{V}(p)$ is the limiting covariance matrix of $(n^{-1/2}\sum_{i=1}^n \{s(X_i, p) - E[s(X, p)]\}, \sqrt{n}(\hat{\nu} - \nu)', \sqrt{n}(\hat{\gamma} - \gamma))'$, and $\Gamma(p;\gamma)$ is a function such that

$$|E[s(X(\hat{\gamma}), p)] - E[s(X(\gamma), p)] - \Gamma(p;\gamma)'(\hat{\gamma} - \gamma)| = o_p(n^{-1/2}).$$

To obtain a critical value for testing, we can adapt the bootstrap procedure presented in Proposition 1 (by replacing X_i^* and $\bar{s}(p)$ in (2.7) with $X_i^*(\hat{\gamma}^*)$ and $n^{-1}\sum_{i=1}^n s(X_i(\hat{\gamma}), p)$, respectively). The asymptotic validity of this bootstrap procedure can be shown under the additional condition: $\sup_{p\in\mathbb{S}^d} |\bar{s}(X_i(\hat{\gamma}^*), p) - \bar{s}(X_i(\hat{\gamma}), p) - \Gamma(p;\gamma)'(\hat{\gamma}^* - \hat{\gamma})| = o_{p^*}(n^{-1/2})$ with probability approaching 1.

It is also possible to employ the sieve empirical likelihood statistic by replacing X_i in (2.8) with the estimated set $X_i(\hat{\gamma})$. Recall that in Section 2.2.2 we were able to incorporate nuisance parameters into the sieve statistic by linearizing the term $s(\Theta_0(\hat{\nu}), p) - s(\Theta_0(\nu), p)$ and incorporating the effect of the resulting additional terms via penalization (see Appendix B.1 for more details). We can proceed similarly for the case of estimated sets if we impose the following assumption enabling linearization of $\bar{s}(X_i(\hat{\gamma}), p) - \bar{s}(X_i(\gamma), p)$ as

$$\sup_{p\in\mathbb{S}^d} |\bar{s}(X_i(\hat{\gamma}), p) - \bar{s}(X_i(\gamma), p) - \bar{\Gamma}(p;\gamma)'(\hat{\gamma} - \gamma)| = o_p(n^{-1/2}),$$

where $\bar{\Gamma}(.;.)$ is the derivative of $\bar{s}(X_i(\gamma), p)$ with respect to γ satisfying some regularity properties akin to Assumption S (iii) (i.e., (i) $\bar{\Gamma}(p;\gamma)$ converges uniformly in both p and ν to a non-stochastic $\Gamma(p;\gamma)$ satisfying $\sup_{p\in\mathbb{S}^d} \|\Gamma(.,\gamma)\| < \infty$ and (ii) for all $\tilde{\gamma}$ in some neighborhood of γ, $\sup_{p\in\mathbb{S}^d} \left\|\bar{\Gamma}(p;\tilde{\gamma}) - \bar{\Gamma}(p;\gamma)\right\| \leq M \|\tilde{\gamma} - \gamma\|^\alpha$ for some $\alpha \geq 2/3$ and $M < \infty$ independent of $\tilde{\gamma}$). By a straightforward modification of the penalty term in (2.9), we can obtain a corresponding result to Theorem 5 for the case of estimated random sets.

Alternatively, it is possible to employ a profile likelihood approach as in section (2.3.3); this is particularly attractive for tests on low dimensional projections of the set $\Theta_0(\nu)$.

2.3.5 Measurements on support function

In medical imaging and robotic vision, researchers sometimes directly observe measurements of the support function of a convex set of interest (see, Fisher *et al.*, 1997). When noiseless measurements of $\{s(X_i, \cdot)\}_{i=1}^n$ are available, the marked empirical likelihood method can be applied immediately to hypothesis testing. Another common statistical question in image analysis of convex shaped data is to recover a set of interest from noisy measurements of its support function. In this problem, we observe the pairs $\{s_i, p_i\}_{i=1}^n$, where $s_i = s(\Theta, p_i) + \epsilon_i$ with error ϵ_i and $p_i \in \mathbb{S}^d$. Fisher *et al.* (1997) developed an estimation method for Θ by estimating the support function $s(\Theta, \cdot)$ nonparametrically. Our empirical likelihood approach can be adapted to test the hypothesis that Θ takes a particular shape Θ_0, such as a circle or ellipse. The marked empirical likelihood function under the restriction $E[s_i|p_i = p] = s(\Theta_0, p)$ may be constructed as

$$\tilde{\ell}_n(p) = \max\left\{ \prod_{i=1}^n n w_i \middle| \sum_{i=1}^n w_i K_b(p_i - p)\{s_i - s(\Theta_0, p)\} = 0,\ w_i \geq 0,\ \sum_{i=1}^n w_i = 1 \right\}, \quad (2.12)$$

where $K_b(\cdot)$ is a kernel function depending on the smoothing parameter b. For example, the Cramér-von Mises type statistic, given by $T_n = \int_{p\in\mathbb{S}^d} -2\log\tilde{\ell}_n(p)dp$, can be shown to be asymptotically normal under the null after certain normalizations as in Chen, Härdle and Li (2003). Alternatively, following Härdle and Mammen (1993), a wild bootstrap method (i.e., resampling $s_i^* = s(\Theta_0, p_i) + v_i^* \hat{\epsilon}_i$ with $\hat{\epsilon}_i = s_i - s(\Theta_0, p)$ and $v_i^* \sim$two-point distribution) can be applied to obtain the critical value.

Simulation results, presented in Appendix B.5, demonstrate reasonable size and power properties for our empirical likelihood test.

2.4 Examples

2.4.1 Best linear prediction with interval valued dependent variable

We first consider the issue of best linear prediction with interval valued dependent variables. In particular, we employ the setup of Beresteanu and Molinari (2008), follow their argument, and use the characterization they provide. See also Bontemps, Magnac and Maurin (2012) for an extension to instrumental variable regression.

In usual regression models, we are mostly interested in the best linear relationship between a dependent variable y and independent variables x, which can be estimated by the least squares method. On the other hand, if y is unobservable but we observe the interval $[y_L, y_U]$ to which y belongs almost surely, it would be of interest to conduct inference on the set of the least squares coefficients $\Upsilon = \{\arg\min_\theta \int \{y - (1, x')\theta\}^2 d\mu \text{ for some } \mu \in \mathcal{M}\}$, where $\mathcal{M}$ is the set of distributions of (y, x) compatible with $y \in [y_L, y_U]$ almost surely. There are numerous examples of interval data, including data on wealth (e.g., the Health and Retirement Study) and income (e.g., the Current Population Survey), top coding in surveys, and ordered categorical measurements (e.g., age, expenditure, GPA, and so on). By using the Aumann expectation for the random set $W = \begin{pmatrix} [y_L, y_U] \\ [xy_L, xy_U] \end{pmatrix} \subset \mathbb{R}^{\dim(x)+1}$, the set of least square coefficients may be written as $\Upsilon = \Sigma^{-1}\mathbb{E}[W]$, where $\Sigma = E\begin{pmatrix} 1 & x' \\ x & xx' \end{pmatrix}$ (see, Beresteanu and Molinari, 2008, Proposition 4.1).[7]

We note that if there is no intercept in the regression and x is scalar (or there is only an intercept), then the set of best linear predictors is the interval $\Upsilon = [E[xy_L]/E[x^2], E[xy_U]/E[x^2]]$. Thus, inference on Υ may be conducted by the conventional empirical likelihood for the vector of parameters $(E[xy_L], E[xy_U], E[x^2])$ or via regressions of y_L and y_U on the scalar x. However, if the regression model contains an intercept or x is a vector, then the set Υ is multi-dimensional and neither the conventional empirical likelihood for $(E[(1, x')y_L], E[(1, x')y_U], \Sigma)$ nor regressions of y_L and y_U on $(1, x')$ are sufficient for characterizing it completely. Intu-

[7] Chandrasekhar *et al.* (2012) extended this model further to allow for y_L and y_U to be nonparametrically estimable functions. Although it is beyond the scope of this chapter, it would be interesting to extend our empirical likelihood approach to such situations.

itively this is because, as can be seen from the characterization of the support function of Υ given below, we also need to consider situations where some observations of y take the value y_L while the others take y_U.

For the following theoretical results we shall suppose that x is a continuous random variable which ensures Υ is strictly convex. Regarding the support function, the null hypothesis $\mathrm{H}_0 : \Upsilon = \Upsilon_0$ for a strictly convex Υ_0 can be written as $\mathrm{H}_0 : E[s(W,p)] = s(\Sigma\Upsilon_0, p)$ for all $p \in \mathbb{S}^d$, where $s(W,p) = [y_L + (y_U - y_L)\mathbb{I}\{(1,x')p \geq 0\}](1,x')p$ and $d = \dim(x) + 1$. This is equivalent to the general setup of Section 2 if one defines $\Theta_0(\nu) = \Sigma\Upsilon_0$, where the nuisance parameter $\nu = \mathrm{vec}(\Sigma)$ is estimated by its sample counterpart $\mathrm{vec}(\hat{\Sigma})$. Furthermore, since $s(\Sigma\Upsilon_0, p) = s(\Upsilon_0, \Sigma p)$, the support function of the set $\Sigma\Upsilon_0$ can be computed from that of Υ_0. Let $\nabla s(\Upsilon_0, p)' = [y_L + (y_U - y_L)\mathbb{I}\{(1,x')p \geq 0\}](1,x')$ be the Fréchet derivative of $s(\Upsilon_0, p)$ with respect to p, and define $G(p;\nu) = p \otimes \nabla s(\Upsilon_0, \Sigma p)$, where $\otimes$ represents the Kronecker product. Note that $G(p;\nu)'$ is the pointwise derivative of $s(\Sigma\Upsilon_0, p)$ ($s(\Theta_0(\nu_0))$ in the terminology of Section 2) with respect to $\nu = \mathrm{vec}(\Sigma)$. In this setup, the null distributions of the empirical likelihood statistics are obtained as follows.

Proposition 2. *Consider the setup of this subsection. Assume that $\{y_{Li}, y_{Ui}, x_i\}_{i=1}^n$ is i.i.d., where the distribution of x_i is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{d-1}$, and Σ is full rank.*

(i) *Suppose $E[\|(y_{Li}, y_{Ui}, x_i'y_{Li}, x_i'y_{Ui})\|^\xi] < \infty$ for some $\xi > 2$, $E[\|x_i\|^4] < \infty$, and $\mathrm{Var}(y_{Li}|x_i), \mathrm{Var}(y_{Ui}|x_i) \geq \underline{\sigma}^2$ a.s. for some $\underline{\sigma}^2 > 0$. Then $K_n \xrightarrow{d} \sup_{p\in\mathbb{S}^d} \frac{\tilde{Z}(p)^2}{\mathrm{Var}(s(W_i,p))}$ under H_0, where $\tilde{Z}(\cdot) = Z(\cdot) - G(\cdot;\nu)'\Gamma$ is the Gaussian process implied from $(Z(p), \Gamma)' \sim N(0, \tilde{V}(p))$ and $\tilde{V}(p)$ is the covariance matrix of the vector $(s(W_i,p), \{z_i - vec(\Sigma)\}')$.*

(ii) *Suppose $E[\|(y_{Li}, y_{Ui}, x_i'y_{Li}, x_i'y_{Ui})\|^\xi] < \infty$ for some $\xi \geq 4$, $E[\|x_i\|^4] < \infty$, and $\nabla s(\Upsilon_0, p)$ is locally Hölder continuous of order $\alpha \geq 2/3$ over the domain $\mathbb{S}^d$. Also assume $k \to \infty$ and $(k^5\dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$, where $\dot{\phi}_k$ is defined in Appendix B.1. Then $\frac{L_n - k}{\sqrt{2k}} \xrightarrow{d} N(0,1)$ under H_0.*

The assumptions are similar to those of Beresteanu and Molinari (2008, Theorem 4.3). These results are obtained by verifying the conditions in Theorems 4 and 5. The critical values for the marked empirical likelihood test may be obtained by the bootstrap procedure presented in Proposition 1.

We now evaluate the finite sample performance of our test statistic by conducting inference on the returns to education on (log) wages using data from the Current Population Survey (CPS). We use data from the March 2009 wave of the CPS on white males aged between 20 and 50 who earn at least \$1000/year. This gives 18017 observations on wages and education. Analogous to the construction in Beresteanu and Molinari (2008), the wage data (in thousands of dollars) is artificially bracketed and top-coded in terms of the following brackets (the top coding value is \$100 million):
[1, 5], [5, 7.5], [7.5, 10], [10, 12.5], [12.5, 15], [15, 20], [20, 25], [25, 30], [30, 35], [35, 40], [40,50], [50, 60], [60, 75], [75, 100], [100, 150], [150, 100000]

Thus, the variables (y_{Li}, y_{Ui}, x_i) correspond to lower and upper bounds of log wages and education, respectively. We draw 5000 samples of size n =100, 200, 500, 1000, and 2000 from the 'true' population (consisting of 18017 observations from the CPS) and conduct inference for Υ, the set of intercept and slope coefficients consistent with the population data. Table 2.1 reports the rejection frequencies of the marked empirical likelihood test under the nominal 5% rejection level. This is compared with Wald-type test statistics based on the Hausdorff distances $nd_H\left(\hat{\Sigma}^{-1}\frac{1}{n}\oplus_{i=1}^n W_i, \Upsilon_0\right)^2$ and $nd_H\left(\frac{1}{n}\oplus_{i=1}^n W_i, \hat{\Sigma}\Upsilon_0\right)^2$ (called Wald 1 and 2, respectively). The first Wald-type test was proposed by Beresteanu and Molinari (2008). For both the marked empirical likelihood and Wald tests, the critical values are obtained by the bootstrap calibrations outlined in Section 2.2 with 399 repetitions. In Table 2.1 it is seen that the marked empirical likelihood test has good size control and performs better than both Wald tests for smaller sample sizes. As explained previously, the Wald statistic is not invariant to multiplication of the sets by a constant matrix unlike the empirical likelihood tests; this drawback is evident in the different sizes for the two Wald tests.[8] The statistics vary considerably along p; for some directions $p = (\cos\vartheta, \sin\vartheta)'$ with $\vartheta = \left(0, \frac{\pi}{3}, \frac{\pi}{4}, \frac{2\pi}{3}, \frac{\pi}{2}\right)$, the critical values of Wald 1, marked EL, and $\hat{V}(\hat{\Sigma}p)$ are $(5.3\times10^{-2}, 2.0\times10^{-5}, 1.4\times10^{-4}, 2.5\times10^{-4})$, $(10, 6.8, 4.3, 2.0, 0.14)$, and $(7.5, 337.4, 610.0, 870.8, 1.1\times10^{3})$, respectively.

We can also adapt the construction of the confidence set based on K_n, described in

[8] As expected, however, the marked empirical likelihood test is computationally more expensive than the Wald test. In particular, for sample size $n = 1000$, the marked empirical likelihood test with 399 bootstrap repetitions has an average run time of 5.7 seconds as compared to 0.6 seconds for the Wald test.

Section 2.2, to the present context. We exploit the invariance property of K_n which ensures that with probability α the inequalities $s(\Upsilon, p) \leq n^{-1} \sum_{i=1}^{n} s(\hat{\Sigma}^{-1} X_i, p) + \sqrt{\frac{\hat{c}_\alpha}{n}} \hat{V}(\hat{\Sigma}^{-1} p)^{1/2}$ hold asymptotically for all $p \in \mathbb{S}^d$, where $\hat{c}_\alpha$ estimates the α-th quantile of the limiting distribution of K_n. In particular, we can obtain $\hat{c}_\alpha$ by the bootstrap procedure presented in Section 2.3.4. Figure 2.1 displays the 95% confidence region thus obtained for a sample size of $n = 1000$, along with the 'true' population region and also the confidence region from the Wald-type test proposed in Beresteanu and Molinari (2008). It can be seen that the confidence region based on K_n covers an area that is much less ($< 5\%$) than the one based on the Wald test.

We can also employ our inferential procedures to obtain confidence intervals for the best linear predictor of the (log) wage y given some education x. This is equivalent to providing a confidence region for the projection $R\Upsilon_0$ where $R = (1, x)$. To this end, we can use the results from Section 2.3.4 on estimated random sets by exploiting the fact $E[s(\Sigma^{-1} W, R'q)] = s(\Upsilon, R'q)$, where setting $q = 1$ and -1 gives the upper and lower bounds for the confidence interval. Table 2.2 reports the estimated prediction intervals for the cases when $x = 12$ (corresponding to high school education) and $x = 16$ (corresponding to undergraduate degree). For computational reasons we report the results for profile likelihood using the Euclidean likelihood function (c.f. Section (2.2.1)). The profile likelihood is used to obtain a joint confidence set for the upper and lower bounds of the interval, from which we obtain a necessarily conservative confidence interval by taking the worst possible value for each of the components. Nevertheless, the length of the confidence interval is comparable to, or smaller, than those based on the Marked EL and Wald statistics.

In Appendix B.4, we report additional numerical results to compare the marked empirical likelihood confidence region - displayed in Figure 2.1 - with the one based on the method by Chernozhukov, Kocatulum and Menzel (2015).

2.4.2 Boolean model

In the context of mathematical morphology, geostatistics, and particle statistics, researchers often observe a series of two or three dimensional random sets, such as tumors and sand or rock grains (see, Stoyan, 1998, for a review). One of the most popular models to explain the

n	Size (Marked EL)	Size (Wald 1)	Size (Wald 2)
100	0.038	0.098	0.107
200	0.049	0.073	0.081
500	0.057	0.069	0.059
1000	0.053	0.057	0.059
2000	0.050	0.056	0.058

Table 2.1: Rejection frequencies of the marked empirical likelihood and Wald tests at the nominal 5% level

Education	True Region	Profile Lik.	Marked EL	Wald
High school degree	[3.549, 3.931]	[3.454, 3.999]	[3.456, 3.995]	[3.465, 3.983]
Undergraduate degree	[4.020, 4.915]	[3.967, 5.051]	[3.906, 5.003]	[3.873, 4.976]

Table 2.2: 95% confidence intervals for the best linear predictor of (log) wage y given education x using profile likelihood, marked Empirical Likelihood and Wald statistics

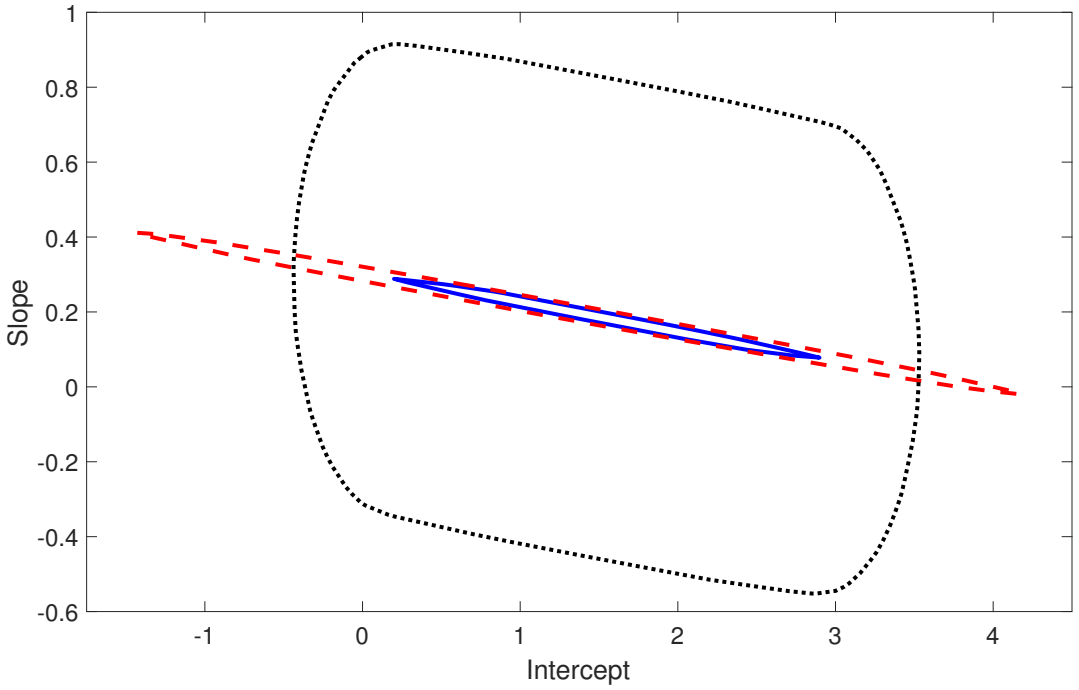

Figure 2.1: The population identification region (solid line) and the corresponding 95% confidence regions using the marked empirical likelihood statistic (dashed line) and the Wald statistic (dotted line) for sample size $n = 1000$.

growth pattern of these shapes is the Boolean model, where the random set is generated as $X = \cup_j \{W_j \oplus \{g_j\} : g_j \in G\}$ based on i.i.d. copies of random sets $W_j \subset \mathbb{R}^d$ $(j = 1, 2, \ldots)$, and a point process G in $\mathbb{R}^d$ for the foci $\{g_j : j \in \mathbb{N}\}$. For example, Cressie and Hulting (1992) developed a Boolean model to describe the growth of tumor shapes by specifying G to be a Poisson process with constant intensity function λ over a unit circle support. For simplicity, we shall assume that $W_j = W$ is a non-random ball with unknown radius R. We note that taking R to be non-stochastic is not too strong a requirement in this instance. Indeed, as seen in Cressie and Hulting (1992), the variance of R is an order of magnitude smaller than its mean. We thus consider $\gamma = (R, \lambda)$ as parameters of the tumor growth process which differ for normal and malignant tissues; consequently, we wish to conduct inference on these joint parameters.

To estimate $\gamma = (R, \lambda)$, Cressie and Hulting (1992) focused on the hitting probability (or capacity functionals). Alternatively, we can conduct inference using the Aumann expectation. More precisely, given the hypothesized parameter value γ_0, we can numerically evaluate the Aumann expectation $\Theta(\gamma_0) = \mathbb{E}[X(\gamma_0)]$. Then based on the sample $\{X_1, \ldots, X_n\}$ of tumor shapes of patients, the hypothesis $\mathrm{H}_0 : \gamma = \gamma_0$ can be tested via our methods for $\mathbb{E}[X] = \Theta(\gamma_0)$, specifically the marked (Section 2.1) and sieve empirical likelihood (Section 2.2) statistics.

We note that X may not be convex in this example. However, as long as X is compact valued, the Aumann expectation $\mathbb{E}[X]$ emerges as the almost sure limit of the Minkowski average of the sample $\{X_1, \ldots, X_n\}$. Therefore, the Aumann expectation can be intuitively interpreted as the 'average' shape of the observed sets. Furthermore, since the underlying probability measure is non-atomic in this example, it holds that $\mathbb{E}[X] = \mathbb{E}[\mathrm{co}(X)]$ (see, the discussion in Section 2.2). So, even though X is non-convex, our inferential procedures continue to hold after applying the convex hull operation (note: the support function remains unchanged since $s(X, p) = s(\mathrm{co}(X), p)$ for any compact X).

We present some Monte Carlo simulation based on Cressie and Hulting (1992) to evaluate the finite sample performance of our test statistics. In particular, we simulate the data from the estimated parameter values for γ obtained in Cressie and Hulting (1992, Table 3) with 5000 Monte Carlo replications for the sample sizes n =100, 200, and 500. Numerical

evaluation of $\Theta(\gamma_0)$ is achieved by averaging over 5000 draws of the process generated using the parameter value γ_0.

Table 2.3 reports the rejection frequencies of the marked empirical likelihood test under the nominal 5% rejection level. The null hypothesis is $\mathrm{H}_0 : \gamma_0 = (1.342, 4.046)$. We consider three types of alternatives $\mathrm{H}_1^a : (1.342, 4.5)$, $\mathrm{H}_1^b : (1.320, 4.046)$, and $\mathrm{H}_1^c : (1.320, 4.5)$. The critical values for this test are obtained by implementing the bootstrap procedure outlined in Section 2.3.4 with 99 repetitions. With respect to CPU seconds, the average computing time to obtain the bootstrap critical values is 4.85 for 399 repetitions and 1.84 for 99 repetitions. With additional parallel processing, we expect that these times may be further reduced. The first column indicates the test statistic has good size control over the sample sizes. The second and third columns show that the statistic is sensitive to slight changes in R and, to a lesser extent, changes in λ. This is consistent with the standard deviations of the estimates in Cressie and Hulting (1992, Table 3) which are large for λ compared to R. The fourth column reports the power properties of the test when changing both R and λ. In this case, these changes somewhat cancel each other out in the net effect (lower radius vs. higher number of foci), which explains why the alternative H_1^c is harder to reject.

Table 2.4 reports analogous results for the sieve empirical likelihood test. We construct the sieve from a grid of equidistant angle values corresponding to directions of the support function. We report outcomes for different values of sieve size k =3, 5, and 10. The critical values for the test are based on a χ^2_k calibration since, for the sample sizes and values of k considered, the theoretical normal approximation is found to be too rough. We see that the sieve empirical likelihood dominates the marked empirical likelihood in terms of power for all values of k while having comparable size control for smaller values of k.

So far we have considered inference for the joint hypothesis involving both parameters R and λ. By using our empirical likelihood tests with nuisance parameters, it is also possible to test the single parameter hypothesis $\mathrm{H}_0 : \lambda = \lambda_0$ by plugging-in an estimated value for R (e.g. the one in Cressie and Hulting, 1992).

n	H_0	H_1^a	H_1^b	H_1^c
100	0.059	0.259	0.389	0.074
200	0.063	0.461	0.679	0.099
500	0.071	0.856	0.973	0.173

Table 2.3: Rejection frequencies of the marked empirical likelihood test at the nominal 5% level

n, k	H_0	H_1^a	H_1^b	H_1^c
100, 3	0.058	0.328	0.475	0.078
100, 5	0.074	0.383	0.584	0.088
100, 10	0.102	0.393	0.521	0.121
200, 3	0.059	0.569	0.789	0.095
200, 5	0.066	0.648	0.890	0.101
200,10	0.085	0.581	0.847	0.110
500, 3	0.070	0.944	0.993	0.176
500, 5	0.082	0.974	0.999	0.202
500, 10	0.090	0.940	0.999	0.174

Table 2.4: Rejection frequencies of the sieve empirical likelihood test at the nominal 5% level

2.4.3 Treatment effect

We consider the problem of inference for nonparametric bounds on average treatment effects in the presence of imperfect compliance. In particular, we conduct a simulation study based on the Vitamin A supplementation example in Balke and Pearl (1997, Section 4.1). Briefly, the study consisted of administering doses of Vitamin A in a randomized trial to check for the effect on mortality. While the assignment to control and treatment groups was random, there were a substantial number of subjects who did not consume the treatment even when assigned to the treatment group. In the absence of any further assumptions on the relationship between compliance and response, Balke and Pearl (1997) obtained the sharpest possible bounds on the average treatment effect, which are of the form described in (2.11). Using the marked empirical likelihood statistic with estimated random sets proposed in Section 2.3.4, we can provide ways to conduct inference and construct confidence intervals for such bounds.

We use data simulated from the estimated joint probability distributions obtained in Balke and Pearl (1997, Tables 1 and 2) with 5000 Monte Carlo replications for each of the sample sizes n =500, 1000, 2500, and 5000. Note that the numerical example in Balke

n	H_0	H_1^a	H_1^b	H_1^c
500	0.053	0.193	0.122	0.197
1000	0.054	0.342	0.288	0.335
2500	0.051	0.891	0.967	0.940
5000	0.055	0.998	1.000	0.998

Table 2.5: Rejection frequencies of the marked empirical likelihood test at the nominal 5% level

and Pearl (1997) is based on over 20000 observations. We look at the size and power properties of the marked empirical likelihood test statistic under the null of the identified set $\mathrm{H}_0 : \Theta_0 = [-0.1946, 0.0054]$ and the alternative hypotheses obtained by expanding, contracting, and shifting Θ_0 to the left by a value of 0.025 (i.e., $\mathrm{H}_1^a : [-0.2196, 0.0304]$, $\mathrm{H}_1^b : [-0.1696, -0.0196]$, and $\mathrm{H}_1^c : [-0.2196, -0.0196]$, respectively). The critical value for the test is obtained by implementing the bootstrap procedure outlined in Section 2.3.4 with 399 repetitions. The tuning parameter ϱ for the 'smooth-max' approximation (cf. Section 2.3.4) employed in this test is chosen to be $\varrho = 1000$.

Table 2.5 reports the rejection frequencies of the marked empirical likelihood test under the nominal 5% rejection level. We can see that the our testing procedure has excellent size properties across all sample sizes (which are much smaller than the numerical example in Balke and Pearl, 1997). Also, our test has reasonable power properties against the three types of alternatives when the sample size is large enough.

A comparison with the Wald statistic of Beresteanu and Molinari (2008) shows that both statistics have similar size and power properties. The marked empirical likelihood test appears on average to have higher power, but the difference is marginal; in particular, the confidence regions are around 3.5% shorter. Because the results are so similar, we do not report the additional simulations here.

Chapter 3

Inference on distribution functions under measurement error

3.1 Introduction

This chapter is concerned with inference on the cumulative distribution function (cdf) F_{X^*} in the classical measurement error model $X = X^* + \epsilon$. Here, we observe X instead of X^* and ϵ is a measurement error. There is a rich literature on using density deconvolution for estimating the probability density function (pdf) f_{X^*} (see, Meister, 2009, for a review). By contrast, the literature on estimation and inference for the cdf F_{X^*} is relatively thin. Fan (1991) proposed a cdf estimator by integrating the deconvolution density estimator with some truncation. This truncation for the integral is circumvented in Hall and Lahiri (2008) (for the case where the pdf f_ϵ of ϵ is symmetric) and Dattner, Goldenshluger and Juditsky (2011) (for the case where f_ϵ is possibly asymmetric). Hall and Lahiri (2008) studied the L_2-risk properties of the cdf estimator. Dattner, Goldenshluger and Juditsky (2011) considered minimax rate optimal estimation of F_{X^*}. Both Hall and Lahiri (2008) and Dattner, Goldenshluger and Juditsky (2011) focused on the risk properties of the estimator $\hat{F}_{X^*}(t_0)$ at a given t_0 and assumed ordinary smooth densities for f_ϵ. These papers demonstrate that in contrast to the no measurement error case, the cdf estimator $\hat{F}_{X^*}(t_0)$ typically converges to $F_{X^*}(t_0)$ at a nonparametric rate. On the other hand, Söhl and Trabs (2012) established a uniform central limit theorem for linear functionals of the deconvolution estimator that can be applied to

derive a Donsker-type theorem, i.e., the weak convergence of $\sqrt{n}\{\hat{F}_{X^*}(\cdot) - F_{X^*}(\cdot)\}$ to a Gaussian process. Söhl and Trabs (2012) considered the case of ordinary smooth f_ϵ, and for the Donsker-type result obtained therein, it is demanded that the Fourier transform f_ϵ^{ft} satisfy $|f_\epsilon^{\mathrm{ft}}(\cdot)| \leq C|\cdot|^{-\beta}$ for some $\beta < 1/2$ and $C > 0$. The latter excludes the Laplace distribution, for instance. It must be emphasized that (except for Fan, 1991, on the truncated estimator) all these papers concentrate on the case of ordinary smooth and known f_ϵ, so the cases of super smooth and unknown f_ϵ (with repeated measurements) are not covered.

In this chapter, we investigate validity of asymptotic and bootstrap approximations for the distribution of the maximal deviation $T_n = \sup_{t\in\mathcal{T}} |\hat{F}_{X^*}(t) - F_{X^*}(t)|$ in the sup-norm over some set $\mathcal{T}$ between the deconvolution cdf estimator $\hat{F}_{X^*}$ of Hall and Lahiri (2008), and F_{X^*}. Our analysis allows f_ϵ to be ordinary or super smooth, or to be unknown and estimated by repeated measurements. We also characterize the convergence rate of the bootstrap approximation error and find that it is of polynomial order under ordinary smooth errors, and logarithmic order under super smooth errors. Our approximation results on the distribution of T_n are applicable to various contexts, such as confidence bands for F_{X^*} and its quantiles, and for performing various cdf-based tests such as goodness-of-fit tests for parametric models of densities, two sample homogeneity tests, and tests for stochastic dominance. We emphasize that some inference problems, such as testing for stochastic dominance, are cumbersome to be handled by density-based methods. Also, even in cases where density-based methods are applicable (e.g., goodness-of-fit testing), the cdf-based methods are expected to have desirable power properties.

In the context of density deconvolution, Bissantz, Dümbgen, Holzmann and Munk (2007) extended Bickel and Rosenblatt's (1973) construction of uniform confidence bands for densities to the classical measurement error model with the ordinary smooth f_ϵ. A recent paper by Kato and Sasaki (2016) considered confidence bands of the pdf f_{X^*} with unknown f_ϵ. In contrast to the above papers, this chapter is concerned with inference on the cdf F_{X^*}. Dattner, Reiß and Trabs (2016) proposed a quantile estimator of X^* and obtained the uniform convergence rate. This chapter provides a confidence band for the quantile function of X^*.

This chapter is organized as follows. In Section 3.2, we focus on the case of known f_ϵ and present the asymptotic and bootstrap approximations for T_n. Section 3.3 considers the case where f_ϵ is unknown but repeated measurements on X^* are available, and studies validity of a bootstrap approximation for the distribution of T_n. Section 3.4 contains four applications of the main results: a confidence band for quantiles (Section 3.4.1), goodness-of-fit test for parametric models of F_{X^*} (Section 3.4.2), homogeneity test for two samples (Section 3.4.3), and test for stochastic dominance (Section 3.4.4). Section 3.5 presents some simulation evidences. In Section 3.6, we consider a real data example. In particular, we employ the new test of stochastic dominance to study welfare changes of different population subgroups using potentially mis-measured income data from Korea. All proofs are contained in Appendix C.

3.2 Case of known measurement error distribution

3.2.1 Setup

We first introduce our basic setup. Suppose we observe a random sample $\{X_i\}_{i=1}^n$ generated from

$$X = X^* + \epsilon, \tag{3.1}$$

where X^* is an unobservable variable of interest and ϵ is its measurement error. Throughout the chapter, ϵ is assumed to be independent of X^* (i.e., ϵ is the classical measurement error). Let $\mathrm{i} = \sqrt{-1}$ and f^{ft} be the Fourier transform of a function f. If the pdf f_ϵ of ϵ is known, the pdf f_{X^*} of X^* can be estimated by the so-called deconvolution kernel density estimator (see, e.g., Stefanski and Carroll, 1990)

$$\hat{f}_{X^*}(t) = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{K}\left(\frac{t - X_i}{h}\right), \quad \text{where } \mathbb{K}(u) = \frac{1}{2\pi}\int_{-1}^{1} e^{-\mathrm{i}\omega u}\frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)}d\omega, \tag{3.2}$$

where h is a bandwidth and K is a kernel function with K^{ft} supported on $[-1, 1]$. Furthermore, if f_ϵ is symmetric, integration of $\hat{f}_{X^*}$ yields the following estimator for the cdf F_{X^*}

of X^* (see, Hall and Lahiri, 2008)

$$\hat{F}_{X^*}(t) = \frac{1}{2} + \frac{1}{n}\sum_{i=1}^{n} \mathbb{L}\left(\frac{t - X_i}{h}\right), \quad \text{where } \mathbb{L}(u) = \frac{1}{2\pi}\int_{-1}^{1} \frac{\sin(\omega u)}{\omega} \frac{K^{\text{ft}}(\omega)}{f_\epsilon^{\text{ft}}(\omega/h)} d\omega. \tag{3.3}$$

For the general case of possibly asymmetric f_ϵ, an estimator for F_{X^*} is obtained by replacing $\mathbb{L}(u)$ with $\mathbb{L}_a(u) = \frac{1}{\pi}\int_0^1 \frac{1}{\omega} \text{Im}\left[e^{-\text{i}\omega u} \frac{K^{\text{ft}}(\omega)}{f_\epsilon^{\text{ft}}(\omega/h)}\right] d\omega$ (Dattner, Goldenshluger and Juditsky, 2011), where Im$[\cdot]$ stands for the imaginary part. Although we hereafter focus on the cdf estimator in (3.3), our results can be extended to the general asymmetric case.

This section is concerned with approximation for the distribution of the maximal deviation

$$T_n = \sup_{t \in \mathcal{T}} |\hat{F}_{X^*}(t) - F_{X^*}(t)|, \tag{3.4}$$

under the sup-norm, where $\mathcal{T}$ is a compact interval specified by the researcher. A direct use of such approximation is construction of the confidence band for F_{X^*} over $\mathcal{T}$. Several other ways to use this approximation are presented in Section 3.4. In Section 3.2.2 below, we consider a bootstrap approximation for the distribution of T_n. In Section 3.2.3, we also present an asymptotic approximation based on the Gumbel distribution for ordinary smooth measurement error densities.

3.2.2 Bootstrap approximation

Consider a nonparametric bootstrap resample $\{X_i^\#\}_{i=1}^n$ from $\{X_i\}_{i=1}^n$ with equal weights. The bootstrap counterpart of T_n is given by $T_n^\# = \sup_{t \in \mathcal{T}} |\hat{F}_{X^*}^\#(t) - \hat{F}_{X^*}(t)|$, where $\hat{F}_{X^*}^\#$ is defined as in (3.3) using $X_i^\#$. To establish validity of the bootstrap approximation, we impose the following assumptions.

Assumption C. (i) $\{X_i\}_{i=1}^n$ is an i.i.d. sample from $X = X^* + \epsilon$. X^* and ϵ are independent. (ii) The densities f_X, f_{X^*}, and f_ϵ are bounded and continuous on $\mathbb{R}$, and $\inf_{t \in \mathcal{T}^\delta} f_X(t) > c$ for some $c > 0$ and δ-expansion $\mathcal{T}^\delta$ of $\mathcal{T}$. Also, $E|X^*| < \infty$ and $E|\epsilon| < \infty$. (iii) $\sup_{\omega \in \mathbb{R}}\{(1 + |\omega|)^\gamma |f_{X^*}^{\text{ft}}(\omega)|\} < C$ for some $\gamma, C > 0$. (iv) $f_\epsilon^{\text{ft}}(\omega) \neq 0$ for all $\omega \in \mathbb{R}$, $f_\epsilon^{\text{ft}}(\omega)$ is differentiable at all $\omega \in \mathbb{R}$, and f_ϵ is an even function.

Assumption C (i) is on the setup wherein we assume that ϵ is a classical measurement

error.[1] Assumption C (ii) is mild but excludes the Cauchy measurement error. This assumption is required for characterizing the bias of the estimator (see, e.g., Hall and Lahiri, 2008). The Cauchy measurement error is also ruled out in van Es and Uh (2005) who show pointwise asymptotic normality of the deconvolution density estimator. Assumption C (iii), analogous to the so-called Sobolev condition, is used to characterize the rate for the bias term (cf. Hall and Lahiri, 2008). Assumption C (iv) contains conditions on f_ϵ. The first condition is common in the density deconvolution literature but may be relaxed by taking a ridge approach as in Hall and Meister (2007). The last condition is used to derive the cdf estimator in (3.3) as in Hall and Lahiri (2008). Also when we consider estimation of f_ϵ using repeated measurements, symmetry of f_ϵ gives us a simple estimator (Delaigle, Hall and Meister, 2008).

We now present two classes of assumptions on the tail behavior of f_ϵ. The first is the class of ordinary smooth densities.

Assumption OS. (i) There exist $\beta > 1/2$ and $c, C, \omega_0 > 0$ such that

$$c|\omega|^{-\beta} \leq |f_\epsilon^{\mathrm{ft}}(\omega)| \leq C|\omega|^{-\beta},$$

for all $|\omega| \geq \omega_0$. (ii) K is an even function with $K^{\mathrm{ft}}(\omega) = (1-\omega^q)^r \mathbb{I}\{|\omega| \leq 1\}$ for some $q, r \geq 2$. There exist $c_1, C_1 > 0$ such that

$$n^{-1/4} h^{\beta-1/2} \int |\mathbb{K}(u)| du < C_1 n^{-c_1}, \tag{3.5}$$

for all n large enough. Also, letting $\bar{\mathbb{K}}(u) = \frac{1}{\pi}\int_0^1 \cos(\omega u) \frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)} \mathbb{I}\{|\omega| \geq h\omega_0\} d\omega$, it holds that

$$h^{\beta-1/2} \int |\mathbb{K}(u) - \bar{\mathbb{K}}(u)| du = O(h^s), \tag{3.6}$$

for some $s > 0$. (iii) As $n \to \infty$, it holds $h \to 0$, $\sqrt{n} h^{\beta-1/2} \to \infty$, $n^\nu h \to 0$ for some $\nu \in (0, 1/2)$, and $n^{1+2\xi} h^{2(\beta+\gamma)-1} \to 0$ for some $\xi > 0$.

Assumption OS (i) is a standard condition to characterize ordinary smooth densities.

[1] The independence assumption between X^* and ϵ is standard but, if necessary, can be relaxed to the sub-independence assumption, see Schennach (2013).

Note that we focus on the case of $\beta > 1/2$, where the cdf estimator $\hat{F}_{X^*}$ converges at a nonparametric rate (Dattner, Goldenshluger and Juditsky, 2011). For the case of $\beta < 1/2$, the estimator $\hat{F}_{X^*}$ typically converges at the $\sqrt{n}$-rate and a Donsker-type theorem applies (Söhl and Trabs, 2012). Assumption OS (ii) contains conditions for the kernel function. The first condition specifies a particular form for K that is commonly used in the literature (e.g., Delaigle and Hall, 2006). The second condition ensures that the deconvolution kernel $\mathbb{K}$ is L_1-integrable. The term $n^{-1/4}$ in (3.5) is required to ensure that the bootstrap counterpart $T_n^{\#}$ convergences to a Gaussian process at a polynomial rate in n (see, Lemma 2). If f_ϵ^{ft} is twice differentiable, applying the integration by parts formula twice gives

$$\mathbb{K}(u) = \frac{1}{u^2} \int_0^1 \cos(\omega u) \left\{ \frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)} \right\}'' d\omega,$$

and a sufficient condition for (3.5) is

$$n^{-1/4} h^{\beta - 1/2} \sup_{|\omega| \leq 1} \left| \left\{ \frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)} \right\}'' \right| = O(n^{-c_1}),$$

for some $c_1 > 0$. The third condition assures that $\mathbb{K}$ is well approximated by its trimmed version $\bar{\mathbb{K}}$. Since

$$\int |\mathbb{K}(u) - \bar{\mathbb{K}}(u)| du = \frac{1}{\pi} \int \left| \int_0^{h\omega_0} \cos(\omega u) \frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)} d\omega \right| du,$$

applying the integration by parts formula twice again implies that a sufficient condition for (3.6) is given by

$$h^{\beta+1/2} \sup_{|\omega| \leq h\omega_0} \max \left\{ \left| \left(\frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)} \right)' \right|, \left| \left(\frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)} \right)'' \right| \right\} = O(h^s),$$

for some $s > 0$. Based on the above sufficient conditions, it is possible to show that Assumption OS (ii) is satisfied by a large class of ordinary smooth error distributions including Laplace and its convolutions. Intuitively these conditions mean that f_ϵ^{ft} should not oscillate too wildly around its trend implied by the ordinary smooth density. Finally, Assumption OS (iii) contains conditions for the bandwidth h.

