NONPARAMETRIC ESTIMATION AND INFERENCE IN ECONOMETRICS

A Dissertation

by

TA-CHENG HUANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Qi Li
Committee Members,	Yonghong An
	Dennis W. Jansen
	Ursula Müller-Harknett
	Ke-Li Xu
Head of Department,	Timothy J. Gronberg

May 2018

Major Subject: Economics

ABSTRACT

This dissertation includes two essays: The first one is on nonparametric inference in causal effect models, and the second one is on nonparametric estimation in financial economics.

In the first essay, we propose a nonparametric test for unobserved heterogeneous treatment effects in a general framework, allowing for self-selection to the treatment. The proposed modified Kolmogorov-Smirnov-type test is consistent and simple to implement. Monte Carlo simulations show that our test performs well in finite samples. For illustration, we apply our test to study heterogeneous treatment effects of the Job Training Partnership Act on earnings and the impacts of fertility on family income.

In the second essay, we provide an alternative to the existing estimations of implied volatility in option pricing. The use of state price densities to gather information about market sentiment and other empirical characteristics that describe important phenomena is popular in literature and in practice. The estimation of the implied volatility surface to extract these densities is a crucial intermediate step in the process, and the methods to do so are varied in literature. This essay proposes an estimation procedure that is relative new in nonparametric literature: ℓ_1 trend filtering. We show its advantages over typically used nonparametric and parametric methods, commonly used in literature and in practice, to deal with this particular estimation problem. Additionally, the method maintains smaller prediction errors than the comparison models across different number of observations and levels of noise.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Qi Li (Chair) of the Department of Economics, Professor Yonghong An of the Department of Economics, Professor Dennis W. Jansen of the Department of Economics, Professor Ursula Müller-Harknett of the Department of Statistics, and Professor Ke-Li Xu of the Department of Economics. All committee members but Professor Xu, who is at Indiana University, Bloomington, are at Texas A&M University.

The theoretical analyses depicted in the first essay were conducted in part with Professor Yu-Chin Hsu of the Institute of Economics at Academia Sinica and Professor Haiqing Xu of the Department of Economics at University of Texas at Austin. In addition, the estimation procedure illustrated in the second essay were conducted in part with Pablo Crespo of Department of Economics at the Graduate Center of the City University of New York.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by the Department of Economics at Texas A&M University and the Lynde and Harry Bradley Fellowship from Private Enterprise Research Center at Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

This dissertation includes two independent essays. In the first essay, I propose a nonparametric test for unobserved heterogeneity in treatment effects. The test can answer an crucial policy-relevant research question: Whether the measured treatment effects are also valid for other subpopluations? The second essay provides an alternative to the existing estimations of implied volatility in option pricing when usable data is scarce on daily basis.

1.1 The First Essay

Heterogeneous treatment effects due to unobserved latent variables has been emphasized in the policy evaluation literature. See e.g. Heckman *et al.* (1997), Heckman and Vytlacil (2001, 2005), Abadie *et al.* (2002), Abadie (2003), Blundell and Powell (2003) , Matzkin (2003), Chesher (2003, 2005), Chernozhukov and Hansen (2005), Florens *et al.* (2008), Imbens and Newey (2009), Frölich and Melly (2013), D'Haultfœuille and Février (2015), Torgovitsky (2015), and among many others. The interpretation and credibility of the instrumental variable (IV) approach rely on the hypothesis that treatment effects are homogeneous across individuals, after controlling for covariates. In this paper, we develop a nonparametric test for unobserved heterogeneous treatment effects under the standard instrumental variable framework.

In this essay, we use a nonseparable equation for the structural relationship to model unobserved heterogeneous treatment effects. Given the nonseparability of the structural relationship, treatment effects vary across individuals, even after controlling for all observed covariates. (See e.g. Matzkin, 2003; Chesher, 2003, 2005). In the presence of endogeneity, it is well known that such heterogeneity of treatment effects brings challenges to estimating e.g. *average treatment effects* (ATE). In particular, Imbens and Angrist (1994) show that the conventional IV estimation only recovers the *"Local Average Treatment Effects"*(LATE), rather than the ATE. On the other hand, the homogenous treatment effects assumption substantially simplifies identification and estimation of ATE, since it implies the ATE is the same as the LATE, after controlling for observed

covariates. For instance, Angrist and Krueger (1991) use a two–stage least square approach to estimate treatment effects. Therefore, providing evidence for homogeneous treatment effects justifies the results and interpretations from the transitional IV estimation.

Though important, there are only a handful of papers on testing for such unobserved heterogeneity. In the context of ideal social experiments, Heckman *et al.* (1997) develop a lower bound for the variance of heterogeneous treatment effects, thereby providing a test for whether or not the data are consistent with the identical treatment effects model. Moreover, Hoderlein and Mammen (2009) discuss specification tests for endogeneity as well as unobserved heterogeneity in nonseparable triangular models. Recently, Lu and White (2014) and Su *et al.* (2015) establish nonparametric tests for unobserved heterogeneous treatment effects under the unconfoundedness assumption. In particular, Lu and White (2014) test unobserved heterogeneity in treatments effects via testing an equivalent independence condition on observables. Mainly motivated by Lu and White (2014), we show that in the presence of endogeneity, model restrictions arising from the homogeneous treatment effects hypothesis can also be characterized by an alternative set of independence conditions, which is constructed by using the LATE estimator.

Another closely related paper is Heckman *et al.* (2010) who test the absence of self-selection on the gain to treatment in the generalized Roy model framework, allowing for (unobserved) heterogeneous treatment effects. Similar to their work, our testing problem is formulated in a model allowing for both unobserved heterogeneity and selection into treatment, called as the "essential heterogeneity" in Heckman *et al.* (2006).

Nonparametric tests for (conditional) independence restrictions have been well studied in different contexts. See e.g. Andrews (1997); Su and White (2007, 2008, 2014); Bouezmarni *et al.* (2012); Hoderlein and White (2012); Linton and Gozalo (2014); Huang *et al.* (2016), among many others. When one considers testing independence restrictions of variables that are nonparametrically constructed, however, a key technical issue arises, especially in the case where the nonparametric components are functions of continuous covariates (see e.g. Lu and White, 2014). Motivated by Stinchcombe and White (1998), we modify the classic Kolmogorov–Smirnov tests by using the

primitive function of CDF's to represent the independence condition. Such a modification is novel and plays a key role for our approach. Moreover, we establish the asymptotic properties of the proposed tests under the null and alternative hypotheses.