The second class of measurement error densities, called the super smooth densities, is presented as follows.

Assumption SS. (i) There exist $\mu, c, C, \omega_0, \lambda > 0$ and $\lambda_0 \in \mathbb{R}$ such that

$$c|\omega|^{\lambda_0} \exp(-|\omega|^{\lambda}/\mu) \leq |f_{\epsilon}^{\mathrm{ft}}(\omega)| \leq C|\omega|^{\lambda_0} \exp(-|\omega|^{\lambda}/\mu),$$

for all $|\omega| \geq \omega_0$. (ii) K is an even function with $K^{\mathrm{ft}}(\omega) = (1-\omega^q)^r \mathbb{I}\{|\omega| \leq 1\}$ for some $q, r \geq 2$. There exist $\mu_1 > 2\mu$ and $c_1, C_1 < \infty$ such that

$$\frac{1}{\varsigma(h)} \int |\mathbb{K}(u)| du < C_1 h^{-c_1} \exp\left(\frac{1}{\mu_1 h^{\lambda}}\right), \tag{3.7}$$

for all n large enough, where

$$\varsigma(h) = h^{\vartheta} \exp\left(\frac{1}{\mu h^{\lambda}}\right) \tag{3.8}$$

with $\vartheta = \lambda(r+1/2) + \lambda_0 + 1/2$. Also, letting $\bar{\mathbb{K}}(u) = \frac{1}{\pi} \int_0^1 \cos(\omega u) \frac{K^{\mathrm{ft}}(\omega)}{f_{\epsilon}^{\mathrm{ft}}(\omega/h)} \mathbb{I}\{|\omega| \geq h\omega_0\} d\omega$, it holds that

$$\frac{1}{\varsigma(h)} \int |\mathbb{K}(u) - \bar{\mathbb{K}}(u)| du = O(n^{-s}), \tag{3.9}$$

for some $s > 0$. (iii) $h = \left(\frac{\mu}{2} \log n + \mu\vartheta_1 \log\log n\right)^{-1/\lambda}$ for some $\vartheta_1 \in ((\vartheta - \gamma)/\lambda + 1, \vartheta/\lambda)$.

Assumption SS (i) a standard condition to characterize super smooth densities. Assumption SS (ii) contains conditions for the kernel function, and similar comments apply as the ordinary smooth case. The condition $\mu_1 > 2\mu$ is required to guarantee that the bootstrap counterpart $T_n^{\#}$ convergences to a Gaussian process at a polynomial rate in n (see, Lemma 18 in Appendix C). If $f_{\epsilon}^{\mathrm{ft}}$ is twice differentiable, a sufficient condition for (3.7) is

$$\frac{1}{\varsigma(h)} \sup_{|\omega| \leq 1} \left| \left(\frac{K^{\mathrm{ft}}(\omega)}{f_{\epsilon}^{\mathrm{ft}}(\omega/h)} \right)'' \right| = O\left(h^{-a} \exp\left(\frac{1}{\mu_1 h^{\lambda}}\right) \right),$$

for some $a > 0$. Also, a sufficient condition for (3.9) is

$$\exp\left(-\frac{1}{\mu h^{\lambda}}\right) \sup_{|\omega| \leq h\omega_0} \max\left\{ \left| \left(\frac{K^{\mathrm{ft}}(\omega)}{f_{\epsilon}^{\mathrm{ft}}(\omega/h)} \right)' \right|, \left| \left(\frac{K^{\mathrm{ft}}(\omega)}{f_{\epsilon}^{\mathrm{ft}}(\omega/h)} \right)'' \right| \right\} = O(n^{-a_1}),$$

for some $a_1 > 0$. For instance, these conditions are satisfied if

$$\sup_{|\omega| \leq 1} \max\{|A'(\omega/h)|, |A''(\omega/h)|\} = O\left(h^{-a_1} \exp\left(\frac{1}{\mu_1 h^{\lambda}}\right)\right), \tag{3.10}$$

for some $a_1 > 0$, where $A(\omega) = \frac{\exp(-|\omega|^{\lambda}/\mu)}{f_{\epsilon}^{\text{ft}}(\omega)}$. Based on (3.10), we can see that Assumption SS (ii) is satisfied by a large class of super smooth error distributions including Gaussian and its convolutions. Since the function $A(.)$ inherits the differentiability properties of f_{ϵ}^{ft}, the condition (3.10) intuitively means that f_{ϵ}^{ft} should not oscillate too wildly around its trend implied by the super smooth density. Assumption SS (iii) is on the bandwidth h. Note that this condition implicitly requires $\gamma > \lambda$.

Let $\hat{c}_{\alpha}$ denote the $(1-\alpha)$-th quantile of the bootstrap statistic $T_n^{\#}$. Under these assumptions, validity of the bootstrap approximation is established as follows.[2]

Theorem 6. *Suppose that Assumption C holds true. Then*

$$P\{T_n \leq \hat{c}_{\alpha}\} \geq 1 - \alpha - \delta_n, \tag{3.11}$$

for some positive sequence $\delta_n = O(n^{-c})$ (under Assumption OS) or $\delta_n = O((\log n)^{-c})$ (under Assumption SS) with $c > 0$.

Remark 5. Based on this theorem, we can construct an asymptotic confidence band for F_{X^*} over $\mathcal{T}$ with level α as $\mathcal{C}_n(t) = [\hat{F}_{X^*}(t) \pm \hat{c}_{\alpha}]$ for $t \in \mathcal{T}$ in the sense that

$$P\{F_{X^*}(t) \in \mathcal{C}_n(t) \text{ for all } t \in \mathcal{T}\} \geq 1 - \alpha - \delta_n,$$

for $\delta_n = O(n^{-c})$ (under Assumption OS) or $\delta_n = O((\log n)^{-c})$ (under Assumption SS) with some $c > 0$. Note that the approximation error δ_n is of polynomial order under Assumption OS (the ordinary smooth case) and of logarithmic order under Assumption SS (the super smooth case). We also note from the proof of the theorem that the slower approximation rate for the super-smooth case is solely due to the bias; if bias correction were possible, the

[2]Here, we present bootstrap approximation results for the statistic T_n which decays to zero. Alternatively, we could have normalized T_n without affecting any of the conclusions. This is analogous to whether we present the bootstrap approximation for the non-normalized object $\hat{\theta} - \theta$ or the normalized one $\sqrt{n}(\hat{\theta} - \theta)$, where θ is some parameter and $\hat{\theta}$ its estimator.

bootstrap approximation error would be of polynomial order in both cases.

Remark 6. To implement the bootstrap approximation in Theorem 6, we need to choose the bandwidth h. For estimation of the cdf $F_{X^*}(t_0)$ at a given t_0, Hall and Lahiri (2008) suggested choosing h that minimizes the approximate integrated MSE based on the normal reference distribution on X^*. For estimation of the quantile function of X^*, Dattner, Reiß and Trabs (2016) developed an adaptive method to choose h based on Lepski (1990). In Section 3.5, for the simulations, we suggest a bandwidth selection rule based on Bissantz, Dümbgen, Holzmann and Munk (2007). The basic idea is to estimate the ideal bandwidth that minimizes the maximal deviation between $\hat{F}_{X^*}$ and F_{X^*} under the sup-norm by utilizing a series of estimates $\hat{F}_{X^*}$ based on different values of h.

3.2.3 Asymptotic Gumbel approximation for ordinary smooth case

For the ordinary smooth case, it is also possible to characterize the asymptotic distribution of the standardized object

$$t_n = \sup_{t \in \mathcal{T}} |f_X(t)^{-1/2} \{\hat{F}_{X^*}(t) - F_{X^*}(t)\}|, \tag{3.12}$$

using the Gumbel distribution. Under additional assumptions, listed in Assumption G in Appendix C.3, we can follow similar steps in Bickel and Rosenblatt (1973) and Bissantz, Dümbgen, Holzmann and Munk (2007) to show the following result.

Theorem 7. *Suppose that Assumptions C, OS, and G hold, and* $(nh)^{-1}(\log n)^3 \to 0$ *as* $n \to \infty$. *Then*

$$P\left\{(-2\log h)^{1/2}(B^{-1/2}t_n - b_n) \le c\right\} \to \exp(-2\exp(-c)), \tag{3.13}$$

for all $c \in \mathbb{R}$, *where the constant* B *and sequence* b_n *are defined in Appendix C.3 (eq. (C.25)).*

See Appendix C.3 for a detailed statement and discussion of Assumption G, and for the proof of this theorem.

Remark 7. As shown in (3.13), the limiting behavior of t_n is characterized by the Gumbel distribution. Based on (3.13) and the conventional kernel density estimator $\hat{f}_X$ for f_X, we can also obtain an asymptotically valid critical value to conduct inference. For example, the asymptotic confidence band at level α for F_{X^*} is given by

$$\mathcal{C}_n^G(t) = [\hat{F}_{X^*}(t) \pm B^{1/2}\hat{f}_X(t)^{1/2}\{c_\alpha^G(-2\log h)^{-1/2} + b_n\}],$$

for $t \in \mathcal{T}$, where c_α^G solves $\exp(-2\exp(-c_\alpha^G)) = \alpha$. However, as discussed in the next remark, the asymptotic Gumbel approximation requires additional assumptions and tends to be less accurate than the bootstrap approximation.

Remark 8. Compared to the bootstrap approximation, the asymptotic Gumbel approximation has two drawbacks. First, the Gumbel approximation requires additional assumptions (Assumption G). Second, as indicated by Bissantz, Dümbgen, Holzmann and Munk (2007), the approximation error (i.e., δ_n in (3.11) for the bootstrap approximation) by (3.13) is typically a logarithmic rate even under Assumption OS, and therefore tends to be less accurate than the bootstrap approximation in (3.11). This contrast between the asymptotic Gumbel and bootstrap approximations was first clarified by Chernozhukov, Chetverikov and Kato (2014) for construction of confidence bands on the density with no measurement error. Kato and Sasaki (2016) extended their results for confidence bands on the pdf f_{X^*} with unknown f_ϵ. We obtain analogous results for confidence bands on the cdf F_{X^*}. We also note that in contrast to Chernozhukov, Chetverikov and Kato (2014) and Kato and Sasaki (2016) who employed Gaussian multiplier bootstrap methods, Theorem 6 shows validity of the conventional nonparametric bootstrap. Accordingly the techniques used in the proof of Theorem 6 are quite different: in particular, we employ Komlós, Major and Tusnády's (1975) coupling along with anti-concentration inequalities for Gaussian processes (Chernozhukov, Chetverikov and Kato, 2015) while the latter employ the Slepian-Stein type coupling for suprema of empirical processes constructed in Chernozhukov, Chetverikov and Kato (2014). Finally, we also obtain deterministic bounds on the approximation error of the bootstrap; to the best of our knowledge this is new in the literature on deconvolution.

Remark 9. We note that the asymptotic Gumbel approximation in (3.13) is available only

for the ordinary smooth case. It remains an open question whether we can establish such an asymptotic approximation for the super smooth case. As discussed in Bissantz, Dümbgen, Holzmann and Munk (2007, p. 486) for the density deconvolution, the main difficulty is that the limiting form of the deconvolution kernel (eq. (C.24) in Appendix C.3) is not available for the super smooth case. On the other hand, as shown in Theorem 6, we emphasize that the bootstrap approximation is valid even for the super smooth case.

3.3 Case of unknown measurement error distribution

The assumption of known measurement error density f_ϵ is unrealistic in many applications. In this section, we consider the situation where f_ϵ is unknown and needs to be estimated. In general, f_ϵ cannot be identified by a single measurement. Identification of f_ϵ can be restored however if we have two or more independent noisy measurements of the variable X^*. More specifically, suppose that we observe

$$X_{i,j} = X_i^* + \epsilon_{i,j} \quad \text{for } j = 1, \ldots, N_i \text{ and } i = 1, \ldots, n,$$

where X_i^* is the error-free variable and $\epsilon_{i,j}$'s are independently distributed measurement errors from the density f_ϵ. We thus have N_i repeated measurements of each variable X_i^*. We shall assume that the number of repeated observations is bounded above (i.e., $N_i \leq C < \infty$ for all i). This assumption is not critical for our theory but allows us to simplify the proofs considerably. Since in practice the number of repeated measurements is small anyway, we do not pursue the generalization to growing C. Under the assumption that f_ϵ is symmetric (Assumption C (iv)), its Fourier transform f_ϵ^{ft} can be estimated by (Delaigle, Hall and Meister, 2008)

$$\hat{f}_\epsilon^{\text{ft}}(\omega) = \left| \frac{1}{N} \sum_{i=1}^{n} \sum_{(j_1,j_2)\in\mathcal{J}_i}^{N_i} \cos\{\omega(X_{i,j_1} - X_{i,j_2})\} \right|^{1/2}, \tag{3.14}$$

where $N = \frac{1}{2}\sum_{i=1}^{n} N_i(N_i - 1)$, $\mathcal{J}_i$ is the set of $\frac{1}{2}N_i(N_i - 1)$ distinct pairs (j_1, j_2) with $1 \leq j_1 < j_2 \leq N$, and we ignore all the observations with $N_i = 1$. By plugging this

estimator into (3.3), we can estimate the cdf F_{X^*} by

$$\tilde{F}_{X^*}(t) = \frac{1}{2} + \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\tilde{\mathbb{L}}\left(\frac{t - X_{i,j}}{h}\right), \quad \text{where } \tilde{\mathbb{L}}(u) = \frac{1}{2\pi}\int_{-1}^{1}\frac{\sin(\omega u)}{\omega}\frac{K^{\text{ft}}(\omega)}{\hat{f}_{\epsilon}^{\text{ft}}(\omega/h)}d\omega. \quad (3.15)$$

In this section, we consider bootstrap approximation of the distribution of the maximal deviation $\tilde{T}_n = \sup_{t\in\mathcal{T}}|\tilde{F}_{X^*}(t) - F_{X^*}(t)|$. To construct the bootstrap counterpart of $\tilde{T}_n$, we suggest resampling from the set of observed variables $\{X_{i,j}\}$ while keeping the estimated measurement error density $\hat{f}_{\epsilon}^{\text{ft}}$ the same. More precisely, the bootstrap version of $\tilde{F}_{X^*}$ is given by

$$\tilde{F}^{\#}_{X^*}(t) = \frac{1}{2} + \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\tilde{\mathbb{L}}\left(\frac{t - X^{\#}_{i,j}}{h}\right),$$

where $X^{\#}_{i,j}$ is randomly drawn from the pooled observations $\{X_{i,j}\}$. The bootstrap counterpart of $\tilde{T}_n$ is obtained as $\tilde{T}^{\#}_n = \sup_{t\in\mathcal{T}}|\tilde{F}^{\#}_{X^*}(t) - \tilde{F}_{X^*}(t)|$.

To establish validity of the bootstrap approximation by $\tilde{T}^{\#}_n$, we first show that the cdf estimator $\tilde{F}_{X^*}$ under repeated measurements converges fast enough under the sup-norm to $\hat{F}_{X^*}$ so that the distributional properties of the latter would continue to hold. Previously, for the case of density deconvolution, Delaigle, Hall and Meister (2008) showed that under certain conditions, the deconvolution pdf estimator $\tilde{f}_{X^*}$ using $\hat{f}_{\epsilon}^{\text{ft}}$ enjoys the same first-order asymptotic properties as the estimator $\hat{f}_{X^*}$ in (3.2) for the case of known f_{ϵ}. Also, this result was obtained in terms of the uniform MSE metric, $\sup_t E|\tilde{f}_{X^*}(t) - \hat{f}_{X^*}(t)|^2$. Since validity of the confidence bands rests on controlling the sup-norm, we derive a corresponding result for the cdf estimators under the sup-norm. To this end, we add the following conditions.

Assumption B. (i) There exist $c \in (0,1)$ and $C > 0$ such that $P\{|\epsilon| \geq M\} \leq C(\log M)^{-1/c}$ for all $M > 0$. (ii) As $n \to \infty$, it holds $\log n/(nh^{4\beta}) \to 0$ and $nh^{4\beta+1} \to \infty$.

Based on these conditions, we are able to prove the following theorem.

Theorem 8. *Suppose that Assumptions C, OS, and B hold with $\gamma > \beta + 1$. Then for some $c > 0$,*

$$\sqrt{n}h^{\beta-1/2}\sup_{t\in\mathcal{T}}|\tilde{F}_{X^*}(t) - \hat{F}_{X^*}(t)| = o_p(n^{-c}).$$

Let $\tilde{c}_{\alpha}$ be the $(1-\alpha)$-th quantile of the bootstrap statistic $\tilde{T}^{\#}_n$. Based on the above

theorem, validity of the bootstrap approximation is established as follows.

Theorem 9. *Suppose that Assumptions C, OS, and B hold with $\gamma > \beta + 1$. Then*

$$P\{\tilde{T}_n \leq \tilde{c}_\alpha\} \geq 1 - \alpha - o_p(1). \tag{3.16}$$

Remark 10. Based on this theorem, we can construct an asymptotic confidence band for F_{X^*} over $\mathcal{T}$ with level α as $[\tilde{F}_{X^*}(t) \pm \tilde{c}_\alpha]$ for $t \in \mathcal{T}$. The key additional requirement $\gamma > \beta + 1$ says that f_{X^*} is smoother than f_ϵ by up to a derivative. As shown in Theorem 8, this ensures that the error from estimating f_ϵ^{ft} is asymptotically negligible. Also, we note that the conditions $nh^{4\beta+1} \to \infty$ in Assumption B (ii) and $n^{1+2\xi}h^{2(\beta+\gamma)-1} \to 0$ for some $\xi > 0$ in Assumption OS (iii) hold simultaneously only if $\gamma > \beta + 1$.

Remark 11. Note that the above theorems are presented only for the ordinary smooth case. A similar result can be derived for the super smooth case under the assumption that f_{X^*} is smoother than f_ϵ, i.e. the former is also super smooth. Sufficient conditions for the super smooth case are: in addition to Assumptions C, SS, and B (i), that $\vartheta_1 < 2\lambda_0/\lambda$ in the expression for h in Assumption SS (iii), and

$$\int \left| \frac{\omega^a f_{X^*}^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega)} \right|^2 d\omega < \infty,$$

for some $a > \vartheta - \lambda_0 + 1$.

Thus far we have focused on the case where repeated measurements on X^* are available and f_ϵ^{ft} can be estimated by (3.14) under the symmetry assumption on f_ϵ. If f_ϵ is not necessarily symmetric, but repeated measurements are available, then we can employ the estimator by Li and Vuong (1998) or Comte and Kappus (2015) based on Kotlarski's identity.

In some applications, a separate independent experiment may give us observations from f_ϵ (see, e.g., Efromovich, 1997, and Neumann, 1997). We refer to Meister (2009, Section 2.6) for an overview on estimation of f_ϵ in such cases. Suppose that we have m independent observations $(\epsilon_1, \ldots, \epsilon_m)$ of the error terms. Using these, the Fourier transform f_ϵ^{ft} can be estimated as

$$\tilde{f}_\epsilon^{\mathrm{ft}}(t) = \frac{1}{m} \sum_{i=1}^{m} \exp(\mathrm{i}t\epsilon_i).$$

We may use the above to obtain an estimator of the CDF by plugging-in the value of $\tilde{f}_\epsilon^{\text{ft}}(t)$ in (3.3). A bootstrap confidence band can thus be constructed by simply replacing $\hat{f}_\epsilon^{\text{ft}}$ with $\tilde{f}_\epsilon^{\text{ft}}$, and following the same procedure laid out earlier in this section. The asymptotic validity of the procedure can be demonstrated by replacing Assumption B (ii) with the following:

Assumption B1. As $n \to \infty$, it holds $m \to \infty$, $\sqrt{(n/m)}h^{\beta-1/2}\log(1/h) \to 0$ and $\log n/(mh^{4\beta}) \to 0$.

The formal proof that the resulting bootstrap is consistent follows by analogous arguments as that used to prove Theorems 8 and 9, and is therefore omitted.

3.4 Applications

3.4.1 Confidence band for quantile function

In addition to the confidence band for F_{X^*}, the results in the previous sections can be utilized to obtain the confidence band for the quantile function of X^*. Hall and Lahiri (2008) proposed estimating the u-th quantile $Q(u) = F_{X^*}^{-1}(u)$ by

$$\hat{Q}(u) = \sup\{t : \hat{F}_{X^*}^{\text{m}}(t) \leq u\},$$

where $\hat{F}_{X^*}^{\text{m}}(t) = \sup_{y \leq t} \hat{F}_{X^*}(y)$ is a monotone version of $\hat{F}_{X^*}(t)$. To obtain the confidence band for the quantile function $Q(u)$ over some interval $[u_1, u_2]$, we impose the following assumptions.

Assumption Q. (i) $F_{X^*}^{-1}(u)$ exists and is unique for all $u \in [u_1, u_2]$ such that $0 < u_1 < u_2 < 1$. There exists an interval $\mathcal{H}$ satisfying $F_{X^*}^{-1}[u_1 - \varepsilon, u_2 + \varepsilon] \subset \mathcal{H}$ for some $\varepsilon > 0$, $\inf_{x \in \mathcal{H}} f_X(x) > 0$, and $0 < \inf_{x \in \mathcal{H}} f_{X^*}(x) \leq \sup_{x \in \mathcal{H}} f_{X^*}(x) < \infty$. (ii) $\sup_{x \in \mathcal{H}} |f_{X^*}(x + \delta) - f_{X^*}(x)| \leq M|\delta|^a$ for all δ sufficiently small, with $a > 0$ (under Assumption OS) and $a = 1$ (under Assumption SS).

Based on these assumptions, we can obtain the asymptotic confidence bands for the quantile function as follows.

Theorem 10. *Suppose that Assumptions C, Q, and either OS or SS hold true. Then,*

$$P\left\{\hat{Q}(u) - \frac{\hat{c}_\alpha}{\hat{f}_{X^*}(\hat{Q}(u))} \leq Q(u) \leq \hat{Q}(u) + \frac{\hat{c}_\alpha}{\hat{f}_{X^*}(\hat{Q}(u))} \text{ for all } u \in [u_1, u_2]\right\} \geq 1 - \alpha - o(1).$$

Remark 12. Dattner, Reiß and Trabs (2016) have obtained the uniform convergence rate of their quantile estimator, say $\bar{Q}(u)$, based on the M-estimation method. In particular, Dattner, Reiß and Trabs (2016, Proposition 2.6) obtained that under an MSE optimal choice of the bandwidth,

$$\sup_{u\in[u_1,u_2]} |\bar{Q}(u) - Q(u)| = O_p\left(\left(\frac{\log n}{n}\right)^{\frac{\gamma}{2(\beta+\gamma)-1}}\right).$$

Thus, Theorem 10 is complementary in that it provides a confidence band for $Q(u)$ over $u \in [u_1, u_2]$. Note that as with the case of the cdf, we require under-smoothing to obtain the asymptotically valid confidence band, which excludes the MSE optimal bandwidth.

3.4.2 Goodness-of-fit testing

Another useful application of our results is goodness-of-fit testing on parametric models for F_{X^*}. Consider a parametric model $\{G_{X^*}(\cdot, \theta) : \theta \in \Theta\}$ for the distribution of the error-free variable X^* of interest. For simplicity, suppose the measurement error density f_ϵ is known as in Section 3.2. Our method can be adapted to the case of unknown f_ϵ. The goodness-of-fit testing problem of our interest is

$$H_0 : F_{X^*}(t) = G_{X^*}(t, \theta) \text{ over } t \in \mathcal{T} \text{ for some } \theta \in \Theta,$$

against negation of H_0. Let $\hat{\theta}$ be some $\sqrt{n}$-consistent estimator of the true parameter θ_0 under H_0. A typical example of $\hat{\theta}$ is the maximum likelihood estimator using the density function $\int g_{X^*}(t - a, \theta) f_\epsilon(a) da$ on the observable X, where g_{X^*} is the density of G_{X^*}.

To test H_0, we can employ the Kolmogorov-type statistic

$$K_n = \sup_{t\in\mathcal{T}} |\hat{F}_{X^*}(t) - G_{X^*}(t, \hat{\theta})|,$$

and its bootstrap counterpart is given by

$$K_n^{\#} = \sup_{t \in \mathcal{T}} |\hat{F}_{X^*}^{\#}(t) - G_{X^*}(t, \hat{\theta}^{\#})|,$$

where $\hat{F}_{X^*}^{\#}$ and $\hat{\theta}^{\#}$ are computed by the (parametric) bootstrap resample $\{X_i^{\#}\}_{i=1}^n$ from $X^{\#} = X_*^{\#} + \epsilon^{\#}$ with $X_*^{\#} \sim G_{X^*}(\cdot, \hat{\theta})$ and $\epsilon^{\#} \sim f_\epsilon$. In contrast to the no measurement error case, the cdf estimator $\hat{F}_{X^*}$ converges at a slower rate than $\sqrt{n}$. Therefore, if $\hat{\theta}$ is $\sqrt{n}$-consistent, then the estimation error of $\hat{\theta}$ is negligible under H_0, and the validity of the bootstrap critical value follows by a modification of the proof of Theorem 6. The result is summarized in the following corollary. Let $\hat{c}_\alpha^K$ be the $(1-\alpha)$-th quantile of $K_n^{\#}$.

Corollary 2. *Suppose that Assumption C holds true, the null H_0 is satisfied at θ_0, $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$, and the density of $G_{X^*}(\cdot, \theta)$ is bounded for all θ in a neighborhood of θ_0. Then*

$$P\{K_n > \hat{c}_\alpha^K\} \leq \alpha + \delta_n,$$

for some positive sequence $\delta_n = O(n^{-c})$ (under Assumption OS) or $\delta_n = O((\log n)^{-c})$ (under Assumption SS) with $c > 0$.

Consistency of the test can be shown analogously. If f_ϵ is unknown but repeated measurements on X^* are available, a similar result holds true by replacing $\hat{F}_{X^*}$ and $\hat{F}_{X^*}^{\#}$ with $\tilde{F}_{X^*}$ and $\tilde{F}_{X^*}^{\#}$, respectively.

3.4.3 Homogeneity test

Our bootstrap and asymptotic approximation results can be extended to two sample problems. Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$ be two independent samples of X and Y. X is generated as in (3.1). Also Y is generated as

$$Y = Y^* + \delta,$$

where Y^* is the unobservable error-free variable with the distribution function F_{Y^*} and δ is its measurement error. We assume δ is independent of Y^*. Suppose we wish to test the

homogeneity hypothesis

$$H_0 : F_{X^*}(t) = F_{Y^*}(t) \quad \text{for all } t \in \mathcal{T},$$

against the negation of H_0. The Kolmogorov-type statistic presented in the last subsection can be modified as follows

$$S_{n,m} = \sup_{t\in\mathcal{T}} |\hat{F}_{X^*}(t) - \hat{F}_{Y^*}(t)|,$$

where $\hat{F}_{Y^*}$ is the estimator for F_{Y^*}as in (3.3) using the sample $\{Y_i\}_{i=1}^m$. In this case, the bootstrap counterpart of $S_{n,m}$ is given by

$$S^{\#}_{n,m} = \sup_{t\in\mathcal{T}} \left| \hat{F}^{\#}_{X^*}(t) - \hat{F}^{\#}_{Y^*}(t) - \{\hat{F}_{X^*}(t) - \hat{F}_{Y^*}(t)\} \right|,$$

where $\hat{F}^{\#}_{Y^*}$ using the sample $\{Y_i\}_{i=1}^m$ is defined in the same manner as $\hat{F}^{\#}_{X^*}$. The $(1-\alpha)$-th quantile $\hat{c}^S_\alpha$ of $S^{\#}_{n,m}$ provides an asymptotically valid critical value as follows.

Corollary 3. *Suppose that Assumption C holds true for both $X = X^* + \epsilon$ and $Y = Y^* + \delta$, and that $n/(n+m) \to \tau \in (0,1)$ as $n, m \to \infty$. Then under H_0*

$$P\{S_{n,m} > \hat{c}^S_\alpha\} \le \alpha + \delta_{n,m},$$

for some positive sequence $\delta_{n,m} = O(n^{-c})$ (under Assumption OS for both ϵ and δ) or $\delta_{n,m} = O((\log n)^{-c})$ (under Assumption SS for both ϵ and δ) with $c > 0$.

An analogous result is available for the case of unknown f_ϵ by replacing $\hat{F}_{X^*}$ and $\hat{F}_{Y^*}$ with their repeated measurements versions. Also, if we wish to test the homogeneity hypothesis H_0 but Y has no measurement error (i.e., $Y = Y^*$), we can replace $\hat{F}_{Y^*}$ with the empirical distribution function of the sample $\{Y_i\}_{i=1}^m$.

3.4.4 Stochastic dominance test

Another intriguing application of our main results is testing the hypothesis of the (first-order) stochastic dominance

$$H_0 : F_{X^*}(t) \leq F_{Y^*}(t) \quad \text{for all } t \in \mathcal{T}, \tag{3.17}$$

against the negation of H_0. By modifying the Kolmogorov-type test in Section 3.4.3, the test statistic for (3.17) and its bootstrap counterpart are given by

$$\begin{aligned} D_{n,m} &= \sup_{t\in\mathcal{T}}\{\hat{F}_{X^*}(t) - \hat{F}_{Y^*}(t)\}, \\ D^{\#}_{n,m} &= \sup_{t\in\mathcal{T}}\left\{\hat{F}^{\#}_{X}(t) - \hat{F}^{\#}_{Y}(t) - \{\hat{F}_{X}(t) - \hat{F}_{Y}(t)\}\right\}, \end{aligned}$$

where $\hat{F}^{\#}_X$ and $\hat{F}^{\#}_Y$ are computed as in (3.3) using nonparametric bootstrap resamples $\{X^{\#}_i\}_{i=1}^n$ and $\{Y^{\#}_i\}_{i=1}^m$ from $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$, respectively.

Let $\hat{c}^D_\alpha$ denote the $(1-\alpha)$-th quantile of the bootstrap statistic $D^{\#}_{n,m}$. The bootstrap validity of our stochastic dominance test is established as follows.

Theorem 11. *Suppose that Assumption C holds true for both $X = X^* + \epsilon$ and $Y = Y^* + \delta$, and that $n/(n+m) \to \tau \in (0,1)$ as $n, m \to \infty$.*

(i) *Under H_0,*

$$P\{D_{n,m} > \hat{c}^D_\alpha\} \leq \alpha + \varrho_{n,m},$$

for some positive sequence $\varrho_{n,m} = O(n^{-c})$ (under Assumption OS for both ϵ and δ) or $\varrho_{n,m} = O((\log n)^{-c})$ (under Assumption SS for both ϵ and δ) with $c > 0$.

(ii) *Let $\mathcal{P}_0$ be the set of probability measures of (X, Y) satisfying H_0 (but f_δ and f_ϵ are*

fixed) and

$$
\begin{aligned}
&0 < c_X \le \inf_{t\in\mathcal{T}} f_X(t) \le \sup_{t\in\mathcal{T}} f_X(t) \le C_X < \infty,\\
&0 < c_Y \le \inf_{t\in\mathcal{T}} f_Y(t) \le \sup_{t\in\mathcal{T}} f_Y(t) \le C_Y < \infty,\\
&\sup_{\omega\in\mathbb{R}}\{(1+|\omega|)^{\gamma_X}|f^{\mathrm{ft}}_{X^*}(\omega)|\} \le M_X < \infty,\\
&\sup_{\omega\in\mathbb{R}}\{(1+|\omega|)^{\gamma_Y}|f^{\mathrm{ft}}_{Y^*}(\omega)|\} \le M_Y < \infty,
\end{aligned}
$$

for some $c_X, c_Y, \gamma_X, \gamma_Y, C_X, C_Y, M_X, M_Y > 0$ that are independent of (f_X, f_Y). Then

$$\sup_{P\in\mathcal{P}_0} P\{D_{n,m} > \hat{c}^D_\alpha\} \le \alpha + \varrho_{n,m},$$

for some positive sequence $\varrho_{n,m} = O(n^{-c})$ (under Assumption OS) or $\varrho_{n,m} = O((\log n)^{-c})$ (under Assumption SS) with $c > 0$.

(iii) *Under the alternative H_1 (i.e., H_0 is false) and Assumption OS or SS,*

$$P\{D_{n,m} > \hat{c}^D_\alpha\} \to 1.$$

Remark 13. Based on the proof of Theorem 11 (iii), we can characterize some local power properties. Suppose that both measurement errors are ordinary smooth. For any sequence $M_n \to \infty$, $D_{n,m}$ is consistent (i.e., $P\{D_{n,m} > \hat{c}^D_\alpha\} \to 1$) against local alternatives of the form

$$H_{1n} : F_{Y^*}(t) > F_{X^*}(t) + M_n\gamma_n \text{ for some } t \in \mathcal{T},$$

where

$$\gamma_n = n^{-1/2}\max\left\{h_X^{1/2-\beta_X}\sqrt{\log(1/h_X)},\ h_Y^{1/2-\beta_Y}\sqrt{\log(1/h_Y)}\right\},$$

and h_X and h_Y are (possibly different) bandwidths for the estimators $\hat{F}_{X^*}$ and $\hat{F}_{Y^*}$, respectively. A similar expression is available for γ_n in the super smooth case with $h_X^{\beta_X-1/2}$, $h_Y^{\beta_Y-1/2}$ replaced by $\varsigma_X^{-1}(h_X)$, $\varsigma_Y^{-1}(h_Y)$ respectively. Finally in the mixed error case, i.e when one of the errors is ordinary smooth while the other is super-smooth, the value of γ_n is determined by the super-smooth error (e.g $\gamma_n = n^{-1/2}\varsigma_X(h_X)\sqrt{\log(1/h_X)}$ if ϵ is super-smooth).

3.5 Simulation

In this section, we investigate the finite sample performance of the bootstrap uniform confidence band discussed in Theorem 6 using simulation experiments.

3.5.1 Simulation designs

We generate data from the model (3.1), where the unobserved variable of interest X^* is drawn from the normal distribution $N(0, \sigma^2_{X^*})$ and the measurement error ε is drawn from the Laplace distribution $L(0, \sigma^2_\varepsilon)$ or the normal distribution $N(0, \sigma^2_\varepsilon)$. We fix $\sigma_{X^*} = 1$ and choose σ_ε so that 'signal-to-noise ratio (SNR)' is given by $\sigma_{X^*}/\sigma_\varepsilon = 2, 3, 4$. We use the kernel function K defined by

$$K(\omega) = \frac{48\cos\omega}{\pi\omega^4}\left(1 - \frac{15}{\omega^2}\right) - \frac{144\sin\omega}{\pi\omega^5}\left(2 - \frac{5}{\omega^2}\right),$$

whose Fourier transformation is given by $K^{\mathrm{ft}}(\omega) = (1-\omega^2)^3 \cdot \mathbb{I}\{|\omega| \leq 1\}$. We consider four different sample sizes $n = 100, 250, 500, 1000$ and three different confidence levels $1 - \alpha = 0.80, 0.90, 0.95$. The number of simulation and bootstrap repetitions are 2000 and 1000, respectively. We compute the coverage probabilities of our confidence bands for F_{X^*} over the interval $[-2\sigma_{X^*},\ 2\sigma_{X^*}]$.

3.5.2 Bandwidth choice

We adapt the bandwidth selection method of Bissantz, Dümbgen, Holzmann and Munk (2007, Section 5.2) to the cdf estimation. First we consider J different bandwidths: $h_j = h_0 j/J$ for $j = 1, 2, \ldots, J$, where h_0 is a pilot bandwidth. A pilot bandwidth is an oversmoothing bandwidth obtained by multiplying $\gamma > 1$ to the normal reference rule of Hall and Lahiri (2008, Section 4.2). The normal reference rule was originally suggested by Delaigle and Gijbels (2004) to estimate density functions and was modified by Hall and Lahiri (2008) to the setting of estimating distribution functions. For $j = 2, \ldots, J$, define the distances

$$L_\infty(\hat{F}_{X^*},\ F_{X^*}) = ||\hat{F}_{X^*} - F_{X^*}||_\infty, \quad d^{(\infty)}_{j-1,j} = ||\hat{F}_{X^*,j-1} - \hat{F}_{X^*,j}||_\infty,$$

where $\hat{F}_{X^*,j}$ denotes the deconvolution estimator (3.3) with bandwidth $h = h_j$ and $||\cdot||_\infty$ denotes the supremum norm. For over-smoothing bandwidths, $L_\infty(\hat{F}_{X^*}, F_{X^*})$ changes only moderately with increasing bandwidth, while with undersmoothing bandwidth the distance suddenly increases with decreasing bandwidth. Based on this observation, Bissantz, Dümbgen, Holzmann and Munk (2007) suggest choosing the bandwidth to be the largest one at which $d^{(\infty)}_{j-1,j}$ is more than τ (for some $\tau > 1$) times greater than $d^{(\infty)}_{J-1,J}$. In our simulations, we choose $J = 20$ (number of possible bandwidths), $\tau = 3$ and $\gamma = 1.5$. We find that the simulation results are insensitive to the precise choice of the parameters.

Figures 3.1 and 3.2 illustrate the distances over different bandwidths for three different random samples with the measurement error drawn from the Laplace and normal distributions respectively. A comparison of two plots in the figures indicates that the bandwidth at which $d^{(\infty)}_{j-1,j}$ changes suddenly (marked by a circle, a square, or star) is a good indicator of the bandwidth at which the true distance L_∞ is about to stagnate.

3.5.3 Simulation results

Table 3.1 presents the empirical coverage probabilities of our bootstrap confidence bands. The simulated probabilities are generally close to the nominal confidence levels. As we expected, the coverage errors tend to be smaller when the sample size is larger or when the signal-to-ratio is larger.

Figures 3.3 and 3.4 depict some typical examples for the true cdf (CDF, F_{X^*}), deconvolution cdf estimate (ECDF, $\hat{F}_{X^*}$), and uniform confidence bands (CB), when the latent true distribution is standard normal and the measurement errors are drawn from Laplace and normal distributions. The figures demonstrate that the uniform confidence bands perform reasonably well even for small sample size $n = 100$ and the widths of the bands shrink substantially as the sample size increases from $n = 100$ to $n = 500$.

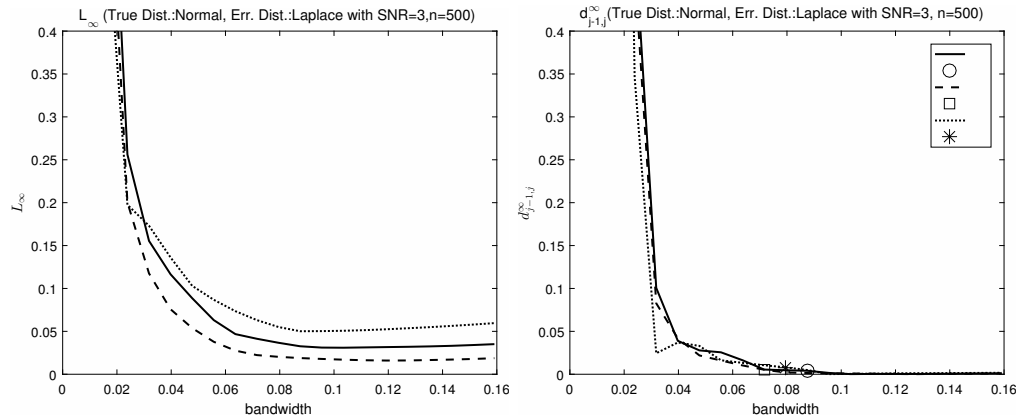

Figure 3.1: L_∞ and $d^\infty_{j-1,j}$ distances under Laplace error

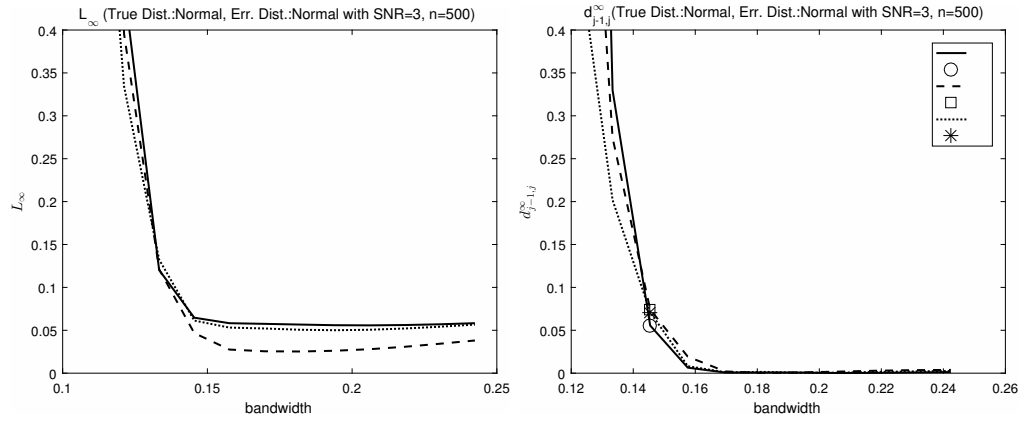

Figure 3.2: L_∞ and $d^\infty_{j-1,j}$ distances under Normal error

3.6 Real data example

3.6.1 Data description

In this section, we apply the stochastic dominance test to the Korea Household Income and Expenditure Survey data to investigate welfare changes of different population subgroups between 2006 and 2012. We use the data because the OECD report (2008) shows that, among OECD countries, Korea has the most significant variations in within-age group inequality and, compared to the inequality within the working age group, the relative inequality within the retirement age group is the worst. The data fit into our framework because it is well known that survey data are inherently affected by various sources of measurement errors, see Deaton (1997) and Bound, Brown and Mathiowetz (2000) for potential

Level	n	Laplace Error			Normal Error		
		SNR=2	SNR=3	SNR=4	SNR=2	SNR=3	SNR=4
0.80	100	0.818	0.828	0.828	0.780	0.833	0.826
	250	0.811	0.818	0.823	0.790	0.803	0.810
	500	0.807	0.812	0.830	0.793	0.805	0.817
	1000	0.811	0.824	0.836	0.763	0.789	0.812
0.90	100	0.911	0.919	0.924	0.882	0.920	0.924
	250	0.897	0.913	0.916	0.888	0.899	0.903
	500	0.902	0.915	0.921	0.880	0.892	0.911
	1000	0.898	0.907	0.919	0.883	0.886	0.903
0.95	100	0.963	0.961	0.961	0.943	0.956	0.967
	250	0.957	0.958	0.963	0.938	0.947	0.956
	500	0.953	0.959	0.962	0.936	0.949	0.959
	1000	0.951	0.955	0.958	0.932	0.945	0.955

Table 3.1: Simulated uniform coverage probabilities for F_{X^*} under Laplace and Normal errors.

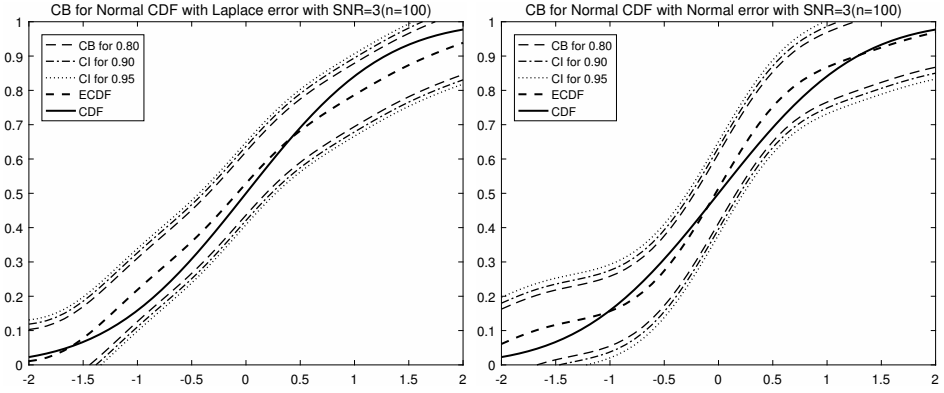

Figure 3.3: Uniform confidence bands under Laplace (left) and Normal (right) errors with $n = 100$

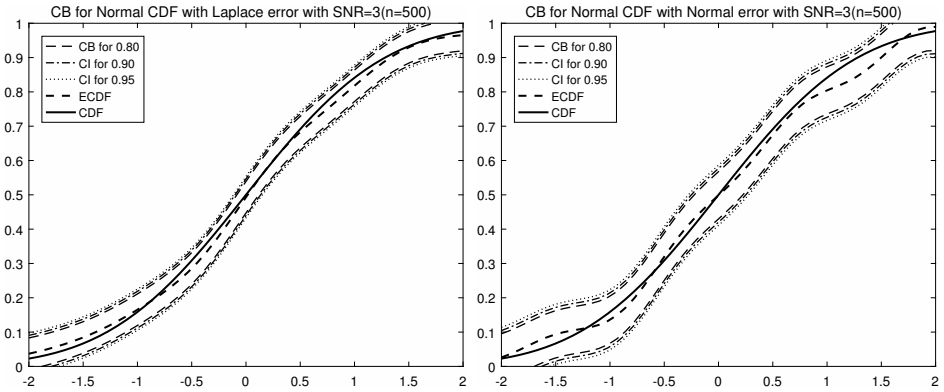

Figure 3.4: Uniform confidence bands under Laplace (left) and Normal (right) errors with $n = 500$

Year	Age Group	Sample Size	Mean	S.D.
2006	25-45	12045	1,650	910
	45-65	8512	1,575	1,034
	60+	4605	1,047	862
	65+	3250	968	823
	70+	2050	944	823
2012	25-45	8722	1,800	910
	45-65	7653	1,814	1,106
	60+	5166	1,105	934
	65+	3700	974	879
	70+	2439	891	857

Table 3.2: Descriptive Statistics (Income unit: 1,000 won)

sources of measurement errors in household-based survey data. The survey reports incomes from various sources and consumption of goods and services for each household. We first compute the real household disposable income by adding all incomes, public pension, social benefits and transfers, minus tax, public pension premium and social security fees, after adjusting for inflation using the 2010 consumer price index. We then obtain the individualized data by adjusting the total household disposable income using the square-root equivalization scale, which is a common practice to approximate individual welfare.

Table 3.2 shows the descriptive statistics for the data. It shows that average real incomes of individuals in all age groups except those over 70 have increased from 2006 to 2012. Standard deviations of all incomes have also increased slightly over the same period. The results are consistent with the finding of OECD (2008). However, unless the income distributions are normal, comparison of only the first two moments is not sufficient to draw a conclusion on the uniform ordering of nonparametric income distributions that does not depend on a specific social welfare function. This motivates us to consider a stochastic dominance criterion (see, e.g., Levy (2016)).

3.6.2 Results

We consider two different null hypotheses for each age group: (i) The 2006 income distribution first-order stochastically dominates that the 2012 income distribution (abbreviated to 06 FSD 12) (ii) The 2012 income distribution first-order stochastically dominates the 2006 income distribution (abbreviated to 12 FSD 06). As a benchmark test, we consider

Age Group	Null Hypothesis	BD	Laplace Error			Normal Error		
			SNR=2	SNR=3	SNR=4	SNR=2	SNR=3	SNR=4
25-45	06 FSD 12	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	12 FSD 06	1.000	0.998	1.000	1.000	1.000	1.000	1.000
45-65	06 FSD 12	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	12 FSD 06	1.000	1.000	1.000	1.000	1.000	1.000	1.000
60+	06 FSD 12	0.000	0.037	0.023	0.013	0.000	0.000	0.000
	12 FSD 06	0.039	0.000	0.000	0.000	0.305	0.234	0.054
65+	06 FSD 12	0.353	0.400	0.652	0.704	0.143	0.240	0.189
	12 FSD 06	0.000	0.001	0.000	0.000	0.027	0.013	0.003
70+	06 FSD 12	0.928	0.501	0.934	0.988	0.664	0.698	0.715
	12 FSD 06	0.000	0.000	0.000	0.000	0.000	0.000	0.001

Table 3.3: Bootstrap P-values from BD and our tests

Barrett and Donald (2003, BD)'s test based on the observed incomes, neglecting the presence of measurement errors. We choose the bandwidth as in our simulation experiments and assume Laplace and normal measurement errors. The variance of measurement errors is determined so that the signal-to-noise ratio (SNR) is 2,3, or 4.[3]

Table 3.3 reports the bootstrap p-values of the tests. The BD test implies that, for age groups 25-45 and 45-65, the 2012 income significantly dominates the 2006 income and, for age group 60+, there appears to be no dominance relationship (i.e. the two distributions cross), while for age group 70+ the 2006 income dominates the 2012 income. Similar results hold when we apply our test assuming Laplace measurement errors. However, when the measurement errors are normal, our test shows that, for age group 60+, there is a significant evidence that the 2012 income dominates the 2006 income. This implies that the ambiguous result (crossing of two distribution functions) for the age-group 60+ might be due to the presence of measurement errors in the observed data.

[3]In practice, as mentioned in Section 3.3, the error variance is generally not identified unless repeated measurements or extraneous information is available. However, in the case of the CPS income survey data, Bound and Krueger(1991) mentioned that "the error variance represents 27.6% of the total variance in CPS earnings for men and 8.9% for women." According to their remark, the signal-to-ratios are 1.9 for men and 3.35 for women, both of which lie in the range we considered.

Appendix A

Supplementary material and proofs for Chapter 1

A.1 Proofs of Main Results

Let $\mathbf{Z} = (\mathbf{Y}, \mathbf{W}, \mathbf{X})$ denote the observed data. Let $\tilde{P}_0$ denote the joint probability distribution of the observed data $\mathbf{Z} = (\mathbf{Y}, \mathbf{W}, \mathbf{X})$ together with $\mathbf{M}$; and $\tilde{E}_0[.]$, the corresponding expectation over $\tilde{P}_0$. I shall reserve the notation $\xrightarrow{p}$ for convergence in probability with respect to $\tilde{P}_0$. I shall also use the notation a.s-$\tilde{P}_0$ for 'almost surely under $\tilde{P}_0$'. As defined in the man text, let P^*_θ denote the joint distribution of $\mathbf{W}^*, \mathbf{S}^*$ conditional on $\mathbf{Z}, \mathbf{M}$, when $W^* \sim$ Bernoulli$(F(X_i^{*\prime}\theta))$. In other words, this is equivalent to the distribution of the bootstrap sequence of observations (conditional on the data and $\mathbf{M}$) when the treatments, $\mathbf{W}^*$, are constructed using θ instead of $\bar{\hat{\theta}}$. I shall use P^* as a shorthand for $P^*_{\bar{\hat{\theta}}}$.

In the proof I consider local sequences of the form $\theta_N = \bar{\hat{\theta}} + h/\sqrt{N}$ for some vector h. This in turn indexes a local sequence of bootstrap probability distributions $P^*_{\theta_N}$, or P^*_N for simplicity of notation. Let $\mathbf{Z}^*_N = (\mathbf{W}^*_N, \mathbf{X}^*) = f(\mathbf{W}^*_N, \mathbf{S}^*)$ denote the bootstrap observations obtained under P^*_N. I index the observations with N to reflect the fact that the distribution of $\mathbf{Z}^*_N$ as a function of the data depends on θ_N, which varies with N. I shall denote by $\mathcal{L}(.)$ the (unconditional) probability law of some random variable, and by $\mathcal{L}^*_N(.)$ the probability law of a random variable under the bootstrap distribution P^*_N conditional on the data and $\mathbf{M}$. Let $E^*_N[.]$ be the expectation of a random variable with respect to P^*_N.

Let $\Lambda_N^*(\theta|\theta') \equiv \log(dP_\theta^*/dP_{\theta'}^*)$ denote the difference in log-likelihood of the bootstrap probability distributions evaluated at θ and θ', i.e.

$$\Lambda_N^*(\theta|\theta') = L(\theta|\mathbf{Z}_N^*) - L(\theta'|\mathbf{Z}_N^*).$$

The bootstrap estimator of θ under P_N^* is represented by $\hat{\theta}_N^*$. Denote by $\psi_{N,i}^*(\theta_N)$, the influence function for $\hat{\theta}_N^*$ under P_N^*, i.e.

$$\psi_{N,i}^*(\theta_N) = X_i^* \frac{W_{N,i}^* - F(X_i^{*\prime}\theta_N)}{F(X_i^{*\prime}\theta_N)(1 - F(X_i^{*\prime}\theta_N))} f(X_i^{*\prime}\theta_N),$$

and let $S_N^*(\theta_N) = N^{-1/2} \sum_{i=1}^N \psi_{N,i}^*(\theta_N)$ denote the corresponding normalized score function.

Suppose that one had access to $e_{1i}(\theta), e_{2i}(W_i;\theta)$ instead of $\hat{e}_{1i}(\theta), \hat{e}_{2i}(W_i;\theta)$. Then denote

$$\tilde{\varepsilon}_i^*(\theta) = e_{1S_i^*}(\theta_N) + W_{N,i}^* \nu_{S_i^*}(1;\theta) - (1 - W_{N,i}^*)\nu_{S_i^*}(0;\theta),$$

where

$$\nu_i(w;\theta) = \left(1 + \frac{\tilde{K}(i;w,\theta)}{M}\right) e_{2\mathcal{J}_w(i)}(w;\theta).$$

Additionally set

$$\begin{aligned}\tilde{\Xi}^*(\theta) &\equiv E_\theta^*[\tilde{\varepsilon}_i^*(\theta)] \\ &= \frac{1}{N} \sum_{k=1}^N \left\{ e_{1k}(\theta) + F(X_k'\theta)\nu_i(1;\theta) - \left(1 - F(X_k'\theta)\right) \nu_i(1;\theta) \right\}.\end{aligned}$$

Finally define the bootstrap estimator with the 'true' error terms $e_{1i}(\theta), e_{2i}(W_i;\theta)$ as

$$\tilde{T}_N^*(\theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \left\{ \tilde{\varepsilon}_i^*\left(\hat{\theta}^*\right) - \tilde{\Xi}^*\left(\hat{\theta}^*\right) \right\}. \tag{A.1}$$

A.1.1 Proof of Theorem 1

My proof of the bootstrap consistency builds on the method of proof of Abadie and Imbens (2016, Theorem 1). I aim to show that

$$\mathcal{L}_N^* \begin{pmatrix} T_N^*(\theta_N) \\ \sqrt{N}(\hat{\theta}_N^* - \theta_N) \\ \Lambda_N^*\left(\bar{\hat{\theta}}|\theta_N\right) \end{pmatrix} \xrightarrow{p} \mathcal{L}(\mathbf{V}); \tag{A.2}$$

$$\mathbf{V} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ -h' I_{\theta_0} h/2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' I(\theta_0)^{-1} & -c'h \\ I(\theta_0)^{-1} c & I(\theta_0)^{-1} & -h \\ -h'c & -h & h' I(\theta_0) h \end{pmatrix}\right).$$

Given (A.2), the claim follows by similar arguments in Abadie and Imbens (2016) involving the use of Le Cam's third lemma together with a Le Cam skeleton or discretization argument as in Andreou and Werker (2011). Subsequently, I focus on proving (A.2).

To simplify notation I shall assume that $\bar{\hat{\theta}} \to \theta_0$ and $\theta_N \to \theta_0$ almost surely in $\tilde{P}_0$. This is without loss of generality as one can always convert convergence in probability (wrt $\tilde{P}_0$) to almost sure convergence (wrt $\tilde{P}_0$) using a subsequence argument.[1] Henceforth, in all of the proofs it is implicitly assumed that I am working within such a subsequence.

Lemma 1 in Appendix A.2 implies that with probability approaching one under $\tilde{P}_0$,

$$\begin{aligned} \Lambda_N^*\left(\bar{\hat{\theta}}|\theta_N\right) &= -h' S_N^*(\theta_N) - \frac{1}{2} h' I(\theta_0) h + o_{P_N^*}(1); \\ \sqrt{N}(\hat{\theta}_N^* - \theta_N) &= I(\theta_0)^{-1} S_N^*(\theta_N) + o_{P_N^*}(1). \end{aligned}$$

Consequently by the above it suffices for (A.2) to show

$$\mathcal{L}_N^* \begin{pmatrix} T_N^*(\theta_N) \\ S_N^*(\theta_N) \end{pmatrix} \xrightarrow{p} \mathcal{L}(\mathbf{V}_2); \qquad \mathbf{V}_2 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & c' \\ c & I(\theta_0) \end{pmatrix}\right). \tag{A.3}$$

[1] That $\bar{\hat{\theta}} \xrightarrow{p} \theta_0$ is simply a consequence of $\hat{\theta} \xrightarrow{p} \theta_0$, since the grid size $d/\sqrt{N}$ also goes to 0 as $N \to \infty$.

Now by Lemma 2 in Appendix A.2, it follows

$$\left\| \; T_N^*(\theta_N) - \tilde{T}_N^*(\theta_N) \; \right\| = o_{P_N^*}(1),$$

with probability approaching one under $\tilde{P}_0$. Hence to prove (A.3), it is enough to show

$$\mathcal{L}_N^* \begin{pmatrix} \tilde{T}_N^*(\theta_N) \\ S_N^*(\theta_N) \end{pmatrix} \xrightarrow{p} \mathcal{L}(\mathbf{V}_2). \tag{A.4}$$

Consider the linear combination $C_N = t_1 \tilde{T}_N^*(\theta_N) + t_2' S_N^*(\theta_N)$. For any value of θ let

$$h(x;\theta) \equiv E[X_i | F(X_i'\theta) = x].$$

I can then write $C_N = N^{-1/2} \sum_{i=1}^N \delta_{N,i}^*$, where $\delta_{N,i}^* = t_2' \alpha_{N,i}^* + t_2' \beta_{N,i}^* + t_1 \gamma_{N,i}^*$, with

$$\alpha_{N,i}^* = h(X_i^{*\prime}\theta_N;\theta_N) \frac{W_{N,i}^* - F\left(X_i^{*\prime}\theta_N\right)}{F\left(X_i^{*\prime}\theta_N\right)\left(1 - F\left(X_i^{*\prime}\theta_N\right)\right)} f\left(X_i^{*\prime}\theta_N\right),$$

$$\beta_{N.i}^* = \left\{X_i^* - h(X_i^{*\prime}\theta_N;\theta_N)\right\} \frac{W_{N,i}^* - F\left(X_i^{*\prime}\theta_N\right)}{F\left(X_i^{*\prime}\theta_N\right)\left(1 - F\left(X_i^{*\prime}\theta_N\right)\right)} f\left(X_i^{*\prime}\theta_N\right),$$

and

$$\gamma_{N,i}^* = \tilde{\varepsilon}_i^*\left(\theta_N\right) - \tilde{\Xi}^*\left(\theta_N\right).$$

Observe that under the bootstrap DGP, $E_N^*[\alpha_{N,i}^* | X_i^*] = 0$ and $E_N^*[\beta_{N,i}^* | X_i^*] = 0$ by the construction of $W_{N,i}^*$; and $E_N^*[\gamma_{N,i}^*] = 0$ by the definition of $\tilde{\Xi}^*\left(\theta_N\right)$. Hence $\{\delta_{N,i}^*, i = 1 \ldots N\}$ are iid zero mean random variables under P_N^*, and by the Lindberg-Feller central limit theorem for triangular arrays with iid sequences I obtain

$$\mathcal{L}_N^*(C_N) \xrightarrow{p} \mathcal{L}(\mathbf{v}); \quad \mathbf{v} \sim N(0, \sigma_B^2),$$

with

$$\sigma_B^2 = \text{plim } E_N^*[\delta_{N,i}^{*2}],$$

where the plim is taken over $\tilde{P}_0$.

I characterize the variance by expanding $\delta^{*2}_{N,i} = \left(t_2'\alpha^*_{N,i} + t_2'\beta^*_{N,i} + t_1\gamma^*_{N,i}\right)^2$ and considering the bootstrap expectation of each term in turn. Under the definition of $h(.)$, it follows by the usual algebra involving Assumptions 3 & 4 that

$$E^*_N\left[\left(t_2'\alpha^*_{N,i} + t_2'\beta^*_{N,i}\right)^2\right] \xrightarrow{p} E\left[\frac{f^2(X'\theta_0)}{F(X'\theta_0)(1-F(X'\theta_0))}\left(t_2'X\right)^2\right] = t_2'I(\theta_0)t_2.$$

Thus it only remains to obtain the probability limit under $\tilde{P}_0$ of

$$\begin{aligned} V_1(\theta_N) &\equiv E^*_N\left[\left(t_1\gamma^*_{N,i}\right)\left(t_2'\alpha^*_{N,i}\right)\right], \\ V_2(\theta_N) &\equiv E^*_N\left[(t_1\gamma^*_{N,i})^2\right], \qquad \text{and} \\ V_3(\theta_N) &\equiv E^*_N\left[(t_1\gamma^*_{N,i})\cdot(t_2'\beta^*_{N,i})\right]. \end{aligned}$$

A difficulty with proving the above is that within the matching function, $K_M(i;\theta_N)$, the treatments in the original sample are distributed as $W_i \sim \text{Bernoulli}(F(X_i'\theta_0))$, whereas the matches are evaluated in terms of the proximity with respect to $F(X_i'\theta_N)$. Thus, to obtain the probability limits, I make a second use of the skeleton argument of Le Cam. This exploits the discretization $\bar{\hat{\theta}}$ of $\hat{\theta}$ defined previously, and involves replacing $\bar{\hat{\theta}}$ with the local asymptotic sequence $\breve{\theta}_N = \theta_0 + \breve{h}/\sqrt{N}$, for some $\breve{h} \in \mathbb{R}$. To this end, I employ the following notation:

Parametrize the multinomial random variables $\mathbf{M}$ (Section 1.3.2) as $\mathbf{M}(\theta)$ for the case when the estimated propensity score is given by θ (rather than $\bar{\hat{\theta}}$). Denote by $\mathbf{U} = (U_1, \ldots, U_N)$ a vector of N independent uniform random variables corresponding to each observation, and drawn independently of $\mathbf{W}, \mathbf{X}, \mathbf{Y}$. Then it is possible to couple $\mathbf{M}(\theta) = \mathcal{H}(\mathbf{U}; F(\mathbf{X}'\theta))$, where $\mathcal{H}(.; F(\mathbf{X}'\theta))$ is some transformation indexed by the parameter θ.[2] I represent by $\bar{P}_\theta$ the probability law for $\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{U}$ with $\mathbf{W} \sim \text{Bernoulli}(F(\mathbf{X}'\theta))$, and let $\bar{E}_\theta[.]$ denote the corresponding expectation over $\bar{P}_\theta$. A convenient feature of $\bar{P}_\theta$ (as compared

[2] $\mathcal{H}(,; F(\mathbf{X}'\theta))$ can be interpreted as a function that transforms a uniformly distributed random variable into a single-draw multinomial random variable. Note that knowledge of $F(\mathbf{X}'\theta)$ uniquely pins down the quantiles $\left\{\pi_1(\theta), \ldots, \pi_{q_N-1}(\theta)\right\}$ and the number of treated and untreated populations denoted by $N_1(l;\theta)$, $N_0(l;\theta)$ in each partition. Thus the uniform random variable can be transformed into the multinomial random variable, $M(i;\theta)$, for each observation i, by partitioning the unit interval into $N_w(l;\theta)$ equi-spaced segments.

to $\tilde{P}_\theta$) is that it doesn't depend on the value of $\breve{\theta}_N$; indeed, this is the reason for employing the coupling. Given $\breve{\theta}_N$, I construct a local asymptotic sequence for the bootstrap indexed by $\bar{\theta}_N = \breve{\theta}_N + h/\sqrt{N}$. Let $\bar{P}_N^* \equiv P_{\bar{\theta}_N}^*$ denote the bootstrap probability indexed by $\bar{\theta}_N$, and $\bar{E}_N^*[.]$ the bootstrap expectation under $\bar{P}_N^*$. For convenience set $\bar{P}_N \equiv \bar{P}_{\bar{\theta}_N}$ and $\bar{P}_0 \equiv \bar{P}_{\theta_0}$, with the corresponding expectation operators $\bar{E}_N[.] \equiv \bar{E}_{\bar{\theta}_N}[.]$ and $\bar{E}_0[.] \equiv \bar{E}_{\theta_0}[.]$. Finally, I also introduce the quantities

$$
\begin{aligned}
V_1(h,\theta) &\equiv E^*_{\theta+h/\sqrt{N}}\left[\left(t_1\gamma_i^*\left(\theta+\frac{h}{\sqrt{N}};\theta\right)\right)\left(t_2'\alpha_i^*\left(\theta+\frac{h}{\sqrt{N}}\right)\right)\right];\\
V_2(h,\theta) &\equiv E^*_{\theta+h/\sqrt{N}}\left[\left(t_1\gamma_i^*\left(\theta+\frac{h}{\sqrt{N}};\theta\right)\right)^2\right]; \quad \text{and}\\
V_3(h,\theta) &\equiv E^*_{\theta+h/\sqrt{N}}\left[\left(t_1\gamma_i^*\left(\theta+\frac{h}{\sqrt{N}};\theta\right)\right)\left(t_2'\beta_i^*\left(\theta+\frac{h}{\sqrt{N}}\right)\right)\right],
\end{aligned}
$$

where, for any θ_1, θ_2,

$$
\alpha_i^*(\theta_1) = h^*(X_i^{*\prime}\theta_1;\theta_1)\frac{W_i^* - F\left(X_i^{*\prime}\theta_1\right)}{F\left(X_i^{*\prime}\theta_1\right)\left(1-F\left(X_i^{*\prime}\theta_1\right)\right)}f\left(X_i^{*\prime}\theta_1\right),
$$

$$
\beta_i^*(\theta_1) = \left\{X_i^* - h^*(X_i^{*\prime}\theta_1;\theta_1)\right\}\frac{W_i^* - F\left(X_i^{*\prime}\theta_1\right)}{F\left(X_i^{*\prime}\theta_1\right)\left(1-F\left(X_i^{*\prime}\theta_1\right)\right)}f\left(X_i^{*\prime}\theta_1\right); \text{ and}
$$

$$
\gamma_i^*(\theta_1;\theta_2) = \tilde{\varepsilon}_i^*\left(\theta_1;\theta_2\right) - E^*_{\theta_1}[\tilde{\varepsilon}_i^*\left(\theta_1;\theta_2\right)].
$$

Here $\tilde{\varepsilon}_i^*\left(\theta_1;\theta_2\right)$ is defined analogously to $\tilde{\varepsilon}_i^*\left(\theta_1\right)$ but with $\bar{\hat{\theta}}$ replaced by θ_2; in particular, this involves replacing $\mathbf{M}$ in the definition of $\tilde{\varepsilon}_i^*\left(\theta_1\right)$ with $\mathbf{M}(\theta_2) = \mathcal{H}(\mathbf{U}; F(\mathbf{X}'\theta_2))$. It is useful to observe that $V_k(\theta_N) = V_k\left(h;\bar{\hat{\theta}}\right)$ for $k = 1, 2, 3$.

In Lemmas 3 - 5 in Appendix A.2, I show that for any bounded $\breve{h}$ within the definition

of $\breve{\theta}_N$,[3]

$$\begin{aligned} V_1(h, \breve{\theta}_N) &= o_{\bar{P}_N}(1); \\ V_2(h, \breve{\theta}_N) &= t_1^2\sigma^2 + o_{\bar{P}_N}(1); \text{ and} \\ V_3(h, \breve{\theta}_N) &= 2t_1 c' t_2 + o_{\bar{P}_N}(1). \end{aligned} \tag{A.5}$$

Then, employing a version of Le Cam's skeleton argument, I show that

$$\begin{aligned} V_1\left(h, \bar{\hat{\theta}}\right) &= o_{\bar{P}_0}(1); \\ V_2\left(h, \bar{\hat{\theta}}\right) &= t_1^2\sigma^2 + o_{\bar{P}_0}(1); \text{ and} \\ V_3\left(h, \bar{\hat{\theta}}\right) &= 2t_1 c' t_2 + o_{\bar{P}_0}(1). \end{aligned}$$

I illustrate the reasoning for the case of $V_2\left(h, \bar{\hat{\theta}}\right)$; the others can be argued similarly. Note that $\bar{P}_N$ and $\bar{P}_0$ are mutually contiguous by the usual arguments involving Le Cam's first lemma. Thus by (A.5) and contiguity, I have $V_2(h, \breve{\theta}_N) = t_1^2\sigma^2 + o_{\bar{P}_0}(1)$. Let $\mathbf{v}$ denote the asymptotic normal limit of $\sqrt{N}(\hat{\theta} - \theta_0)$ under $\bar{P}_0$. Then for any $j \in \mathbb{Z}^d$,

$$\mathcal{L}_{\bar{P}_0}\begin{pmatrix} V_2\left(h, \theta_0 + dj/\sqrt{N}\right) - t_1^2\sigma^2 \\ \sqrt{N}(\hat{\theta} - \theta_0) - dj \end{pmatrix} \to \mathcal{L}\begin{pmatrix} 0 \\ \mathbf{v} - dj \end{pmatrix}.$$

Additionally, the following events are equivalent for each $j \in \mathbb{Z}^d$:

$$\left\{\sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right) = dj\right\} \equiv \left\{-\frac{d}{2}i < \sqrt{N}(\hat{\theta} - \theta_0) - dj \leq -\frac{d}{2}i\right\},$$

where i denotes a vector of ones of dimension d. Combining the above gives that for each $j \in \mathbb{Z}^d$, and any $\epsilon > 0$,

$$\bar{P}_0\left\{\left|V_2\left(h, \theta_0 + dj/\sqrt{N}\right) - t_1^2\sigma^2\right| > \epsilon \cap \sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right) = dj\right\} \to 0$$

[3]For equation (A.5), note that the matches are now evaluated in terms of proximity wrt $F(X_i'\bar{\theta}_N)$, which is also the propensity score characterizing the distribution of the treatments since $W_i \sim \text{Bernoulli}(F(X_i'\bar{\theta}_N))$ under $\bar{P}_N$.

as $N \to \infty$. Hence for each $C < \infty$,

$$\begin{aligned}
&\bar{P}_0 \left\{ \left| V_2\left(h, \bar{\hat{\theta}}\right) - t_1^2\sigma^2 \right| > \epsilon \cap \left| \sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right) \right| \leq C \right\} \\
&= \sum_{j \in \mathbb{Z}^d : d|j| \leq C} \bar{P}_0 \left\{ \left| V_2\left(h, \theta_0 + dj/\sqrt{N}\right) - t_1^2\sigma^2 \right| > \epsilon \cap \sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right) = dj \right\} \to 0.
\end{aligned}$$

Since $\sqrt{N}\left(\bar{\hat{\theta}} - \theta_0\right)$ is $O_{\bar{P}_0}(1)$, letting $C \to \infty$ above implies $V_2\left(h, \bar{\hat{\theta}}\right) = t_1^2\sigma^2 + o_{\bar{P}_0}(1)$, as claimed.

By definition, the probability distribution of $V_2(\theta_N)$ under $\tilde{P}_0$ is equivalent to that of $V_2\left(h, \bar{\hat{\theta}}\right)$ under $\bar{P}_0$; and similarly for the distribution of $V_1(\theta_N), V_3(\theta_N)$ under $\tilde{P}_0$. Combining the above results, I have thus shown

$$\sigma_B^2 = t_1^2\sigma^2 + 2t_1 c' t_2 + t_2' I(\theta_0) t_2.$$

This proves (A.3), which completes the proof of the theorem.

A.1.2 Proof of Corollary 1

Let $F(.)$ denote the cdf of $\mathbf{v} \sim N\left(0, \sigma^2 - c' I_{\theta_0}^{-1} c\right)$. By taking L (cf Step 7 in Section 1.3.3) sufficiently large, the claim follows if I show that

$$E_{\mathbf{M}}\left[F_n^*(t)|\mathbf{Z}\right] \xrightarrow{p} F(t) + O(d) \tag{A.6}$$

uniformly over $t \in \mathbb{R}$ under P_0 (here $E_{\mathbf{M}}[.|\mathbf{Z}]$ denotes the expectation over $\mathbf{M}$ conditional on the data). But by the Glivenko-Cantelli theorem, pointwise convergence implies uniform convergence, hence it suffices to show (A.6) holds for each $t \in \mathbb{R}$ under P_0. So I fix some arbitrary $t \in \mathbb{R}$.

Recall the definitions of $\bar{P}_0$ and $\mathbf{U}$ from the proof of Theorem 1. By Theorem 1, $F_n^*(t) \xrightarrow{p} F(t)+O(d)$ under $\bar{P}_0$. By employing a subsequence argument, the convergence in probability (wrt $\bar{P}_0$) can be converted to almost sure convergence (wrt $\bar{P}_0$). By this construction,

$$F_n^*(t) \to F(t) + O(d), \text{ a.s} - \bar{P}_0. \tag{A.7}$$

Note that by independence (of $\mathbf{Z}$,$\mathbf{U}$), $\bar{P}_0$ is equivalent to the product measure, $P_0 \times P_U$, of the respective marginal measures, P_0, P_U, of $\mathbf{Z}$ and $\mathbf{U}$. Denote by Ω the set of all realizations, z, of $\mathbf{Z}$ for which $F_n^*(t) \to F(t) + O(d)$, a.s $- P_U$. By the independence of $\mathbf{Z}$ and $\mathbf{U}$, it must be that $P_0(\Omega) = 1$ for (A.7) to hold. At the same time, the dominated convergence theorem gives

$$E_{\mathbf{U}}\left[F_n^*(t)\right] \to F(t) + O(d) \tag{A.8}$$

for each $z \in \Omega$; hence (A.8) holds almost surely over P_0. Since $E_{\mathbf{M}}\left[F_n^*(t)|\mathbf{Z}\right] \equiv E_{\mathbf{U}}\left[F_n^*(t)\right]$, this immediately proves (A.6).

A.2 Lemmas

Hereafter, I shall use the notation wpa1-$\tilde{P}_0$ as a shorthand for 'with probability approaching one under $\tilde{P}_0$'.

I also introduce the following notation: For $w = 0, 1$ let

$$e_{3i}(w;\theta) = \bar{\mu}(w; X_i) - \mu(w; F(X_i'\theta)).$$

Also let

$$e_{4i}(W_i;\theta) = Y_i - \bar{\mu}(W_i, X_i).$$

Note that it is possible to decompose $e_{2i}(W_i;\theta) = e_{3i}(W_i;\theta) + e_{4i}(W_i;\theta)$.

In Lemmas 3-5, I work with the local asymptotic sequence $\breve{\theta}_N = \theta_0 + \breve{h}/\sqrt{N}$ in place of $\bar{\hat{\theta}}$. To this end, I employ the notation introduced in Appendix A.1. Represent by $\left\{\pi_1(\breve{\theta}_N), \ldots, \pi_{q_N-1}(\breve{\theta}_N)\right\}$ the sample q_N-quantiles of $F(X'\breve{\theta}_N)$ with $\pi_0(\breve{\theta}_N) = 0$ and $\pi_{q_N}(\breve{\theta}_N) = 1$. I introduce $l(i)$ as the block index of observation i wrt $F(X_i'\breve{\theta}_N)$, i.e $l(i) = k$ if $\pi_{l-1}(\breve{\theta}_N) \leq F(X_i'\breve{\theta}_N) < \pi_l(\breve{\theta}_N)$. Also, denote by $S_w(l;\theta)$ the set of all observations with $W_i = w$ whose propensity scores evaluated at θ - i.e $F(X_i'\theta)$ - lie in the l-th block (even as the blocks themselves are obtained from quantiles of $F(X'\breve{\theta}_N)$):

$$S_w(l;\theta) \equiv \left\{i : \ \pi_{l-1}(\breve{\theta}_N) \leq F(X_i'\theta) < \pi_l(\breve{\theta}_N) \ \cap \ W_i = w\right\}.$$

Based on the above, I set $S(l;\theta) = S_1(l;\theta) \cup S_0(l;\theta)$. Furthermore, I also denote

$$N_0(l;\theta) = \#S_0(l;\theta); \quad N_1(l;\theta) = \#S_1(l;\theta); \quad N(l;\theta) = N_0(l;\theta) + N_1(l;\theta),$$

where $\#A$ denotes the cardinality of any set A.

For $w = 0, 1$, the average matching function, defined as the expectation of $\tilde{K}_M(i; w, \theta)$ over $\mathbf{U}$ given $(\mathbf{X}, \mathbf{W})$, is represented by

$$\bar{K}_M(i; w, \theta) = \begin{cases} K_M(i;\theta) & \text{if } w = W_i \\ \frac{1}{N_w\left(l(i);\breve{\theta}_N\right)} \sum_{j \in S_w\left(l(i);\breve{\theta}_N\right)} K_M(j;\theta) & \text{if } w \neq W_i \ . \end{cases}$$

Slightly abusing notation, I suppress indexing the quantities $\tilde{K}_M(.), \bar{K}_M(.), \nu(.), l(.)$ with the additional label $\breve{\theta}_N$. However it should be understood implicitly that these quantities are now constructed by replacing $\bar{\hat{\theta}}$ with $\breve{\theta}_N$. Finally, I also define (again suppressing the index with respect to $\breve{\theta}_N$),

$$\nu_{(3)i}(w;\theta) = \left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M}\right) e_{3\mathcal{J}_w(i)}(w;\theta);.$$
$$\nu_{(4)i}(w;\theta) = \left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M}\right) e_{4\mathcal{J}_w(i)}(w;\theta).$$

Lemma 1. *Suppose that $\bar{\hat{\theta}} \to \theta_0$ a.s-$\tilde{P}_0$. Then under Assumptions 1-5, wpa1-$\tilde{P}_0$,*

$$\Lambda_N^*\left(\bar{\hat{\theta}} | \theta_N\right) = -h' S_N^*(\theta_N) - \frac{1}{2} h' I(\theta_0) h + o_{P_N^*}(1), \tag{A.9}$$

and

$$\sqrt{N}(\hat{\theta}_N^* - \theta_N) = I(\theta_0)^{-1} S_N^*(\theta_N) + o_{P_N^*}(1). \tag{A.10}$$

Proof. Define

$$\hat{I}_N^*(\theta) = \frac{1}{N} \frac{d^2 L(\theta | \mathbf{Z}_N^*)}{d\theta d\theta'}; \quad \breve{I}_N^*(\theta) = \frac{1}{N} \sum_{i=1}^N \psi_{N,i}^*(\theta_N) \psi_{N,i}^{*\prime}(\theta_N).$$

Under Assumptions 3(i)-(ii), I can show that $\sup_{\theta \in \mathcal{N}} E_N^* \left\| \hat{I}_N^*(\theta) - \breve{I}_N^*(\theta) \right\|^2 \xrightarrow{p} 0$. The same

assumptions also suffice to show $\sup_{\theta\in\mathcal{N}} \left\|\hat{I}_N^*(\theta) - I^*(\theta_N)\right\| \xrightarrow{p} 0$, where

$$I^*(\theta_N) \equiv E_N^*\left[\psi_{N,i}^*(\theta_N)\psi_{N,i}^{*\prime}(\theta_N)\right] = \frac{1}{N}\sum_{i=1}^{N} X_i X_i' \frac{f^2(X_i'\theta_N)}{F(X_i'\theta_N)(1-F(X_i'\theta_N))}.$$

The term inside the summation is non-negative and uniformly bounded for all N sufficiently large (by Assumptions 3(i)-(ii)). Consequently, by Assumption 4 and standard arguments, $\sup_{\theta\in\mathcal{N}} \|I_N^*(\theta) - I(\theta)\| \xrightarrow{p} 0$. Combining the above proves that wpa1-$\tilde{P}_0$,

$$\sup_{\theta\in\mathcal{N}} \left\|\hat{I}_N^*(\theta) - I(\theta)\right\| = o_{P_N^*}(1). \tag{A.11}$$

Under Assumptions 3(i)-(ii), standard second order Taylor expansion arguments assure that for any $\epsilon > 0$,

$$\sup_{\theta\in\mathcal{N}} P_\theta^*\left(\left|\Lambda_N^*\left(\theta + h/\sqrt{N}|\theta\right) - h'S_N^*(\theta) + \frac{1}{2}h'\hat{I}_N^*(\theta)h\right| > \epsilon\right) \xrightarrow{p} 0. \tag{A.12}$$

The above implies

$$P_N^*\left(\left|\Lambda_N^*\left(\bar{\hat{\theta}}|\theta_N\right) + h'S_N^*(\theta_N) + \frac{1}{2}h'\hat{I}_N^*(\theta_N)h\right| > \epsilon\right) \xrightarrow{p} 0. \tag{A.13}$$

Combined with (A.11), I have thus shown the following: wpa1-$\tilde{P}_0$,

$$\Lambda_N^*\left(\bar{\hat{\theta}}|\theta_N\right) = -h'S_N^*(\theta_N) - \frac{1}{2}h'I(\theta_0)h + o_{P_N^*}(1). \tag{A.14}$$

This proves the first claim of the lemma.

The limiting distribution of $S_N^*(\theta_N)$ under P_N^* can be ascertained using the Lindberg-Feller central limit theorem for triangular arrays. Indeed, $\psi_{N,i}^*(.)$ is mean zero and uniformly bounded by Assumptions 3(i)-(ii), which implies the Lyapunov condition is trivially satisfied. The bootstrap variance of $\psi_{N,i}^*(.)$ is also simply $I^*(\theta_N)$. Thus by the arguments leading to (A.11), I obtain

$$\mathcal{L}_N^*(S_N^*(\theta_N)) \xrightarrow{p} \mathcal{L}(\mathbf{v_2}) \tag{A.15}$$

with $\mathbf{v_2} \sim N(0, I(\theta_0))$. From (A.14) and (A.15), it follows by an application of Le Cam's

first lemma that P_N^* and P^* are mutually contiguous, wpa1-$\tilde{P}_0$.

I shall now prove that wpa1-$\tilde{P}_0$,

$$\left\|\hat{\theta}^* - \bar{\hat{\theta}}\right\| = o_{P^*}(1). \tag{A.16}$$

I shall show $P^*\left(\left\|\hat{\theta}^* - \theta_0\right\| > \epsilon\right) \xrightarrow{p} 0$ for any $\epsilon > 0$. Since $\bar{\hat{\theta}} \to \theta_0$ a.s-$\tilde{P}_0$, this proves (A.16). To this end it suffices to verify the conditions for the consistency result of Newey and McFadden (1994, Theorem 2.7). Note that each summand within $L(\theta|\mathbf{W}^*, \mathbf{X}^*)$ is uniformly bounded wpa1-$\tilde{P}_0$ (due to Assumptions 3(i)-(ii) and 5(ii)); hence standard arguments using Markov's inequality assure that wpa1-$\tilde{P}_0$,

$$\frac{1}{N} L(\theta|\mathbf{W}^*, \mathbf{X}^*) - \frac{1}{N} E^*\left[L(\theta|\mathbf{W}^*, \mathbf{X}^*)\right] = o_{P^*}(1).$$

Now it is possible to expand

$$\frac{1}{N} E^*\left[L(\theta|\mathbf{W}^*, \mathbf{X}^*)\right] = \frac{1}{N}\sum_{i=1}^{N} A_{1N,i}(\theta),$$

where

$$A_{1N,i}(\theta) = F\left(X_i'\bar{\hat{\theta}}\right) \ln F\left(X_i'\theta\right) + \left(1 - F\left(X_i'\bar{\hat{\theta}}\right)\right) \ln\left(1 - F\left(X_i'\theta\right)\right).$$

The uniform law of large numbers, together with the fact $\bar{\hat{\theta}} \to \theta_0$ a.s-$\tilde{P}_0$, assures

$$\frac{1}{N}\sum_{i=1}^{N} A_{1N,i}(\theta) \xrightarrow{p} E_0\left[F\left(X_i'\theta_0\right) \ln F\left(X_i'\theta\right) + \left(1 - F\left(X_i'\theta_0\right)\right) \ln\left(1 - F\left(X_i'\theta\right)\right)\right] \equiv M(\theta).$$

I have thus shown that pointwise for each θ,

$$\frac{1}{N} L(\theta|\mathbf{W}^*, \mathbf{X}^*) = M(\theta) + o_{P^*}(1),$$

wpa1-$\tilde{P}_0$. Clearly $M(\theta)$ is concave. Furthermore, since $E_0[X_i X_i']$ is positive definite, θ_0 is the unique maximiser of $M(\theta)$ (see Newey and McFadden, 1994, Example 2.1 in p.2125). Combining the above, it can be noted that all the conditions for applying Theorem 2.7 of

Newey and McFadden (1994) are verified. This proves (A.16).