The essay is organized as follows. In Section 2.1, we introduce the model and derive testable model restrictions. Section 2.2 discusses our test statistics and their asymptotic results. We distinguish the cases whether covariates include continuous variables. In Section 2.3, we conduct Monte Carlo experiments to study the finite-sample performance of the proposed test. Section 3.3 illustrates our testing approach by two empirical applications. Section 2.5 extends our approach to the Regression Discontinuity design. All proofs are collected in the Appendix.

1.2 The Second Essay

The use of market prices as given to perform analysis on empirical phenomena in securities has become increasingly prominent in research. In terms of using prices to measure market sentiment, the state price density (SPD), also known as risk neutral density (RND) has become the tool of choice. The moments of this distribution possess information about market beliefs on the evolution of prices for the underlying security. As such, extracting the SPD has importance in practice. Breeden and Litzenberger (1978) show that SPDs can be extracted from option prices theoretically via a second derivative with respect to strike prices. For this extraction, using existing theoretical pricing models for options can reveal good analytical solutions, but it seldom matches data. Thus, it has become common practice to use a version of the Black and Scholes (1973) and Merton (1973) pricing formula (BSM) as a transformation from the space of prices to the space of "implied volatilities." Estimating implied volatilities can yield to feasible, practical results that can be used to extract the SPD. However, the usefulness of this exercise depends directly on the magnitude of the prediction error. The data on option prices tend to be noisy and scarce, which poses a threat to the accuracy of predicted values. The estimation of implied volatility is thus a non-trivial problem under which misspecification can yield misleading results and incorrect interpretations.

The literature on the methodology for estimation of implied volatility can be divided into two broad categories: parametric and nonparametric. Parametric methods carry the advantage of being

easy to use, have good inference capabilities and possess extrapolation features beyond the scope of the dataset. However, the underlying assumption of functional form required for a parametric method is strong. This is problematic, as there is no guarantee (particularly in the cases of sudden changes in beliefs of the behavior of a price) that the estimates will work for all days under the same specification. The nonparametric methods produce flexible fits and do not require for the researcher to make any specification assumption. The limitations of these methods lie on the fact that they are slow converging, meaning that the prediction error depends heavily on the number of observations in the dataset. Goodness of fit suffers in the case of a low number of observations. Due to the large dataset requirements , particularly in multidimensional models, daily option data is often inappropriate and at risk of oversmoothing when using it for SPD extraction. It is then common while using kernel estimators, to also use aggregated data. While this is a reasonable idea, it becomes problematic when addressing the fact that interpretability is somewhat lost with aggregated data. Namely, since we want to use the main features of the SPD to assess beliefs about risk on the price of a security, the day to day variation can be important for the researcher. Furthermore, when using aggregated data, regime changes become a concern as they are no longer identifiable. We propose to use a nonparametric estimator, ℓ_1-trend filtering, which provides a flexible alternative that can operate with smaller prediction errors than other nonparametric methodology across different sample sizes. Our review is focused on methods that are comparable to trend filtering. Curve fitting and kernel smoothing methods are what we use for benchmarks. Jackwerth (2004) provides an extensive list on methodologies, including those which extract the SPDs directly in his monograph.

Shimko (1993) takes advantage of the fact that option prices can be separated into cross sections according to tenor and fits a second order polynomial with a simple regression on a measure of moneyness at each tenor. Malz (2014) fits quadratic polynomials via interpolating splines. Aparicio and Hodges (1998) fit cubic B-Splines, and Hayes *et al.* (2003) use interpolated cubic splines to fit implied volatilities. Aït-Sahalia and Lo (1998) use a kernel estimator (Nadaraya-Watson), and this paper was later used as the base for an extension to fit local linear kernel estimators.

In order to perform our comparison, we pick Shimko (1993) as its performance is notable among parametric methods, and two local linear kernel estimators with different bandwidth parameters as our nonparametric representatives. The choice of the local linear kernel estimators comes from the popularity of Aït-Sahalia and Lo (1998) methodology, with a simple correction to avoid poor fits in the extremes of the datasets.[1] ℓ_1-trend filtering can be compared to a smoothing spline in which knots are selected adaptively. This estimation procedure can handle relatively smaller datasets than the kernel estimators, but is still limited. In our application at least seven observations at each tenor cross-section are required to perform the estimation. In addition, note that as in all nonparametric methods, extrapolation beyond the scope of the dataset is not possible. This issue limits our ability to observe tail behavior in the SPDs, to solve this problem we use Figlewski (2008) suggested method for extrapolating tails by grafting them from an extreme value (GEV) distribution.

In our results we show via Monte Carlo that trend filtering provides better fits than the kernel estimators across different sample sizes and levels of noise. The gap in the prediction error is reduced under high noise and small sample sizes, but trend filtering continues to show a distinct advantage over its nonparametric competitiors across the distribution of simulations. These results are important, as daily observations are likely to be few, thus rendering trend filtering as a viable estimation strategy. In addition, trend filtering outperforms the Shimko (1993) specification regardless of the size of the dataset. This parametric specification is a comfortable choice for practicioners for small sample sizes. As a final simulation check, we fit the specified simulation model and show that the nonparametric flexibility of trend filtering overcomes the slight overfitting prediction errors than even the correct specification can cause.

Moreover, we perform a comparison of trend filtering against Shimko (1993) and a fourth order polynomial stringwise regression on S&P 500, Apple Inc., Kellog Company and Gap Inc. quotes for the time period from January 1, 2007 to January 1, 2010. The data is separated randomly into "training" and "testing" (out of sample) datasets. We see that trend filtering outperforms the

[1]The Nadaraya-Watson estimator is known to have issues around the boundaries of the dataset. Local linear performs significantly better in this aspect.

parametric specifications consistently.

The essay is organized as follows. Section 3.1 presents the background theory on SPDs and implied volatility surfaces as well as introducing ℓ_1-trend filtering formally. The Monte Carlo experiment and its results are demonstrated in Section 3.2. Section 3.3 summarizes empirical work and performance assesment. Section 3.4 concludes.

2. TESTING FOR UNOBSERVED HETEROGENEOUS TREATMENT EFFECTS IN A NONSEPARABLE MODEL WITH ENDOGENOUS SELECTION

2.1 Model and Testable Restrictions

We consider the following nonseparable treatment effect model:

$$Y = g(D, X, \epsilon) \tag{2.1}$$

where $Y \in \mathbb{R}$ is outcome variable, $D \in \{0, 1\}$ denotes treatment status, $X \in \mathbb{R}^{d_X}$ are covariates, ϵ is an unobserved random disturbance of general form (e.g. without invoking any restriction on the dimensionality of ϵ), and g is an unknown but smooth function defined on $\{0, 1\} \times \mathscr{S}_{X\epsilon}$.[1] In particular, the treatment variable D is allowed to be correlated with ϵ so as to allow for selection to the treatment; see e.g. Heckman *et al.* (1997). To deal with endogeneity, we introduce a binary instrumental variable $Z \in \{0, 1\}$. Throughout the paper, we use upper case letters to denote random variables, and their corresponding lower case letters to stand for realizations of random variables.