I can now prove the second claim of the lemma. Using (A.16) and Assumption 3(ii) (finite second derivatives for $F(.)$), the usual linearization arguments can be applied show that wpa1-$\tilde{P}_0$,

$$\sqrt{N}\left(\hat{\theta}^* - \bar{\hat{\theta}}\right) = \hat{I}_N^*\left(\bar{\hat{\theta}}\right)^{-1} S_N^*\left(\bar{\hat{\theta}}\right) + o_{P^*}(1).$$

Contiguity, proven earlier, then gives

$$\sqrt{N}\left(\hat{\theta}_N^* - \theta_N\right) = -h + \hat{I}_N^*\left(\bar{\hat{\theta}}\right)^{-1} S_N^*\left(\bar{\hat{\theta}}\right) + o_{P_N^*}(1), \tag{A.17}$$

wpa1-$\tilde{P}_0$. Using (A.12) and (A.11), together with Assumption 4 (which implies $I(.)$ is continuous on $\mathcal{N}$), I adapt the arguments of Bickel et al (1998, Proposition 2.1.2) to show that wpa1-$\tilde{P}_0$,

$$\left\| S_N^*(\theta_N) - S_N^*\left(\bar{\hat{\theta}}\right) - \hat{I}_N^*\left(\bar{\hat{\theta}}\right) h \right\| = o_{P_N^*}(1).$$

Substituting the above in (A.17), and using (A.11) proves (A.10), the second claim of the lemma. □

Lemma 2. *Under Assumptions 1-5 and $\theta_N \to \theta_0$ a.s-$\tilde{P}_0$, it holds $\left|T_N^*(\theta_N) - \tilde{T}_N^*(\theta_N)\right| = o_{P_N^*}(1)$, wpa1-$\tilde{P}_0$.*

Proof. Define $\varrho_{N,i}^*(\theta_N) = \varepsilon_i^*(\theta_N) - \tilde{\varepsilon}_i^*(\theta_N)$, and observe that

$$T_N^*(\theta_N) - \tilde{T}_N^*(\theta_N) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ \varrho_{N,i}^*(\theta_N) - E^*[\varrho_{N,i}^*(\theta_N)] \right\}.$$

Hence, I obtain

$$E^*\left|T_N^*(\theta_N) - \tilde{T}_N^*(\theta_N)\right|^2 \leq E^*\left|\varrho_{N,i}^*(\theta_N)\right|^2 \equiv A_N.$$

I can further bound $A_N \leq 2(A_{1N} + A_{2N}^{(0)} + A_{2N}^{(1)})$, where, for $w = 0, 1$

$$A_{1N} = E^*\left|\hat{e}_{1S_i^*}(\theta_N) - e_{1S_i^*}(\theta_N)\right|^2, \quad \text{and}$$
$$A_{2N}^{(w)} = E^*\left|\hat{\nu}_{S_i^*}(w;\theta_N) - \nu_{S_i^*}(w;\theta_N)\right|^2.$$

By Assumption 5, standard arguments assure $A_{1N} \xrightarrow{p} 0$ under $\tilde{P}_0$. Consequently I focus

on the term $A_{2N}^{(1)}$. By the definition of $\tilde{K}_M(i;w,\theta)$, there exists some constant $C<\infty$ for which

$$\begin{aligned} A_{2N}^{(1)} &\leq C\left\{1+\sup_{1\leq i\leq N} K_M^2(i;\theta_N)\right\} \times \frac{1}{N}\sum_{i=1}^{N}\left\{e_{2\mathcal{J}_1(i)}(1;\theta_N)-\hat{e}_{2\mathcal{J}_1(i)}(1;\theta_N)\right\}^2 \\ &\equiv \Gamma_{1N}\times\Gamma_{2N}. \end{aligned}$$

By Lemma 6 in Appendix A.3, $\Gamma_{1N}=o_p(N^{\xi/2})$ under $\tilde{P}_0$ for any ξ arbitrarily small. Next consider the term Γ_{2N}: The maximum number of times an observation i is used as a secondary match is bounded by the matching function for the nearest neighbor matching, given by $K_{\mathrm{NN}}(i)$. Consequently,

$$\Gamma_{2N}\leq\left\{1+\sup_{1\leq i\leq N} K_{\mathrm{NN}}(i)\right\}\sup_{\theta\in\mathcal{N}}\frac{1}{N}\sum_{i=1}^{N}\left\{\hat{e}_{2i}(1;\theta)-e_{2i}(1;\theta)\right\}^2.$$

Now, by Abadie and Imbens (2006, Lemma 3), $\sup_{1\leq i\leq N} K_{\mathrm{NN}}(i)=o_p(N^{\xi/2})$ under $\tilde{P}_0$ for any ξ arbitrarily small. Combined with Assumption 5, this assures $\Gamma_{2N}=O_p(N^{-\xi/2})$ under $\tilde{P}_0$. Taken together, the above imply $A_{2N}^{(1)}\xrightarrow{p}0$. Analogous arguments for $w=0$ similarly imply $A_{2N}^{(0)}\xrightarrow{p}0$. This completes the proof of the lemma. □

Lemma 3. *Under Assumptions 1-5 and $\bar{\theta}_N\to\theta_0$, it holds $V_1(h,\breve{\theta}_N)=o_{\bar{P}_N}(1)$.*

Proof. I first note that

$$\bar{E}_N^*\left[\left(t_1\gamma_i^*(\bar{\theta}_N;\breve{\theta}_N\right)\left(t_2'\alpha_i^*(\theta_N)\right)\right]=\bar{E}_N^*\left[\left(t_1\tilde{\varepsilon}_{N,i}^*(\bar{\theta}_N;\breve{\theta}_N)\right)\left(t_2'\alpha_i^*(\bar{\theta}_N)\right)\right]$$

since α_i^* is mean zero under $\bar{E}_N^*[.]$. Decompose

$$\tilde{\varepsilon}_{N,i}^*(\bar{\theta}_N;\breve{\theta}_N)=e_{1S_i^*}(\bar{\theta}_N)+W_i^*\nu_{S_i^*}(1;\bar{\theta}_N)-(1-W_i^*)\nu_{S_i^*}(0;\bar{\theta}_N).$$

Now based on the bootstrap DGP it is straightforward to verify $\bar{E}_N^*\left[e_{1S_i^*}(\bar{\theta}_N)\left(t_2'\alpha_i^*(\bar{\theta}_N)\right)\right]=$

0. Hence the claim follows if I show that

$$Q_N^{(1)}(\bar{\theta}_N) \equiv \bar{E}_N^* \left[\left(W_i^* \nu_{S_i^*}(1; \bar{\theta}_N) \right) \left(t_2' \alpha_i^*(\bar{\theta}_N) \right) \right] = o_{\bar{P}_N}(1);$$
$$Q_N^{(0)}(\bar{\theta}_N) \equiv \bar{E}_N^* \left[\left((1 - W_i^*) \nu_{S_i^*}(0; \bar{\theta}_N) \right) \left(t_2' \alpha_i^*(\bar{\theta}_N) \right) \right] = o_{\bar{P}_N}(1).$$

I show $Q_N^{(1)}(\bar{\theta}_N) \xrightarrow{p} 0$ under $\bar{P}_N$; that $Q_N^{(0)}(\bar{\theta}_N) \xrightarrow{p} 0$ follows by similar reasoning. To this end, first define the quantity

$$\tau(X_i' \bar{\theta}_N) = t_1 t_2' h \left(X_i \bar{\theta}_N; \bar{\theta}_N \right) f \left(X_i' \bar{\theta}_N \right).$$

Due to Assumption 3(i), which implies X_i^* is bounded, it follows $h(.; \bar{\theta}_N)$ is uniformly bounded over its domain for all $\bar{\theta}_N$. Combined with Assumption 3(ii) (boundedness of $f(.)$), this implies $\tau(X_i' \bar{\theta}_N) \leq C < \infty$ uniformly in both i and N .

Taking the bootstrap expectations, I obtain after some algebra

$$Q_N^{(1)}(\bar{\theta}_N) = \frac{1}{N} \sum_{i=1}^{N} \tau(X_i' \bar{\theta}_N) \left(1 + \frac{\tilde{K}_M(i; 1, \bar{\theta}_N)}{M} \right) e_{2\mathcal{J}_1(i)} \left(1; \bar{\theta}_N \right).$$

I can decompose $Q_N^{(1)}(\bar{\theta}_N)$ further as

$$Q_N^{(1)}(\bar{\theta}_N) = \frac{1}{N} \sum_{i=1}^{N} \vartheta_{(3)N,i} + \frac{1}{N} \sum_{i=1}^{N} \vartheta_{(4)N,i} \equiv Q_{3N}^{(1)}(\bar{\theta}_N) + Q_{4N}^{(1)}(\bar{\theta}_N),$$

where

$$\vartheta_{(3)N,i} = \tau(X_i' \bar{\theta}_N) \left(1 + \frac{\tilde{K}_M(i; 1, \bar{\theta}_N)}{M} \right) e_{3\mathcal{J}_1(i)} \left(1; \bar{\theta}_N \right);$$
$$\vartheta_{(4)N,i} = \tau(X_i' \bar{\theta}_N) \left(1 + \frac{\tilde{K}_M(i; 1, \bar{\theta}_N)}{M} \right) e_{4\mathcal{J}_1(i)} \left(1; \bar{\theta}_N \right).$$

First consider the term $Q_{4N}^{(1)}(\bar{\theta}_N)$: For each i, $\bar{E}_N[\vartheta_{(4)N,i} | \mathbf{W}, \mathbf{X}, \mathbf{U}] = 0$ due to the definition of $e_{4i}(.)$. Furthermore, I also have $\bar{E}_N[\vartheta_{(4)N,i} \vartheta_{(4)N,,j} | \mathbf{W}, \mathbf{X}, \mathbf{U}] = 0$ for all i, j for which $\mathcal{J}_1(i) \neq \mathcal{J}_1(j)$. Denoting $\mathcal{S}_k = \{i \in \{1, \ldots, N\} : \mathcal{J}_1(i) = k\}$, I note that the cardinality of $\mathcal{S}_k$ is bounded by $K_{\text{NN}}(k)$. Hence it follows that the number of pairs (i, j) for which $\bar{E}_N[\vartheta_{(4)N,i} \vartheta_{(4)N,,j} | \mathbf{W}, \mathbf{X}, \mathbf{U}] \neq 0$ is bounded above by $N \sup_{1 \leq k \leq N} K_{\text{NN}}(k)$. Now by

Assumption 3(v) (which assures $\sup_x E_0[Y^4|X=x] \leq C < \infty$), it follows $\sup_{1\leq i\leq N} \bar{E}_N \left[|e_{4i}(1;\theta)|^2 |\mathbf{W},\mathbf{X},\mathbf{U} \right] \leq C < \infty$ uniformly over $\theta \in \mathcal{N}$ (note that $\mathbf{U}$ is independent of $\mathbf{Y}_1, \mathbf{Y}_0$ by definition). Thus, by the Markov inequality and the boundedness of $\tau(.)$, there exists some $C_1 < \infty$for which

$$\begin{aligned}\bar{E}_N \left[\left\{ Q_{4N}^{(1)}(\bar{\theta}_N) \right\}^2 |\mathbf{W},\mathbf{X},\mathbf{U} \right] &\leq C_1 N^{-1} \left\{ 1 + \sup_{1\leq i\leq N} K_{\mathrm{NN}}(i) \right\} \left\{ 1 + \sup_{1\leq i\leq N} \tilde{K}_M^2(i;1,\bar{\theta}_N) \right\} \\ &= C_1 N^{-1} \left\{ 1 + \sup_{1\leq i\leq N} K_{\mathrm{NN}}(i) \right\} \left\{ 1 + \sup_{1\leq i\leq N} K_M^2(i;\bar{\theta}_N) \right\}.\end{aligned}$$

Using the result of Abadie and Imbens (2006, Lemma 3), $\bar{E}_N[\sup_{1\leq i\leq N} K_{\mathrm{NN}}^r(i)] = O(N^{\xi})$ for any finite r, and some $\xi > 0$ arbitrarily small. Taking a further expectation on both sides of the above equation and employing Lemma 6, together with Holder's inequality, gives $\bar{E}_N \left[\left\{ Q_{4N}^{(1)}(\bar{\theta}_N) \right\}^2 \right] = O(N^{-(1-\xi)})$ for some $\xi > 0$ arbitrarily small. This proves $Q_{4N}^{(1)}(\bar{\theta}_N) = o_{\bar{P}_N}(1)$.

Next consider the term $Q_{3N}^{(1)}(\bar{\theta}_N)$. First I successively approximate this term by the quantities $Q_{31N}^{(1)}(\bar{\theta}_N)$, $Q_{32N}^{(1)}(\bar{\theta}_N)$, where

$$\begin{aligned}Q_{31N}^{(1)}(\bar{\theta}_N) &= \frac{1}{N}\sum_{i=1}^N \vartheta_{(31)N,i}; \quad \vartheta_{(31)N,i} = \tau(X_i'\bar{\theta}_N) \left(1 + \frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M} \right) e_{3i}\left(1;\bar{\theta}_N\right); \\ Q_{32N}^{(1)}(\bar{\theta}_N) &= \frac{1}{N}\sum_{i=1}^N \vartheta_{(32)N,i}; \quad \vartheta_{(31)N,i} = \tau(X_i'\bar{\theta}_N) \left(1 + \frac{\bar{K}_M(i;1,\bar{\theta}_N)}{M} \right) e_{3i}\left(1;\bar{\theta}_N\right);\end{aligned}$$

In the first case, by Lemma 6,

$$\begin{aligned}\left| Q_{3N}^{(1)}(\bar{\theta}_N) - Q_{31N}^{(1)}(\bar{\theta}_N) \right| &\leq C \left\{ 1 + \sup_{1\leq i\leq N} K_M(i;\bar{\theta}_N) \right\} \max_{1\leq i\leq N} \left| e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N) - e_{3i}(1;\bar{\theta}_N) \right| \\ &= O_{\bar{P}_N}(N^{\xi}) \cdot \max_{1\leq i\leq N} \left| e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N) - e_{3i}(1;\bar{\theta}_N) \right|.\end{aligned}$$

The last term can in turn be bounded as

$$\begin{aligned}
&\max_{1\leq i\leq N}\left|e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N)-e_{3i}(1;\bar{\theta}_N)\right| \\
&\leq \max_{1\leq i\leq N}\left|\bar{\mu}(1;X_{\mathcal{J}_1(i)})-\bar{\mu}(1;X_i)\right|+\max_{1\leq i\leq N}\left|\mu\left(1;F(X'_{\mathcal{J}_1(i)}\bar{\theta}_N);\bar{\theta}_N\right)-\mu\left(1;F(X_i'\bar{\theta}_N);\bar{\theta}_N\right)\right| \\
&\leq \max_{1\leq i\leq N}\left\|X_{\mathcal{J}_1(i)}-X_i\right\|=O_{\bar{P}_N}(N^{-1/k}),
\end{aligned}$$

where the first inequality follows by Assumption 3(i)-(ii); the third by Assumption 3(v) (which implies Lipschitz continuity of $\bar{\mu}(1;.)$ and $\mu(1;.;\bar{\theta}_N)$ uniformly over $\bar{\theta}_N \in \mathcal{N}$); and the final step follows by the results of Abadie and Imbens (2006, Lemma 2) on the bias of nearest neighbor matching. This proves $\left|Q^{(1)}_{31N}(\bar{\theta}_N)-Q^{(1)}_{32N}(\bar{\theta}_N)\right| \xrightarrow{p} 0$ under $\bar{P}_N$. I now argue $\left|Q^{(1)}_{31N}(\bar{\theta}_N)-Q^{(1)}_{32N}(\bar{\theta}_N)\right| \xrightarrow{p} 0$ under $\bar{P}_N$: Observe that $Q^{(1)}_{32N}(\bar{\theta}_N) = \bar{E}_N[Q^{(1)}_{31N}(\bar{\theta}_N)|\mathbf{W},\mathbf{X}]$ (the expectation being taken over $\mathbf{U}$, conditional on $\mathbf{W},\mathbf{X}$). But conditional on $\mathbf{W},\mathbf{X}$, the random variables $\{U_i : 1\leq i\leq N\}$ are all independent of each other. Hence, by standard arguments involving the Markov inequality, together with Lemma 6 (i.e, $\bar{E}_N\left[\sup_{1\leq i\leq N} K^2_M(i;\bar{\theta}_N)\right]=o(N^\delta)$) and Assumption 3, (which implies $\tau(X_i'\bar{\theta}_N)<\infty$ and $|e_{3i}(1;\bar{\theta}_N)|<\infty$ uniformly in i and N), it follows $\left|Q^{(1)}_{31N}(\bar{\theta}_N)-Q^{(1)}_{32N}(\bar{\theta}_N)\right| = o_{\bar{P}_N}(1)$.

It now remains to obtain the probability limit wrt $\bar{P}_N$ of $Q^{(1)}_{32N}(\bar{\theta}_N)$. Exploiting the definition of $\bar{K}_M(i;1,\bar{\theta}_N)$ and reordering the variables in the summation gives

$$\begin{aligned}
Q^{(1)}_{32N}(\bar{\theta}_N) &= \frac{1}{N}\sum_{W_j=1}\tau(X_j'\bar{\theta}_N)\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)e_{3j}\left(1;\bar{\theta}_N\right) \\
&\quad+\frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{N_1(l(j))}\left\{\sum_{i\in S_0\left(l(j);\breve{\theta}_N\right)}\tau(X_i'\bar{\theta}_N)e_{3i}\left(1;\bar{\theta}_N\right)\right\} \\
&\equiv A^{(1)}_N(\bar{\theta}_N)+B^{(1)}_N(\bar{\theta}_N).
\end{aligned}$$

Conditional on $\mathbf{W},\mathbf{X}'\bar{\theta}_N$, the summands within $A^{(1)}_N(\bar{\theta}_N)$ are mean zero and uncorrelated. Hence using Assumption 3 and Lemma 6, standard arguments assure $A^{(1)}_N(\bar{\theta}_N)=o_{\bar{P}_N}(1)$. Next, consider the term $B^{(1)}_N(\bar{\theta}_N)$: Suppose for simplicity that N/q_N is integer valued so that $N(l)=N/q_N$ for all l. I shall successively approximate $B^{(1)}_N(\bar{\theta}_N)$ by $B^{(1)}_{1N}(\bar{\theta}_N)$ and

$B_{2N}^{(1)}(\bar{\theta}_N)$,[4] where

$$B_{1N}^{(1)}(\bar{\theta}_N) = \frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{F(X_j'\bar{\theta}_N)}\left\{\frac{q_N}{N}\sum_{i\in S_0\left(l(j);\breve{\theta}_N\right)}\tau(X_i'\bar{\theta}_N)e_{3i}\left(1;\bar{\theta}_N\right)\right\};$$

and

$$B_{2N}^{(1)}(\bar{\theta}_N) = \frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{F(X_j'\bar{\theta}_N)}\left\{\frac{q_N}{N}\sum_{i\in S_0\left(l(j);\bar{\theta}_N\right)}\tau(X_i'\bar{\theta}_N)e_{3i}\left(1;\bar{\theta}_N\right)\right\}.$$

I first show that

$$\left|B_N^{(1)}(\bar{\theta}_N) - B_{1N}^{(1)}(\bar{\theta}_N)\right| \xrightarrow{p} 0. \tag{A.18}$$

Indeed a straightforward consequence of Lemma 9 and Assumption 3 (which implies $F^{-1}(.)$ is Lipschitz continuous and X_i is uniformly bounded) is that

$$\begin{aligned}&\sup_{1\le j\le N}\left|\frac{N(l(j))}{N_1(l(j))} - F^{-1}(X_j'\bar{\theta}_N)\right| \\ &\le \sup_{1\le j\le N}\left|\frac{N(l(j))}{N_1(l(j))} - F^{-1}\left(X_j'\breve{\theta}_N\right)\right| + \sup_{1\le j\le N}\left|F^{-1}(X_j'\bar{\theta}_N) - F^{-1}\left(X_j'\breve{\theta}_N\right)\right| = o_{\bar{P}_N}(1).\end{aligned}$$

Combining the above result with Lemma 6, and the fact $\tau(X_i'\bar{\theta}_N), |e_{3i}(1;\bar{\theta}_N)|$ are uniformly bounded, proves (A.18). Next, I show that

$$\left|B_{1N}^{(1)}(\bar{\theta}_N) - B_{2N}^{(1)}(\bar{\theta}_N)\right| = o_{\bar{P}_N}(1). \tag{A.19}$$

Let

$$\Delta(l;\bar{\theta}_N) \equiv S_0(l;\breve{\theta}_N) \triangle S_0(l;\bar{\theta}_N),$$

where $C \triangle D$ denotes the symmetric difference between any two sets C, D. Lemma 10 assures

$$\bar{P}_N\left(\max_{1\le l\le q_N}\#\Delta(l;\bar{\theta}_N) \ge N^{(1+\delta)/2}\right) \le q_N\exp(-N^\delta) \to 0 \tag{A.20}$$

for any $\delta > 0$ arbitrarily small (where $\#C$ denotes the cardinality of a set C). Combined

[4] Note the difference in summation between the two terms.

with the boundedness property of $\tau(X_i'\bar{\theta}_N)$ and $|e_{3i}(1;\bar{\theta}_N)|$, (A.20) implies

$$\begin{aligned}\left|B_{1N}^{(1)}(\bar{\theta}_N) - B_{2N}^{(1)}(\bar{\theta}_N)\right| &= O_{\bar{P}_N}\left(\frac{q_N}{N^{(1-\delta)/2}}\right) \times \frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{F(X_j'\bar{\theta}_N)} \\ &= O_{\bar{P}_N}\left(\frac{q_N}{N^{(1-\delta)/2}}\right) \times O_{\bar{P}_N}(1) = o_{\bar{P}_N}(1),\end{aligned}$$

where the first equality follows from Lemma 6 and Assumption 3; and the final equality follows by Assumption 6. I have thus shown (A.19).

To complete the proof of the Lemma it remains to show

$$B_{2N}^{(1)}(\bar{\theta}_N) = o_{\bar{P}_N}(1). \tag{A.21}$$

Let $\rho_{N,i} = \tau(X_i'\bar{\theta}_N)e_{3i}(1;\bar{\theta}_N)$. For each l, the collection of random variables $\{\rho_{N,i} : i \in S_0(l;\bar{\theta}_N)\}$ are mean zero and uncorrelated conditional on $\mathbf{X}'\bar{\theta}_N$. Furthermore, wpa1-$\bar{P}_N$,

$$\#S_0(l(j);\bar{\theta}_N) \leq \frac{N}{q_N} + \max_{1\leq l\leq q_N} \#\Delta(l;\bar{\theta}_N) \leq \frac{N}{q_N} + N^{(1+\delta)/2}.$$

Hence for each $\epsilon > 0$, by the Markov inequality

$$\begin{aligned}\bar{P}_N\left(\max_{1\leq l\leq q_N}\left|\frac{q_N}{N}\sum_{i\in S_0(l;\bar{\theta}_N)}\rho_{N,i}\right| \geq \epsilon\right) &\leq \sum_{l=1}^{q_N}\bar{P}_N\left(\left|\frac{q_N}{N}\sum_{i\in S_0(l;\bar{\theta}_N)}\rho_{N,i}\right| \geq \epsilon\right) \\ &= O\left(\frac{q_N^2}{N} + \frac{q_N^2}{N^{(3-\delta)/2}}\right) = o(1).\end{aligned}$$

This, combined with the fact

$$\frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)\frac{1}{F(X_j'\bar{\theta}_N)} = O_{\bar{P}_N}(1),$$

immediately proves (A.21). □

Lemma 4. *Under Assumptions 1-5 and $\bar{\theta}_N \to \theta_0$, it holds $V_2(h,\breve{\theta}_N) = t_1^2\sigma^2 + o_{\bar{P}_N}(1)$.*

Proof. For the remainder of this proof I shall denote $p_{i,N} = F(X_i'\bar{\theta}_N)$. Additionally, for $a = 3, 4$, I set

$$\phi_{(a)i}(w;\theta) = (2w-1)\nu_{(a)i}(w;\theta).$$

First, note that $\bar{E}_N^*\left[\tilde{\varepsilon}_i^*(\bar{\theta}_N;\breve{\theta}_N)\right] = o_{\bar{P}_N}(1)$. Indeed this follows by a similar argument as in the proof of Lemma 3. Hence it suffices to show that $\bar{E}_N^*\left[\tilde{\varepsilon}_i^{*2}\left(\bar{\theta}_N;\breve{\theta}_N\right)\right] = \sigma^2 + o_{\bar{P}_N}(1)$. To this end, I decompose

$$\tilde{\varepsilon}_i^*(\bar{\theta}_N;\breve{\theta}_N) = e_{1S_i^*}(\bar{\theta}_N) + \phi_{(3)S_i^*}(W_i^*;\bar{\theta}_N) + \phi_{(4)S_i^*}(W_i^*;\bar{\theta}_N), \tag{A.22}$$

and determine the probability limits of all the squared and cross product terms in (A.22), after taking the bootstrap expectation.

I begin with the probability limit of $\bar{E}_N^*\left[\phi_{(4)S_i^*}^2\left(W_i^*;\bar{\theta}_N\right)\right]$ under $\bar{P}_N$. Note that

$$\begin{aligned}
\bar{E}_N^*\left[\phi_{(4)S_i^*}^2\left(W_i^*;\bar{\theta}_N\right)\right] &= \frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 e_{4\mathcal{J}_1(i)}^2\left(1;\bar{\theta}_N\right) \\
&\quad + \frac{1}{N}\sum_{i=1}^N (1-p_{i,N})\left(1+\frac{\tilde{K}_M(i;0,\bar{\theta}_N)}{M}\right)^2 e_{4\mathcal{J}_0(i)}^2\left(0;\bar{\theta}_N\right) \\
&\equiv \Gamma_N^{(1)} + \Gamma_N^{(0)},
\end{aligned}$$

I shall characterize probability limit of $\Gamma_N^{(1)}$. That for $\Gamma_N^{(0)}$ follows by a similar argument.

Recall the definition $\bar{\sigma}^2(w,X) = E[Y^2|W=w,X]$. I shall first successively approximate $\Gamma_N^{(1)}$ by $\Gamma_{1N}^{(1)}$, $\Gamma_{2N}^{(1)}$, where

$$\begin{aligned}
\Gamma_{1N}^{(1)} &= \frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 \bar{\sigma}^2(1;X_{\mathcal{J}_w(i)}); \\
\Gamma_{2N}^{(1)} &= \frac{1}{N}\sum_{i=1}^N p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 \bar{\sigma}^2(1;X_i).
\end{aligned}$$

In the first instance, I can expand

$$\Gamma_N^{(1)} - \Gamma_{1N}^{(1)} = \frac{1}{N}\sum_{i=1}^N \zeta_{N,i},$$

where

$$\zeta_{N,i} = p_{i,N}\bar{\Psi}_i(1,\bar{\theta}_N)\left\{e_{4\mathcal{J}_w(i)}^2(W_i,\bar{\theta}_N) - \bar{\sigma}^2(1;X_{\mathcal{J}_w(i)})\right\}.$$

Clearly $\bar{E}_N[\zeta_{N,i}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ and $\bar{E}_N[\zeta_{N,i}\zeta_{N,j}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ for all i,j such that $\mathcal{J}_1(i) \neq$

$\mathcal{J}_1(j)$. Consequently by similar arguments[5] as in the proof of Lemma 3, it follows $\left|\Gamma_N^{(1)} - \Gamma_{1N}^{(1)}\right| = o_{\bar{P}_N}(1)$. Additionally, I can also show $\left|\Gamma_{1N}^{(1)} - \Gamma_{2N}^{(1)}\right| = o_{\bar{P}_N}(1)$ by similar arguments[6] as that used in the proof of Lemma 3 (note that by Assumption 3(v), $\bar{\sigma}^2(1;.)$ is Lipschitz continuous and uniformly bounded).

It now remains to obtain the probability limit wrt $\bar{P}_N$ of $\Gamma_{1N}^{(1)}$. By paralleling some of the steps[7] in the proof of Lemma 3, it follows

$$\left|\Gamma_{2N}^{(1)} - \Gamma_{3N}^{(1)}\right| = o_{\bar{P}_N}(1),$$

where

$$\Gamma_{3N}^{(1)} = \frac{1}{N}\sum_{W_j=1} p_{j,N}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 \bar{\sigma}^2(1;X_j) + \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 \frac{1}{p_{j,N}}\left\{\frac{q_N}{N}\sum_{i\in S_0(l(j);\bar{\theta}_N)} p_{i,N}\bar{\sigma}^2(1;X_i)\right\}.$$

Define

$$\Gamma_{4N}^{(1)} = \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 m_1\left(p_{j,N};\bar{\theta}_N\right),$$

where

$$m_1\left(p;\bar{\theta}_N\right) = \bar{E}_N\left[\bar{\sigma}^2(1;X)|F(X_i'\bar{\theta}_N) = p\right].$$

I now show

$$\left|\Gamma_{3N}^{(1)} - \Gamma_{4N}^{(1)}\right| = o_{\bar{P}_N}(1). \tag{A.23}$$

To this end, I define an intermediate variable:

$$\Gamma_{31N}^{(1)} = \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 p_{j,N}\cdot m_1\left(p_{j,N};\bar{\theta}_N\right) + \frac{1}{N}\sum_{W_j=1}\left(1 + \frac{K_M(j;\bar{\theta}_N)}{M}\right)^2 \frac{1-p_{j,N}}{p_{j,N}}\left\{\frac{1}{N_0\left(l(j);\bar{\theta}_N\right)}\sum_{i\in S_0(l(j);\bar{\theta}_N)} p_{i,N}\cdot m_1\left(p_{i,N};\bar{\theta}_N\right)\right\}.$$

[5]Specifically, the ones used to prove $Q_{4N}^{(1)}(\bar{\theta}_N) \xrightarrow{p} 0$.

[6]Specifically, the ones used to prove $\left|Q_{3N}^{(1)}(\bar{\theta}_N) - Q_{31N}^{(1)}(\bar{\theta}_N)\right| = o_{\bar{P}_N}(1)$.

[7]Precisely, the steps leading to $\left|Q_{31N}^{(1)}(\bar{\theta}_N) - Q_{32N}^{(1)}(\bar{\theta}_N)\right| = o_{\bar{P}_N}(1)$, followed by reordering of the terms in $Q_{32N}^{(1)}(\bar{\theta}_N)$, and finally succesive approximations of $B_N^{(1)}(\bar{\theta}_N)$ with $B_{1N}^{(1)}(\bar{\theta}_N)$ and $B_{2N}^{(1)}(\bar{\theta}_N)$.

By Lemmas 9 and 10, and Assumption 6, there exists some $c > 0$ for which it holds

$$\min_{1\leq l\leq q_N} N_0\left(l(j);\bar{\theta}_N\right) \geq \min_{1\leq l\leq q_N} N_0(l) - N^{(1+\delta)/2} \geq cN/q_N, \tag{A.24}$$

with probability approaching one under $\bar{P}_N$. The same lemmas together with Assumptions 3,6 also assure

$$\begin{aligned}\sup_{1\leq j\leq N}\left|\frac{q_N N_0\left(l(j);\bar{\theta}_N\right)}{N} - (1-p_{j,N})\right| &\leq \sup_{1\leq j\leq N}\left|\frac{q_N N_0\left(l(j);\bar{\theta}_N\right)}{N} - \left(1-F(X_j'\breve{\theta})\right)\right| + o_{\bar{P}_N}(N^{-\frac{1}{2}}) \\ &\leq \sup_{1\leq j\leq N}\left|\frac{q_N N_0\left(l(j);\breve{\theta}_N\right)}{N} - \left(1-F(X_j'\breve{\theta})\right)\right| + o_{\bar{P}_N}\left(\frac{q_N}{N^{(1-\delta)/2}} + N^{-\frac{1}{2}}\right) = o_{\bar{P}_N}(1).\end{aligned}$$

Additionally, by the usual arguments based on the Markov inequality, and employing (A.24) together with Assumption 6, it follows

$$\bar{P}_N\left(\max_{1\leq l\leq q_N}\left|\frac{1}{N_0\left(l;\bar{\theta}_N\right)}\sum_{i\in S_0(l;\bar{\theta}_N)} p_{i,N}\bar{\sigma}^2(1;X_i) - \frac{1}{N_0\left(l;\bar{\theta}_N\right)}\sum_{i\in S_0(l;\bar{\theta}_N)} p_{i,N}m_1\left(p_{i,N};\bar{\theta}_N\right)\right| \geq \epsilon\right) \to 0.$$

Combining the above results with the fact

$$\frac{1}{N}\sum_{W_j=1}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)^2\frac{1}{p_{j,N}} = O_{\bar{P}_N}(1),$$

proves that $\left|\Gamma_{3N}^{(1)} - \Gamma_{31N}^{(1)}\right| = o_{\bar{P}_N}(1)$. Now, define $w_1\left(p;\bar{\theta}_N\right) \equiv p\cdot m_1(p;\bar{\theta}_N)$. I can bound

$$\begin{aligned}&\left|\Gamma_{31N}^{(1)} - \Gamma_{4N}^{(1)}\right| \\ &\leq \left\{\frac{1}{N}\sum_{W_j=1}\frac{1-p_{j,N}}{p_{j,N}}\left(1+\frac{K_M(j;\bar{\theta}_N)}{M}\right)^2\right\}\max_{1\leq l\leq q_N}\max_{i,j\in S_0(l;\bar{\theta}_N)}\left|w_1\left(p_{i,N};\bar{\theta}_N\right) - w_1\left(p_{j,N};\bar{\theta}_N\right)\right| \\ &= O_{\bar{P}_N}(1)\cdot\max_{1\leq l\leq q_N}\max_{i,j\in S_0(l;\bar{\theta}_N)}\left|w_1\left(p_{i,N};\bar{\theta}_N\right) - w_1\left(p_{j,N};\bar{\theta}_N\right)\right| \\ &\leq O_{\bar{P}_N}(1)\cdot\max_{1\leq l\leq q_N}\max_{i,j\in S_0(l;\bar{\theta}_N)}|p_{i,N}-p_{j,N}| \\ &\leq O_{\bar{P}_N}(1)\cdot\max_{1\leq l\leq q_N}\left|\pi_{l-1}(\breve{\theta}_N) - \pi_l(\breve{\theta}_N)\right| = o_{\bar{P}_N}(1),\end{aligned}$$

where the first equality follows by Assumption 3(i)-(iii) together with Lemma 6; the second

inequality follows by the uniform Lipschitz continuity of $m_1\left(.;\bar{\theta}_N\right)$ (Assumption 3(v)); the third inequality follows by the definition of $S_0(l;\bar{\theta}_N)$; and the final equality follows by Lemma 8. I have thus shown (A.23).

Now, the probability limit of $\Gamma_{4N}^{(1)}$ under $\bar{P}_N$ can be obtained by the techniques of Abadie and Imbens (2016) (See also Lemmas (14)-(16) in Appendix A.4). The probability limit of $\Gamma_{4N}^{(0)}$ under $\bar{P}_N$ is obtained analogously. Combining the expressions gives the probability limit of $\bar{E}_N^*\left[\phi^2_{(4)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right]$, which is equivalent to that obtained in Abadie and Imbens (2016).

Next consider the term $\bar{E}_N^*\left[\phi^2_{(3)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right]$. As before I can decompose

$$\begin{aligned}\bar{E}_N^*\left[\phi^2_{(3)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right] &= \frac{1}{N}\sum_{i=1}^{N} p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 e^2_{3\mathcal{J}_1(i)}\left(1;\bar{\theta}_N\right)\\ &\quad+\frac{1}{N}\sum_{i=1}^{N}(1-p_{i,N})\left(1+\frac{\tilde{K}_M(i;0,\bar{\theta}_N)}{M}\right)^2 e^2_{3\mathcal{J}_0(i)}\left(0;\bar{\theta}_N\right)\\ &\equiv \Delta_N^{(1)}+\Delta_N^{(0)}.\end{aligned}$$

Consider the term $\Delta_N^{(1)}$: By similar arguments as in Lemma 3,

$$\max_{1\le i\le N}\left|e_{3\mathcal{J}_1(i)}(1;\bar{\theta}_N)-e_{3i}(1;\bar{\theta}_N)\right| = O_{\bar{P}_N}(N^{-1/k}).$$

Together with Lemma 6, the above assures $\left|\Delta_N^{(1)}-\Delta_{1N}^{(1)}\right| = o_{\bar{P}_N}(1)$, where

$$\Delta_{1N}^{(1)} = \frac{1}{N}\sum_{i=1}^{N} p_{i,N}\left(1+\frac{\tilde{K}_M(i;1,\bar{\theta}_N)}{M}\right)^2 e^2_{3i}\left(1;\bar{\theta}_N\right).$$

Now the probability limit of $\Delta_{1N}^{(1)}$ can be analyzed the same arguments as that employed for $\Gamma_{1N}^{(1)}$. Doing so gives the probability limit for $\bar{E}_N^*\left[\phi^2_{(3)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right]$ under $\bar{P}_N$, which is again equivalent to the corresponding expression in Abadie and Imbens (2016).

Finally, it is straightforward to obtain the probability limit of $\bar{E}_N^*[e^2_{1S_i^*}(\bar{\theta}_N)]$ under $\bar{P}_N$ using standard methods. Taken together I can show

$$\bar{E}_N^*[e^2_{1S_i^*}(\bar{\theta}_N)]+\bar{E}_N^*\left[\phi^2_{(4)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right]+\bar{E}_N^*\left[\phi^2_{(3)S_i^*}\left(W_i^*;\bar{\theta}_N\right)\right] = \sigma^2+o_{\bar{P}_N}(1).$$

It only remains to verify that the bootstrap expectation of the cross product terms in (A.22) converge in probability to 0 under $\bar{P}_N$. Consider, for instance,

$$\Phi_N \equiv \bar{E}_N^* \left[\phi_{(3)i}(W_i^*; \bar{\theta}_N) \cdot \phi_{(4)i}(W_i^*; \bar{\theta}_N) \right].$$

Taking the bootstrap expectations, I observe $\Phi_N = \Phi_N^{(1)} + \Phi_N^{(0)}$, where

$$\Phi_N^{(1)} = \frac{1}{N} \sum_{i=1}^{N} p_{i,N} \left(1 + \frac{\tilde{K}_M(i; 1, \bar{\theta}_N)}{M} \right)^2 e_{3\mathcal{J}_1(i)}(1; \bar{\theta}_N) e_{4\mathcal{J}_1(i)}(1; \bar{\theta}_N),$$

and a similar expression holds for $\Phi_N^{(0)}$. Denoting

$$\varrho_{N,i} = p_{i,N} \left(1 + \frac{\tilde{K}_M(i; 1, \bar{\theta}_N)}{M} \right)^2 e_{3\mathcal{J}_1(i)}(1; \bar{\theta}_N) e_{4\mathcal{J}_1(i)}(1; \bar{\theta}_N),$$

I note that $\bar{E}_N[\varrho_{N,i} | \mathbf{W}, \mathbf{X}, \mathbf{U}] = 0$ and $\bar{E}_N[\varrho_{N,i}\varrho_{N,j} | \mathbf{W}, \mathbf{X}, \mathbf{U}] = 0$ for all i, j such that $\mathcal{J}_1(i) \neq \mathcal{J}_1(j)$. Consequently, by similar arguments as in the proof of Lemma 3, it follows $\Phi_N^{(1)} \xrightarrow{p} 0$ under $\bar{P}_N$. By symmetry, I also have $\Phi_N^{(0)} \xrightarrow{p} 0$ under $\bar{P}_N$, implying $\Phi_N = o_{\bar{P}_N}(1)$. Some more algebra on the usual lines shows that the remaining cross product terms also converge in probability to 0 under $\bar{P}_N$. This completes the proof of the lemma. □

Lemma 5. *Under Assumptions 1-5 and $\bar{\theta}_N \to \theta_0$, it holds $V_3(h, \breve{\theta}_N) = 2t_1 c' t_2 + o_{\bar{P}_N}(1)$.*

Proof. For the course of this proof set $h(.; \bar{\theta}_N)$ as $h_N(.)$. Furthermore, to simplify the algebra I again employ the notation (first introduced in the proof of Lemma 4)

$$\phi_{(a)i}(w; \theta) = (2w - 1)\nu_{(a)i}(w; \theta).$$

By the construction of the bootstrap DGP, it follows $V_3(h, \breve{\theta}_N) = \bar{E}_N^* \left[\left(t_1 \varepsilon_i^*(\bar{\theta}_N; \breve{\theta}_N) \right) \left(t_2' \beta_i^*(\bar{\theta}_N) \right) \right]$ since $\bar{E}_N^*[\beta_i^*(\bar{\theta}_N)] = 0$. I then decompose the term $\varepsilon_i^*(\bar{\theta}_N; \breve{\theta}_N)$ as in equation (A.22) and determine the probability limits of the bootstrap expectations of the resulting terms.

First, taking the bootstrap expectations it can be verified $\bar{E}_N^* \left[e_{1S_i^*}(\bar{\theta}_N) \cdot t_2' \beta_i^*(\bar{\theta}_N) \right] = 0$. At the end of the proof I show that

$$\bar{E}_N^* \left[\phi_{(4)i}(W_i^*; \bar{\theta}_N) \cdot t_2' \beta_i^*(\bar{\theta}_N) \right] = o_{\bar{P}_N}(1). \tag{A.25}$$

Hence it suffices for the claim to prove

$$\bar{E}_N^* \left[\phi_{(3)i}(W_i^*; \bar{\theta}_N) \cdot t_2' \beta_i^*(\bar{\theta}_N) \right] = t_2' c + o_{\bar{P}_N}(1).$$

Taking the bootstrap expectations, I obtain

$$\bar{E}_N^* \left[\phi_{(3)i}(W_i^*; \bar{\theta}_N) \cdot t_2' \beta_i^*(\bar{\theta}_N) \right] = T_N^{(1)} + T_N^{(0)},$$

where for $w = 0, 1$,

$$T_N^{(w)} = \frac{1}{N} \sum_{i=1}^{N} f\left(X_i' \bar{\theta}_N \right) t_2' \left\{ X_i - h_N \left(X_i' \bar{\theta}_N \right) \right\} \left(1 + \frac{\tilde{K}_M(i; w, \theta)}{M} \right) e_{3\mathcal{J}_w(i)}(w; \theta).$$

Let me now denote

$$T_{1N}^{(w)} = \frac{1}{N} \sum_{i=1}^{N} f\left(X_i' \bar{\theta}_N \right) t_2' \left\{ X_i - h_N \left(X_i' \bar{\theta}_N \right) \right\} \left(1 + \frac{\tilde{K}_M(i; w, \bar{\theta}_N)}{M} \right) e_{3i}(w; \bar{\theta}_N).$$

Using the properties of nearest neighbor matching, I can employ similar arguments as in the proof of Lemma (3) to show that for $w = 0, 1$,

$$\left| T_N^{(w)} - T_{1N}^{(w)} \right| = o_{\bar{P}_N}(1).$$

Thus the probability limit under $\bar{P}_N$ of $\bar{E}_N^* \left[\phi_{(3)i}(W_i^*; \bar{\theta}_N) \cdot t_2' \beta_i^*(\bar{\theta}_N) \right]$ is equivalent to that of $T_{1N}^{(1)} + T_{1N}^{(0)}$. The latter in turn can be obtained by following similar arguments as in the proof of Lemma 4. Hence, after some algebra I obtain $T_{1N}^{(1)} + T_{1N}^{(0)} = t_2' c + o_{\bar{P}_N}(1)$.

It only remains now to show (A.25). Taking the bootstrap expectation gives

$$\bar{E}_N^* \left[\phi_{(4)i}(W_i^*; \bar{\theta}_N) \cdot t_2' \beta_i^*(\bar{\theta}_N) \right] = V_N^{(1)} + V_N^{(0)},$$

where for $w = 0, 1$,

$$V_N^{(w)} = \frac{1}{N} \sum_{i=1}^{N} f\left(X_i' \bar{\theta}_N \right) t_2' \left\{ X_i - h_N \left(X_i' \bar{\theta}_N \right) \right\} \nu_{(4)i}(w; \bar{\theta}_N) \equiv \frac{1}{N} \sum_{i=1}^{N} \sigma_{N,i}.$$

By law of iterated expectations $\bar{E}_N[\sigma_{N,i}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ and $\bar{E}_N[\sigma_{N,i}\sigma_{N,j}|\mathbf{W},\mathbf{X},\mathbf{U}] = 0$ for all i,j such that $\mathcal{J}_w(i) \neq \mathcal{J}_w(j)$. Consequently by similar arguments as in the proof of Lemma 3, it follows $V_N^{(w)} = o_{\bar{P}_N}(1)$ for $w = 0, 1$. This concludes the proof of the lemma. □

A.3 Additional Lemmas

I use the same notation as in Appendices A.1 and A.2.

Lemma 6. *Suppose that Assumptions 1-3 hold . Then for any $q < \infty$ and δ arbitrarily small, it holds that uniformly in N,*

$$\sup_{\theta\in\mathcal{N}} \bar{E}_\theta \left[|K_M(i;\theta)|^q\right] < \infty,$$

and

$$\sup_{\theta\in\mathcal{N}} \bar{E}_\theta \left[\sup_{1\leq i\leq N} |K_M(i;\theta)|^q\right] = o(N^\delta).$$

Proof. The first claim follows by similar arguments as in Abadie and Imbens (2016, Lemma S.8), after employing Lemma 11 (in particular the second statement) and Lemma 12. The second claim follows by paralleling the arguments of Abadie and Imbens (2006, Additional proofs p.23). □

Let $\mathcal{N}$ denote some neighborhood of θ_0 such that Assumptions 1-5 hold for each $\theta \in \mathcal{N}$. Additionally let $G_{w,\theta}(.)$ denote the CDF of the sample propensity score $F(X'\theta)$ conditional on $W = w$; and $g_{w,\theta}(.)$ the corresponding density function (where it exists). At the same time $G_\theta(.)$ denotes the unconditional CDF of the propensity score $F(X'\theta)$, $Q_\theta(.) \equiv G_\theta^{-1}(.)$ its corresponding quantile function, and $g_\theta(.)$ its density function. The empirical CDF of $F(X'\theta)$ is denoted as

$$\hat{G}_\theta(t) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{F(X_i'\theta) \leq t),$$

and the corresponding empirical quantile function as

$$\hat{Q}_\theta(p) = \inf\{t : \hat{G}_\theta(t) \geq p\}.$$

Note that by construction $\hat{G}_\theta(\hat{Q}_\theta(p)) = p$ for any $p \in (0,1)$. To simplify notation I shall

employ the convention $G_{w,N}(.) \equiv G_{w,\breve{\theta}_N}(.)$, $G_N(.) \equiv G_{\breve{\theta}_N}(.)$, $G_w(.) \equiv G_{w,\theta_0}(.)$ and $G(.) \equiv G_{\theta_0}(.)$ The other terms $g_{w,N}(.), g_w(.), g_N(.), g(.)$ and $\hat{G}_N(.), \hat{Q}_N(.)$ for $w = 0, 1$ are defined analogously. As in Appendix A.2, in what follows I suppress indexing the quantities with the additional label $\breve{\theta}_N$. However it should be implicitly understood that I have replaced $\bar{\hat{\theta}}$ with $\breve{\theta}_N$.

Lemma 7. *Suppose that Assumptions 3 hold. Then for any sequence $\breve{\theta}_N$ such that $\breve{\theta}_N \to \theta_0$,*

$$\sup_{p\in(0,1)} \left|\hat{Q}_N(p) - Q(p)\right| = o_{\bar{P}_N}(1).$$

Proof. I first show that

$$\sup_{t\in[0,1]} \left|\hat{G}_\theta(t) - G(t)\right| = o_{\bar{P}_N}(1). \tag{A.26}$$

Consider the class of functions $\mathcal{G} \equiv \{x'\theta; \theta \in \mathcal{N}\}$ (here x denotes the functional argument). Observe that $\mathcal{G}$ is finite dimensional, being a subset of the space of all linear combinations of $e_1(x), \ldots, e_k(x)$: the (linear) functions corresponding to each axis in the Euclidean $\mathbb{R}^k$ space. By the results of Pollard (2012), this implies that the class of all sets of the form $\{x : x'\theta \leq t\}$ for $\theta \in \mathcal{N}$ and $t \in \mathbb{R}$ is a VC class; equivalently, so is the class of sets $\{x : F(x'\theta) \leq t\}$ a VC class for $\theta \in \mathcal{N}$ and $t \in \mathbb{R}$, since $F(.)$ is strictly monotone. Hence, by the uniform law of large numbers for VC class sets (see Pollard, 2012; also Vapnik and Chervonenkis, 1971), I obtain

$$\sup_{\theta\in\mathcal{N};\ t\in[0,1]} \left|\hat{G}_\theta(t) - G_\theta(t)\right| = o_{\bar{P}_N}(1).$$

By the fact $\breve{\theta}_N \to \theta_0$, together with Assumption 3(i)-(ii),

$$\sup_{t\in[0,1]} |G_N(t) - G(t)| \to 0.$$

Combining the above immediately proves (A.26).

Using (A.26), and recalling that $\hat{G}_N\left(\hat{Q}_N(q)\right) = q$, I have

$$\sup_{q\in(0,1)} \left|q - G\left(\hat{Q}_N(q)\right)\right| = \sup_{q\in(0,1)} \left|\hat{G}_N\left(\hat{Q}_N(q)\right) - G\left(\hat{Q}_N(q)\right)\right| \to 0.$$

Now $Q(.) \equiv G^{-1}(.)$ is uniformly continuous on $(0, 1)$ by virtue of the fact - implied by

Assumption 3(iii) - that $G(.)$ is strictly increasing and continuous on its interval valued support. Hence it follows from the previous display equation that

$$\sup_{q\in(0,1)} \left|Q(q) - \hat{Q}_N(q)\right| = o_{\bar{P}_N}(1),$$

as claimed in the Lemma. □

Lemma 8. *Suppose that Assumptions 3,7 hold. Then for any sequence $\breve{\theta}_N$ such that $\breve{\theta}_N \to \theta_0$,*

$$\max_{1\le l\le q_N} \left|\pi_{l-1}(\breve{\theta}_N) - \pi_l(\breve{\theta}_N)\right| = o_{\bar{P}_N}(1).$$

Proof. Note that $\pi_1(\breve{\theta}_N), \ldots, \pi_{q_N}(\breve{\theta}_N)$ are obtained by evaluating the quantile function $\hat{Q}_N(.)$ at the values $\{1/q_N, 2/q_N, \ldots, q_N - 1/q_N\}$. The claim is thus a straightforward consequence of the previous lemma together with uniform continuity of $Q(.)$ and $q_N \to \infty$. □

Lemma 9. *Suppose that Assumptions 3,6 hold. Then for any sequence $\breve{\theta}_N$ such that $\breve{\theta}_N \to \theta_0$, there exists some universal constant $c > 0$ for which*

$$\bar{P}_N\left(\min_{1\le l\le q_N} N_w(l) \ge c\frac{N}{q_N}\right) \ge 1 - o\left(\frac{q_N^2}{N}\right)$$

for $w = 0, 1$. Furthermore,

$$\max_{1\le j\le N} \left|\frac{N_1(l(j))}{N(l(j))} - F(X_j'\bar{\theta}_N)\right| = o_{\bar{P}_N}(1),$$

and

$$\max_{1\le j\le N} \left|\frac{N(l(j))}{N_1(l(j))} - \frac{1}{F(X_j'\bar{\theta}_N)}\right| = o_{\bar{P}_N}(1).$$

Proof. Assume for simplicity that N/q_N is an integer. Then $N(l) = N/q_N$ for all l. Now,

$$\begin{aligned}
&\bar{P}_N\left(\min_{1\le l\le q_N} N_w(l) \ge c\frac{N}{q_N}\right) = \bar{P}_N\left(N_w(l) \ge c\frac{N}{q_N} \text{ for } l = 1, \ldots q_N\right) \\
&= \prod_{l=1}^{q_N} \bar{P}_N\left(N_w(l) \ge c\frac{N}{q_N}\right) = \prod_{l=1}^{q_N} \bar{P}_N\left(\frac{q_N}{N}\sum_{i\in S_w(l)} W_i \ge c\right) \\
&\ge \left(1 - \underline{\eta}\frac{q_N}{N}\right)^{q_N} = 1 - o\left(\frac{q_N^2}{N}\right),
\end{aligned}$$

where the second equality follows by the iid property of the observations; and the inequality is based on an application of the Markov inequality after noting $\bar{E}_N[W_i] = F(X_i'\bar{\theta}_N)$ with $\min_{1\leq i\leq N} F(X_i'\bar{\theta}_N) \geq \underline{\eta}$ for some $\underline{\eta} > 0$ by Assumption 3(i). This proves the first claim of the lemma.

For each l, let $\dot{p}_{l,N} \equiv \bar{E}_N[q_N N_1(l)/N]$. Since both $\dot{p}_{l(j),N}$ and $F(X_j'\bar{\theta}_N)$ lie within $\left[\pi_{l-1}(\breve{\theta}_N) - \pi_l(\breve{\theta}_N)\right]$ for some l, by Lemma 8 it suffices for the second claim to show that

$$\max_{1\leq l\leq q_N} \left| \frac{q_N N_1(l)}{N} - \dot{p}_{l,N} \right| = o_{\bar{P}_N}(1). \tag{A.27}$$

Fix some $\epsilon > 0$. By the Markov inequality, for each $1 \leq l \leq q_N$,

$$\bar{P}_N \left(\left| \frac{q_N N_1(l)}{N} - \dot{p}_{l,N} \right| > \epsilon \right) \leq \frac{q_N}{N\epsilon}.$$

Hence, by Assumption 6 it follows

$$\bar{P}_N \left(\max_{1\leq l\leq q_N} \left| \frac{q_N N_1(l)}{N} - \dot{p}_{l,N} \right| > \epsilon \right) \leq \frac{q_N^2}{N\epsilon} \to 0.$$

This proves (A.27), which completes the proof of the second claim of the lemma. The third claim follows immediately from the second, since by the previous arguments in this proof the events

$$\min_{1\leq l\leq q_N} \frac{N(l)}{N_1(l)} \geq c > 0; \quad \text{and} \quad \min_{1\leq j\leq N} F(X_j'\bar{\theta}_N) \geq \underline{\eta} > 0$$

occur with probability greater than or equal to $1 - o\left(q_N^2/N\right)$ under $\bar{P}_N$. □

For $w = 0, 1$ let $\Delta_w(l; \bar{\theta}_N) \equiv S_w(l; \breve{\theta}_N) \triangle S_w(l; \bar{\theta}_N)$. Also for any set A, let $\#A$ denote the cardinality of that set.

Lemma 10. *Suppose that Assumptions 3,7 hold. Then for any sequence $\breve{\theta}_N$ such that $\breve{\theta}_N \to \theta_0$, it holds, for $w = 0, 1$ and some $\delta > 0$ arbitrarily small, that*

$$\bar{P}_N \left(\max_{1\leq l\leq q_N} \#\Delta_w(l; \bar{\theta}_N) \geq N^{(1+\delta)/2} \right) \leq q_N \exp(-N^\delta).$$

Proof. Without loss of generality I consider the case when $w = 1$. Define

$$\delta_N = \max_{1 \le i \le N} \left| F(X_i'\bar{\theta}_N) - F(X_i'\breve{\theta}_N) \right|.$$

By Assumption 3(i)-(iii), $\delta_N \le C/\sqrt{N}$ for some $C < \infty$. Also, let $\mathcal{C}_{l,N}$ denote the set

$$\begin{aligned} \mathcal{C}_{l,N} \equiv \Big\{ i : \pi_{l-1}(\breve{\theta}_N) - \delta_N \le F(X_i'\breve{\theta}_N) \le \pi_{l-1}(\breve{\theta}_N) + \delta_N \\ \cup\ \pi_l(\breve{\theta}_N) - \delta_N \le F(X_i'\breve{\theta}_N) \le \pi_l(\breve{\theta}_N) + \delta_N \Big\}. \end{aligned}$$

Clearly $\#\Delta_w(l;\bar{\theta}_N) \le \#\mathcal{C}_{l,N}$. Represent by $\varpi_{i,l,N}$ the random variable $\mathbb{I}\{i \in \mathcal{C}_{l,N}\}$. By the bound on δ_N and the fact $g_{1,N} \le C_2 < \infty$ uniformly in N (in turn due to Assumption 3(iii), see Lemma 11), it follows $\bar{E}_N[\varpi_{i,l,N}] \le C_3/\sqrt{N}$ for some $C_3 < \infty$ independent of l, N. Hence for each l, and some sequence $M_N \asymp N^{\delta}$ independent of l, I obtain

$$\begin{aligned} &\bar{P}_N\left(\#\Delta_w(l;\bar{\theta}_N) \ge N^{(1+\delta)/2}\right) \le \bar{P}_N\left(\#\mathcal{C}_{l,N} \ge N^{(1+\delta)/2}\right) \\ &= \bar{P}_N\left(\frac{1}{N}\sum_{i=1}^{N} \varpi_{i,l,N} \ge \sqrt{N^{\delta-1}}\right) \\ &\le \bar{P}_N\left(\left|\frac{1}{N}\sum_{i=1}^{N} \varpi_{i,l,N} - \bar{E}_N\left[\varpi_{i,l,N}\right]\right| \ge \sqrt{\frac{M_N}{N}}\right) \le \exp(-M_N), \end{aligned}$$

where the final step follows by Hoeffding's inequality. But

$$\bar{P}_N\left(\max_{1 \le l \le q_N} \#\Delta_w(l;\bar{\theta}_N) \ge N^{(1+\delta)/2}\right) \le \sum_{l=1}^{q_N} \bar{P}_N\left(\#\Delta_w(l;\bar{\theta}_N) \ge N^{(1+\delta)/2}\right);$$

hence the claim follows immediately through the above arguments. □

A.4 Uniform statements of the results in Abadie and Imbens (2016)

The lemmas in this appendix are based on Abadie and Imbens (2016), which are extended to apply uniformly over all θ in a neighborhood of θ_0. I thus modify the proofs of Abadie and Imbens (2016) accordingly.

I use the same notation as in Appendices A.1, A.2 and A.3. In addition, I employ the following: Let $p_i(\theta)$ denote $p(X;\theta) \equiv F(X'\theta)$ and $p_{i,N} = p(X_i;\bar{\theta}_N)$. Also let $q_{0,\theta} = E_0[1 - F(X_i'\theta)]$ and $q_{1,\theta} = E_0[F(X_i'\theta)]$ denote the unconditional probabilities that $W_i = 0$ and $W_i = 1$ respectively when the propensity score is $F(X'\theta)$. To simplify notation I shall employ the convention $q_{w,N}(.) \equiv q_{w,\bar{\theta}_N}$ and $q_w \equiv q_{w,\theta_0}$ for $w = 0, 1$. Finally for $w = 0, 1$, let $N_{w,\theta} = \left\{\sum_{i=1}^{N} \mathbb{I}_{W_i=w};\ W_i \sim \text{Bernoulli}(F(X_i'\theta)\right\}$. Per convention, let $N_w \equiv N_{w,\theta_0}$ and $N_{w,N} \equiv N_{w,\bar{\theta}_N}$.

Lemma 11. *Suppose that Assumptions 3 hold. Then: (i) the support of $g_\theta(.)$ and $g(.)$ lies within the interval $[\underline{p}, \bar{p}]$ for some $0 < \underline{p} < \bar{p} < 1$; (ii) there exist universal constants $\underline{c}$ and $\bar{C}$ such that $\underline{c} < \sup_{\theta\in\mathcal{N}} (g_{1,\theta}(p)/g_{0,\theta}(p)) < \bar{C}$ uniformly over all p such that $g_\theta(p) \neq 0$; and (iii) there exist universal constants $1 > \bar{\eta} \geq \underline{\eta} > 0$ such that $q_{w,\theta} \in [\underline{\eta}, \bar{\eta}]$ uniformly in $\theta \in \mathcal{N}$.*

Proof. That the support of $g_\theta(.)$ and $g(.)$ is within some interval $[\underline{p}, \bar{p}]$ follows from the bounded support assumption for X (Assumption 3(i)), and the fact $f(.)$ is strictly positive and bounded (Assumption 3(ii)). Additionally, the support condition on X together with Assumption 3(ii) also ensures existence of universal constants $1 > \bar{\eta} \geq \underline{\eta} > 0$ such that $q_{w,\theta} \in [\underline{\eta}, \bar{\eta}]$ uniformly in $\theta \in \mathcal{N}$. By the Bayes theorem, $g_{0,\theta}(p) = (1-p)g_\theta(p)/q_{0,\theta}$ and $g_{1,\theta}(p) = pg_\theta(p)/q_{1,\theta}$. This proves the existence of $g_{w,\theta}(.)$ for $w = 0, 1$. Given the support condition for $g_\theta(.)$ proved already, the claim $\underline{c} < \sup_{\theta\in\mathcal{N}} (g_{1,\theta}(p)/g_{0,\theta}(p)) < \bar{C}$ follows by similar arguments as in the proof of Abadie and Imbens (2016, Lemma S.2). □

Lemma 12. *Suppose that for $w = 0, 1$, $N_{w,\theta}$ are truncated for values smaller than M and greater than $N - M$ where $N > 2M$. Then for any $q < \infty$ and $w = 0, 1$ there exists $M_q < \infty$ such that,*

$$\sup_{\theta\in\mathcal{N}} E_\theta \left[\left|\frac{N}{N_{w,\theta}}\right|^q\right] \leq M_q.$$

Proof. Observe that $N_{w,\theta}$ is a binomial variable with parameters $(N, q_{w,\theta})$ where $q_{w,\theta} \in [\underline{\eta}, \bar{\eta}]$ uniformly in $\theta \in \mathcal{N}$ by Lemma 11. Hence the claim follows by similar arguments as in the proof of Abadie and Imbens (2016, Lemma S.3). □

Let $\xi_{1:N_w}, \ldots, \xi_{N_w:N_w}$ denote the order statistics for a set of N_w random variables drawn

from the uniform distribution. Denote the interval support of $F(X'\bar{\theta}_N)$ by $[a_N, b_N]$.

Lemma 13. *Suppose that Assumptions 1-4 hold. Then for any sequence $\{\bar{\theta}_N\}$ satisfying $\bar{\theta}_N \to \theta_0$ it holds that under $\bar{P}_N$*

$$\max_{i=1,\ldots,N} \left| G_{w,N}^{-1}(\xi_{i:N_w}) - G_{w,N}^{-1}(i/N_w) \right| = o_p(1). \tag{A.28}$$

Proof. I first prove (A.28). By the fact $\bar{\theta}_N \to \theta_0$ and Assumptions 3(i),(ii), it follows that $G_{w,N}(.)$ is compactly supported for all N sufficiently large. Furthermore, under the same assumptions, it follows

$$\sup_{p \in \mathbb{R}} |G_{w,N}(p) - G_w(p)| \to 0. \tag{A.29}$$

By Assumption 3(iii), $G_{w,N}^{-1}(.)$ exists for N sufficiently large (since $g_N(.)$ and consequently $g_{w,N}(.)$ are strictly positive within an interval support for $F(X'\theta_N)$)[8]. Then

$$\sup_{q \in (0,1)} \left| G_w\left(G_{w,N}^{-1}(q)\right) - q \right| = \sup_{q \in (0,1)} \left| G_w\left(G_{w,N}^{-1}(q)\right) - G_{w,N}\left(G_{w,N}^{-1}(q)\right) \right| \to 0.$$

Now $G_w^{-1}(.)$ is uniformly continuous on $[0,1]$ by virtue of the fact $G_w(.)$ is strictly increasing and, therefore, continuous on a compact set. Hence, it follows from the above that

$$\sup_{q \in (0,1)} \left| G_{w,N}^{-1}(q) - G_w^{-1}(q) \right| \to 0. \tag{A.30}$$

I thus obtain

$$\max_{i=1,\ldots,N} \left| G_{w,N}^{-1}(\xi_{i:N_w}) - G_{w,N}^{-1}(i/N_w) \right| = \max_{i=1,\ldots,N_w} \left| G_w^{-1}(\xi_{i:N_w}) - G_w^{-1}(i/N_w) \right| + o(1) = o_{\bar{P}_N}(1),$$

where the second equality follows by similar arguments as in Abadie and Imbens (2016, Lemma S.4). This proves (A.28). □

Let $S_{N,k}$ denote the probability that observation k (with W_i equal to w say) will be used as a match for an arbitrary observation from the opposite treatment arm under the propensity score $F(X'\bar{\theta}_N)$, conditional on both $\mathbf{W}$ and all the observations from its own

[8] For the end points I set $G_{w,N}^{-1}(0) = a_N$ and $G_{w,N}^{-1}(1) = b_N$.

treatment status, denoted by $\mathbf{X}_w$.

Lemma 14. *Suppose that Assumptions 1-4 hold. Further suppose that for all $\theta \in \mathcal{N}$, the function $l_w(p;\theta) \leq C$ uniformly in both $p \in \mathbb{R}$ and $\theta \in \mathcal{N}$. Then under $\bar{P}_N$,*

$$\frac{1}{N}\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right) K_M(i;\bar{\theta}_N) - \frac{N_{1-w,N}}{N}\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right) S_{N,i} = o_p(1),$$

and

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right) K_M^2(i;\bar{\theta}_N)\\ &- \frac{1}{N}\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right)\left(N_{1-w,N}^2 S_{N,i}^2 + N_{1-w,N} S_{N,i}(1-S_{N,i})\right) = o_p(1).\end{aligned}$$

Proof. The proof of this result is a straightforward extension of Abadie and Imbens (2016, Lemma S.10) and therefore omitted. □

For the next Lemma, let $p_{w,j:N}$ denote j-th order statistic of $\{p_{i,N} : W_{N,i} = w\}$. I set $p_{w,j:N} = a_N$ if $j < 1$ and $p_{w,j:N} = b_N$ if $j > N$, where $[a_N, b_N]$ denotes the interval support of $F(X'\bar{\theta}_N)$. Also let V_i denote the rank of observation i, in terms of $F(X_i'\bar{\theta}_N)$, within the sample of observations that have the same treatment status as itself.

Additionally, define $\chi_{0,\theta}(p) = \frac{p}{1-p}\frac{q_{0,N}}{q_{1,N}}$ for $p \in [a_\theta, b_\theta]$ and $\chi_{1,\theta}(.) = \chi_{0,\theta}^{-1}(.)$, where $[a_\theta, b_\theta]$ denotes the interval support of $F(X'\theta)$. I also set $\chi_{0,N} \equiv \chi_{0,\bar{\theta}_N}$ and $\chi_{1,N} \equiv \chi_{1,\bar{\theta}_N}$. Note that $\chi_{w,N}(p) = (g_{1-w,N}/g_{w,N})(p)$ except on the set $\{p \in [a_N, b_N] : g_N(p) = 0\}$, which has Lebesgue measure zero by Assumption 3(iii).

Lemma 15. *Suppose that Assumptions 1-4 hold and that $\bar{\theta}_N \to \theta_0$. Further suppose that for all $\theta \in \mathcal{N}$, the function $l_w(p;\theta)$ is uniformly bounded in both $p \in [a_\theta, b_\theta]$, and $\theta \in \mathcal{N}$. Then for each $w = 0, 1$, under $\bar{P}_N$, (i)*

$$\begin{aligned}&\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right)\\ &\quad\times\left(S_{N,i} - \chi_{w,N}(p_{i,N})\frac{G_{w,N}(p_{w,V_i+M:N_{w,N}}) - G_{w,N}(p_{w,V_i-M:N_{w,N}})}{2}\right) = o_p(1).\end{aligned}$$

and (ii)

$$\sum_{i=1}^{N} l_w\left(p_{i,N};\bar{\theta}_N\right) N_{w,N} \times \left(S_{N,i}^2 - \left(\chi_{w,N}(p_{i,N})\frac{G_{w,N}(p_{w,V_i+M:N_{w,N}}) - G_{w,N}(p_{w,V_i-M:N_{w,N}})}{2}\right)^2\right) = o_p(1).$$

Proof. In terms of the method for the proof, I adapt the arguments of Abadie and Imbens (2016, Lemma S.7) to allow for triangular arrays. Without loss of generality I prove the above for the case $w = 0$. Also to simplify notation, I set $N_{0,N} = N_0$ for the duration of this proof.

I first show that for any fixed $K < \infty$,

$$\max_{1\le i\le N_0} |p_{0,V_i+K:N_0} - p_{0,V_i:N_0}| = o_{\bar{P}_N}(1). \tag{A.31}$$

By equation (A.30) in Lemma 13, and the fact $G_0^{-1}(.)$ is uniformly continuous on $[0,1]$, it follows that the sequence $G_{0,N}^{-1}(.)$ is uniformly equicontinuous. Hence for each $\epsilon > 0$, there exists $\delta > 0$ such that

$$\begin{aligned}
&\bar{P}_N\left(\max_{1\le i\le N_0} |p_{0,V_i+K:N_0} - p_{0,V_i:N_0}| > \epsilon\right) \\
&\le \bar{P}_N\left(\max_{1\le i\le N_0} |G_{0,N}\left(p_{0,V_i+K:N_0}\right) - G_{0,N}\left(p_{0,V_i:N_0}\right)| > \delta\right) \\
&\le \Pr\left(\max_{1\le i\le N_0} |\xi_{V_i+K:N_0} - \xi_{V_i:N_0}| > \delta\right) \to 0,
\end{aligned}$$

where the limit follows by standard properties of uniform spacings. This proves (A.31).

Define

$$\Omega_{Ni} \equiv \left[\frac{p_{0,V_i:N_0} + p_{0,V_i+M:N_0}}{2}, \frac{p_{0,V_i:N_0} + p_{0,V_i-M:N_0}}{2}\right].$$

Let

$$Z_{N,i} = l_0\left(p_{i,N};\bar{\theta}_N\right) N_0\left(S_{N,i} - h_{0,N}(p_{i,N})\frac{G_{0,N}\left(p_{0,V_i+M:N_0}\right) - G_{0,N}\left(p_{0,V_i-M:N_0}\right)}{2}\right).$$

As in the proof of Abadie and Imbens (2016, Lemma S.7), note that

$$\begin{aligned} S_{N,i} = &\int_{a_N}^{(p_{0,V_i:N_0}+p_{0,V_i+M:N_0})/2} g_{1,N}(p)dp\ \mathbb{I}_{\{V_i\leq M\}} \\ &+ \int_{\Omega_{Ni}} g_{1,N}(p)dp\ \mathbb{I}_{\{M<V_i\leq N-M\}} \\ &+ \int_{(p_{0,V_i:N_0}+p_{0,V_i-M:N_0})/2}^{b_N} g_{1,N}(p)dp\ \mathbb{I}_{\{V_i>N-M\}}. \end{aligned} \tag{A.32}$$

Then by the properties of uniform spacings I obtain

$$N_0 S_{N,i} - N_0 \int_{\Omega_{Ni}} g_{1,N}(p)dp = o_{\bar{P}_N}(1). \tag{A.33}$$

Now by the proof of Lemma 11, for each $p \in [a_\theta, b_\theta]$,

$$\left(\frac{g_{1,N}}{g_{0,N}}\right)(p) = \frac{p}{1-p}\frac{q_{0,N}}{q_{1,N}}\mathbb{I}\left\{g_N(p)\neq 0\right\} \equiv \chi_{0,N}(p)\mathbb{I}\left\{g_N(p)\neq 0\right\},$$

where $\{\chi_{0,N}\}$ is uniformly equicontinuous on $p \in [a_N, b_N]$ by Lemma 11. Since $g_{1,N}(p) = g_{0,N}(p) = 0$ whenever $g_N(p) = 0$, the mean value theorem for Lebesgue-Steltjes integrals ensures

$$\begin{aligned} &\int_{\Omega_{Ni}} g_{1,N}(p)dp = \int_{\Omega_{Ni}} \chi_{0,N}(p)g_{0,N}(p)dp \ = \int_{\Omega_{Ni}} \chi_{0,N}(p)dG_{0,N}(p) \\ &= \chi_{0,N}\left(\bar{p}_{i,N,M}\right)\left(G_{0,N}\left(\frac{p_{0,V_i:N_0}+p_{0,V_i+M:N_0}}{2}\right) - G_{0,N}\left(\frac{p_{0,V_i:N_0}+p_{0,V_i-M:N_0}}{2}\right)\right). \end{aligned}$$

for some $\bar{p}_{i,N,M} \in \Omega_{Nl}$. Substituting in (A.33), a second application of the mean value theorem then implies

$$N_0 S_{N,i} - N_0 \chi_{0,N}\left(\bar{p}_{i,N,M}\right) g_{0,N}\left(\tilde{p}_{i,N,M}\right)\left(p_{0,V_i+M:N_0} - p_{0,V_i-M:N_0}\right)/2 = o_{\bar{P}_N}(1),$$

for some $\tilde{p}_{i,N,M} \in \Omega_{Ni}$. Substituting the above in the expression for Z_{Ni}, and applying the mean value theorem again on $G_{0,N}\left(p_{0,V_i+M:N_0}\right) - G_{0,N}\left(p_{0,V_i-M:N_0}\right)$, I obtain for some

$\check{p}_{i,N,M} \in [p_{0,V_i+M:N_0}, p_{0,V_i-M:N_0}]$,

$$Z_{N,i} = o_{\bar{P}_N}(1) + l_0\left(p_{i,N}; \bar{\theta}_N\right) N_0 \left\{\chi_{0,N}\left(\bar{p}_{i,N,M}\right) g_{0,N}\left(\tilde{p}_{i,N,M}\right) - \chi_{0,N}\left(p_{i,N}\right) g_{0,N}\left(\check{p}_{i,N,M}\right)\right\}$$
$$\times \left(p_{0,V_i+M:N_0} - p_{0,V_i-M:N_0}\right)/2.$$

Now using (A.31) together with the facts $\{\chi_{0,N}\}$ and $\{g_{0,N}\}$ are uniformly equicontinuous (the latter by Assumption 3-(iii)), it follows $Z_{N,i} = o_p(1)$ under $\bar{P}_N$ for each i.

I now show that for any $r < \infty$, there exists some constant $M_r < \infty$ such that,

$$\bar{E}_N \left|Z_{N,i}\right|^r < M_r \quad \text{for all } 1 \leq i \leq N, \tag{A.34}$$

where the expectation here, and in the rest of the proof, is taken under $\bar{P}_N$. By standard properties of uniform spacings,

$$\begin{aligned}
&\bar{E}_N \left|N_{0,N}\left\{G_{0,N}\left(p_{0,V_i+M:N_0}\right) - G_{0,N}\left(p_{0,V_i-M:N_0}\right)\right\}\right|^r \\
&= \bar{E}_N \left|N_{0,N}\left(\xi_{V_i+M:N_0} - \xi_{V_i-M:N_0}\right)\right|^r \quad < M_{1r}
\end{aligned}$$

for some constant $M_{1r} < \infty$. Hence, by part (ii) of Lemma 11, and the assumption $l_0(p;\theta)$ is uniformly bounded, it suffices for (A.34) to show $\bar{E}_N \left|N_0 S_{N,i}\right|^r$ is uniformly bounded. Let $S_{Ni,(a)}$, $S_{Ni,(b)}$ and $S_{Ni,(c)}$ denote the three terms in that order from the expression for $S_{N,i}$ in equation (A.32). By part (ii) of Lemma 11, and the properties of uniform spacings (see Abadie and Imbens, Lemma S.5; or as applied in their Lemma S.7),

$$\begin{aligned}
\bar{E}_N \left|N_0 S_{Ni,(a)}\right|^r &\leq \bar{C}^r \bar{E}_N \left|N_0 G_{0,N}\left(p_{0,2M:N_0}\right)\right|^r \\
&\leq \bar{C}^r \bar{E}_N \left|N_0 \xi_{2M:N_0}\right|^r < M_{r,(a)}
\end{aligned}$$

for some $M_{r,(a)} < \infty$. A similar argument also shows that $\bar{E}_N \left|N_0 S_{Ni,(c)}\right|^r < M_{r,(c)} < \infty$.

Finally, consider

$$\begin{aligned}\left|S_{Ni,(b)}\right|^r &= \left|\int_{\Omega_{Ni}} \chi_{0,N}(p) g_{0,N}(p) dp\right|^r \leq \bar{C}^r \left|\int_{V_i - M:N_0}^{V_i + M:N_0} g_{0,N}(p) dp\right|^r \\ &= \bar{C}^r \left|\xi_{V_i+M:N_0} - \xi_{V_i-M:N_0}\right|^r,\end{aligned}$$

where the inequality follows from $\sup_p |\chi_{0,N}(p)| < \bar{C}$ due to Lemma 11. Hence by the properties of uniform spacings, $\bar{E}_N \left|N_0 S_{Ni,(b)}\right|^r < M_{r,(b)} < \infty$. By the above I have thus shown (A.34).

Equation (A.34), together with $Z_{N,i} = o_p(1)$ under $\bar{P}_N$, implies $\bar{E}_N |Z_{Ni}| \to 0$. Since the choice of i was arbitrary, the above holds true for all $1 \leq i \leq N_0$. Hence application of the Markov inequality assures $N^{-1} \sum_{i=1}^N Z_{N,i} = o_p(1)$ under $\bar{P}_N$. This proves the first part of the Lemma. Part (ii) follows by analogous arguments. □

Lemma 16. *Suppose that Assumptions 1-4 hold. Further suppose that for all $\theta \in \mathcal{N}$, the function $m_w(.;\theta) : [\underline{p}, \bar{p}] \to \mathbb{R}$ is non-negative, uniformly equicontinuous in $\mathcal{N}$ i.e*

$$\lim_{\delta \to 0} \sup_{p \in \mathbb{R}, \theta \in \mathcal{N}} |m_w(p;\theta) - m_w(p+\delta;\theta)| = 0,$$

and also satisfies $m_w(p;\dot{\theta}_N) \to m_w(p;\theta_0)$ point-wise in each p for any sequence $\dot{\theta}_N \to \theta_0$. Then for any non-negative integer M, and sequence $\{\bar{\theta}_N\}$ satisfying $\bar{\theta}_N \to \theta_0$ a.s-$\bar{P}_N$, it holds that under $\bar{P}_N$,

$$\begin{aligned}&\sum_{i=1}^N m_w\left(G_{w,N}^{-1}(\xi_{i:N_{w,N}}); \bar{\theta}_N\right)\left(\xi_{i+M:N_{w,N}} - \xi_{i-M:N_{w,N}}\right) \\ &= \frac{2M}{N_{w,N}} \sum_{i=1}^N m_w\left(G_{w,N}^{-1}(\xi_{i:N_{w,N}}); \bar{\theta}_N\right) + o_p(1),\end{aligned}$$

and

$$\begin{aligned}&N_{w,N} \sum_{i=1}^N m_w\left(G_{w,N}^{-1}(\xi_{i:N_{w,N}}); \bar{\theta}_N\right)\left(\xi_{i+M:N_{w,N}} - \xi_{i-M:N_{w,N}}\right)^2 \\ &= \frac{2M(2M+1)}{N_{w,N}} \sum_{i=1}^N m_w\left(G_{w,N}^{-1}(\xi_{i:N_{w,N}}); \bar{\theta}_N\right) + o_p(1).\end{aligned}$$

Proof. Using Lemma 6 and Lemma 13, the proof of this result is a straightforward extension of Abadie and Imbens (2016, Lemma S.6), and therefore omitted. □

Appendix B

Supplementary material and proofs for Chapter 2

Section B.1 presents the assumptions and some definitions for the statement of Theorems 4 & 5, and Proposition 1. Sections B.2 and B.3 present proofs for Theorems 4 and 5 from Chapter 2, respectively. Section B.4 reports additional numerical results for Section 2.4.1. Section B.5 provides additional simulation results to illustrate the empirical likelihood test proposed in Section 2.3.5.

B.1 Assumptions and some definitions

Let $\mathbb{G}_n f(\cdot) = n^{-1/2} \sum_{i=1}^n (f(X_i) - E[f(X)])$ be the empirical process. Hereafter "w.p.a.1" means "with probability approaching one". For Theorem 4 and Proposition 1, we impose the following assumptions.

Assumption M.

(i) $\{X_i\}_{i=1}^n$ is an i.i.d. sequence of compact and convex SVRSs. The class $\{s(X,p) : p \in \mathbb{S}^d\}$ is a μ-Donsker class with envelope F such that $E[|F|^\xi] < \infty$ for some $\xi > 2$. Also, $\inf_{p\in\mathbb{S}^d} \mathrm{Var}(s(X,p)) > 0$.

(ii) $\hat{\nu} \xrightarrow{p} \nu$, $\|\Theta_0(\hat{\nu})\|_H = O_p(1)$, and there exists a function $G(p;\nu)$ continuous in $p \in \mathbb{S}^d$ satisfying (2.5).

(iii) For every finite collection of points $\{p_1, \ldots, p_J\} \in \mathbb{S}^d$, the vector $(\mathbb{G}_n s(\cdot, p_1), \ldots, \mathbb{G}_n s(\cdot, p_J), \sqrt{n}(\hat{\nu} - \nu))$ converges in distribution to a Gaussian random vector.

Assumption M'. For the bootstrap probability P^* conditional on the data, it holds w.p.a.1,

$$\sup_{p \in \mathbb{S}^d} |s(\Theta_0(\hat{\nu}^*), p) - s(\Theta_0(\hat{\nu}), p) - G(p; \nu)'(\hat{\nu}^* - \hat{\nu})| = o_{p^*}(n^{-1/2}).$$

For Theorem 5, we restrict attention to the situation where $\nu = f(E[z])$ is a smooth function of means of $z \in \mathbb{R}^{\dim(z)}$. A consistent estimator of ν is given by $\hat{\nu} = f(\bar{z})$. We introduce the following notation: Let $m_k(X_i)$, $\tilde{m}_k(X_i)$, $\dot{m}_k(X_i)$, and $\hat{m}_k(X_i)$ be k-dimensional vectors whose j-th elements are given by

$$\begin{aligned}
m_{k,j}(X_i) &= s(X_i, p_j) - s(\Theta_0(\hat{\nu}), p_j), \quad \tilde{m}_{k,j}(X_i) = s(X_i, p_j) - s(\Theta_0(\nu), p_j), \\
\dot{m}_{k,j}(X_i) &= s(X_i, p_j) - s(\Theta_0(\nu), p_j) - G(p_j; \nu)' \nabla f(E[z])'(z_i - E[z]), \\
\hat{m}_{k,j}(X_i) &= s(X_i, p_j) - s(\Theta_0(\hat{\nu}), p_j) - G(p_j; \hat{\nu})' \nabla f(\bar{z})'(z_i - \bar{z}),
\end{aligned}$$

respectively. Define $\hat{V}_k = n^{-1} \sum_{i=1}^n m_k(X_i) m_k(X_i)'$, $V_k = \mathrm{Var}(\tilde{m}_k(X_i))$, $\dot{V}_k = \mathrm{Var}(\dot{m}_k(X_i))$, $\bar{V}_k = n^{-1} \sum_{i=1}^n \hat{m}_k(X_i) \hat{m}_k(X_i)'$, $\dot{\phi}_k = \lambda_{\min}(\dot{V}_k)$, and $\bar{\phi}_k = \lambda_{\min}(\bar{V}_k)$. The test statistic L_n in (2.9) is defined as the maximum over a shrinking neighborhood $\Lambda_n = \{\gamma \in \mathbb{R}^k : \|\gamma\| \leq C \bar{\phi}_k^{-3/2} \sqrt{k/n}\}$ for some positive constant C. In particular, C is chosen to satisfy $C > \max\{2C' \bar{\phi}_k^{-1/2}, 1\}$ where C' is the positive constant obtained from $\|\bar{m}\| \leq C' \sqrt{k/n}$ w.p.a.1. The condition on C ensures that the local maximum $\hat{\gamma}$ lies in the interior of Λ_n w.p.a.1 even in the case when $\dot{\phi}_k^{-1}$ is bounded. If $\dot{\phi}_k^{-1}$ diverges to infinity, this additional condition on C may be dispensed with. Note that the optimization in (2.9) is well defined only in the region $S_n = \{\gamma \in \mathbb{R}^k : \gamma' m_k(X_i) > -1 \text{ for all } i = 1, \ldots, n\}$. However, since our assumptions guarantee $\max_{1 \leq i \leq n} \sup_{\gamma \in \Lambda_n} |\gamma' m_k(X_i)| = o_p(1)$, it holds that $\Lambda_n \subseteq S_n$ w.p.a.1. For Theorem 5, we impose the following assumptions.

Assumption S.

(i) Assumption M holds with the envelope function F in Assumption M (i) satisfying $E[|F|^\xi] < \infty$ for some $\xi \geq 4$.

(ii) $\nabla f(\cdot)$ is Hölder continuous of order $\alpha \geq 2/3$ in a neighborhood of $E[z]$. Furthermore, $E[\|z\|^4] < \infty$.

(iii) For some neighborhood $\mathcal{N}$ of ν, there exists a function $G(\cdot;.) : \mathbb{S}^d \times \mathcal{N} \to \mathbb{R}^{\dim(\nu)}$ such that $\sup_{p\in\mathbb{S}^d} \|G(p;\nu_m) - G(p;\nu)\| \to 0$ for all $\nu_m \to \nu$, where $G(p;\nu)$ is defined in Assumption M (ii). Furthermore, for all $\tilde{\nu} \in \mathcal{N}$, $\sup_{p\in\mathbb{S}^d} \|G(p;\tilde{\nu}) - G(p;\nu)\| \leq M \|\tilde{\nu} - \nu\|^\alpha$ for some $\alpha \geq 2/3$ and $M < \infty$ independent of $\tilde{\nu}$.

(iv) $k \to \infty$ and $(k^5 \dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$ as $n \to \infty$.

B.2 Proof of Theorem 4

We first derive the limiting distribution of K_n under H_0. By Assumption M (ii),

$$n^{-1/2} \sum_{i=1}^{n} \{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\} = \mathbb{G}_n s(\cdot, p) - G(p;\nu)'(\hat{\nu} - \nu) + o_p(n^{-1/2}),$$

uniformly over $p \in \mathbb{S}^d$. Assumptions M (i) and (iii) guarantee weak convergence of the process $\{\mathbb{G}_n s(\cdot, p), \sqrt{n}(\hat{\nu} - \nu) : p \in \mathbb{S}^d\}$ to $\{Z(p), Z_1 : p \in \mathbb{S}^d\}$. Thus, by continuity of $G(p;\nu)$ (Assumption M(ii)), it follows that $n^{-1/2} \sum_{i=1}^n \{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}$ converges weakly to $Z(p) - G(p;\nu)' Z_1$. Using Assumptions M (i) and (ii) and standard arguments, $\sup_{p\in\mathbb{S}^d} |n^{-1} \sum_{i=1}^n \{s(X_i, p) - s(\Theta_0(\hat{\nu}), p)\}^2 - \mathrm{Var}(s(X, p))| \xrightarrow{p} 0$. From the envelope condition in Assumption M (i) and a Borel-Cantelli lemma argument as in Owen (1988), it holds $\max_{1\leq i\leq n} \sup_{p\in\mathbb{S}^d} |s(X_i, p)| = o(n^{1/2})$ almost surely. This, along with $\|\Theta_0(\hat{\nu})\|_H = O_p(1)$ (Assumption M (ii)), implies $\max_{1\leq i\leq n} \sup_{p\in\mathbb{S}^d} |s(X_i, p) - s(\Theta_0(\hat{\nu}), p)| = o_p(n^{1/2})$. Combining these results, the null distribution of K_n follows by a similar argument as in the proof of Hjort, McKeague and van Keilegom (2009, Theorem 2.1).