As is motivated in the seminal paper by Matzkin (1999), the non-additivity of the structural relationship g in ϵ captures the idea of unobserved heterogeneous treatment effects in that the individual treatment effect, $g(1, X, \epsilon) - g(0, X, \epsilon)$, would depend on the unobserved individual heterogeneity ϵ, even after controlling for covariates X. Therefore, we have the following proposition.

Proposition 2.1. *Suppose* (2.1) *holds, then the homogeneous treatment effects hypothesis, i.e., for some measurable function* $\delta_0(\cdot) : \mathscr{S}_X \mapsto \mathbb{R}$*,*

$$\mathbb{H}_0 : \; g(1, X, \cdot) - g(0, X, \cdot) = \delta_0(X) \tag{2.2}$$

[1] For a generic random vector A, We use $\mathscr{S}_A$ to denote the support of A.

holds if and only if the structural relationship $g(\cdot,\cdot,\cdot)$ is additively separable in ϵ (w.r.t. D), i.e.,

$$g(D, X, \epsilon) = m(D, X) + \nu(X, \epsilon), \tag{2.3}$$

where $m : \mathscr{S}_{DX} \mapsto \mathbb{R}$ and $\nu : \mathscr{S}_{X\epsilon} \mapsto \mathbb{R}$.

Proposition 2.1 follows Lu and White (2014). Note that if (2.3) holds, $\delta_0(x) = m(1, x) - m(0, x)$ in (2.2), which is the homogenous individual treatment effects across individuals with the same value of covariates.

A key insight from Lu and White (2014) is that they further show that the equivalence between the additive separability hypotheses (i.e. eq. (2.3)) and a conditional independence restriction on observables. In the presence of treatment endogeneity, we derive a similar set of model restrictions. For each $x \in \mathscr{S}_X$ and $z = 0, 1$, let $p(x, z) = \Pr(D = 1|X = x, Z = z)$ be the *propensity score*.

Assumption 2.1. *Suppose $Z \perp\!\!\!\perp \epsilon|X$ and $p(x, 0) \neq p(x, 1)$ for all $x \in \mathscr{S}_X$. Without loss of generality, let $p(x, 0) < p(x, 1)$ for all $x \in \mathscr{S}_X$.*

Assumption 2.1 is standard in the literature, which requires the instrumental variable Z to be (conditionally) exogenous and relevant. See e.g. Imbens and Angrist (1994) and Chernozhukov and Hansen (2005). Throughout, we maintain Assumption 2.1.

Moreover, let $\mu(x, z) = \mathbb{E}(Y|X = x, Z = z)$. Under $\mathbb{H}_0$ and Assumption 2.1, we have

$$\mu(x, z) = \mathbb{E}\left[g(0, x, \epsilon)|X = x\right] + \delta_0(x)p(x, z), \text{ for } z = 0, 1.$$

In the above equation system, we treat $\mathbb{E}\left[g(0, X, \epsilon)|X = x\right]$ and $\delta_0(x)$ as two unknowns. Solve the equations, then we identify $\delta_0(x)$ by:

$$\delta(x) \equiv \frac{\mu(x, 1) - \mu(x, 0)}{p(x, 1) - p(x, 0)} = \frac{\text{Cov}(Y, Z|X)}{\text{Cov}(D, Z|X)}. \tag{2.4}$$

See Imbens and Angrist (1994) for the LATE interpretation of (2.4). Note that $\delta(x)$ is well defined

under Assumption 2.1 and identified as well directly from the data regardless the monotonicity of the selection.

Let $W \equiv Y + (1 - D) \cdot \delta(X)$. Note that the null hypothesis $\mathbb{H}_0$ implies that $W = g(1, X, \epsilon)$. Therefore, under Assumption 2.1, W is conditionally independent of Z given X. The next lemma summarizes the above discussion.

Lemma 2.1. *Suppose* (2.1) *and Assumption 2.1 hold. Then,* $\mathbb{H}_0$ *implies that* $W \perp\!\!\!\perp Z \mid X$.

The proof of Lemma 2.1 is straightforward, and hence omitted. To provide a consistent test, we now establish the sufficiency of the conditional independence for testing $\mathbb{H}_0$.

Assumption 2.2 (Single–index error term)**.** *There exist measurable functions* $\tilde{g} : \mathscr{S}_{DX} \times \mathbb{R} \mapsto \mathbb{R}$ *and* $\nu : \mathscr{S}_{X\epsilon} \mapsto \mathbb{R}$ *such that*

$$g(D, X, \epsilon) = \tilde{g}(D, X, \nu(X, \epsilon)).$$

Moreover, $\tilde{g}(d, x, \cdot)$ *is strictly increasing in the scalar–valued index* ν *for* $d = 0, 1$ *and all* $x \in \mathscr{S}_X$.

Assumption 2.2 imposes the monotonicity of the single–index error term, for which various simplified assumptions have also been made in the literature for identification and estimation of nonseparable functions. For instance, among many others, Matzkin (2003) and Chesher (2003) assume the structural function g is strictly increasing in the scalar–valued error term ϵ. Note that Assumption 2.2 holds under the null hypothesis $\mathbb{H}_0$, represented in terms of (2.3). Hence, Assumption 2.2 narrows down the space of alternatives such that the model restrictions derived in Lemma 2.1 is sufficient to distinguish the null and alternative hypotheses.

Assumption 2.3 (Monotone selection)**.** *The selection to the treatment is given by*

$$D = \mathbb{1}\left[\theta(X, Z) - \eta \geq 0\right], \tag{2.5}$$

where θ *is an unknown function, and* $\eta \in \mathbb{R}$ *is an error term satisfying* $Z \perp\!\!\!\perp (\epsilon, \eta) \mid X$.

Imbens and Angrist (1994) first introduce the monotone selection assumption, which is essentially the "no defier" condition. Moreover, Vytlacil (2002) shows that such a monotonicity condition

is observationally equivalent to the weak monotonicity of (2.5) in the error term η. Vuong and Xu (2016) point out Assumption 2.3 can be relaxed to the strict monotonicity of $\mathbb{P}(Y \leq y; D = 1|X, Z = 1) - \mathbb{P}(Y \leq y; D = 1|X, Z = 0)$ in $y \in \mathscr{S}^{\circ}_{Y|X,D=1}$, the interior region of $\mathscr{S}_{Y|X,D=1}$.