We now prove the second assertion, $K_n \to \infty$ under H_1. Let $g_i(p, t) = s(X_i, p) - s(\Theta_0(t), p)$ for $t = \nu$ or $\hat{\nu}$. Under H_1, there exists $p^* \in \mathbb{S}^d$ such that $E[g_i(p^*, \nu)] \neq 0$. We prove the case of $E[g_i(p^*, \nu)] > 0$ only; the case of $E[g_i(p^*, \nu)] < 0$ can be shown in the

same manner. Pick any $\delta \in (0, 1/2)$. Observe that

$$\begin{aligned} -\log \ell_n(p^*) &= \sup_{\lambda \in \mathbb{R}} \sum_{i=1}^n \log(1 + \lambda g_i(p^*, \hat{\nu})) \geq \sum_{i=1}^n \log(1 + n^{-(1/2+\delta)} g_i(p^*, \hat{\nu})) \\ &= n^{1/2-\delta} \left\{ \frac{1}{n} \sum_{i=1}^n g_i(p^*, \hat{\nu}) \right\} + n^{-2\delta} \left\{ \frac{1}{2n} \sum_{i=1}^n g_i(p^*, \hat{\nu})^2 \right\} + O_p(n^{-2\delta}), \end{aligned}$$

where the first equality follows from the convex duality and the second equality follows from a Taylor expansion. Since the first term diverges to infinity and the other terms are negligible under Assumptions M (i)-(iii), the conclusion is obtained.

B.3 Proof of Theorem 5

We first derive the limiting distribution of $(L_n - k)/\sqrt{2k}$ under H_0. Define $\dot{g}_i(p) = s(X_i, p) - s(\Theta_0(\nu)) - G(p;\nu)'\nabla f(E[z])'(z_i - E[z])$, $\bar{m}_k = n^{-1}\sum_{i=1}^n m_k(X_i)$, and $\bar{\dot{m}}_k = n^{-1}\sum_{i=1}^n \dot{m}_k(X_i)$. Note that by the mean value theorem (applicable here by Assumption S (iii)), for each $p \in \mathbb{S}^d$ there exists some $\tilde{\nu}_p$ satisfying $\|\tilde{\nu}_p - \nu\| \leq \|\hat{\nu} - \nu\|$ and $s(\Theta_0(\hat{\nu}), p) - s(\Theta_0(\nu), p) = G(p; \tilde{\nu}_p)'(\hat{\nu} - \nu)$. Thus by Assumption S (ii) and the asymptotic expansion $\hat{\nu} - \nu = \nabla f(E[z])' n^{-1} \sum_{i=1}^n (z_i - E[z]) + O_p(n^{-(1+\alpha)/2})$, we have

$$\begin{aligned} \left\|\bar{m}_k - \bar{\dot{m}}_k\right\| &\leq \sqrt{k} \sup_{p \in \mathbb{S}^d} \left\| s(\Theta_0(\hat{\nu}), p) - s(\Theta_0(\nu), p) - G(p;\nu)'(\hat{\nu} - \nu) \right\| + O_p(\sqrt{k/n^{1+\alpha}}) \\ &\leq \sqrt{k}\, \|\hat{\nu} - \nu\| \sup_{p \in \mathbb{S}^d} \|G(p; \tilde{\nu}_p) - G(p;\nu)\| + O_p(\sqrt{k/n^{1+\alpha}}) = O_p(\sqrt{k/n^{1+\alpha}}) \end{aligned} \quad \text{(B.1)}$$

Also note that

$$\bar{\dot{m}}_k = O_p(\sqrt{k/n}), \qquad \bar{m}_k = O_p(\sqrt{k/n}), \quad \text{(B.2)}$$

where the first statement follows from the fact that the process $\{\dot{g}_i(p); p \in \mathbb{S}^d\}$ is μ-Donsker by Assumption S (i), and the second statement follows by (B.1). Next, observe that

$$\left\|\hat{V}_k - V_k\right\| \leq k \sup_{p,q \in \mathbb{S}^d} \left| \frac{1}{n} \sum_{i=1}^n \{\dot{g}_i(p)\dot{g}_i(q) - E[\dot{g}_i(p)\dot{g}_i(q)]\} \right| + O_p\left(\sqrt{\frac{k}{n}}\right) = O_p(k/\sqrt{n}) \quad \text{(B.3)}$$

where the inequality follows from $\sup_{p \in \mathbb{S}^d} |s(\Theta(\hat{\nu}), p) - s(\Theta(\nu), p)| = O_p(n^{-1/2})$ and the equality follows from the fact that the process $\{\dot{g}_i(p)\dot{g}_i(q); p, q \in \mathbb{S}^d\}$ is μ-Donsker. Further-

more, using Assumptions S (i) and (ii) combined with $\|\bar{z} - E[z]\| = O_p(n^{-1/2})$, straightforward algebra ensures that

$$\|\dot{m}_k(X_i) - \hat{m}_k(X_i)\| = O_p(\sqrt{k/n^\alpha})\,\|z_i - E[z]\| + O_p(\sqrt{k/n}).$$

We can now see that $\bar{V}_k - n^{-1}\sum_{i=1}^n \dot{m}_k(X_i)\dot{m}_k(X_i)'$ is bounded by $2n^{-1}\sum_{i=1}^n\{k^{1/2}\dot{g}_i\delta_i + \delta_i^2\}$, where $\delta_i = \|\dot{m}_k(X_i) - \hat{m}_k(X_i)\|$. Substituting the expression for the latter from the previous equation and noting that our assumptions guarantee $E[\dot{g}_i^2] < \infty$, we obtain $\left\|\bar{V}_k - n^{-1}\sum_{i=1}^n \dot{m}_k(X_i)\dot{m}_k(X_i)'\right\| = O_p(\sqrt{k^2/n^\alpha})$ using the law of large numbers. Moreover, $\left\|n^{-1}\sum_{i=1}^n \dot{m}_k(X_i)\dot{m}_k(X_i)' - \dot{V}_k\right\| = O_p(k/\sqrt{n})$ by analogous weak convergence arguments as used to show (B.3). Combining these results proves

$$\left\|\bar{V}_k - \dot{V}_k\right\| = O_p(\sqrt{k^2/n^\alpha}). \tag{B.4}$$

We also make frequent use of the following fact implied by (B.4) and the rate condition $(k^5\dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$:

$$|\bar{\phi}_k^c - \dot{\phi}_k^c| = o_p(\dot{\phi}_k^c) \quad \text{for each } c \in \mathbb{R}. \tag{B.5}$$

For the conclusion of this theorem, it is sufficient to show the followings:

$$\frac{L_n(\hat{\nu}) - n\bar{m}_k'\bar{V}_k^{-1}\bar{m}_k}{\sqrt{2k}} \xrightarrow{p} 0, \tag{B.6}$$

$$\frac{n\bar{m}_k'\bar{V}_k^{-1}\bar{m}_k - k}{\sqrt{2k}} \xrightarrow{d} N(0,1). \tag{B.7}$$

We first show (B.6). Let $\hat{\gamma} \in \arg\max_{\gamma\in\Lambda_n} G_n(\gamma)$ and $D_n = \max_{1\le i\le n}\|m_k(X_i)\|$. Also define $G_n^*(\gamma) = n(2\gamma'\bar{m}_k - \gamma'\bar{V}_k\gamma)$, which is maximized at $\gamma^* = \bar{V}_k^{-1}\bar{m}_k$. For (B.6), it is sufficient to show that $\hat{\gamma}, \gamma^* = O_p(\dot{\phi_k}^{-1}\sqrt{k/n})$, and $\sup_{\gamma\in\Omega_n\subseteq\Lambda_n} k^{-1/2}|G_n(\gamma) - G_n^*(\gamma)| \xrightarrow{p} 0$ where $\Omega_n = \{\gamma \in \mathbb{R}^k : \|\gamma\| \le c\dot{\phi_k}^{-1}\sqrt{k/n}\}$ with $c > 0$ chosen to ensure Ω_n contains both $\hat{\gamma}$ and γ^* w.p.a.1 and $\Omega_n \subseteq \Lambda_n$ (such a c exists by the definition of Λ_n). Indeed, these are shown by an argument similar to the proof of Hjort, McKeague and van Keilegom (2009,

Proposition 4.1) if the following requirements are satisfied under $(k^5\dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$:

$$(n^{-1/2}k^{3/2}\dot{\phi}_k^{-3})D_n = o_p(1), \tag{B.8}$$

$$\|\gamma^*\| = O_p(\dot{\phi}_k^{-1}\sqrt{k/n}), \tag{B.9}$$

$$\lambda_{\max}(\hat{V}_k) = O_p(k), \tag{B.10}$$

$$\hat{\gamma} \text{ exists w.p.a.1 and } \|\hat{\gamma}\| = O_p(\dot{\phi}_k^{-1}\sqrt{k/n}). \tag{B.11}$$

We first show (B.8). Using the envelope condition in Assumption S (i) which implies $\sup_{k\in\mathbb{N}} E[\|k^{-1/2}\tilde{m}_k(X_i)\|^\xi] < \infty$, an argument similar to the proof of Hjort, McKeague and van Keilegom (2009, Lemma 4.1) guarantees $(n^{-1/2}k^{3/2}\dot{\phi}_k^{-3})\max_{1\le i\le n}\|\tilde{m}_k(X_i)\| = o_p(1)$ under the rate condition $(k^5\dot{\phi}_k^{-6})^{\frac{\xi}{\xi-2}}/n \to 0$. Furthermore, $\max_{1\le i\le n}\|\tilde{m}_k(X_i) - m_k(X_i)\| \le \sup_{p\in\mathbb{S}^d}|s(\Theta(\hat{\nu}),p) - s(\Theta(\nu),p)| = O_p(n^{-1/2})$, and (B.8) follows. Next, (B.9) follows from (B.2) and (B.5). To show (B.10), observe that $\left\|\hat{V}_k - n^{-1}\sum_{i=1}^n \tilde{m}_k(X_i)\tilde{m}_k(X_i)'\right\| = O_p(k/\sqrt{n})$ by Assumption S (ii) and
$\|n^{-1}\sum_{i=1}^n \tilde{m}_k(X_i)\tilde{m}_k(X_i)'\| = O_p(k)$ by $E[\|X_i\|_H^2] < \infty$. Hence, using $\lambda_{\max}(\hat{V}_k) \le \left\|\hat{V}_k\right\|$ and the triangle inequality, (B.10) is verified. Finally, for (B.11), we first note that $\hat{\gamma}$ exists w.p.a.1 since $\Lambda_n \subseteq S_n$ w.p.a.1 and Λ_n is a compact set. Thus, letting $b_n = \max_{1\le i\le n}\sup_{\gamma\in\Lambda_n}\{1-(1+\gamma' m_k(X_i))^{-2}\}$, an expansion around $\gamma = 0$ yields $0 \le G_n(\hat{\gamma}) \le n\{2\hat{\gamma}'\bar{m}_k - \hat{\gamma}'(\bar{V}_k - b_n\hat{V}_k)\hat{\gamma}\}$. Note that

$$b_n = O_p\left(\max_{1\le i\le n}\sup_{\gamma\in\Lambda_n}|\gamma' m_k(X_i)|\right) = O_p\left(D_n \sup_{\gamma\in\Lambda_n}\|\gamma\|\right) = o_p(\dot{\phi}_k^{3/2}k^{-1}),$$

where the last equality follows from (B.5), (B.8) and the definition of Λ_n. Consequently, $\lambda_{\min}(\bar{V}_k - b_n\hat{V}_k) \ge \bar{\phi}_k - |b_n|\lambda_{\max}(\hat{V}_k) = \dot{\phi}_k(1+o_p(1))$, where the equality also uses (B.10) and (B.5). Thus $\hat{\gamma}'(\bar{V}_k + b_n\hat{V}_k)\hat{\gamma} \ge \|\hat{\gamma}\|^2\dot{\phi}_k(1+o_p(1))$, which implies $\|\hat{\gamma}\| \le 2\dot{\phi}_k^{-1}\|\bar{m}_k\|(1+o_p(1))$. Therefore, by (B.2) it must be the case that $\hat{\gamma}$ is an interior solution w.p.a.1. (by the choice of C in the definition of Λ_n) and that $\|\hat{\gamma}\| = O_p(\dot{\phi}_k^{-1}\sqrt{k/n})$. This proves (B.11). Combining these results, the claim in (B.6) follows.

We now show (B.7). We can decompose

$$\begin{aligned}\frac{n\bar{m}_k'\bar{V}_k^{-1}\bar{m}_k - k}{\sqrt{2k}} &= \frac{n\bar{m}_k'(\bar{V}_k^{-1} - \dot{V}_k^{-1})\bar{m}_k}{\sqrt{2k}} + \frac{n(\bar{m}_k - \bar{\dot{m}}_k)'\dot{V}_k^{-1}\bar{m}_k}{\sqrt{2k}} \\ &\quad + \frac{n\bar{\dot{m}}_k'\dot{V}_k^{-1}(\bar{m}_k - \bar{\dot{m}}_k)}{\sqrt{2k}} + \frac{n\bar{\dot{m}}_k'\dot{V}_k^{-1}\bar{\dot{m}}_k - k}{\sqrt{2k}}. \end{aligned} \tag{B.12}$$

By de Jong and Bierens (1994, Lemma 4a), the first term of (B.12) is bounded by $nk^{-1/2}\|\bar{m}_k\|^2\bar{\phi}_k^{-1}\dot{\phi}_k^{-1}\left\|\bar{V}_k - \dot{V}_k\right\|$ and is thus negligible using (B.2),(B.4) and (B.5). Next, by (B.1),(B.2) and (B.5) the second term of (B.12) is bounded by $n\dot{\phi}_k^{-1}\left\|\bar{m}_k - \bar{\dot{m}}_k\right\|\|\bar{m}_k\|/\sqrt{2k} = O_p(\dot{\phi}_k^{-1}\sqrt{k/n^\alpha})$ which is negligible for $\alpha \geq 1/3$. Negligibility of the third term of (B.12) follows by a similar argument. Finally, note that $E[\dot{m}_k(X_i)] = 0$ and $\mathrm{Var}(\dot{m}_k(X_i)) = \dot{V}_k$. Therefore, arguing as in the proof of de Jong and Bierens (1994, Theorem 1), the last term of (B.12) converges in distribution to $N(0,1)$ under the rate condition $\dot{\phi}_k^{-4}k^2/n \to 0$. Thus the result in (B.7) follows.

We now prove the second assertion, $(L_n - k)/\sqrt{2k} \to \infty$ under H_1. Since in the limit the points $\{p_1, \ldots, p_k\}$ form a dense subset of $\mathbb{S}^d$ and the support function is continuous, under H_1 there exists an integer N such that for all $n \geq N$ the set of points includes a direction p^* for which $E[s(X_i, p^*) - s(\Theta_0(\nu), p^*)] \neq 0$. Without loss of generality we prove the case of $E[s(X_i, p^*) - s(\Theta_0(\nu), p^*)] > 0$. Define $g_i(p) = s(X_i, p) - s(\Theta_0(\hat{\nu}), p)$ and $\bar{g}_i(p) = s(X_i, p) - s(\Theta_0(\hat{\nu}), p) - G(p; \hat{\nu})'\nabla f(\bar{z})'(z_i - \bar{z})$. Pick any $\delta \in (0, 0.3)$ and observe

$$\begin{aligned} L_n &\geq 2\sum_{i=1}^n \log(1 + n^{-(1/2+\delta)}g_i(p^*)) + n^{-2\delta}\Big\{\frac{1}{n}\sum_{i=1}^n g_i(p^*)^2 - \frac{1}{n}\sum_{i=1}^n \bar{g}_i(p^*)^2\Big\} \\ &= 2n^{1/2-\delta}\left\{\frac{1}{n}\sum_{i=1}^n g_i(p^*)\right\} - n^{-2\delta}\left\{\frac{1}{n}\sum_{i=1}^n \bar{g}_i(p^*)^2\right\} + O_p(n^{-2\delta}), \end{aligned}$$

for all $n \geq N$, where the inequality follows by setting $\gamma = n^{-(1/2+\delta)}e^* \in \Gamma_n$ w.p.a.1, where e^* is the unit vector that selects the component of $m_k(X_i)$ containing p^*, and the equality follows from a Taylor expansion. Now, $n^{-1}\sum_{i=1}^n g_i(p^*) \xrightarrow{p} E[s(X_i, p^*)] - s(\Theta_0(\nu), p^*) \neq 0$ by a suitable law of large numbers and $n^{-1}\sum_{i=1}^n \bar{g}_i(p^*)^2 \xrightarrow{p} E[\dot{g}_i(p^*)^2] < \infty$ by a similar argument used to show (B.4). Thus, L_n diverges to infinity at the rate $n^{1/2-\delta}$ which implies that $(L_n - k)/\sqrt{2k}$ diverges.

B.4 Additional numerical results for Section 2.4.1

In this section we report additional numerical results to compare the marked empirical likelihood confidence region obtained in Section 2.4.1 with the one based on the method by Chernozhukov, Kocatulum and Menzel (2015) (hereafter CKM). As in Section 2.4.1, we consider the relationship between the unobservable dependent variable y and regressors x, where we observe the interval $[y_L, y_U]$ satisfying $y_L \leq y \leq y_U$ almost surely.

Consider the set of coefficients characterized by the conditional moment inequalities

$$\Xi = \{\theta : \ E[y_L|x] \leq (1, x')\theta \leq E[y_U|x]\}.$$

The set Ξ would be the identified region of interest if we assume $E[y|x] = (1, x')\theta$. It is important to note that the set Ξ is a subset of

$$\Upsilon = \{\arg\min_{\theta} \int \{y - (1, x')\theta\}^2 d\mu \text{ for some } \mu \in \mathcal{M}\},$$

which is the identified region of interest in Section 2.4.1. Indeed, this can be seen from the fact that Υ is obtained as the set of parameters satisfying $E[(1, x')\{y - (1, x')\theta\}] = 0$.

If all the regressors x are discrete, then Ξ is characterized by a finite number of moment inequalities (see, Andrews and Shi, 2015, for a general case). CKM suggest a general approach to obtain confidence regions in this context by combining the moment inequalities into a single one using the smooth-max approximation (see, Section 2.3.4). In our numerical example with log wages and education, the education variable takes 13 values and thus provides 26 moment inequalities. Since it is computationally difficult to work with a smooth-max approximation with such a large number of moments, we simplify the problem by partitioning the regressor values into four bins and utilizing the moment inequalities within each bin (corresponding to a total of 8 moment inequalities). In particular, we partition the education variable into the following broad categories: Less than 10^{th} grade ($x \leq 10$); High school graduate ($x \in [11, 12]$); some college or associate degree including vocational training ($x \in [13, 14]$); and Bachelor's degree or higher ($x \geq 15$).

Figure B.1 compares the 95% confidence region of CKM for Ξ with that from the marked

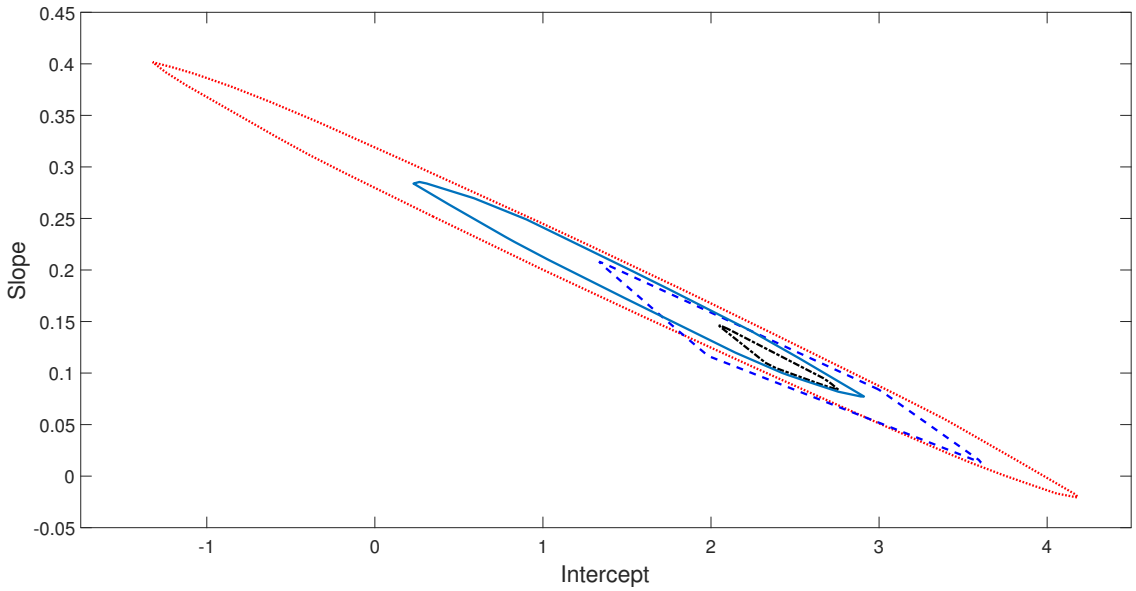

Figure B.1: The population identification regions for regression with interval outcomes Ξ (dash-dotted line) and for the best linear prediction Υ (solid line) as well as the corresponding 95% confidence regions via CKM (dashed line) and the marked empirical likelihood statistic (dotted line). The sample size is $n = 1000$.

empirical likelihood for Υ. The sample size is $n = 1000$. The tuning parameter ϱ for the 'smooth-max' approximation is chosen to be $\varrho = 100$. The critical values in both cases were obtained using bootstrap with 999 repetitions. Unsurprisingly, the CKM confidence region is smaller than that obtained by the marked empirical likelihood. This is due to the fact that the region Ξ is considerably smaller than Υ as can be seen from Figure B.1. From Figure B.1, we can thus infer the following: If it is possible to impose additional assumptions to satisfy the conditional moment restriction $E[y|x] = x'\theta$, then characterizing the set using moment inequalities leads to a much smaller confidence region. At the same time, the best linear predictor is more robust to possible misspecification and thus, is applicable more generally, albeit at the expense of a larger confidence set.

B.5 Simulation results for Section 2.3.5

We consider the problem of testing the shape of a set based on noisy measurements of the support function, as discussed in Section 2.3.5. We employ the simulation design of Fisher *et al.* (1997), where the underlying set is an ellipse relative to the origin with the

support function taking the form $s(\Theta, p) = (\theta_1^2 \cos^2 p + \theta_2^2 \sin^2 p)^{1/2}$ for $p \in [-\pi, \pi]$. Noisy measurements $\{s_i, p_i\}_{i=1}^n$ of the support function are generated using $s_i = s(\Theta, p_i) + \epsilon_i$ with $p_i \sim$ Uniform$[-\pi, \pi]$ and $\epsilon_i \sim N(0, 0.16)$.

We consider two types of testing problems here. First, we test whether the set Θ takes a particular shape. In the first four columns of Table B.1, we report the rejection frequencies of the marked empirical likelihood test based on eq. (2.12) of Chapter 2 for the null hypotheses H_0^a :'Θ is a circle with $(\theta_1, \theta_2) = (1, 1)$' and H_0^b :'Θ is an ellipse with $(\theta_1, \theta_2) = (1, 2)$'. To compute the test statistic we follow Fisher *et al.* (1997) in employing the von Mises density function $K_b(z) = e^{b \cos z} / \int_{-\pi}^{\pi} e^{b \cos z} dz$ on the circle as the kernel and set the smoothing parameter to be $b = 8$ (which corresponds to the inverse of the square of the bandwidth for the conventional kernel density estimator). In the last two rows of Table B.1 we present the results for different values of the bandwidth by setting $b = 4$ and 16 when $n = 200$. The critical value of the test is computed using the wild bootstrap based on Härdle and Mammen (1993). We consider sample sizes of n =100, 200, and 500. The number of Monte Carlo replications is 1000 for all cases. The first and third columns of Table B.1 indicate that the marked empirical likelihood test based on eq. (2.12) of Chapter 2 has reasonable size properties for both null hypotheses and over all sample sizes. The second and fourth columns evaluate power properties of the test against the alternatives $\mathrm{H}_1^a : (\theta_1, \theta_2) = (1.1, 1)$ and $\mathrm{H}_1^b : (\theta_1, \theta_2) = (1.1, 2)$, respectively. In both cases, the power of the empirical likelihood test increases with the sample size at a reasonably fast rate.

Second, we conduct a goodness-of-fit test for the null H_0^c :'Θ is a ellipse with $s(\Theta, p) = (\theta_1^2 \cos^2 p + \theta_2^2 \sin^2 p)^{1/2}$ for some (θ_1, θ_2)'. For this testing problem, (θ_1, θ_2) are nuisance parameters to be estimated. The marked empirical likelihood statistic is modified by replacing $\{s_i - s(\Theta_0, p)\}$ in eq. (2.12) of Chapter 2 with its estimated counterpart $\{s_i - (\hat{\theta}_1^2 \cos^2 p + \hat{\theta}_2^2 \sin^2 p)^{1/2}\}$, where $(\hat{\theta}_1, \hat{\theta}_2)$ is the nonlinear least squares estimator. Under the null H_0^c, the measurements on the support function are generated by $(\theta_1, \theta_2) = (1, 2)$. Under the alternative H_1^c, the data are generated by $s(\Theta, p) = (\cos^2 p + \cos p \sin p + 4 \sin^2 p)^{1/2}$. The critical value is again computed using the wild bootstrap. The last two columns of Table B.1 report the rejection frequencies of this test. Although the test is slightly undersized, it shows good size and power performance.

n, b	H_0^a :circle	H_1^a	H_0^b :ellipse	H_1^b	H_0^c	H_1^c
100, 8	0.022	0.425	0.079	0.413	0.049	0.182
200, 8	0.026	0.851	0.029	0.608	0.020	0.409
500, 8	0.039	0.999	0.025	0.958	0.013	0.991
200, 4	0.036	0.854	0.025	0.557	0.016	0.332
200, 16	0.014	0.774	0.037	0.668	0.012	0.443

Table B.1: Rejection frequencies of the marked empirical likelihood test at the nominal 5% level

Finally, the last two rows of Table B.1 show that the rejection frequencies are not very sensitive to the choice of the smoothing parameter b under the null and alternative hypotheses.

Appendix C

Supplementary material and proofs for Chapter 3

C.1 Proofs of Theorems

Hereafter, let $P^{\#}$ and $E^{\#}$ be the conditional probability and expectation under the bootstrap distribution given $\{X_i\}_{i=1}^n$, respectively. Also, denote

$$\begin{aligned}
\bar{\mathbb{L}}(u) &= \frac{1}{\pi}\int_0^1 \frac{\sin(\omega u)}{\omega}\frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)}\mathbb{I}\{|\omega|\geq\omega_0\}d\omega,\\
\mathcal{G}_n(t) &= r(h)\int \bar{\mathbb{L}}\left(\frac{t-a}{h}\right) f_X(a)^{1/2}dW(a),\\
p_\varepsilon(\mathcal{G}_n) &= \sup_x P\left\{\left|\sup_{t\in\mathcal{T}}\mathcal{G}_n(t)-x\right|\leq\varepsilon\right\},
\end{aligned}$$

where W is a (two-sided) Wiener process on $\mathbb{R}$, f_X is the pdf of X, and

$$r(h)=\begin{cases} h^{\beta-\frac{1}{2}} & \text{under Assumption OS}\\ 1/\varsigma(h) \text{ in eq. (8)} & \text{under Assumption SS}\end{cases}.$$

Note that analogous to $\bar{\mathbb{K}}$ (defined in Assumptions OS (ii) and SS (ii)), $\bar{\mathbb{L}}$ is considered as a trimmed version of $\mathbb{L}$. Due to the trimming, properties of the Fourier transform guarantee $\bar{\mathbb{L}}\in L_2(\mathbb{R})$ for each h under the assumption $f_\epsilon^{\mathrm{ft}}\neq 0$, and this guarantees existence of the stochastic integral in the definition of $\mathcal{G}_n$.

Also, for any $a \in (0,1)$, let c_a denote the constant such that $\sqrt{n}h^{\beta-\frac{1}{2}}c_a$ is the $(1-a)$-th quantile of $\sup_{t\in\mathcal{T}}|\mathcal{G}_n(t)|$.

C.1.1 Proof of Theorem 6

We only prove the statement under Assumption OS (i.e., the ordinary smooth case). The statement under Assumption SS is shown by a similar argument using Lemmas 20-22 in Appendix C.2.

First, we prove

$$P\left\{\sqrt{n}h^{\beta-\frac{1}{2}}\hat{c}_\alpha > \sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}} - \epsilon_{1n}\right\} \geq 1-\delta_{2n}, \tag{C.1}$$

for some $\epsilon_{1n}, \delta_{1n}, \delta_{2n} = O(n^{-c})$ with $c > 0$. Lemma 18 in Appendix C.2 implies that with probability greater than $1-\delta_{2n}$,

$$\begin{aligned} 1-\alpha &= P^{\#}\left\{\sqrt{n}h^{\beta-\frac{1}{2}}\sup_{t\in\mathcal{T}}\left|\hat{F}^{\#}_{X^*}(t) - \hat{F}_{X^*}(t)\right| \leq \sqrt{n}h^{\beta-\frac{1}{2}}\hat{c}_\alpha\right\} \\ &\leq P^{\#}\left\{\sup_{t\in\mathcal{T}}\left|\tilde{\mathcal{G}}_n(t)\right| \leq \sqrt{n}h^{\beta-\frac{1}{2}}\hat{c}_\alpha + \epsilon_{1n}\right\} + \delta_{1n}, \end{aligned}$$

for some $\epsilon_{1n}, \delta_{1n}, \delta_{2n} = O(n^{-c})$ with $c > 0$, where $\tilde{\mathcal{G}}_n$ has the same distribution as $\mathcal{G}_n$ under $P^{\#}$. Since $\sqrt{n}h^{\beta-\frac{1}{2}}c_a$ is also the $(1-a)$-th quantile of $\sup_{t\in\mathcal{T}}\left|\tilde{\mathcal{G}}_n(t)\right|$ under $P^{\#}$, the above inequality implies

$$P^{\#}\left\{\sup_{t\in\mathcal{T}}\left|\tilde{\mathcal{G}}_n(t)\right| \leq \sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}\right\} \leq P^{\#}\left\{\sup_{t\in\mathcal{T}}\left|\tilde{\mathcal{G}}_n(t)\right| \leq \sqrt{n}h^{\beta-\frac{1}{2}}\hat{c}_\alpha + \epsilon_{1n}\right\},$$

with probability greater than $1-\delta_{2n}$. Thus, we obtain (C.1).

The main result is thus obtained from the following sequence of inequalities

$$
\begin{aligned}
P\{T_n \le \hat{c}_\alpha\} &\ge P\left\{\sup_{t\in\mathcal{T}} |\mathcal{G}_n(t)| \le \sqrt{n}h^{\beta-\frac{1}{2}}\hat{c}_\alpha - \epsilon_n\right\} - \delta_n \\
&\ge P\left\{\sup_{t\in\mathcal{T}} |\mathcal{G}_n(t)| \le \sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}} - \epsilon_{1n} - \epsilon_n\right\} - \delta_n - \delta_{2n} \\
&\ge P\left\{\sup_{t\in\mathcal{T}} |\mathcal{G}_n(t)| \le \sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}\right\} - 2p_{\bar{\epsilon}_n}(\mathcal{G}_n) - \delta_n - \delta_{2n} \\
&= 1 - \alpha - \delta_{1n} - 2p_{\bar{\epsilon}_n}(\mathcal{G}_n) - \delta_n - \delta_{2n} \\
&\ge 1 - \alpha - \delta_{1n} - M\bar{\epsilon}_n\sqrt{\log(1/h)} - \delta_n - \delta_{2n},
\end{aligned}
$$

where the first inequality follows from Lemma 17, the second inequality follows from (C.1), the third inequality follows from the definitions of $\bar{\epsilon}_n = \epsilon_{1n} + \epsilon_n$ and $p_\varepsilon(\mathcal{G}_n)$, along with the fact $\mathcal{G}_n$ and $-\mathcal{G}_n$ have the same distribution (which ensures $p_\varepsilon(|\mathcal{G}_n|) \le 2p_\varepsilon(\mathcal{G}_n)$), the equality follows from the definition that $\sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}$ is the $(1-\alpha-\delta_{1n})$-th quantile of $\sup_{t\in\mathcal{T}}|\mathcal{G}_n(t)|$, and the last inequality follows from Lemma 19. Therefore, letting $\delta_{3n} = \delta_{1n} + M\bar{\epsilon}_n\sqrt{\log(1/h)} + \delta_n + \delta_{2n}$, we have

$$
P\{T_n \le \hat{c}_\alpha\} \ge 1 - \alpha - \delta_{3n}.
$$

Since $\delta_n, \delta_{1n}, \delta_{2n}, \bar{\epsilon}_n$ are all positive sequences of order $O(n^{-a})$ with some $a > 0$ and $\sqrt{\log(1/h)}$ is a log-rate, we obtain the conclusion.

C.1.2 Proof of Theorem 8

For simplicity, we restrict attention to the case of $N_i = 2$. For more general situations where N_i is arbitrary but bounded above by C, the proof follows by similar arguments after accounting for the dependence structure in $\hat{f}_\epsilon^{\mathrm{ft}}$.

We first make the following preliminary observations. Note that $\tilde{F}_{X^*}(t)$ can be alternatively written as

$$
\tilde{F}_{X^*}(t) = \frac{1}{2\pi}\int_{-1/h}^{1/h} \frac{\mathrm{Im}\{e^{\mathrm{i}\omega t}\hat{f}_X^{\mathrm{ft}}(\omega)\}}{-\omega}\frac{K^{\mathrm{ft}}(h\omega)}{\hat{f}_\epsilon^{\mathrm{ft}}(\omega)}d\omega. \tag{C.2}
$$

where $\hat{f}_X^{\mathrm{ft}}(\omega) = N^{-1}\sum_{i,j} e^{\mathrm{i}\omega X_{i,j}}$ denotes the empirical characteristic function. A similar expression holds for $\hat{F}_{X^*}$. Let $\xi = (f_\epsilon^{\mathrm{ft}})^2$ and $\hat{\xi} = (\hat{f}_\epsilon^{\mathrm{ft}})^2$. We note the following properties

for $\hat{\xi}$

$$E\left[\int_{\omega_0}^{h^{-1}} \omega^{-a}|\hat{\xi}(\omega)-\xi(\omega)|^2 d\omega\right] = \begin{cases} n^{-1}h^{-(1-a)} & \text{if } a<1 \\ n^{-1} & \text{if } a>1 \\ n^{-1}\log(1/h) & \text{if } a=1 \end{cases} \tag{C.3}$$

$$\sup_{|\omega|\le h^{-1}} |\xi/\hat{\xi}| \le 1+o_p(1). \tag{C.4}$$

The results in (C.3) can be shown by expanding the expectations. To show (C.4), we use Yukich (1987, Theorem 6.3) which assures that under Assumption B (i), $\sup_{|\omega|\le h^{-1}}|\hat{\xi}-\xi| = O_p(\sqrt{\log n/n})$ for $h = O(n^{-c})$ with some $c>0$. Combined with Assumption B (ii), this implies $\{\min_{|\omega|\le h^{-1}}|\hat{\xi}|\}^{-1} = O_p(h^{-2\beta})$. Thus we obtain

$$\sup_{|\omega|\le h^{-1}} |\xi/\hat{\xi}| \le 1+\sup_{|\omega|\le h^{-1}}|(\hat{\xi}-\xi)/\hat{\xi}| = 1+O_p\left(\left(\frac{\log n}{nh^{4\beta}}\right)^{1/2}\right) = 1+o_p(1),$$

thereby proving (C.4).

Pick any $\eta \in (1/2, \gamma-\beta)$. Under Assumptions C (iii) and OS (i), it can be verified that

$$\int_{-1/h}^{1/h}\left|\frac{\omega^\eta f^{\mathrm{ft}}_{X^*}(\omega)}{\xi(\omega)^{1/2}}\right|^2 d\omega = O(1). \tag{C.5}$$

We shall also make frequent use of the following algebraic inequality:

$$|\hat{\xi}^{1/2}-\xi^{1/2}| \le \xi^{-1/2}|\hat{\xi}-\xi|. \tag{C.6}$$

We now proceed to the main part of the proof. By (C.2), we can expand

$$\begin{aligned} \tilde{F}_{X^*}(t)-\hat{F}_{X^*}(t) &= \frac{1}{\pi}\int_0^{\omega_0}\frac{\mathrm{Im}\{e^{-\mathrm{i}\omega t}\hat{f}^{\mathrm{ft}}_X(\omega)\}}{-\omega}\{\hat{\xi}(\omega)^{-1/2}-\xi(\omega)^{-1/2}\}K^{\mathrm{ft}}(h\omega)d\omega \\ &\quad +\frac{1}{\pi}\int_{\omega_0}^{1/h}\frac{\mathrm{Im}\{e^{-\mathrm{i}\omega t}\hat{f}^{\mathrm{ft}}_X(\omega)\}}{-\omega}\{\hat{\xi}(\omega)^{-1/2}-\xi(\omega)^{-1/2}\}K^{\mathrm{ft}}(h\omega)d\omega \\ &= B_{1n}(t)+B_{2n}(t). \end{aligned}$$

For the term $B_{1n}(t)$, using (C.6), we have

$$|B_{1n}(t)| \leq \frac{1}{\pi}\int_0^{\omega_0}\left|\frac{\mathrm{Im}\{e^{-\mathrm{i}\omega t}\hat{f}_X^{\mathrm{ft}}(\omega)\}}{-\omega}\right|\left|\frac{\xi(\omega)}{\hat{\xi}(\omega)}\right|^{1/2}\frac{|\hat{\xi}(\omega)-\xi(\omega)|}{\xi(\omega)^{3/2}}d\omega.$$

By the fact $\sup_{|\omega|\leq\omega_0}|\hat{\xi}-\xi| = O_p(n^{-1/2})$ and (C.4), we obtain

$$\sup_{t\in\mathcal{T}}|B_{1n}(t)| = O_p(n^{-1/2})\sup_{t\in\mathcal{T}}I(t),$$

where

$$\begin{aligned}
I(t) &= \int_0^{\omega_0}\left|\frac{\mathrm{Im}\{e^{-\mathrm{i}\omega t}\hat{f}_X^{\mathrm{ft}}(\omega)\}}{-\omega}\right|d\omega \\
&\leq \int_0^{\omega_0}\left|\frac{\sin(\omega t)}{\omega}\mathrm{Re}\{\hat{f}_X^{\mathrm{ft}}(\omega)\}\right|d\omega + \int_0^{\omega_0}\left|\frac{\cos(\omega t)}{\omega}\mathrm{Im}\{\hat{f}_X^{\mathrm{ft}}(\omega)\}\right|d\omega \\
&\leq \int_0^{\omega_0}\left|\frac{\sin(\omega t)}{\omega}\right|d\omega + \int_0^{\omega_0}\left|\frac{\mathrm{Im}\{\hat{f}_X^{\mathrm{ft}}(\omega)\}}{\omega}\right|d\omega \\
&= I_1(t) + I_2.
\end{aligned}$$

Since $\mathcal{T}$ is a compact set, it holds $\sup_{t\in\mathcal{T}}I_1(t) < \infty$. By the definition of $\hat{f}_X^{\mathrm{ft}}$, the random variable I_2 can be bounded as

$$I_2 \leq \frac{1}{N}\sum_{i,j}\int_0^{\omega_0}\left|\frac{\sin(\omega X_{i,j})}{\omega}\right|d\omega \equiv \frac{1}{N}\sum_{i,j}T_{i,j}.$$

Since

$$E[T_{i,j}] = E\int_0^{\omega_0|X_{i,j}|}\left|\frac{\sin(t)}{t}\right|dt \leq C_1 + E[\log|X_{i,j}|] < \infty$$

for some $C_1 > 0$, it holds $I_2 = O_p(1)$. Combining these results, we obtain $\sup_{t\in\mathcal{T}}|B_{1n}(t)| = O_p(n^{-1/2})$.

For the term $B_{2n}(t)$, we further expand

$$\begin{aligned}
B_{2n}(t) &= -\frac{1}{\pi}\int_{\omega_0}^{1/h}\frac{\mathrm{Im}\{e^{-\mathrm{i}\omega t}f_X^{\mathrm{ft}}(\omega)\}}{-\omega\xi(\omega)}\{\hat{\xi}(\omega)^{1/2}-\xi(\omega)^{1/2}\}K^{\mathrm{ft}}(h\omega)\frac{\xi(\omega)^{1/2}}{\hat{\xi}(\omega)^{1/2}}d\omega \\
&\quad + \frac{1}{\pi}\int_{\omega_0}^{1/h}\frac{\mathrm{Im}\left\{e^{-\mathrm{i}\omega t}\{\hat{f}_X^{\mathrm{ft}}(\omega)-f_X^{\mathrm{ft}}(\omega)\}\right\}}{-\omega}\{\hat{\xi}(\omega)^{-1/2}-\xi(\omega)^{-1/2}\}K^{\mathrm{ft}}(h\omega)d\omega \\
&= B_{21n}(t) + B_{22n}(t).
\end{aligned}$$

For the term $B_{21n}(t)$, we have

$$\begin{aligned}\sup_{t\in\mathcal{T}}|B_{21n}(t)| &\leq \frac{1}{\pi}\int_{\omega_0}^{1/h}\left|\frac{\omega^{\eta}f^{\mathrm{ft}}_{X^*}(\omega)}{\xi(\omega)^{1/2}}\right|\left|\frac{\hat{\xi}(\omega)-\xi(\omega)}{\omega^{1+\eta}\xi(\omega)^{1/2}}\right|\left|\frac{\xi(\omega)}{\hat{\xi}(\omega)}\right|^{1/2}d\omega \\ &\leq C_2(1+o_p(1))\left(\int_{\omega_0}^{1/h}\left|\frac{\omega^{\eta}f^{\mathrm{ft}}_{X^*}(\omega)}{\xi(\omega)^{1/2}}\right|^2 d\omega\right)^{1/2}\left(\int_{\omega_0}^{1/h}\omega^{2(\beta-\eta-1)}|\hat{\xi}(\omega)-\xi(\omega)|^2 d\omega\right)^{1/2} \\ &= O(n^{-1/2}h^{(\eta-\beta+1/2)\wedge 0}),\end{aligned}$$

for some $C_2 > 0$, where the first inequality follows from the fact $|\mathrm{Im}\{e^{-\mathrm{i}\omega t}f^{\mathrm{ft}}_X(\omega)\}| \leq |f^{\mathrm{ft}}_X(\omega)| = |f^{\mathrm{ft}}_{X^*}(\omega)|\xi(\omega)^{1/2}$ and (C.6), the second inequality follows from (C.4) and Assumption OS (i), and the equality follows from (C.3) and (C.5).