Note that the second half of Assumption 2.1 implies $\theta(x, 0) < \theta(x, 1)$ for all $x \in \mathscr{S}_X$. Let $\mathcal{C}_x \equiv \{\eta \in \mathbb{R} : \theta(x, 0) < \eta \leq \theta(x, 1)\}$ be the "complier group" given $X = x$ (see Imbens and Angrist, 1994, for the concept "complier group").

Assumption 2.4. *The support of $g(d, x, \epsilon)$ given $X = x$ and the complier group $\mathcal{C}_x$ equals to the support of $g(d, x, \epsilon)$ given $X = x$, i.e.,*

$$\mathscr{S}_{g(d,x,\epsilon)|X=x,\ \eta\in\mathcal{C}_x} = \mathscr{S}_{g(d,x,\epsilon)|X=x}.$$

Assumption 2.4 is a support condition, first introduced by Vuong and Xu (2016) as the effectiveness of the instrument variable. It implies $\mathscr{S}_{g(d,x,\epsilon)|X=x,\ \eta\in\mathcal{C}_x} = \mathscr{S}_{Y|D=d,X=x}$. Note that the distribution of $g(d, x, \epsilon)$ given $X = x$ and $\eta \in \mathcal{C}_x$ can be identified, see, e.g., Imbens and Rubin (1997). Thus, Assumption 2.4 is testable. Specifically, for all $t \in \mathbb{R}$,

$$\begin{aligned}&\mathbb{P}[g(d, x, \epsilon) \leq t|X = x, \eta \in \mathcal{C}_x] = \\ &\quad \frac{\Pr(Y \leq t, D = d|X = x, Z = 1) - \Pr(Y \leq t, D = d|X = x, Z = 0)}{\Pr(D = d|X = x, Z = 1) - \Pr(D = d|X = x, Z = 0)},\end{aligned}$$

from which we can identify the support $\mathscr{S}_{g(d,x,\epsilon)|X=x,\eta\in\mathcal{C}_x}$.

Assumption 2.4 allows one to use the data to address questions involving counterfactuals of outcomes of the "always takers" and the "never–takers" groups. It is possible to provide sufficient primitive conditions for Assumption 2.4. For instance, if one assumes $\mathscr{S}_{\epsilon|X=x,\ \eta\in\mathcal{C}_x} = \mathscr{S}_{\epsilon|X=x}$, or even a stronger condition that (ϵ, η) has a rectangular support conditional on $X = x$, then Assumption 2.4 holds. It is also worth noting that without imposing Assumption 2.4, our methodology can be used to test the null hypothesis of (2.2) holding with respect to the subset $\epsilon \in \mathscr{S}_{\epsilon|X=x,\ \eta\in\mathcal{C}_x}$.

Theorem 2.1. *Suppose* (2.1)*, and Assumptions 2.1 to 2.4 hold. Then $\mathbb{H}_0$ holds if and only if*

$W \perp\!\!\!\perp Z|X$.

From now on, we maintain Assumptions 2.1 to 2.4. By Theorem 2.1, testing the null hypothesis $\mathbb{H}_0$ is equivalent to testing the conditional independence condition $W \perp\!\!\!\perp Z|X$. It is worth pointing out that Theorem 2.1 is related to Lu and White (2014), who show that $\mathbb{H}_0$ holds if and only if $[Y - \mathbb{E}(Y|D, X)] \perp\!\!\!\perp D|X$ under the unconfoundedness condition (i.e. $D \perp\!\!\!\perp \epsilon|X$) and Assumption 2.2.

2.2 Consistent Tests

Based on Theorem 2.1, we now propose tests for unobserved treatment effect heterogeneity via testing the conditional independence restriction. Because Z is binary, the conditional independence restriction in Theorem 2.1 is equivalent to

$$F_{W|XZ}(\cdot|x, 0) = F_{W|XZ}(\cdot|x, 1), \ \forall \ x \in \mathscr{S}_X.$$

Note that the variable W needs to be nonparametrically constructed from the data. In the following discussion, we distinguish the cases whether the covariates X are continuous random variables because the continuous–covariates case is more difficult to deal with due to the nonparametric function $\delta(\cdot)$ in the construction of W. For expositional simplicity, we assume $X \in \mathbb{R}$ in the following discussion. It is straightforward to generalize our results to vector-valued covariates.

2.2.1 Case 1: Discrete Covariates

We first discuss the case where X takes only a finite number of values. Let $\{(Y_i, D_i, X_i', Z_i)' : i \leq n\}$ be a random sample of $(Y, D, X', Z)'$. By Theorem 2.1, we test $\mathbb{H}_0$ via the following model restrictions:

$$F_{W|XZ}(\cdot \ |x, 0) = F_{W|XZ}(\cdot \ |x, 1), \ \forall \ x \in \mathscr{S}_X,$$

where $W = Y + (1 - D)\delta(X)$ is generated from the observables.

We estimate $\delta(X_i)$ as follows

$$\hat{\delta}(X_i) = \frac{\sum_{j\neq i} Y_j Z_j \mathbb{1}(X_j = X_i) \times \sum_{j\neq i} \mathbb{1}(X_j = X_i) - \sum_{j\neq i} Y_j \mathbb{1}(X_j = X_i) \times \sum_{j\neq i} Z_j \mathbb{1}(X_j = X_i)}{\sum_{j\neq i} D_j Z_j \mathbb{1}(X_j = X_i) \times \sum_{j\neq i} \mathbb{1}(X_j = X_i) - \sum_{j\neq i} D_j \mathbb{1}(X_j = X_i) \times \sum_{j\neq i} Z_j \mathbb{1}(X_j = X_i)}.$$

Let further $\hat{W}_i = Y_i + (1 - D_i) \times \hat{\delta}(X_i)$. We are now ready to define our test statistic:

$$\hat{\mathcal{T}}_n = \sup_{w\in\mathbb{R};\ x\in\mathscr{S}_X} \sqrt{n}\ \left|\hat{F}_{\hat{W}|XZ}(w|x,0) - \hat{F}_{\hat{W}|XZ}(w|x,1)\right|,$$

where $\hat{F}_{\hat{W}|XZ}(w|x,z) = \frac{\sum_{i=1}^n \mathbb{1}(\hat{W}_i\leq w)\mathbb{1}(X_i=x,Z_i=z)}{\sum_{i=1}^n \mathbb{1}(X_i=x,Z_i=z)}$.