Now consider the term $B_{22n}(t)$. Applying (C.6) and Assumption OS (i), we can write

$$\begin{aligned}\sup_{t\in\mathcal{T}}|B_{22n}(t)| &\leq \frac{1}{\pi}\int_{\omega_0}^{1/h}|\hat{f}^{\mathrm{ft}}_X(\omega)-f^{\mathrm{ft}}_X(\omega)||\hat{\xi}(\omega)-\xi(\omega)||\xi(\omega)/\hat{\xi}(\omega)|^{1/2}\frac{1}{\omega\xi(\omega)^{3/2}}d\omega \\ &\leq \frac{1}{c^3\pi}\int_{\omega_0}^{1/h}|\hat{f}^{\mathrm{ft}}_X(\omega)-f^{\mathrm{ft}}_X(\omega)||\hat{\xi}(\omega)-\xi(\omega)||\xi(\omega)/\hat{\xi}(\omega)|^{1/2}\omega^{3\beta-1}d\omega,\end{aligned}$$

for some $c > 0$. As in (C.5), it can be shown after expanding the expectation that

$$E\left[\int_{\omega_0}^{1/h}\omega^{-a}|\hat{f}^{\mathrm{ft}}_X(\omega)-f^{\mathrm{ft}}_X(\omega)|^2 d\omega\right] = O((nh^{1-a})^{-1}),$$

for all $a < 1$. Thus, by (C.4) and (C.5), it follows

$$\begin{aligned}\sup_{t\in\mathcal{T}}|B_{22n}(t)| &\leq \frac{1+o_p(1)}{\pi}\int_{\omega_0}^{1/h}\omega^{3\beta-1}|\hat{f}^{\mathrm{ft}}_X(\omega)-f^{\mathrm{ft}}_X(\omega)||\hat{\xi}(\omega)-\xi(\omega)|d\omega \\ &= \frac{1+o_p(1)}{\pi}\left(\int_{\omega_0}^{1/h}\omega^{3\beta-1}|\hat{f}^{\mathrm{ft}}_X(\omega)-f^{\mathrm{ft}}_X(\omega)|^2 d\omega\right)^{1/2}\left(\int_{\omega_0}^{1/h}\omega^{3\beta-1}|\hat{\xi}(\omega)-\xi(\omega)|^2 d\omega\right)^{1/2} \\ &= O((nh^{3\beta})^{-1}).\end{aligned}$$

Combining these results, we obtain

$$\sqrt{n}h^{\beta-1/2}\sup_{t\in\mathcal{T}}|\tilde{F}_{X^*}(t)-\hat{F}_{X^*}(t)| = O_p\left(h^{\eta\wedge(\beta-1/2)}+\frac{1}{\sqrt{n}h^{2\beta+1/2}}\right) = o_p(1),$$

under Assumption B (ii) and the condition $\eta > 1/2$.

C.1.3 Proof of Theorem 9

Define

$$\hat{D}_n^{\#}(t) = \sqrt{n}h^{\beta-1/2}\{\hat{F}_{X^*}^{\#}(t) - \hat{F}_{X^*}(t)\}, \qquad \tilde{D}_n^{\#}(t) = \sqrt{n}h^{\beta-1/2}\{\tilde{F}_{X^*}^{\#}(t) - \tilde{F}_{X^*}(t)\}.$$

Also, let $\hat{f}_X^{\mathrm{ft}\#}(\omega) = N^{-1}\sum_{i,j} e^{\mathrm{i}\omega X_{i,j}^{\#}}$ be the bootstrap counterpart of the empirical characteristic function $\hat{f}_X^{\mathrm{ft}}(\omega) = N^{-1}\sum_{i,j} e^{\mathrm{i}\omega X_{i,j}}$.

We first show that there exist $c, C > 0$ such that

$$P^{\#}\left\{\sup_{t\in\mathcal{T}} |\tilde{D}_n^{\#}(t) - \hat{D}_n^{\#}(t)| \geq Cn^{-c}\right\} = o_p(1). \tag{C.7}$$

By Theorem 8, it is enough for (C.7) to guarantee that there exist $c, C > 0$ satisfying

$$P^{\#}\left\{\sqrt{n}h^{\beta-1/2}\sup_{t\in\mathcal{T}} |\tilde{F}_{X^*}^{\#}(t) - \hat{F}_{X^*}^{\#}(t)| \geq Cn^{-c}\right\} = o_p(1).$$

To this end, note that

$$\begin{aligned}
\tilde{F}_{X^*}^{\#}(t) - \hat{F}_{X^*}^{\#}(t) &= \frac{1}{2\pi}\int_{-1/h}^{1/h} \frac{\mathrm{Im}\left\{e^{-\mathrm{i}\omega t}\left\{\hat{f}_X^{\mathrm{ft}\#}(\omega) - \hat{f}_X^{\mathrm{ft}}(\omega)\right\}\right\}}{-\omega}\{\hat{\xi}(\omega)^{-1/2} - \xi(\omega)^{-1/2}\}K^{\mathrm{ft}}(h\omega)d\omega \\
&\quad + \frac{1}{2\pi}\int_{-1/h}^{1/h} \frac{\mathrm{Im}\{e^{-\mathrm{i}\omega t}\hat{f}_X^{\mathrm{ft}}(\omega)\}}{-\omega}\{\hat{\xi}(\omega)^{-1/2} - \xi(\omega)^{-1/2}\}K^{\mathrm{ft}}(h\omega)d\omega \\
&= C_{1n}(t) + C_{2n}(t).
\end{aligned}$$

The second term $C_{2n}(t)$ equals to $\tilde{F}_{X^*}(t) - \hat{F}_{X^*}(t)$ whose bound is given in Theorem 8. Thus, we only need to consider the first term $C_{1n}(t)$. By expanding the expectations, it can be shown

$$E^{\#}\left[\int_{\omega_0}^{1/h} \omega^{-a}|\hat{f}_X^{\mathrm{ft}\#}(\omega) - \hat{f}_X^{\mathrm{ft}}(\omega)|^2 d\omega\right] = O_p((nh^{1-a})^{-1}),$$

for all $a < 1$, and analogous arguments as in the proof of Theorem 8 yield $\sup_{t\in\mathcal{T}}|C_{1n}(t)| = O_{p^{\#}}((nh^{3\beta})^{-1})$ with probability approaching one. Therefore, by paralleling the arguments in the proof of Theorem 8, we obtain (C.7).

We now proceed by verifying the conditions in the proof of Theorem 6. Lemma 17 and

Theorem 8 ensure existence of a sequence $\epsilon_n = O(n^{-c})$ with some $c > 0$ such that

$$P\left\{\sup_{t\in\mathcal{T}}\left|\sqrt{n}h^{\beta-1/2}\{\tilde{F}_X(t) - F_X(t)\} - \mathcal{G}_n(t)\right| > \epsilon_n\right\} = o_p(1). \tag{C.8}$$

Furthermore by Lemma 18, combined with (C.7), we have that

$$P^{\#}\left\{\sup_{t\in\mathcal{T}}\left|\sqrt{n}h^{\beta-1/2}\{\tilde{F}^{\#}_{X^*}(t) - \tilde{F}_{X^*}(t)\} - \tilde{\mathcal{G}}_n(t)\right| > \epsilon_n\right\} = o_p(1). \tag{C.9}$$

Therefore, by (C.8) and (C.9), the conclusion follows by paralleling the arguments in the proof of Theorem 6.

C.1.4 Proof of Theorem 10

We only prove the theorem under Assumption OS (i.e., the ordinary smooth case). The proof under Assumption SS follows by a similar argument using Lemmas 20-22.

We make the following preliminary observations. First, by the techniques employed in Lemmas 17-19, we can show[1]

$$\sup_{t\in\mathcal{H}}|\hat{f}_{X^*}(t) - f_{X^*}(t)| = O_p(n^{-c}). \tag{C.10}$$

Next by Dattner, Reiß and Trabs (2016, Proposition 2.1), $\left\|\hat{f}_{X^*}\right\|_1 < \infty$ and $\int_{-\infty}^{\infty}\hat{f}_{X^*}(t)dt = 1$ under Assumption C. Thus, we have $\hat{F}_{X^*}(t) = \int_{-\infty}^{t}\hat{f}_{X^*}(v)dv$ or equivalently $\hat{F}'_{X^*}(t) = \hat{f}_{X^*}(t)$. The latter ensures $\hat{F}_{X^*}$ is continuous.

We now show that[2]

$$\sup_{u\in[u_1,u_2]}|\hat{Q}(u) - Q(u)| = o_p(n^{-c_1}), \tag{C.11}$$

for some $c_1 > 0$. By Hall and Lahiri (2008, Theorem 3.7), $\hat{Q}(u)$ converges to $Q(u)$ for each $u \in [u_1, u_2]$. Now $Q_n(u)$ is monotone at each n by construction while $Q(u)$ is continuous by Assumption Q (i). Hence we can modify the proof of the Glivenko-Cantelli theorem (see,

[1] An analogous result applies for the super smooth case by Lemmas 20-22 with the rate replaced by $O_p((\log n)^{-c})$ for some $c > 1$ under the assumption $\gamma > \lambda$ and an MSE optimal bandwidth choice.

[2] For the super smooth case, we can employ similar arguments to show that $\sup_{u\in[u_1,u_2]}|\hat{Q}(u) - Q(u)| = o_p((\log n)^{-c_1})$ for some $c_1 > 1$.

Billingsley, 1995, p. 233), to strengthen the pointwise consistency to a uniform one, i.e.,

$$\sup_{u\in[u_1,u_2]} |\hat{Q}(u) - Q(u)| = o_p(1), \tag{C.12}$$

(see also, Bassett and Koenker, 1986, Theorem 3.1). As $\hat{F}_{X^*}$ is continuous, it follows that $\hat{F}_{X^*}(\hat{Q}(u)) = u$ for all $0 < u < 1$. Consequently,

$$\hat{F}_{X^*}(\hat{Q}(u)) = F_{X^*}(Q(u)) = F_{X^*}(\hat{Q}(u)) + f_{X^*}(\tilde{Q}(u))(\hat{Q}(u) - Q(u)),$$

for some $\tilde{Q}(u)$ such that $|\tilde{Q}(u) - Q(u)| \leq |\hat{Q}(u) - Q(u)|$, and we obtain

$$\sup_{u\in[u_1,u_2]} |\hat{Q}(u) - Q(u)| \leq \left(\inf_{u\in[u_1,u_2]} |f_{X^*}(\tilde{Q}(u))| \right)^{-1} \sup_{u\in[u_1,u_2]} |\hat{F}_{X^*}(\hat{Q}(u)) - F_{X^*}(\hat{Q}(u))|$$

By (C.12) and Assumption Q (i) ($\inf_{x\in\mathcal{H}} f_{X^*}(x) > 0$), we can verify $\inf_{u\in[u_1,u_2]} |f_{X^*}(\tilde{Q}(u))| > 0$ with probability approaching one. Furthermore, we have

$$\sup_{u\in[u_1,u_2]} |\hat{F}_{X^*}(\hat{Q}(u)) - F_{X^*}(\hat{Q}(u))| \leq n^{-\frac{1}{2}} h^{-\beta+\frac{1}{2}} \sup_{t\in\mathcal{H}} |\mathcal{G}_n(t)| + o_p(1) = O_p\left(\left(\frac{\log(1/h)}{nh^{2\beta-1}} \right)^{1/2} \right),$$

where the inequality follows from Lemma 17 after employing the fact $\{\hat{Q}(u) : u \in [u_1, u_2]\} \subset \mathcal{H}$ with probability approaching one due to Assumption Q (i) and (C.12). The equality follows from $E[\sup_{t\in\mathcal{H}} |\mathcal{G}_n(t)|] = O(\sqrt{\log(1/h)})$ (by the proof of Lemma 19). Combining these results, we obtain (C.11) under Assumptions OS (iii) and B (ii).

We now proceed to the main part of the proof. Noting that $\hat{Q}(u) - Q(u) = f_{X^*}(\tilde{Q}(u))^{-1}\{\hat{F}_{X^*}(\hat{Q}(u)) - F_{X^*}(\hat{Q}(u))\}$, we have

$$\begin{aligned} & P\left\{ \hat{Q}(u) - \frac{\hat{c}_\alpha}{\hat{f}_{X^*}(\hat{Q}(u))} \leq Q(u) \leq \hat{Q}(u) + \frac{\hat{c}_\alpha}{\hat{f}_{X^*}(\hat{Q}(u))} \quad \text{for all } u \in [u_1, u_2] \right\} \\ = \; & P\left\{ \sup_{u\in[u_1,u_2]} |\hat{f}_{X^*}(\hat{Q}(u))\{\hat{Q}(u) - Q(u)\}| \leq \hat{c}_\alpha \right\} \geq P\left\{ \sup_{t\in\mathcal{H}} |\hat{F}_{X^*}(t) - F_{X^*}(t)| \leq \hat{c}_\alpha(1 - \Delta_n) \right\} - o_p(1), \end{aligned}$$

where $\Delta_n = \sup_{u\in[u_1,u_2]} \left| \frac{\hat{f}_{X^*}(\hat{Q}(u)) - f_{X^*}(\tilde{Q}(u))}{\hat{f}_{X^*}(\hat{Q}(u))} \right|$ and the inequality follows from the fact $P\left\{ \{\hat{Q}(u) : u \in [u_1, u_2]\} \subset \mathcal{H} \right\} \to 1$ by Assumption Q (i) and (C.11). Also note that $\Delta_n = O_p(n^{-c})$ by Assumption Q (i)-(ii), (C.10), and (C.11). We now have the following sequence

of inequalities

$$
\begin{aligned}
& P\left\{\sup_{t\in\mathcal{H}}|\hat{F}_{X^*}(t)-F_{X^*}(t)|\le\hat{c}_\alpha(1-\Delta_n)\right\}\ge P\left\{\sup_{t\in\mathcal{H}}|\mathcal{G}_n(t)|\le\sqrt{n}h^{\beta-\frac{1}{2}}\hat{c}_\alpha(1-\Delta_n)-\epsilon_n\right\}-\delta_n \\
\ge\ & P\left\{\sup_{t\in\mathcal{H}}|\mathcal{G}_n(t)|\le\left(\sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}-\epsilon_{1n}\right)(1-\Delta_n)-\epsilon_n\right\}-\delta_n-\delta_{2n} \\
\ge\ & P\left\{\sup_{t\in\mathcal{H}}|\mathcal{G}_n(t)|\le\sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}\right\}-2p_{\bar{\epsilon}_n}(\mathcal{G}_n)-\delta_n-\delta_{2n}\ge 1-\alpha-\delta_{1n}-\delta_n-\delta_{2n}-2p_{\bar{\epsilon}_n}(\mathcal{G}_n),
\end{aligned}
$$

where the first inequality follows from Lemma 17, the second inequality can be derived by Lemma 18 and a similar argument in the proof of Theorem 6, the third inequality follows from the definitions of $\bar{\epsilon}_n=\epsilon_n+\epsilon_{1n}(1-\Delta_n)+\sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}\Delta_n$ and the concentration function. Note that Lemma 19 implies $p_{\bar{\epsilon}_n}(\mathcal{G}_n)\le C\bar{\epsilon}_n\sqrt{\log n}$. Recalling that $\sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}$ is the $(\alpha+\delta_{1n})$-th quantile of $\sup_{t\in\mathcal{H}}|\mathcal{G}_n(t)|$, by Chernozhukov, Chetverikov and Kato (2014, Lemma B1),

$$
\sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}\le E\left[\sup_{t\in\mathcal{H}}|\mathcal{G}_n(t)|\right]+\sqrt{2|\log(\alpha+\delta_{1n})|}.
$$

Since $E[\sup_{t\in\mathcal{H}}|\mathcal{G}_n(t)|]=O(\sqrt{\log(1/h)})$, this implies $\sqrt{n}h^{\beta-\frac{1}{2}}c_{\alpha+\delta_{1n}}=O(\sqrt{\log n})$ under Assumptions OS (iii) and B (ii). By the above and the rates of ϵ_n,ϵ_{1n}, it follows $p_{\bar{\epsilon}_n}(\mathcal{G}_n)=O_p(n^{-c_2})$ for some $c_2>0$. Furthermore, by Lemmas 17 and 18, δ_n, δ_{1n}, and δ_{2n} are also $O(n^{-c_3})$ for some $c_3>0$. Combining these results, the conclusion follows.

C.1.5 Proof of Theorem 11

We shall assume for simplicity that $f_\epsilon=f_\delta$, and consequently that the bandwidth choices for both estimators are the same. We only prove for the case of ordinary smooth error density as the proof for super-smooth density follows by the same arguments. Assume that that the smoothness parameter in the former case is β. Let

$$
\mathcal{G}^D_{n,m}(t)=h^{\beta-1/2}\left\{\int\bar{\mathbb{L}}\left(\frac{t-a}{h}\right)f_X(a)^{1/2}dW_1(a)-\sqrt{\frac{n}{m}}\int\bar{\mathbb{L}}\left(\frac{t-a}{h}\right)f_Y(a)^{1/2}dW_2(a)\right\},
$$

where W_1 and W_2 are two independent (two-sided) Wiener processes on $\mathbb{R}$ (for $f_\epsilon\neq f_\delta$ or unequal bandwidths, the $\bar{\mathbb{L}}$ functions in the above integrals would also be different). Also

define

$$\Psi_{n,m}(t) = \{\hat{F}_{X^*}(t) - F_{X^*}(t)\} - \{\hat{F}_{Y^*}(t) - F_{Y^*}(t)\},$$
$$\Psi^{\#}_{n,m}(t) = \{\hat{F}^{\#}_{X^*}(t) - \hat{F}_{X^*}(t)\} - \{\hat{F}^{\#}_{Y^*}(t) - \hat{F}_{Y^*}(t)\}.$$

C.1.5.1 Proof of (i)

Since the samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$ are independent of each other, by the arguments of Lemmas (17)-(19), we can show the following: For some sequences $\epsilon_n, \delta_n = O(n^{-c})$,

$$P\left\{\sup_{t\in\mathcal{T}}\left|\sqrt{n}h^{\beta-1/2}\Psi_{n,m}(t) - \mathcal{G}^D_{n,m}(t)\right| > \epsilon_n\right\} < \delta_n. \tag{C.13}$$

Furthermore with probability greater than $1-\delta_{2n}$, $\delta_{2n} = O(n^{-c})$, there exist sequences $\epsilon_{1n}, \delta_{1n} = O(n^{-c})$ such that

$$P^{\#}\left\{\sup_{t\in\mathcal{T}}\left|\sqrt{n}h^{\beta-1/2}\Psi^{\#}_{n,m}(t) - \tilde{\mathcal{G}}^{D\#}_{n,m}(t)\right| > \epsilon_{1n}\right\} < \delta_{1n}, \tag{C.14}$$

where $\tilde{\mathcal{G}}^{D\#}_{n,m}$ is a tight Gaussian process with the same distribution as $\mathcal{G}^D_{n,m}$ under $P^{\#}$. Finally it also holds that

$$p_{\epsilon_n}(\mathcal{G}^D_{n,m}) \leq M\epsilon_n\sqrt{\log(1/h)}, \tag{C.15}$$

for any sequence $\epsilon_n = O(n^{-c})$ and some $M < \infty$. Now

$$P\left\{D_{n,m} \leq \hat{c}^D_\alpha\right\} \geq P\left\{\sup_{t\in\mathcal{T}}\Psi_{n,m}(t) - \sup_t\{F_{X^*}(t) - F_{Y^*}(t)\} \leq \hat{c}^D_\alpha\right\} \geq P\left\{\sup_{t\in\mathcal{T}}\Psi_{n,m}(t) \leq \hat{c}^D_\alpha\right\},$$

where the last equality follows from $\sup_t\{F_{X^*}(t) - F_{Y^*}(t)\} \leq 0$ under H_0. Using equations (C.13)-(C.15), by paralleling the arguments in the proof of Theorem 6, we can show that

$$P\left\{\sup_{t\in\mathcal{T}}\Psi_{n,m}(t) \leq \hat{c}^D_\alpha\right\} \geq 1 - \alpha - \varrho_{n,m}.$$

Hence the claim follows immediately.

C.1.5.2 Proof of (ii)

It is enough to show that $\rho_{n,m}$ does not depend on $P \in \mathcal{P}_0$. To this end, it is enough to show uniform validity of equations (C.13)-(C.15). Since these equations are essentially two-sample counterparts of Lemmas (17)-(19), it suffices to check uniform validity of the latter.

Note that for Lemma (17), uniformity of the bias term follows by the argument in Hall and Lahiri (2008, Theorem 3.2) using the uniform version of the Sobolev condition (i.e. the constants M_X and M_Y do not depend of (F_{X^*}, F_{Y^*})). For the stochastic term, the constants appearing in the KMT coupling in the proof of Lemma (17) are universal, and constants and sequences in other parts do not depend on $P \in \mathcal{P}_0$. Thus, δ_n in Lemma (17) does not depend on $P \in \mathcal{P}_0$. Similarly, uniformity of Lemma (18) is also verified.

For Lemma (19), it is enough to guarantee that $\sigma_n(t)$ is bounded away from zero and above by universal constants that do not depend on $P \in \mathcal{P}_0$. This is guaranteed by the assumption that f_X and f_Y are bounded away from zero and above by universal constants that do not depend on $P \in \mathcal{P}_0$.

C.1.5.3 Proof of (iii)

Let c_a^D be a constant such that $\sqrt{n}h^{\beta-1/2}c_a^D$ is the $(1-a)$-th quantile of $\sup_{t\in\mathcal{T}} \mathcal{G}_{n,m}^D(t)$. Using equation (C.14) and mirroring the arguments in the proof of Theorem 6, we have that

$$P\left\{\sqrt{n}h^{\beta-1/2}\hat{c}_\alpha^D < \sqrt{n}h^{\beta-1/2}c_{\alpha-\delta_{1n}}^D + \epsilon_{1n}\right\} \geq 1 - \delta_{2n}. \tag{C.16}$$

Under H_1, there exists $t^* \in \mathcal{T}$ such that $\mu = F_{X^*}(t^*) - F_{Y^*}(t^*) > 0$. Then we obtain

$$\begin{aligned} P\{D_{n.m} > \hat{c}_\alpha^D\} &\geq P\left\{\sqrt{n}h^{\beta-1/2}D_{n,m} > \sqrt{n}h^{\beta-1/2}c_{\alpha-\delta_{1n}}^D + \epsilon_{1n}\right\} - \delta_{2n} \\ &\geq P\left\{\mathcal{G}_{n,m}^D(t^*) > \sqrt{n}h^{\beta-1/2}c_{\alpha-\delta_{1n}}^D - \sqrt{n}h^{\beta-\frac{1}{2}}\mu + \epsilon_{1n} + \epsilon_n\right\} - \delta_{2n} - \delta_n, \end{aligned}$$

for some $\epsilon_n, \delta_n = O(n^{-c'})$ with some $c' > 0$, where the first inequality follows from (C.16) and the second inequality follows from (C.13). By analogous arguments as in the proof of Theorem 10, we can show $\sqrt{n}h^{\beta-1/2}c_{\alpha-\delta_{1n}}^D = O(\sqrt{\log(1/h)})$. However under Assumption

OS (iii), $\sqrt{n}h^{\beta-1/2}\sqrt{\log(1/h)}\mu \to +\infty$; hence the conclusion follows immediately.

C.2 Lemmas

Hereafter we use the following notation. By the Ito isometry, the variance function of the Gaussian process $\mathcal{G}_n$ can be shown to be

$$\sigma_n(t) = hr^2(h) \int \bar{\mathbb{L}}^2(a)\, f_X(t - ha)da.$$

Let $\bar{\sigma}_n = \sup_t \sigma_n(t)$ and $\underline{\sigma}_n = \inf_t \sigma_n(t)$. Assumption C (i) ($\inf_{t\in\mathcal{T}} f_X(t) > c > 0$) guarantees that $\underline{\sigma}_n > 0$ for all $n \in \mathbb{N}$.

Also, define the variance sub-metric $d_n(s,t) = Var(\mathcal{G}_n(s) - \mathcal{G}_n(t))$ on $\mathcal{T}$.

C.2.1 Lemmas for Theorem 6 under Assumption OS

Lemma 17. *Under Assumptions C and OS, there exist sequences $\epsilon_n, \delta_n = O(n^{-c})$ for some $c > 0$ such that*

$$P\left\{\sup_{t\in\mathcal{T}} \left|\sqrt{n}h^{\beta-1/2}\{\hat{F}_{X^*}(t) - F_{X^*}(t)\} - \mathcal{G}_n(t)\right| > \epsilon_n\right\} < \delta_n.$$

Proof. By applying the argument in Hall and Lahiri (2008), the bias of the estimator $\hat{F}_{X^*}$ satisfies $\sup_{t\in\mathcal{T}} |E[\hat{F}_{X^*}(t)] - F_{X^*}(t)| = O(h^\gamma)$. Thus, Assumption OS (iii) guarantees

$$\sqrt{n}h^{\beta-1/2} \sup_{t\in\mathcal{T}} |E[\hat{F}_{X^*}(t)] - F_{X^*}(t)| = o(n^{-\xi}).$$

So, the bias term is negligible and it is enough to show that

$$P\left\{\sup_{t\in\mathcal{T}} \left|\sqrt{n}h^{\beta-1/2}\{\hat{F}_{X^*}(t) - E[\hat{F}_{X^*}(t)]\} - \mathcal{G}_n(t)\right| > \epsilon_n\right\} < \delta_n, \tag{C.17}$$

for some $\epsilon_n, \delta_n = O(n^{-c})$ with $c > 0$. Let $F_{X,n}^{EDF}$ be the empirical distribution function by $\{X_i\}_{i=1}^n$, $\alpha_n(x) = \sqrt{n}\{F_{X,n}^{EDF}(x) - F_X(x)\}$ be the empirical process, and

$$D_n(t) = \sqrt{n}h^{\beta-1/2}\{\hat{F}_{X^*}(t) - E[\hat{F}_{X^*}(t)]\} = h^{\beta-1/2} \int \mathbb{L}\left(\frac{t-a}{h}\right) d\alpha_n(a).$$

Then (C.17) is rewritten as

$$P\left\{\sup_{t\in\mathcal{T}}|D_n(t)-\mathcal{G}_n(t)|>\epsilon_n\right\}<\delta_n, \tag{C.18}$$

for some $\epsilon_n, \delta_n = O(n^{-c})$ with $c > 0$.

First, we approximate $D_n(t)$ by

$$D_{n,0}(t) = h^{\beta-1/2}\int \bar{\mathbb{L}}\left(\frac{t-a}{h}\right)d\alpha_n(a),$$

Note that both $D_n(t)$ and $D_{n,0}(t)$ are well defined as Lebesgue-Steltjes integrals.[3] From integration by parts,

$$\begin{aligned} D_n(t) &= h^{\beta-3/2}\int \mathbb{K}\left(\frac{t-a}{h}\right)\alpha_n(a)da \\ &\quad +h^{\beta-1/2}\lim_{a\to\infty}\left\{\mathbb{L}\left(\frac{t-a}{h}\right)\alpha_n(a)\right\}-h^{\beta-1/2}\lim_{a\to-\infty}\left\{\mathbb{L}\left(\frac{t-a}{h}\right)\alpha_n(a)\right\} \\ &= h^{\beta-3/2}\int \mathbb{K}\left(\frac{t-a}{h}\right)\alpha_n(a)da, \end{aligned} \tag{C.19}$$

for all $n \in \mathbb{N}$, where the second equality follows from the facts $\lim_{a\to\pm\infty}\alpha_n(a) = 0$ and $\sup_u|\mathbb{L}(u)| < \infty$ for each h. Since a similar expression applies for $D_{n,0}(t)$, there exists $C > 0$ such that

$$D_n(t)-D_{n,0}(t) = h^{\beta-1/2}\int\{\mathbb{K}(u)-\bar{\mathbb{K}}(u)\}\alpha_n(u-th)du \le Ch^s\sup_u|\alpha_n(u)|,$$

for all n large enough and $t \in \mathcal{T}$, where the inequality follows from Assumption OS (ii). Now by the strong approximation (Komlós, Major and Tusnády, 1975), there exists a tight Brownian bridge $B(t) = W(t) - tW(1)$ and universal constants $C_1, C_2 > 0$ such that

$$P\left\{\sup_u|\alpha_n(u)|\le\sup_u|B(F_X(u))|+C_1\frac{\log n}{\sqrt{n}}\right\}\ge 1-\frac{C_2}{n},$$

for all $n \in \mathbb{N}$. Combining theses results and using the properties of $\sup_u|B(F_X(u))|$ (in

[3]This is verified as follows: By the definition $\mathbb{L}(u) = \int_0^u \mathbb{K}(v)dv$ and Assumption SS (ii), we have $\sup_u|\mathbb{L}(u)| < \infty$. Also, by $\bar{\mathbb{L}}(u) = \int_0^u \bar{\mathbb{K}}(v)dv$ (follows from Fubini's theorem) and Assumption SS (ii), we have $\sup_u|\bar{\mathbb{L}}(u)| < \infty$. Therefore, bounded variation of the empirical process α_n guarantees that both $D_n(t)$ and $D_{n,0}(t)$ are well defined.

particular, $P\{\sup_u |B(F_X(u))| \geq x\} \leq 2\exp(-2x^2)$ for $x > 0$), there exists $C_3 > 0$ such that

$$P\left\{\sup_{t\in\mathcal{T}} |D_n(t) - D_{n,0}(t)| > h^{s/2}\right\} \leq C_3 \exp(-2h^{-s}) + \frac{C_2}{n},$$

for all n large enough. Note that $h^{s/2} = O(n^{-c_1})$ for some $c_1 > 0$ due to Assumption OS (iii) ($n^\nu h \to 0$). Thus, it is enough for (C.18) to show that

$$P\left\{\sup_{t\in\mathcal{T}} |D_{n,0}(t) - \mathcal{G}_n(t)| > \epsilon_n\right\} < \delta_n,$$

for some $\epsilon_n, \delta_n = O(n^{-c})$ with $c > 0$.

Second, we approximate $D_{n,0}(t)$ by

$$D_{n,1}(t) = h^{\beta-1/2} \int \bar{\mathbb{L}}\left(\frac{t-a}{h}\right) dB(F_X(a)).$$

Since $\bar{\mathbb{L}} \in L_2(\mathbb{R})$, this integral exists for all $t \in \mathbb{R}$. Analogous to the integration by parts formula in (C.19), a similar result applies for $D_{n,1}(t)$ based on stochastic integration by parts using the facts $\lim_{u\to\pm\infty} \bar{\mathbb{L}}(u) = 0$ and $\sup_a |B(F_X(a))| < \infty$ almost surely. Thus, we have

$$\begin{aligned} D_{n,0}(t) - D_{n,1}(t) &= h^{\beta-3/2} \int \bar{\mathbb{K}}\left(\frac{t-a}{h}\right) \{\alpha_n(a) - B(F_X(a))\} da \\ &\leq h^{\beta-1/2} \sup_a |\alpha_n(a) - B(F_X(a))| \int |\bar{\mathbb{K}}(u)| du, \end{aligned}$$

for all $n \in \mathbb{N}$, almost surely. Now by Komlós, Major and Tusnády (1975), there exist Brownian bridge B with continuous sample paths and universal constants $C_4, C_5 > 0$ such that

$$P\left\{\sup_{a\in\mathbb{R}} |\alpha_n(a) - B(F_X(a))| > C_4 \frac{\log n}{\sqrt{n}}\right\} \leq \frac{C_5}{n},$$

for all $n \in \mathbb{N}$. Combining this with Assumption OS (ii), there exist $c_2, C_6 > 0$ such that

$$P\left\{\sup_{t\in\mathcal{T}} |D_{n,0}(t) - D_{n,1}(t)| > C_6 n^{-c_2}\right\} \leq \frac{C_5}{n},$$

for all n large enough. Thus, it is enough for (C.18) to show that

$$P\left\{\sup_{t\in\mathcal{T}}|D_{n,1}(t)-\mathcal{G}_n(t)|>\epsilon_n\right\}<\delta_n,$$

for some $\epsilon_n,\delta_n=O(n^{-c})$ with $c>0$.

Third, we approximate $D_{n,1}(t)$ by

$$D_{n,2}(t)=h^{\beta-1/2}\int\bar{\mathbb{L}}\left(\frac{t-a}{h}\right)dW(F_X(a)).$$

By the definition $B(t)=W(t)-tW(1)$, we have

$$|D_{n,1}(t)-D_{n,2}(t)| \quad\leq\quad h^{\beta-1/2}|W(1)|\left|\int\bar{\mathbb{L}}\left(\frac{t-a}{h}\right)f_X(a)da\right|, \tag{C.20}$$

for all $n\in\mathbb{N}$. Therefore, for the rate of $\sup_{t\in\mathcal{T}}|D_{n,1}(t)-D_{n,2}(t)|$, we need to characterize the order of $I_{n1}(t)=\int\bar{\mathbb{L}}\left(\frac{t-a}{h}\right)f_X(a)da$. By the definition of $\bar{\mathbb{L}}$ and

$$\int_{-\infty}^{\infty}\sin(\omega(t-a))f_X(a)da=\frac{1}{2\mathrm{i}}\{e^{\mathrm{i}\omega t}f_X^{\mathrm{ft}}(-\omega)-e^{-\mathrm{i}\omega t}f_X^{\mathrm{ft}}(\omega)\},$$

an application of Fubini's theorem assures

$$\begin{aligned}|I_{n1}(t)| &= \left|\frac{1}{2\mathrm{i}\pi}\int_{\omega_0}^{1/h}\{e^{\mathrm{i}\omega t}f_X^{\mathrm{ft}}(-\omega)-e^{-\mathrm{i}\omega t}f_X^{\mathrm{ft}}(\omega)\}\frac{K^{\mathrm{ft}}(h\omega)}{\omega f_\epsilon^{\mathrm{ft}}(\omega)}d\omega\right| \\ &\leq \frac{1}{\pi}\int_{\omega_0}^{1/h}\omega^{-1}d\omega=O(\log(1/h)).\end{aligned}$$

where the inequality follows from $|f_X^{\mathrm{ft}}|=|f_{X^*}^{\mathrm{ft}}||f_\epsilon^{\mathrm{ft}}|\leq|f_\epsilon^{\mathrm{ft}}|$ and $f_\epsilon^{\mathrm{ft}}(\omega)=f_\epsilon^{\mathrm{ft}}(-\omega)$. Substituting this bound for $I_{n1}(t)$ into (C.20), we obtain

$$P\left\{\sup_{t\in\mathcal{T}}|D_{n,1}(t)-D_{n,2}(t)|>M_nh^{\beta-1/2}\log(1/h)\right\}=O(n^{-c_3}),$$

for some $c_3>0$ and sequence $M_n=\log n$. By Assumption OS (i) ($\beta>1/2$), it holds $M_nh^{\beta-1/2}\log(1/h)=O(n^{-c_4})$ for some $c_4>0$. Therefore, it is enough for (C.18) to show

that

$$P\left\{\sup_{t\in\mathcal{T}}|D_{n,2}(t)-\mathcal{G}_n(t)|>\epsilon_n\right\}<\delta_n,$$

for some $\epsilon_n,\delta_n=O(n^{-c})$ with $c>0$. But we can see that the process $D_{n,2}(t)$ has the same finite dimensional distributions as the process $\mathcal{G}_n(t)$. Therefore, this trivially holds true and the conclusion is obtained. □

Lemma 18. *Under Assumptions C and OS, there exist sequences $\epsilon_{1n},\delta_{1n},\delta_{2n}=O(n^{-c})$ for some $c>0$ such that with probability greater than $1-\delta_{2n}$,*

$$P^{\#}\left\{\sup_{t\in\mathcal{T}}\left|\sqrt{n}h^{\beta-1/2}\{\hat{F}^{\#}_{X^*}(t)-\hat{F}_{X^*}(t)\}-\tilde{\mathcal{G}}_n(t)\right|>\epsilon_{1n}\right\} \quad < \quad \delta_{1n},$$

where $\tilde{\mathcal{G}}_n$ is a tight Gaussian process with the same distribution as $\mathcal{G}_n$ under $P^{\#}$.

Proof. The proof is essentially a reformulation of that of Bissantz, Dümbgen, Holzmann and Munk (2007, Theorem 2.1). Let $\alpha^{\#}_n(t)=\sqrt{n}\{F^{EDF}_{X^{\#},n}-F^{EDF}_{X,n}(t)\}$ denote the bootstrap empirical process. As shown in the proof of Bissantz, Dümbgen, Holzmann and Munk (2007, eq. (21)), based on Shorack (1982), there exists a Brownian bridge $B^{\#}_n$ and universal constants $C,C_1>0$ such that for all $n\in\mathbb{N}$,

$$P^{\#}\left\{\sup_{t\in\mathbb{R}}|\alpha^{\#}_n(t)-B^{\#}_n(F^{EDF}_{X,n}(t))|>C\frac{\log n}{\sqrt{n}}\right\}\leq\frac{C_1}{n},$$

almost surely. Now it is known that the Brownian bridge is Hölder continuous for every exponent $b\in(0,1/2)$ almost surely. Furthermore, by Komlós, Major and Tusnády's (1975) coupling, along with the fact $P\{\sup_t|B(F_X(t))|\geq\log n\}\leq 2\exp(-2(\log n)^2)$, there exist universal constants $C_2,C_3>0$ such that

$$P\left\{\sup_{t\in\mathbb{R}}|F^{EDF}_{X,n}(t)-F_X(t)|>C_2\frac{\log n}{\sqrt{n}}\right\}\leq\frac{C_3}{n},$$

for all $n\in\mathbb{N}$, which consequently implies

$$P\left\{\sup_{t\in\mathbb{R}}|B^{\#}_n(F^{EDF}_{X,n}(t))-B^{\#}_n(F_X(t))|>C_4\frac{\log n}{n^{b/2}}\right\}\leq\frac{C_5}{n},$$

for some universal constants $C_4, C_5 > 0$. Combining these results, there exist universal constants $C_6, C_7, C_8 > 0$ such that with probability greater than $1 - C_6/n$, it holds

$$P^{\#}\left\{\sup_{t\in\mathbb{R}} |\alpha_n^{\#}(t) - B_n^{\#}(F_X(t))| > C_7 \frac{\log n}{n^{b/2}}\right\} \leq \frac{C_8}{n},$$

for all $n \in \mathbb{N}$. Based on this, the conclusion follows by similar arguments as in the proof of Lemma 17. □

Lemma 19. *Suppose that Assumptions C and OS hold true. Then for any sequence $\epsilon_n = O(n^{-c})$ with $c > 0$, there exists a constant $M > 0$ such that*

$$p_{\epsilon_n}(\mathcal{G}_n) \leq M\epsilon_n\sqrt{\log(1/h)},$$

for all n large enough.

Proof. Pick any $\varepsilon > 0$. By Chernozhukov, Chetverikov and Kato (2015, Theorem 3) and separability of $\mathcal{G}_n$, there exists $C > 0$ such that

$$p_\varepsilon(\mathcal{G}_n) \leq C\varepsilon\left\{\underline{\sigma}_n^{-1} E\left[\sup_{t\in\mathcal{T}} |\mathcal{G}_n(t)|\right] + \sqrt{1 \vee \log(\underline{\sigma}_n/\varepsilon)}\right\},$$

for all $n \in \mathbb{N}$. Thus, it is enough to show that

$$E\left[\sup_{t\in\mathcal{T}} |\mathcal{G}_n(t)|\right] = O(\sqrt{\log(1/h)}).$$

Now,

$$d_n^2(s,t) = h^{2\beta}\int\left\{\bar{\mathbb{L}}\left(\frac{s}{h} - a\right) - \bar{\mathbb{L}}\left(\frac{t}{h} - a\right)\right\}^2 f_X(ha)da$$

by the Ito isometry. Note that $\bar{\mathbb{L}}$ is Lipschitz continuous because its derivative $\bar{\mathbb{K}}$ is uniformly bounded on $\mathbb{R}$ (because $h^\beta \sup_u |\bar{\mathbb{K}}(u)| \leq C$ for some $C > 0$ by Assumption OS (i)). Thus, it holds

$$d_n(s,t) \leq C_1 h^{-3/2}|s - t|, \tag{C.21}$$

for some $C_1 > 0$ that is independent of s and t.

Let $D(\varepsilon, d_n)$ be the ε-packing number for the set $\mathcal{T}$ under the sub-metric d_n. By (C.21),

it holds $D(\varepsilon, d_n) \leq 2C_1 h^{-3/2}/\varepsilon$. Pick any $\delta \in (0,1)$. By van der Vaart and Wellner (1996, Corollary 2.2.8), there exist universal constants $C_2, C_3 > 0$ such that

$$\begin{aligned} & E\left[\sup_{d_n(s,t)\leq\delta} |\mathcal{G}_n(s) - \mathcal{G}_n(t)|\right] \\ \leq \quad & C_2 \int_0^\delta \sqrt{\log D(\varepsilon, d_n)} d\varepsilon \leq C_2 \delta \sqrt{\log(2C_1 h^{-3/2})} + C_2 \int_0^\delta \sqrt{\log(1/\varepsilon)} d\varepsilon \leq C_3 \sqrt{\log(1/h)}, \end{aligned}$$

for all $n \in \mathbb{N}$. Thus, there exists a collection of Gaussian random variables $\{\mathcal{G}_n(t_j)\}_{i=1}^{p_n}$ with $p_n = \left\lceil \frac{1}{h^{3/2}\delta} \right\rceil$ such that

$$E\left[\sup_{t\in\mathcal{T}} |\mathcal{G}_n(t)|\right] \leq E\left[\max_{1\leq j\leq p_n} |\mathcal{G}_n(t_j)|\right] + C_3\sqrt{\log(1/h)},$$

for all $n \in \mathbb{N}$. Now the properties of the maximum of Gaussian random variables yields

$$E\left[\max_{1\leq j\leq p_n} |\mathcal{G}_n(t_j)|\right] \leq 2\bar{\sigma}_n \sqrt{1 + \log p_n}.$$

Combining these results, the conclusion follows. □

C.2.2 Lemmas for Theorem 6 under Assumption SS

Lemma 20. *Under Assumptions C and SS, there exist sequences $\epsilon_n = O(\log n)^{-c}$ and $\delta_n = O(n^{-c})$ with $c > 0$ such that*

$$P\left\{\sup_{t\in\mathcal{T}} \left| \frac{\sqrt{n}}{\varsigma(h)} \{\hat{F}_{X^*}(t) - F_{X^*}(t)\} - \mathcal{G}_n(t) \right| > \epsilon_n \right\} < \delta_n.$$

Lemma 21. *Under Assumptions C and OS, there exist sequences $\epsilon_{1n}, \delta_{1n}, \delta_{2n} = O(n^{-c})$ with $c > 0$ such that with probability greater than $1 - \delta_{2n}$,*

$$P^{\#}\left\{\sup_{t\in\mathcal{T}} \left| \frac{\sqrt{n}}{\varsigma(h)} \{\hat{F}_X^{\#}(t) - \hat{F}_X(t)\} - \tilde{\mathcal{G}}_n(t) \right| > \epsilon_{1n} \right\} \quad < \quad \delta_{1n},$$

where $\tilde{\mathcal{G}}_n$ is a tight Gaussian process with the same distributions as $\mathcal{G}_n$ under $P^{\#}$.