Next, we establish the limiting distribution of $\hat{\mathcal{T}}_n$. For expositional simplicity, denote $\mathbb{1}_{XZ}(x,z) \equiv \mathbb{1}(X = x, Z = z)$ and $f_{WD|XZ}(w,d|x,z) \equiv f_{W|DXZ}(w|d,x,z) \times \Pr(D = d|X = x, Z = z)$. Moreover, let

$$\kappa(w,x) \equiv -\frac{f_{WD|XZ}(w,0|x,1) - f_{WD|XZ}(w,0|x,0)}{p(x,1) - p(x,0)}.$$

Note that under Assumptions 2.1 and 2.3, $\kappa(w,x) \geq 0$ since it becomes the conditional density of $g(0,x,\epsilon)$ given the complier group and $X = x$. Moreover, let

$$\psi_{wx} \equiv [\mathbb{1}(W \leq w) - F_{W|X}(w|x)] \times \Big[\frac{\mathbb{1}_{XZ}(x,1)}{\Pr(X=x,Z=1)} - \frac{\mathbb{1}_{XZ}(x,0)}{\Pr(X=x,Z=0)}\Big]; \tag{2.6}$$

$$\phi_{wx} \equiv \kappa(w,x)[W - \mathbb{E}(W|X=x)] \times \Big[\frac{\mathbb{1}_{XZ}(x,1)}{\Pr(X=x,Z=1)} - \frac{\mathbb{1}_{XZ}(x,0)}{\Pr(X=x,Z=0)}\Big]. \tag{2.7}$$

By definition, ψ_{wx} and ϕ_{wx} are random processes indexed by (w,x).

Assumption 2.5. *Let X be a discrete random variable with a finite support. Moreover, the probability distribution of Y given (D,X,Z) admits a uniformly continuous density function $f_{Y|DXZ}$ and $\mathbb{E}(Y^2) < \infty$.*

Theorem 2.2. *Suppose Assumptions 2.1 to 2.5 hold. Then, under* $\mathbb{H}_0$*,*

$$\hat{\mathcal{T}}_n \xrightarrow{d} \sup_{w\in\mathbb{R};\ x\in\mathscr{S}_X} |\mathcal{Z}(w,x)|,$$

where $\mathcal{Z}(\cdot,x)$ *is a mean–zero Gaussian process with covariance kernel:*

$$\textit{Cov}\left[\mathcal{Z}(w,x),\mathcal{Z}(w',x)\right] = \mathbb{E}\left[(\psi_{wx}+\phi_{wx})(\psi_{w'x}+\phi_{w'x})\right],\ \forall w,w'\in\mathbb{R}.$$

Moreover, under $\mathbb{H}_1$*, we have*

$$n^{-\frac{1}{2}}\hat{\mathcal{T}}_n \xrightarrow{p} \sup_{w\in\mathbb{R};\ x\in\mathscr{S}_X} \left|F_{W|XZ}(w|x,0) - F_{W|XZ}(w|x,1)\right|.$$

In the covariance kernel Cov $[\mathcal{Z}(w,x),\mathcal{Z}(w',x')]$, ϕ_{wx} and $\phi_{w'x'}$ appear due to the estimation of $\delta(x)$. By Theorem 2.2, our test is one–sided: reject $\mathbb{H}_0$ at significance level α if and only if $\hat{\mathcal{T}}_n \geq c_\alpha$, where c_α is the $(1-\alpha)$-th quantile of $\sup_{w\in\mathbb{R};\ x\in\mathscr{S}_X} |\mathcal{Z}(w,x)|$.

Because the asymptotic distribution of $\sup_{w\in\mathbb{R};\ x\in\mathscr{S}_X} |\mathcal{Z}(w,x)|$ is complicated, then we apply the multiplier bootstrap method to approximate the entire process for the critical value. See e.g. van der Vaart and Wellner (1996), Delgado and Manteiga (2001) and Barrett and Donald (2003). Specifically, we simulate a sequence of i.i.d. pseudo random variables $\{U_i : i = 1,\cdots,n\}$ with $\mathbb{E}(U) = 0$, $\mathbb{E}(U^2) = 1$, and $\mathbb{E}(U^4) < +\infty$. Moreover, the simulated sample $\{U_i : i = 1,\cdots,n\}$ is independent of the random sample $\{(Y_i, X_i, D_i, Z_i) : i = 1,\cdots,n\}$. Then, we obtain the following simulated empirical process:

$$\hat{\mathcal{Z}}^u(w,x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \times (\hat{\psi}_{wx,i} + \hat{\phi}_{wx,i}),$$

where $\hat{\psi}_{wx,i} + \hat{\phi}_{wx,i}$ is the estimated influence function. Namely,

$$\hat{\psi}_{wx,i} = \left[\mathbb{1}(\hat{W}_i \leq w) - \frac{\sum_{j=1}^n \mathbb{1}(\hat{W}_j \leq w; X_j = x)}{\sum_{j=1}^n \mathbb{1}(X_j = x)}\right] \times \left[\frac{\mathbb{1}(X_i = x, Z_i = 0)}{\hat{\Pr}(X = x, Z = 0)} - \frac{\mathbb{1}(X_i = x, Z_i = 1)}{\hat{\Pr}(X = x, Z = 1)}\right];$$

$$\hat{\phi}_{wx,i} = \hat{\kappa}(w, x)\left[\hat{W}_i - \frac{\sum_{j=1}^n \hat{W}_j \mathbb{1}(X_j = x)}{\sum_{j=1}^n \mathbb{1}(X_j = x)}\right] \times \left[\frac{\mathbb{1}(X_i = x, Z_i = 0)}{\hat{\Pr}(X = x, Z = 0)} - \frac{\mathbb{1}(X_i = x, Z_i = 0)}{\hat{\Pr}(X = x, Z = 1)}\right],$$

where $\hat{\Pr}(X = x, Z = z) = \frac{1}{n}\sum_{j=1}^n \mathbb{1}(X_j = x, Z_j = z)$ and

$$\hat{\kappa}(w, x) = -\frac{\hat{f}_{WD|XZ}(w, 0|x, 1) - \hat{f}_{WD|XZ}(w, 0|x, 0)}{\hat{p}(x, 1) - \hat{p}(x, 0)}.$$

In the definition of $\hat{\kappa}(w, x)$, $\hat{f}_{WD|XZ}(w, 0|x, z) = \frac{\sum_{j=1}^n \mathbb{1}(D_j=0, X_j=x, Z_j=z) \times \frac{1}{h} K(\frac{\hat{W}_j - w}{h})}{\sum_{j=1}^n \mathbb{1}(X_j=x, Z_j=z)}$, where K and h be a bounded kernel function and a smoothing bandwidth, respectively, and $\hat{p}(x, z) = \frac{\sum_{j=1}^n \mathbb{1}(D_j=1, X_j=x, Z_j=z)}{\sum_{j=1}^n \mathbb{1}(X_j=x, Z_j=z)}$. For a given significant level α, the critical value $\hat{c}_n(\alpha)$ is obtained as the $(1-\alpha)$–quantile of the simulated distribution of $\sup_{w \in \mathbb{R},\ x \in \mathscr{S}_X} \left|\hat{\mathcal{Z}}^u(w, x)\right|$.