These lemmas can be shown in the same way as Lemmas 17 and 18. The log rate of ϵ_n in Lemma 20 is due to the bias term. Recall that under Assumption C (ii), the bias of the

estimator $\hat{F}_{X^*}$ is given by

$$\sup_{t\in\mathcal{T}} |E[\hat{F}_{X^*}(t)] - F_{X^*}(t)| = O(h^\gamma).$$

Then due to Assumption SS (iii), it holds $\sqrt{n}h^\gamma/\varsigma(h) = C(\log n)^{-c}$ for some $c > 1$.

Lemma 22. *Suppose that Assumptions C and SS hold true. Then for any sequence $\epsilon_n = O(n^{-c})$ with $c > 0$ and any $r > 0$, there exists $M > 0$ such that*

$$p_{\epsilon_n}(\mathcal{G}_n) \leq M\epsilon_n(\log n)^{1+r},$$

for all n large enough.

Proof. Pick any $\varepsilon > 0$. By Chernozhukov, Chetverikov and Kato (2015, Theorem 3) and separability of the Gaussian process $\mathcal{G}_n$, there exists $C > 0$ such that

$$p_\varepsilon(\mathcal{G}_n) \leq C\varepsilon\left\{\underline{\sigma}_n^{-1} E\left[\sup_{t\in\mathcal{T}} |\mathcal{G}_n(t)|\right] + \sqrt{1 \vee \log(\underline{\sigma}_n/\varepsilon)}\right\},$$

for all $n \in \mathbb{N}$. By Lemmas 23 and 24 shown below, the following hold true:

$$\text{there exist } c_1 > 0 \text{ such that } \underline{\sigma}_n \geq c_1 h^{\lambda+\nu} \text{ for all } \nu > 0 \text{ and } n \text{ large enough,} \tag{C.22}$$

$$\text{there exist } C_1 > 0 \text{ such that } \bar{\sigma}_n \leq C_1 \text{ for all } n \text{ large enough.} \tag{C.23}$$

Observe that

$$d_n^2(s,t) = \frac{h}{\varsigma^2(h)} \int \left\{\bar{\mathbb{L}}\left(\frac{s}{h} - a\right) - \bar{\mathbb{L}}\left(\frac{t}{h} - a\right)\right\}^2 f_X(ha)da$$

by the Ito isometry. Note that $\bar{\mathbb{L}}$ is Lipschitz continuous because its derivative $\bar{\mathbb{K}}$ is uniformly bounded on $\mathbb{R}$ (because $\sqrt{h}\varsigma^{-1}(h)\sup_u |\bar{\mathbb{K}}(u)| \leq C_2 h^{-c_2}$ for some $C_2, c_2 > 0$ by Assumption SS (i)). Thus, it holds $d_n(s,t) \leq C_3 h^{-c_2-3/2}|s-t|$ for some $C_3 > 0$ that is independent of s and t. Using (C.23), an analogous argument as in the proof of Lemma 19 shows that $E\left[\sup_{t\in\mathcal{T}} |\mathcal{G}_n(t)|\right] = O(\sqrt{\log(1/h)})$. Combining this with (C.22) and Assumption SS (iii), the conclusion follows. □

Lemma 23. *Under Assumptions C and SS, there exists $c > 0$ such that $\underline{\sigma}_n \geq ch^{\lambda+\nu}$ for all $\nu > 0$ and n large enough.*

Proof. We only prove the case of $\lambda_0 \geq 0$. The proof for the case of $\lambda_0 < 0$ is similar. Pick any $\varepsilon > 0$. By Assumption C (i), we provide a lower bound for $\underline{\sigma}_n$ via

$$\underline{\sigma}_n = \inf_{t\in\mathcal{T}} \frac{h}{\varsigma^2(h)} \int \bar{\mathbb{L}}^2(a) f_X(t - ha)da \geq \frac{c_1 h}{\varsigma^2(h)} \int_{|a|\leq h^\varepsilon} \bar{\mathbb{L}}^2(a)da,$$

for some $c_1 > 0$. Let

$$\Phi_\epsilon(\omega) = f_\epsilon^{\mathrm{ft}}(\omega)^{-1}\mathbb{I}\{|\omega| \geq \omega_0\}.$$

Using the fact $\sin(x) = x + R(x)$ with $|R(x)| \leq c_2|x|^2$ for some $c_2 > 0$, it follows that for all $|a| \leq h^\varepsilon$,

$$|\bar{\mathbb{L}}(a)| \geq \frac{1}{\pi}\left|a\int_0^1 K^{\mathrm{ft}}(\omega)\Phi_\epsilon\left(\frac{\omega}{h}\right)d\omega\right| - \frac{c_2}{\pi}\left|a\int_0^1 |a\omega|K^{\mathrm{ft}}(\omega)\Phi_\epsilon\left(\frac{\omega}{h}\right)d\omega\right| \geq C\{1-O(h^\varepsilon)\}|aI_n|,$$

where $I_n = \int_0^1 K^{\mathrm{ft}}(\omega)\Phi_\epsilon\left(\frac{\omega}{h}\right)d\omega$ and the last inequality follows from the fact $\sup\{|a\omega| : |a| \leq h^\varepsilon, \omega \in [0,1]\} = h^\varepsilon$.

We now provide a lower bound for I_n. Pick any $\delta > 0$. Observe that

$$\begin{aligned}
h^{\frac{1-\lambda}{2}}\varsigma(h)^{-\frac{1}{2}}|I_n| &= \frac{\exp(-1/\mu h^\lambda)}{h^{\lambda(s+1)+\lambda_0}}\int_{h\omega_0}^1 K^{\mathrm{ft}}(\omega)\Phi_\epsilon\left(\frac{\omega}{h}\right)d\omega \\
&\geq c_3\frac{\exp(-1/\mu h^\lambda)}{h^{\lambda(s+1)}}\int_{h\omega_0}^1 K^{\mathrm{ft}}(\omega)\omega^{-\lambda_0}\exp\left(\frac{|\omega|^\lambda}{h^\lambda\mu}\right)d\omega \\
&\geq c_3\frac{\exp(-1/\mu h^\lambda)}{h^{\lambda(s+1)}}\int_{\delta}^1 K^{\mathrm{ft}}(\omega)\exp\left(\frac{|\omega|^\lambda}{h^\lambda\mu}\right)d\omega \\
&= c_3\int_0^{(1-\delta)h^{-\lambda}} \frac{K^{\mathrm{ft}}(1-h^\lambda v)}{(h^\lambda v)^s}v^s\exp\left(\frac{|1-h^\lambda v|^\lambda - 1}{h^\lambda\mu}\right)dv \\
&\to c_3 r^s\int v^s\exp(-\lambda v/\mu)dv > 0,
\end{aligned}$$

for some $c_3 > 0$, where the first inequality follows from the fact $\Phi_\epsilon(\omega) \geq c_3|\omega|^{-\lambda_0}\exp(|\omega|^\lambda/\mu)$, the second inequality holds since all the terms inside the integral are positive and $\omega^{-\lambda_0}\mathbb{I}\{h\omega_0 \leq \omega \leq 1\} \geq 1$ for $\lambda_0 \geq 0$, the second equality follows from a change of variables, and the con-

vergence follows from the dominated convergence theorem after noting

$$\begin{aligned}
&\frac{K^{\mathrm{ft}}(1-h^\lambda v)}{(h^\lambda v)^s} v^s \exp\left(\frac{|1-h^\lambda v|^\lambda - 1}{h^\lambda \mu}\right) \mathbb{I}\{0 \le v \le (1-\delta)h^{-\lambda}\} \\
\le\ & \begin{cases} \sup_{0\le t\le 1}\{t^{-s}K^{\mathrm{ft}}(1-t)\} v^s \exp(-v/\mu) & \text{if } \lambda \ge 1, \\ \sup_{0\le t\le 1}\{t^{-s}K^{\mathrm{ft}}(1-t)\} v^s \exp(-\lambda v/\mu) & \text{if } 0 < \lambda < 1. \end{cases}
\end{aligned}$$

Thus, it holds $h^{1/2}\varsigma(h)^{-1/2}|I_n| > c_3 h^{\lambda/2}$ for all n large enough.

Combining these results, there exists $c > 0$ such that

$$\underline{\sigma}_n \ge ch^\lambda \int_{|a|\le h^\varepsilon} |a|^2 da \ge ch^{\lambda+3\varepsilon},$$

for all n large enough, and the conclusion follows. □

Lemma 24. *Under Assumptions C and SS, there exists $C > 0$ such that $\bar{\sigma}_n \le C$ for all n large enough.*

Proof. We only prove the case of $\lambda_0 \ge 0$. The proof for the case of $\lambda_0 < 0$ is similar. Pick any $\varepsilon \in (0, 2^{-1/\lambda})$. Since f_X is bounded (Assumption C (ii)), there exists $C_1, C_2 > 0$ such that

$$\begin{aligned}
\bar{\sigma}_n &\le C_1 \frac{\exp(-2/\mu h^\lambda)}{h^{\lambda(2s+1)+2\lambda_0}} \int \bar{\mathbb{L}}^2(a) da = C_2 \frac{\exp(-2/\mu h^\lambda)}{h^{\lambda(2s+1)+2\lambda_0}} \int_{h\omega_0}^1 \left| \frac{K^{\mathrm{ft}}(\omega)}{\omega} \Phi_\epsilon\left(\frac{\omega}{h}\right)\right|^2 d\omega \\
&\le C_2 \omega_0^{-2} \frac{\exp(-2/\mu h^\lambda)}{h^{\lambda(2s+1)+2(1+\lambda_0)}} \int_{h\omega_0}^1 \left| K^{\mathrm{ft}}(\omega) \Phi_\epsilon\left(\frac{\omega}{h}\right)\right|^2 d\omega \\
&\le C_2 \omega_0^{-2} \frac{\exp(-2/\mu h^\lambda)}{h^{\lambda(2s+1)+2(1+\lambda_0)}} \int_{h\omega_0}^{\varepsilon} \left| K^{\mathrm{ft}}(\omega) \left(\frac{\omega}{h}\right)^{-(1+\lambda_0)} \exp\left(\frac{|\omega|^\lambda}{h^\lambda \mu}\right)\right|^2 d\omega \\
&\quad + C_2 \omega_0^{-2} \frac{\exp(-2/\mu h^\lambda)}{h^{\lambda(2s+1)}} \int_{|\omega|>\varepsilon} \left| K^{\mathrm{ft}}(\omega) \omega^{-(1+\lambda_0)} \exp\left(\frac{|\omega|^\lambda}{h^\lambda \mu}\right)\right|^2 d\omega \\
&= T_{1n} + T_{2n},
\end{aligned}$$

for all n large enough, where the first equality follows from Plancherel's isometry,[4] and the second inequality follows from $\Phi_\epsilon(\omega) \le C|\omega|^{-\lambda_0} \exp(|\omega|^\lambda/\mu)$. For T_{1n}, Assumption SS (iii)

[4]Note that $\bar{\mathbb{L}}$ is written as $\bar{\mathbb{L}}(u) = \frac{1}{2\pi}\int_{-1}^1 \frac{e^{-\mathrm{i}\omega u}}{\omega} \frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)} \mathbb{I}\{|\omega| \ge \omega_0\} d\omega$. This integral exists due to the truncation.

and the restriction $\varepsilon \in (0, 2^{-1/\lambda})$ guarantee

$$\begin{aligned} T_{1n} &\leq C_3 \omega_0^{-(1+\lambda_0)} \frac{\exp(-2/\mu h^\lambda)}{h^{\lambda(2s+1)+2(1+\lambda_0)}} \int_{h\omega_0}^{\varepsilon} \left| K^{\mathrm{ft}}(\omega) \exp\left(\frac{|\omega|^\lambda}{h^\lambda \mu}\right)\right|^2 d\omega \\ &\leq C_4 \frac{\exp(-1/\mu h^\lambda)}{h^{\lambda(2s+1)+2(1+\lambda_0)}} = O(n^{-c_1}), \end{aligned}$$

for some $C_3, C_4, c_1 > 0$. For T_{2n}, note that

$$T_{2n} \leq C_5 \varepsilon^{-(1+\lambda_0)} \frac{\exp(-2/\mu h^\lambda)}{h^{\lambda(2s+1)}} \int_{|\omega|>\varepsilon} \left| K^{\mathrm{ft}}(\omega) \exp\left(\frac{|\omega|^\lambda}{h^\lambda \mu}\right)\right|^2 d\omega,$$

for some $C_5 > 0$. By an analogous dominated convergence argument used in the proof of Lemma 23, we can show T_{2n} converges to some finite constant. Combining these results, the conclusion follows. □

C.3 Assumptions and proofs for Theorem 7

In this appendix we prove Theorem 7, on the asymptotic distribution of t_n. Basic steps of our proof follow the recipe laid down by Bissantz, Dümbgen, Holzmann and Munk (2007). Importantly, we impose tail conditions on f_ϵ^{ft} of the form $f_\epsilon^{\mathrm{ft}}(\omega)|\omega|^\beta \to C_\epsilon$ as $|\omega| \to \infty$. Based on this, we define

$$\begin{aligned} \mathcal{K}(u) &= \frac{1}{2\pi C_\epsilon} \int_0^\infty e^{-\mathrm{i}\omega u} \omega^\beta K^{\mathrm{ft}}(\omega) d\omega + \frac{1}{2\pi C_\epsilon} \int_{-\infty}^0 e^{-\mathrm{i}\omega u} |\omega|^\beta K^{\mathrm{ft}}(\omega) d\omega, \\ \mathcal{L}(u) &= \frac{1}{2\pi C_\epsilon} \int_0^\infty \sin(\omega u) \omega^{\beta-1} K^{\mathrm{ft}}(\omega) d\omega + \frac{1}{2\pi C_\epsilon} \int_{-\infty}^0 \sin(\omega u) |\omega|^\beta \omega^{-1} K^{\mathrm{ft}}(\omega) d\omega \end{aligned} \quad \text{(C.24)}$$

These are the pointwise limits of $h^\beta \mathbb{K}(u)$ and $h^\beta \mathbb{L}(u)$ as $h \to 0$ under some assumptions on f_ϵ^{ft}. In addition to Assumptions OS, we impose the following conditions.

Assumption G.

(i) $f_\epsilon^{\mathrm{ft}}(\omega)|\omega|^\beta \to C_\epsilon$ as $|\omega| \to \infty$ for some $\beta > 1/2$.

(ii) $h^\beta \int |\mathbb{K}(u)| du < M$ for some $M > 0$ independent of h. $\int |u|^{3/2} \sqrt{\log(\log^+ |u|)} |\mathcal{K}(u)| du < \infty$. For some $\delta > 0$, $\int |h^\beta \bar{\mathbb{K}}(u) - \mathcal{K}(u)| du = O(h^{1/2+\delta})$.

(iii) $\lim_{u\to\pm\infty} |\mathcal{L}(u)\sqrt{|u|\log(\log^+|u|)}| = 0$. For some $\delta_1 \in (0,1)$, $\int |\mathcal{L}(u)|^{2-\delta_1} du < \infty$. For some $\delta > 0$, $\sup_u |h^\beta \bar{\mathbb{L}}(u/h) - \mathcal{L}(u/h)| = O(h^{1/2+\delta})$.

(iv) f_X and its derivative f'_X are bounded and continuous on $\mathbb{R}$ such that $\lim_{x\to\pm\infty} |x f_X(x) \log(\log^+|x|)| = 0$. Also, $\sup_x |f'_X(x) f_X(x)^{-1/2} \sqrt{|x|\log(\log^+|x|)}| < \infty$. Furthermore it holds $\int |f'_X(x) f_X(x)^{-1/2} \sqrt{|x|\log(\log^+|x|)}| dx < \infty$.

These conditions are generalizations and simplifications of the ones in Bissantz, Dümbgen, Holzmann and Munk (2007). Assumption G (i) is stronger than the usual assumption $f_\epsilon^{\mathrm{ft}}(\omega)|\omega|^\beta < C_\epsilon$ as $|\omega| \to \infty$ but is required for explicit derivation of the limiting distribution.

Assumption G (ii) contains conditions for the deconvolution kernel $\mathbb{K}$. The first condition ensures that $\mathbb{K}$ is L_1-integrable. A sufficient condition for this is that $1/f_\epsilon^{\mathrm{ft}}(\omega)$ is a polynomial function in ω. Indeed in this case it can be shown from the properties of the Fourier transform that $|\mathbb{K}(u)| \sim |u|^{-q}$ as $|u| \to \infty$ under some conditions on f_ϵ^{ft}. For instance, the choice $r > 2$ for K assures $|\mathbb{K}(u)| \sim |u|^{-2}$ under the assumption

$$\int \left| \left\{ \frac{K^{\mathrm{ft}}(\omega)}{f_\epsilon^{\mathrm{ft}}(\omega/h)} \right\}'' \right| d\omega = O(h^{-\beta}).$$

A similar condition is given in, for example, Bissantz, Dümbgen, Holzmann and Munk (2007, eq. (13)). $\mathcal{K}$ in (C.24) is the limit of $\bar{\mathbb{K}}$ as $h \to \infty$ obtained by Assumption G (i). Recall that by Assumption OS (ii), $h^{\beta-\frac{1}{2}} \int |\mathbb{K}(u) - \bar{\mathbb{K}}(u)| du = O(h^s)$. Additionally, it can be shown from the previous assumptions and properties of the Fourier transform of $\omega^\beta K^{\mathrm{ft}}(\omega)$ that $\int |h^\beta \mathbb{K}(u) - \mathcal{K}(u)| du < \infty$. To obtain the rate $h^{1/2+\delta}$ for the latter, we need some additional conditions on the decay of f_ϵ^{ft}. Denote $R(\omega) = f_\epsilon^{\mathrm{ft}}(\omega)\omega^\beta - C_\epsilon$. Then a sufficient condition for the third condition in Assumption G (ii) is that $R(\omega) \sim \omega^{-1/2-\delta}$ as $|\omega| \to \infty$.

Assumption G (iii) contains conditions on the integrated kernel function $\mathbb{L}$. On the first two conditions in Assumption G (iii), we can in fact show the stronger statement that for all the commonly used kernel functions, $\mathcal{L}(u) \sim |u|^{-\beta\wedge 1}$ as $u \to \pm\infty$. Regarding the third

condition in Assumption G (iii), note that we can expand

$$h^{\beta}\bar{\mathbb{L}}\left(\frac{u}{h}\right)-L\left(\frac{u}{h}\right)=\frac{h^{\beta}}{\pi C_{\epsilon}}\int_{\omega_1}^{1/h}\frac{\sin(\omega u)}{\omega}K^{\mathrm{ft}}(h\omega)\frac{R(\omega)}{\psi^{\mathrm{ft}}(\omega)}d\omega-\frac{h^{\beta}}{\pi C_{\epsilon}}\int_{0}^{\omega_1}\sin(\omega u)\omega^{\beta-1}K^{\mathrm{ft}}(h\omega)d\omega.$$

Standard arguments show that this is of the order $h^{1/2+\delta}$ under the assumption $R(\omega)\sim\omega^{-1/2-\delta}$ as $|\omega|\to\infty$.

Assumption G (iv) provides conditions on the decay rates of the pdf f_X and its derivative f'_X. Similar assumptions are adopted in the literature (e.g., Bickel and Rosenblatt, 1973).

Based on these conditions, we obtain Theorem 7 with

$$B=\int\mathcal{L}(a)^2da,\qquad b_n=(-2\log h)^{1/2}+(-2\log h)^{-1/2}\log\left(\frac{\int\{\mathcal{L}'(a)\}^2da}{4\pi B}\right). \tag{C.25}$$

Furthermore, if we consider the simple hypothesis

$$H_0: F_{X^*}(t)=F_0(t)\quad\text{for } t\in\mathcal{T},$$

for some F_0, a test statistic for H_0 is $t_n^0=\sup_{t\in\mathcal{T}}|f_X(t)^{-1/2}\{\hat{F}_{X^*}(t)-F_0(t)\}|$. Consider the sequence of local alternatives

$$H_{1n}: F_{X^*}(t)=F_0(t)+\gamma_n\eta(t)\quad\text{for } t\in\mathcal{T},$$

where $\eta(t)$ is a continuous function and $\gamma_n=\sqrt{n}h^{\beta-1/2}(2\log(1/h))^{1/2}$. By an analogous argument, we can obtain

$$P\left\{(-2\log h)^{1/2}(B^{-1/2}t_n^0-b_n)\le c\right\}\to\exp(-s(\eta)\exp(-c)),$$

for all $c\in\mathbb{R}$, where $s(\eta)=\int_0^1\exp((Bf_{X^*}(a))^{-1/2}\eta(a))+\exp(-(Bf_{X^*}(a))^{-1/2}\eta(a))da$.

C.3.1 Proof of Theorem 7

We show that

$$\sup_{t\in\mathcal{T}}|f_X(t)^{-1/2}\{\hat{F}_{X^*}(t)-F_{X^*}(t)\}-\mathcal{Y}_n(t)|=o_p((-\log(h))^{-1/2}), \tag{C.26}$$

where $\mathcal{Y}_n = h^{-1/2} \int \mathcal{L}\left(\frac{t-a}{h}\right) dW(a)$ is a Gaussian process. Once we obtain (C.26), the conclusion follows by applying the arguments of Bickel and Rosenblatt (1973, Theorem A1). The rate $o_p((-\log(h))^{-1/2})$ is required because later we scale by $(-\log(h))^{1/2}$ to obtain the limiting distribution as in Bickel and Rosenblatt (1973).

First, as in the proof of Lemma 17, the bias term in $Q_n(t)$ is negligible and we can restrict attention to the mean zero process

$$D_n(t) \quad = Q_n(t) - E[Q_n(t)] = \quad h^{\beta-1/2} \int \mathbb{L}\left(\frac{t-a}{h}\right) d\alpha_n(a),$$

where $\alpha_n(a) = \sqrt{n}\{F_{X,n}^{EDF}(a) - F_X(a)\}$ is the empirical process, and $F_{X,n}^{EDF}$ is the empirical distribution function by $\{X_i\}_{i=1}^n$. We approximate $D_n(t)$ by

$$D_{n,1}(t) = h^{\beta-1/2} \int \bar{\mathbb{L}}\left(\frac{t-a}{h}\right) dW(F_X(a)).$$

Indeed the arguments in the proof of Lemma 17 allow us to show

$$\sup_{t\in\mathcal{T}} |D_n(t) - D_{n,1}(t)| = O_p((nh)^{-1/2} \log n)).$$

Also, $D_{n,1}(t)$ has the same finite dimensional distribution as

$$D_{n,2}(t) = h^{\beta-1/2} \int \bar{\mathbb{L}}\left(\frac{t-a}{h}\right) f_X(a)^{1/2} dW(a).$$

Next, we approximate $D_{n,2}(t)$ by

$$D_{n,3}(t) = h^{-3/2} \int \mathcal{K}\left(\frac{t-a}{h}\right) f_X(a)^{1/2} W(a) da.$$

To this end, note that for any $h > 0$,

$$\lim_{a\to\pm\infty} \mathcal{K}\left(\frac{t-a}{h}\right) f_X(a)^{1/2} W(a) \leq \sup_u |\mathcal{K}(u)| \lim_{a\to\pm\infty} |a f_X(a) \log(\log^+ |a|)|^{1/2} = 0,$$

where the inequality follows from the law of the iterated logarithm for the Wiener process and the equality follows from the facts $\sup_u |\mathcal{K}(u)| = O(h^{-\beta-1})$ and Assumption G (iv).

Thus, using stochastic integration by parts, we can write

$$D_{n,2}(t) = h^{\beta-1/2}\int\left\{f_X(t-hu)^{1/2}\bar{\mathbb{K}}(u)+hf_X'(t-hu)f_X(t-hu)^{-1/2}\bar{\mathbb{L}}(u)\right\}W(t-hu)du$$

and obtain

$$\begin{aligned}|D_{n,2}(t)-D_{n,3}(t)| &\leq h^{-1/2}\int\{h^{\beta}\bar{\mathbb{K}}(u)-\mathcal{K}(u)\}f_X(t-hu)^{1/2}W(t-hu)du\\ &\quad +h^{1/2}\int h^{\beta}\bar{\mathbb{L}}(u)f_X'(t-hu)f_X(t-hu)^{-1/2}W(t-hu)du\\ &= T_{n,4}(t)+T_{n,5}(t).\end{aligned}$$

Now by the law of the iterated logarithm and Assumption G (ii) and (iv), it follows $\sup_{t\in\mathcal{T}}|T_{n,4}(t)|=O_p(h^{\delta})$. For the term $T_{n,5}(t)$,

$$\begin{aligned}|T_{n,5}(t)| &\leq h^{-1/2}\sup_u|h^{\beta}\bar{\mathbb{L}}(u/h)-\mathcal{L}(u/h)|\int|f_X'(t-z)f_X(t-z)^{-1/2}W(t-z)|dz\\ &\quad +h^{1/2}\left|\int\mathcal{L}(u)f_X'(t-hu)f_X(t-hu)^{-1/2}W(t-hu)du\right|\\ &= T_{n,51}(t)+T_{n,52}(t).\end{aligned}$$

Using Assumption G (iii)-(iv), an application of the law of the iterated logarithm proves $\sup_{t\in\mathcal{T}}T_{n,51}(t)=O(h^{\beta})$. Next, for the term $T_{n,52}(t)$, Hölder's inequality and the law of the iterated logarithm imply

$$T_{n,52}(t) \leq h^{\delta_1/(4-2\delta_1)}\|\mathcal{L}(u)\|_{2-\delta_1}\left\|f_X'(u)f_X(u)^{-1/2}\sqrt{|u|\log(\log^+|u|)}\right\|_{2+\delta_1/(1-\delta_1)}.$$

By this expression and Assumption G (iii)-(iv), we are able to show $\sup_{t\in\mathcal{T}}|T_{n,52}(t)| = o_p((-\log(h))^{-1/2})$. Combining these results, the claim $\sup_{t\in\mathcal{T}}|D_{n,2}(t)-D_{n,3}(t)|=o_p((-\log(h))^{-1/2})$ follows.

Third, we approximate the process $f_X(t)^{-1/2}D_{n,3}(t)$ with the process

$$D_{n,4}(t)=h^{-3/2}\int\mathcal{K}\left(\frac{t-a}{h}\right)W(a)da.$$

Note that

$$f_X(t)^{-1/2} D_{n,3}(t) - D_{n,4}(t) = h^{-1/2} \int \{f_X(t)^{-1/2} f_X(t - hu)^{1/2} - 1\} \mathcal{K}(u) W(t - hu) du.$$

By the law of the iterated logarithm and Assumption G (ii) and (iv), it follows

$$\sup_{t \in \mathcal{T}} |f_X(t)^{-1/2} D_{n,3}(t) - D_{n,4}(t)| = O_p(h^{1/2}).$$

Fourth, let

$$D_{n,5}(t) = h^{-1/2} \int \mathcal{L}\left(\frac{t-a}{h}\right) dW(a).$$

By stochastic integration by parts formula and Assumption G (ii),

$$D_{n,4}(t) - D_{n,5}(t) = \left\{ \lim_{a \to \infty} L\left(\frac{t-a}{h}\right) W(a) \right\} - \left\{ \lim_{a \to -\infty} L\left(\frac{t-a}{h}\right) W(a) \right\} = 0,$$

for each h, which implies that $D_{n,4}(t) = D_{n,5}(t)$ for all $t \in \mathcal{T}$. Since $D_{n,5}(t)$ has the same finite dimensional distributions as the process $\mathcal{Y}_n$, the claim in (C.26) follows.

Bibliography

[1] Alberto Abadie and Guido W Imbens, *Large sample properties of matching estimators for average treatment effects*, Econometrica **74** (2006), no. 1, 235–267.

[2] ———, *On the failure of the bootstrap for matching estimators*, Econometrica **76** (2008), no. 6, 1537–1557.

[3] ———, *Bias-corrected matching estimators for average treatment effects*, Journal of Business & Economic Statistics **29** (2011), no. 1, 1–11.

[4] ———, *A martingale representation for matching estimators*, Journal of the American Statistical Association **107** (2012), no. 498, 833–843.

[5] ———, *Matching on the estimated propensity score*, Econometrica **84** (2016), no. 2, 781–807.

[6] Elena Andreou and Bas JM Werker, *An alternative asymptotic analysis of residual-based statistics*, Review of Economics and Statistics **94** (2012), no. 1, 88–99.

[7] Donald WK Andrews and Xiaoxia Shi, *Inference based on conditional moment inequalities*, Econometrica **81** (2013), no. 2, 609–666.

[8] Alexander Balke and Judea Pearl, *Bounds on treatment effects from studies with imperfect compliance*, Journal of the American Statistical Association **92** (1997), no. 439, 1171–1176.

[9] G. F. Barrett and S. G. Donald, *Consistent tests for stochastic dominance*, Econometrica **71** (2003), 71–104.

[10] G. W. Bassett and R. Koenker, *Strong consistency of regression quantiles and related empirical processes*, Econometric Theory **2** (1986), 191–201.

[11] Arie Beresteanu, Ilya Molchanov, and Francesca Molinari, *Partial identification using random set theory*, Journal of Econometrics **166** (2012), no. 1, 17–32.

[12] Arie Beresteanu and Francesca Molinari, *Asymptotic properties for a class of partially identified models*, Econometrica **76** (2008), no. 4, 763–814.

[13] P. J. Bickel and M. Rosenblatt, *On some global measures of the deviations of density function estimates*, Annals of Statistics **1** (1973), 1071–1095.

[14] Peter J Bickel, Friedrich Götze, and Willem R van Zwet, *Resampling fewer than n observations: gains, losses, and remedies for losses*, Selected Works of Willem van Zwet, Springer, 2012, pp. 267–297.

[15] P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner, *Efficient and adaptive estimation for semiparametric models*, Johns Hopkins series in the mathematical sciences, Springer New York, 1998.

[16] P. Billingsley, *Probability and measure*, 3rd ed., Wiley, -Blackwell, 1995.

[17] N. Bissantz, L. Dï¿œmbgen, H. Holzmann, and A. Munk, *Non-parametric confidence bands in deconvolution density estimation*, Journal of the Royal Statistical Society, B **69** (2007), 483–506.

[18] Christian Bontemps, Thierry Magnac, and Eric Maurin, *Set identified linear models*, Econometrica **80** (2012), no. 3, 1129–1155.

[19] J. Bound, C. Brown, and N. Mathiowetz, *Measurement error in survey data*, vol. 5, Handbook of Econometrics, chapter 59, Elsevier, 2000.

[20] J. Bound and A. B. Krueger, *The extent of measurement error in longitudinal earnings data: Do two wrongs make a right?*, Journal of Labor Economics **9** (1991), 1–24.

[21] Matias Busso, John DiNardo, and Justin McCrary, *New evidence on the finite sample properties of propensity score reweighting and matching estimators*, Review of Economics and Statistics **96** (2014), no. 5, 885–897.

[22] Ivan A Canay, *El inference for partially identified models: Large deviations optimality and bootstrap validity*, Journal of Econometrics **156** (2010), no. 2, 408–425.

[23] Matias D Cattaneo, Michael Jansson, and Whitney K Newey, *Alternative asymptotics and the partially linear model with many regressors*, Econometric Theory (2016), 1–25.

[24] Ngai Hang Chan, Song Xi Chen, Liang Peng, and Cindy L Yu, *Empirical likelihood methods based on characteristic functions with applications to lévy processes*, Journal of the American Statistical Association **104** (2009), no. 488, 1621–1630.

[25] Arun Chandrasekhar, Victor Chernozhukov, Francesca Molinari, and Paul Schrimpf, *Inference for best linear approximations to set identified functions*, arXiv preprint arXiv:1212.5627 (2012).

[26] Song Xi Chen, Wolfgang Härdle, and Ming Li, *An empirical likelihood goodness-of-fit test for time series*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **65** (2003), no. 3, 663–678.

[27] V. Chernozhukov, D. Chetverikov, and K. Kato, *Anti-concentration and honest, adaptive confidence bands*, Annals of Statistics **42** (2014), 1787–1818.

[28] ———, *Comparison and anti-concentration bounds for maxima of Gaussian random vectors*, Probability Theory and Related Fields **162** (2015), 47–70.

[29] Victor Chernozhukov, Emre Kocatulum, and Konrad Menzel, *Inference on sets in finance*, Quantitative Economics **6** (2015), no. 2, 309–358.

[30] William G Cochran, *The effectiveness of adjustment by subclassification in removing bias in observational studies*, Biometrics (1968), 295–313.

[31] F. Comte and J. Kappus, *Density deconvolution from repeated measurements without symmetry assumption on the errors*, Journal of Multivariate Analysis **140** (2015), 31–46.

[32] Noel Cressie and Frederick L Hulting, *A spatial statistical analysis of tumor growth*, Journal of the American Statistical Association **87** (1992), no. 418, 272–283.

[33] I. Dattner, A. Goldenshluger, and A. Juditsky, *On deconvolution of distribution functions*, Annals of Statistics **39** (2011), 2477–2501.

[34] I. Dattner, M. Reiï¿œ, and M. Trabs, *Adaptive quantile estimation in deconvolution with unknown error distribution*, Bernoulli **22** (2016), 143–192.

[35] RM De Jong and Herman J Bierens, *On the limit behavior of a chi-square type test if the number of conditional moments tested approaches infinity*, Econometric Theory **10** (1994), no. 1, 70–90.

[36] A. Deaton, *The analysis of household surveys: A microeconometric approach to development policy*, Johns Hopkins University Press for the World Bank, Baltimore, 1997.

[37] Rajeev H Dehejia and Sadek Wahba, *Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs*, Journal of the American statistical Association **94** (1999), no. 448, 1053–1062.

[38] A. Delaigle and I. Gijbels, *Bootstrap bandwidth selection in kernel density estimation from a contaminated sample*, Annals of the Institute of Statistical Mathematics **56** (2004), 19–47.

[39] A. Delaigle and P. Hall, *On optimal kernel choice for deconvolution*, Statistics & Probability Letters **76** (2006), 1594–1602.

[40] A. Delaigle, P. Hall, and A. Meister, *On deconvolution with repeated measurements*, Annals of Statistics **36** (2008), 665–685.

[41] Stephen G Donald, Guido W Imbens, and Whitney K Newey, *Empirical likelihood estimation and consistent tests with conditional moment restrictions*, Journal of Econometrics **117** (2003), no. 1, 55–93.

[42] S. Efromovich, *Density estimation for the case of supersmooth measurement error*, Journal of the American Statistical Association **92** (1997), 526–535.

[43] Bradley Efron and Charles Stein, *The jackknife estimate of variance*, The Annals of Statistics (1981), 586–596.

[44] J. Fan, *On the optimal rates of convergence for nonparametric deconvolution problems*, Annals of Statistics **19** (1991), 1257–1272.

[45] Jianqing Fan and Li-Shan Huang, *Goodness-of-fit tests for parametric regression models*, Journal of the American Statistical Association **96** (2001), no. 454, 640–652.

[46] Jianqing Fan, Chunming Zhang, and Jian Zhang, *Generalized likelihood ratio statistics and wilks phenomenon*, Annals of statistics (2001), 153–193.

[47] Jianqing Fan and Jian Zhang, *Sieve empirical likelihood ratio tests for nonparametric functions*, Annals of Statistics (2004), 1858–1907.

[48] Nicholas I Fisher, Peter Hall, Berwin A Turlach, and GS Watson, *On the estimation of a convex set from noisy data on its support function*, Journal of the American Statistical Association **92** (1997), no. 437, 84–91.

[49] Evarist Giné and Joel Zinn, *Bootstrapping general empirical measures*, The Annals of Probability (1990), 851–869.

[50] P. Hall and S. N. Lahiri, *Estimation of distributions, moments and quantiles in deconvolution problems*, Annals of Statistics **36** (2008), 2110–2134.

[51] P. Hall and A. Meister, *A ridge-parameter approach to deconvolution*, Annals of Statistics **35** (2007), 1535–1558.

[52] Wolfgang Hardle and Enno Mammen, *Comparing nonparametric versus parametric regression fits*, The Annals of Statistics (1993), 1926–1947.

[53] James Heckman, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, *Characterizing selection bias using experimental data*, Econometrica **66** (1998), no. 5, 1017–1098.

[54] James J Heckman and V Joseph Hotz, *Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training*, Journal of the American statistical Association **84** (1989), no. 408, 862–874.

[55] James J Heckman, Hidehiko Ichimura, and Petra E Todd, *Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme*, The review of economic studies **64** (1997), no. 4, 605–654.

[56] Nils Lid Hjort, Ian W McKeague, Ingrid Van Keilegom, et al., *Extending the scope of empirical likelihood*, The Annals of Statistics **37** (2009), no. 3, 1079–1111.

[57] Joel L Horowitz and Charles F Manski, *Nonparametric analysis of randomized experiments with missing covariate and outcome data*, Journal of the American statistical Association **95** (2000), no. 449, 77–84.

[58] Guido W Imbens and Donald B Rubin, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press, 2015.

[59] Hiroaki Kaido, *A dual approach to inference for partially identified econometric models*, preprint (2012).

[60] Hiroaki Kaido and Andres Santos, *Asymptotically efficient estimation of models defined by convex moment inequalities*, Econometrica **82** (2014), no. 1, 387–413.

[61] Joseph DY Kang and Joseph L Schafer, *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data*, Statistical science (2007), 523–539.

[62] Kengo Kato and Yuya Sasaki, *Uniform confidence bands in deconvolution with unknown error distribution*, arXiv preprint arXiv:1608.02251 (2016).

[63] DG Kendall, *Foundations of a theory of random set*, Stochastic geometry (1974), 322–376.

[64] Shakeeb Khan and Elie Tamer, *Irregular identification, support conditions, and inverse weight estimation*, Econometrica **78** (2010), no. 6, 2021–2042.

[65] János Komlós, Péter Major, and Gábor Tusnády, *An approximation of partial sums of independent rv'-s, and the sample df. i*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **32** (1975), no. 1-2, 111–131.

[66] Soumendra Nath Lahiri, *Bootstrapping m-estimators of a multiple linear regression parameter*, The Annals of Statistics (1992), 1548–1570.

[67] Robert J LaLonde, *Evaluating the econometric evaluations of training programs with experimental data*, The American economic review (1986), 604–620.

[68] Michael Lechner, *Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies*, Review of Economics and Statistics **84** (2002), no. 2, 205–220.

[69] O. V. Lepski, *A problem of adaptive estimation in Gaussian white noise, teor*, Veroyatnost. i Primenen. **35** (1990), 459–470.

[70] H. Levy, *Stochastic dominance*, Springer, 2016.

[71] Gang Li, *Nonparametric likelihood ratio goodness-of-fit tests for survival data*, Journal of Multivariate Analysis **86** (2003), no. 1, 166–182.

[72] T. Li and Q. Vuong, *Nonparametric estimation of the measurement error model using multiple indicators*, Journal of Multivariate Analysis **65** (1998), 139–165.

[73] Charles F Manski, *Partial identification of probability distributions*, Springer Science & Business Media, 2003.

[74] Georges Matheron, Georges Matheron, Georges Matheron, and Georges Matheron, *Random sets and integral geometry*, Wiley New York, 1975.

[75] A. Meister, *Deconvolution problems in nonparametric statistics*, Springer, 2009.

[76] Ilya Molchanov and Francesca Molinari, *Applications of random set theory in econometrics*, Annu. Rev. Econ. **6** (2014), no. 1, 229–251.

[77] Ilya S Molchanov et al., *Theory of random sets*, vol. 19, Springer, 2005.

[78] M. H Neumann, *On the effect of estimating the error density in nonparametric deconvolution*, Journal of Nonparametric Statistics **7** (1997), 307–330.

[79] Whitney K Newey and Daniel McFadden, *Large sample estimation and hypothesis testing*, Handbook of econometrics **4** (1994), 2111–2245.

[80] OECD, *Growing unequal?: Income distribution and poverty in oecd countries*, 2008.

[81] Taisuke Otsu and Yoshiyasu Rai, *Bootstrap inference of matching estimators for average treatment effects*, Journal of the American Statistical Association (2016), no. just-accepted.

[82] Art B Owen, *Empirical likelihood*, Chapman and Hall/CRC, 2001.

[83] Dimitris N Politis and Joseph P Romano, *Large sample confidence regions based on subsamples under minimal assumptions*, The Annals of Statistics (1994), 2031–2050.

[84] David Pollard, *Convergence of stochastic processes*, Springer Science & Business Media, 2012.

[85] Jin Qin and Jerry Lawless, *Empirical likelihood and general estimating equations*, The Annals of Statistics (1994), 300–325.

[86] Paul R Rosenbaum, *Optimal matching for observational studies*, Journal of the American Statistical Association **84** (1989), no. 408, 1024–1032.

[87] ———, *Design of observational studies*, Springer Science & Business Media, 2009.

[88] Paul R Rosenbaum and Donald B Rubin, *The central role of the propensity score in observational studies for causal effects*, Biometrika (1983), 41–55.

[89] ———, *Reducing bias in observational studies using subclassification on the propensity score*, Journal of the American statistical Association **79** (1984), no. 387, 516–524.

[90] Donald B Rubin, *Estimating causal effects of treatments in randomized and nonrandomized studies.*, Journal of educational Psychology **66** (1974), no. 5, 688.

[91] Susanne Schennach, *Convolution without independence*, Tech. report, cemmap working paper, Centre for Microdata Methods and Practice, 2013.

[92] Galen R Shorack, *Bootstrapping robust regression*, Communications in Statistics-Theory and Methods **11** (1982), no. 9, 961–972.

[93] Jeffrey A Smith and Petra E Todd, *Reconciling conflicting evidence on the performance of propensity-score matching methods*, The American Economic Review **91** (2001), no. 2, 112–118.

[94] Jakob Söhl, Mathias Trabs, et al., *A uniform central limit theorem and efficiency for deconvolution estimators*, Electronic Journal of Statistics **6** (2012), 2486–2518.

[95] L. Stefanski and R. J. Carroll, *Deconvoluting kernel density estimators*, Statistics **21** (1990), 169–184.

[96] Dietrich Stoyan, *Random sets: models and statistics*, International Statistical Review **66** (1998), no. 1, 1–27.

[97] W Stute, W González Manteiga, and M Presedo Quindimil, *Bootstrap approximations in model checks for regression*, Journal of the American Statistical Association **93** (1998), no. 441, 141–149.

[98] Winfried Stute, *Nonparametric model checks for regression*, The Annals of Statistics (1997), 613–641.

[99] Elie Tamer, *Partial identification in econometrics*, Annu. Rev. Econ. **2** (2010), no. 1, 167–195.

[100] A. W. van der Vaart and J. Wellner, *Weak convergence and empirical processes*, Springer, 1996.

[101] Aad W Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge university press, 1998.

[102] A. J. van Es and H. w. Uh, *Asymptotic normality for kernel type deconvolution estimators*, Scandinavian Journal of Statistics **32** (2005), 467–483.

[103] VN Vapnik and A Ya Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications **16** (1971), no. 2, 264.

[104] J. E. Yukich, *Some limit theorems for the empirical process indexed by functions*, Probability Theory and Related Fields **74** (1987), 71–90.