2.2.2 Case 2: Continuous Covariates

We now consider the case where $X \in \mathbb{R}$ is continuously distributed with a finite support. To extend the empirical process argument used in the proof of Theorem 2.2 to this case, we propose a modified Kolmogorov–Smirnov test statistic. Such a modification allows the generated variable W to be constructed from the unknown function $\delta(\cdot)$ as an infinite–dimensional parameter.

Let $\lambda(t) = -t \times \mathbb{1}(t \leq 0)$ and $\Pi(w|x, z) = \mathbb{E}[\lambda(W - w)|X = x, Z = z]$. By definition, $\Pi(\cdot|x, z)$ is the primitive function of the $F_{W|XZ}(\cdot|x, z)$, i.e.,

$$\frac{\partial}{\partial w}\Pi(w|x, z) = F_{W|XZ}(w|x, z).$$

Thus, the model restriction $W \perp\!\!\!\perp Z \mid X$ can be equivalently characterized as follows

$$\Pi(w|x,0) = \Pi(w|x,1), \ \ \forall (w,x) \in \mathbb{R} \times \mathscr{S}_X.$$

It should also be noted that $\lambda(\cdot)$ is continuous and has a directional derivative. For simplicity, we assume $\mathscr{S}_W$ is bounded.

We denote $f_{XZ}(x,z) \equiv f_{X|Z}(x|z) \times \mathbb{P}(Z=z)$ and $\mathbb{1}^*_{XZ}(x,z) \equiv \mathbb{1}(X \leq x; Z = z)$. For $z \in \{0,1\}$, let $z' = 1 - z$ and

$$G(w,x,z) = \mathbb{E}\left[\lambda(W-w)\mathbb{1}^*_{XZ}(x,z) f_{XZ}(X,z')\right].$$

Motivated by Stinchcombe and White (1998), we represent the above conditional expectation restrictions by the following unconditional ones:

$$G(w,x,0) = G(w,x,1), \ \ \forall (w,x) \in \mathbb{R} \times \mathscr{S}_X. \tag{2.8}$$

To see the equivalence, first note that

$$G(w,x,z) = \mathbb{E}\left[\lambda(W-w)\mathbb{1}(X \leq x) f_{X|Z}(X|z')|Z=z\right] \mathbb{P}(Z=0)\mathbb{P}(Z=1).$$

Moreover, by the law of iterated expectation,

$$\frac{\partial}{\partial x}\mathbb{E}[\lambda(W-w)\mathbb{1}(X \leq x) f_{X|Z}(X|z')|Z=z] = \Pi(w|x,z) f_{X|Z}(x|0) f_{X|Z}(x|1).$$

Therefore, we obtain the conditional expectation restrictions as the derivative of (2.8). Note that the estimation of $G(w,x,z)$ avoids any denominator issues, which thereafter simplifies our asymptotic analysis.

Let K and h be a bounded kernel function and a smoothing bandwidth, respectively. Then, we

estimate $\delta(X_i)$ by

$$\hat{\delta}(X_i) = \frac{\sum_{j\neq i} Y_j Z_j K\left(\frac{X_j - X_i}{h}\right) \times \sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right) - \sum_{j\neq i} Y_j K\left(\frac{X_j - X_i}{h}\right) \times \sum_{j\neq i} Z_j K\left(\frac{X_j - X_i}{h}\right)}{\sum_{j\neq i} D_j Z_j K\left(\frac{X_j - X_i}{h}\right) \times \sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right) - \sum_{j\neq i} D_j K\left(\frac{X_j - X_i}{h}\right) \times \sum_{j\neq i} Z_j K\left(\frac{X_j - X_i}{h}\right)}.$$

Moreover, let

$$\hat{f}_{XZ}(X_i, z) = \frac{1}{nh} \sum_{j\neq i} K\big(\frac{X_j - X_i}{h}\big) \mathbb{1}(Z_j = z);$$
$$\hat{G}(w, x, z) = \frac{1}{n} \sum_{i=1}^{n} \lambda(\hat{W}_i - w) \mathbb{1}^*_{X_i Z_i}(x, z) \hat{f}_{XZ}(X_i, z').$$

Thus, we define our test statistic as follows:

$$\hat{\mathcal{T}}_n^c = \sup_{(w,x)\in \mathscr{S}_{WX}} \sqrt{n} \left| \hat{G}(w, x, 0) - \hat{G}(w, x, 1) \right|.$$

In above definition, the support $\mathscr{S}_{WX}$ is assumed to be known for simplicity. In practice, this assumption can be relaxed by using a consistent set estimator $\hat{\mathcal{S}}_{WX}$ of $\mathscr{S}_{WX}$.

We show that the proposed test statistic $\hat{\mathcal{T}}_n^c$ converges in distribution at the regular parameter rate. The key step of our proof is to show that

$$\sup_{(w,x)\in \mathscr{S}_{WX}} \left| \hat{G}(w, x, z) - \tilde{G}(w, x, z) \right| = o_p\left(n^{-1/2}\right). \tag{2.9}$$

where $\tilde{G}(w, x, z) = \frac{1}{n} \sum_{i=1}^{n} (w - \hat{W}_i) \mathbb{1}(W_i \leq w) \times \mathbb{1}^*_{X_i Z_i}(x, z) \hat{f}_{XZ}(X_i, z')$. The above result requires that the nonparametric elements in the estimation of $\hat{\delta}(\cdot)$ should converge to the corresponding true values uniformly at a rate faster than $n^{-1/4}$.

Assumption 2.6. *The support $\mathscr{S}_{WX} \subseteq \mathbb{R}$ is compact. For $z = 0, 1$, $\sup_{(x,z)\in \mathscr{S}_{XZ}} f_{X|Z}(x|z) \leq \overline{f} < +\infty$ and $\inf_{x\in \mathscr{S}_X} |f_{XZ}(x, 1) - f_{XZ}(x, 0)| > 0$.*

Assumption 2.7. *For $z \in \{0, 1\}$, functions $f_{X|Z}(x|z)$, $p(x, z)$ and $\mu(x, z)$ are continuous in x.*

Assumption 2.8. *The support of K is a convex (possibly unbounded) subset of $\mathbb{R}$ with nonempty interior, with the origin as an interior point. $K(\cdot)$ is a bounded differentiable function such that $\int K(u) = 1$, $\int uK(u) = 0$, and $K(u) = K(-u)$ holds for all u in the support.*

Assumption 2.9. *Let $\iota > \frac{1}{4}$. As $n \to \infty$, $h \to 0$ and $n^{\iota}/\sqrt{nh} \to 0$.*

Assumption 2.10. *The first-stage estimators satisfy:*

$$\sup_{(x,z)\in\mathscr{S}_{XZ}} \left|\mathbb{E}\Big[\frac{1}{nh}\sum_{j=1}^{n} \mathbb{1}(Z_j = z)K\big(\frac{X_j - x}{h}\big)\Big] - f_{XZ}(x,z)\right| = O_p(n^{-\iota}),$$

$$\sup_{(x,z)\in\mathscr{S}_{XZ}} \left|\mathbb{E}\Big[\frac{1}{nh}\sum_{j=1}^{n} D_j\mathbb{1}(Z_j = z)K\big(\frac{X_j - x}{h}\big)\Big] - p(x,z)f_{XZ}(x,z)\right| = O_p(n^{-\iota}),$$

$$\sup_{(x,z)\in\mathscr{S}_{XZ}} \left|\mathbb{E}\Big[\frac{1}{nh}\sum_{j=1}^{n} Y_j\mathbb{1}(Z_j = z)K\big(\frac{X_j - x}{h}\big)\Big] - \mathbb{E}(Y|X = x, Z = z)f_{XZ}(x,z)\right| = O_p(n^{-\iota}).$$

Assumptions 2.6 to 2.9 are standard in the semiparametric estimation literature, ensuring that the first–stage nonparametric estimators converge to its expectation at a rate faster than $n^{1/4}$. Note that Assumption 2.9 implies that the $h \gg n^{-1/2}$. Moreover, Assumption 2.10 is a high-level condition that requires the nonparametric estimation bias diminishes uniformly at a rate faster than $n^{1/4}$. Such a condition on the bias term can be satisfied under additional primitive conditions on $K(\cdot)$ and h, as well as the smoothness of the underlying structural functions. See e.g. Pagan and Ullah (1999).

Lemma 2.2. *Suppose Assumptions 2.6 to 2.10 hold. Then,* (2.9) *holds for $z = 0, 1$.*

By Lemma 2.2, it suffices to establish the limiting distribution of $\tilde{G}(w,x,1) - \tilde{G}(w,x,0)$ for the asymptotic properties of our test statistics. Note that in the definition of $\tilde{G}(w,x,z)$, there contains no nonparametric elements estimated in the indicate function.

To establish asymptotic properties for inference, we make the following assumption.

Assumption 2.11. $\sup_{x\in\mathscr{S}_X} \left|\mathbb{E}[\hat{\delta}(x)] - \delta(x)\right| = o_p(n^{-\frac{1}{2}})$ *and* $\sup_{xz\in\mathscr{S}_{XZ}} \left|\mathbb{E}[\hat{f}_{XZ}(x,z)] - f_{XZ}(x,z)\right| = o_p(n^{-\frac{1}{2}})$.

Assumption 2.11 strengthens Assumption 2.10 by requiring the bias term in the first–stage non-parametric estimation to be smaller than $o_p(n^{-1/2})$, which can be established by using high order kernels (see e.g. Powell *et al.*, 1989).

Let $F^*_{WD|XZ}(w,d|x,z) \equiv F_{W|DXZ}(w|d,x,z) \times \Pr(D=d|X=x,Z=z)$ and

$$\kappa^c(w,x) = -\frac{F^*_{WD|XZ}(w,0|x,1) - F^*_{WD|XZ}(w,0|x,0)}{p(x,1)-p(x,0)}.$$

Moreover, we define two random process indexed by (w,x) as follows:

$$\psi^c_{wx} = \Big\{\lambda(w-W) - \mathbb{E}[\lambda(w-W)|X]\Big\}\Big[\frac{\mathbb{1}^*_{XZ}(x,1)}{f_{XZ}(X,1)} - \frac{\mathbb{1}^*_{XZ}(x,0)}{f_{XZ}(X,0)}\Big] f_{XZ}(X,0) f_{XZ}(X,1);$$
$$\phi^c_{wx} = \kappa^c(w,X)[W - \mathbb{E}(W|X)]\Big[\frac{\mathbb{1}^*_{XZ}(x,1)}{f_{XZ}(X,1)} - \frac{\mathbb{1}^*_{XZ}(x,0)}{f_{XZ}(X,0)}\Big] f_{XZ}(X,0) f_{XZ}(X,1).$$

By definition, we have $\mathbb{E}(\psi^c_{wx}|X,Z) = \mathbb{E}(\phi^c_{wx}|X,Z) = 0$ under $\mathbb{H}_0$.

Theorem 2.3. *Suppose the assumptions in Lemma 2.2 and Assumption 2.11 hold. Then, under* $\mathbb{H}_0$*,*

$$\hat{\mathcal{T}}^c_n \xrightarrow{d} \sup_{w\in\mathbb{R};\ x\in\mathscr{S}_X} |\mathcal{Z}^c(w,x)|$$

where $\mathcal{Z}^c(\cdot,\cdot)$ *is a mean–zero Gaussian process with the following covariance kernel*

$$Cov\left[\mathcal{Z}^c(w,x), \mathcal{Z}^c(w',x')\right] = \mathbb{E}\left[(\psi^c_{wx} + \phi^c_{wx})(\psi^c_{w'x'} + \phi^c_{w'x'})\right],\ \forall (w,x),(w',x') \in \mathbb{R}\times\mathscr{S}_X.$$

Moreover, under $\mathbb{H}_1$*, we have*

$$n^{-\frac{1}{2}}\hat{\mathcal{T}}^c_n \xrightarrow{p} \sup_{w\in\mathbb{R};\ x\in\mathscr{S}_X} |G(w,x,0) - G(w,x,1)|.$$

Similar to the discrete–covariates case, we reject $\mathbb{H}_0$ at significance level α if and only if $\hat{\mathcal{T}}^c_n \geq c_\alpha$. Moreover, we apply the multiplier bootstrap method to approximate the entire process and therefore to obtain critical values.

2.3 Monte Carlo Simulations

In this section, we investigate the finite sample performance of our tests with a simulation study. The data are simulated as follows:

$$Y = D + X + [\gamma + (1-\gamma)D] \times \epsilon;$$
$$D = \mathbb{1}\left[\Phi(\eta) \leq 0.5 \times Z\right],$$

where (ϵ, η) conforms to a joint normal distribution with zero mean, unit variance and correlation coefficient $\rho = 0.7$, and $Z \sim Bernoulli(p)$ with $p = 0.25, 0.5$ and 0.75 respectively. Note that we also try different values for the correlation coefficient and all the results are qualitatively similar.[2] For simplicity, X, Z and (ϵ, η) are mutually independent. Moreover, X is uniformly distributed on $\{1, 2, 3, 4\}$ and on $[0, 1]$, respectively, in the discrete covariates and the continuous covariates case. Furthermore, parameter $\gamma \in [0, 1]$ describes the degree of unobserved heterogeneous treatment effects in our specification. In particular, $\mathbb{H}_0$ holds if and only if $\gamma = 1$. Intuitively, smaller γ, more power we expect from our tests. To investigate size and power of our tests, we choose $\gamma \in \{1, 0.75, 0.5\}$.

We consider sample size $n = 1000, 2000, 4000$, a nominal level of $\alpha = 5\%$, and $2,000$ Monte Carlo repetitions. To compute the suprema of the simulated stochastic processes, we use $n/10$ grids on the support of $[\min_{i=1}^{n}(\hat{W}_i), \max_{i=1}^{n}(\hat{W}_i)]$. Moreover, we use 500 multiplier bootstrap samples to simulate the p-values. Regarding the estimation of $\kappa(w, x)$, we choose the second order Gaussian kernel function with the bandwidth, $h_n = c \cdot \text{std}(\hat{W}) \cdot n^{-1/5}$, and we set $c \in \{0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3\}$ to study the sensitivity of the test to the bandwidth.

Table 2.1 reports rejection probabilities of our simulations in the discrete-covariates case under the null hypothesis (i.e. $\gamma = 1$) and alternative hypotheses (i.e. $\gamma = 0.75, 0.5$). From Panel A, the level of our test is fairly well behaved: It gets closer to the nominal level as the sample size increases and the rejection probabilities are not sensitive to the constant c for the bandwidth

[2] Additional Monte Carlo simulation results are available upon request.

choice. Panels B and C show that the power of the test is reasonable. In particular, when γ is closer to 1, it is more difficult to detect such a "local" alternative. Therefore, we obtain relatively small power even when sample size reaches $n = 2000$ in Panel B. For relatively "small" sample size, e.g., $n = 1000$, our results show that our test performs better with a larger bandwidth choice. Moreover, when p (i.e. the probability of $Z = 1$) is 0.5, all the results for size and power dominate the other two cases with $p = 0.25, 0.75$, which is expected by our asymptotic theory.

Next, we evaluate the performance of our tests in the case where the covariate X is continuous. To compute the suprema, we calculate the test statistic by using $n/20$ grid points in the support $[\min_{i=1}^n(\hat{W}_i), \max_{i=1}^n(\hat{W}_i)]$, as well as in the support $[\min_{i=1}^n(X_i), \max_{i=1}^n(X_i)]$. Table 2.2 reports the size and power properties of our test, which are qualitatively similar to the results in the discrete-covariates case.

2.4 Empirical Applications

2.4.1 The Effect of Job Training Program on Earnings

We now apply our tests to study the effects of the job training program on earnings, i.e., the *National Job Training Partnership Act* (JTPA), commissioned by the Department of Labor. This program began funding training from 1983 to late 1990's to increase employment and earnings for participants. The major component of JTPA aims to support training for the economically disadvantaged. The effects of JTPA training programs on earnings have also been studied by e.g. Heckman *et al.* (1997); Abadie *et al.* (2002) under a general framework allowing for unobserved heterogeneous treatment effects.[3]

Our sample consists of 11,204 observations from the JTPA, a survey dataset from over $20,000$ adults and out-of-school youths who applied for JTPA in 16 local areas across the country between 1987 and 1989.[4] Each participant was assigned randomly to either a program group or a control group (1 out of 3 on average). Members of the program group are eligible to participate JTPA ser-

[3]The data is publicly available at http://upjohn.org/services/resources/employment-research-data-center/national-jtpa-study.

[4]JTPA services are provided at 649 sites, which might not be randomly chosen. For a given site, the applicants were randomly selected for the JTPA dataset.

Table 2.1: Rejection probabilities ($\alpha = 5\%$) in the discrete–covariates case.

p	n	$c = 0.7$	$c = 0.8$	$c = 0.9$	$c = 1.0$	$c = 1.1$	$c = 1.2$	$c = 1.3$
Panel A: rejection probabilities at null hypothesis with $\gamma = 1$								
0.25	1000	0.0025	0.0045	0.0080	0.0105	0.0140	0.0190	0.0250
	2000	0.0130	0.0160	0.0230	0.0275	0.0330	0.0345	0.0415
	4000	0.0265	0.0320	0.0415	0.0460	0.0490	0.0530	0.0575
0.5	1000	0.0090	0.0120	0.0160	0.0235	0.0300	0.0395	0.0460
	2000	0.0250	0.0300	0.0340	0.0410	0.0415	0.0445	0.0490
	4000	0.0350	0.0430	0.0500	0.0525	0.0565	0.0610	0.0625
0.75	1000	0.0040	0.0075	0.0135	0.0180	0.0270	0.0335	0.0390
	2000	0.0140	0.0210	0.0245	0.0285	0.0360	0.0415	0.0480
	4000	0.0230	0.0280	0.0340	0.0390	0.0455	0.0505	0.0570
Panel B: rejection probabilities at alternative hypothesis with $\gamma = 0.75$								
0.25	1000	0.0125	0.0205	0.0340	0.0490	0.0605	0.0745	0.0885
	2000	0.0810	0.1065	0.1370	0.1610	0.1805	0.1985	0.2120
	4000	0.2610	0.2930	0.3160	0.3385	0.3600	0.3780	0.3935
0.5	1000	0.0390	0.0585	0.0775	0.1005	0.1185	0.1340	0.1405
	2000	0.1590	0.1920	0.2205	0.2485	0.2675	0.2830	0.2970
	4000	0.4360	0.4705	0.4945	0.5240	0.5395	0.5510	0.5730
0.75	1000	0.0230	0.0395	0.0540	0.0700	0.0855	0.1010	0.1100
	2000	0.0970	0.1260	0.1525	0.1710	0.1880	0.2050	0.2175
	4000	0.3035	0.3300	0.3565	0.3775	0.3955	0.4120	0.4245
Panel C: rejection probabilities at alternative hypothesis with $\gamma = 0.50$								
0.25	1000	0.1975	0.2760	0.3515	0.4145	0.4490	0.4790	0.5030
	2000	0.7335	0.8010	0.8445	0.8705	0.8870	0.8985	0.9045
	4000	0.9985	0.9990	0.9990	0.9990	0.9990	0.9990	0.9990
0.5	1000	0.5215	0.5915	0.6445	0.6860	0.7065	0.7155	0.7255
	2000	0.9630	0.9715	0.9750	0.9780	0.9825	0.9820	0.9835
	4000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.75	1000	0.3600	0.4330	0.4815	0.5135	0.5295	0.5370	0.5370
	2000	0.8645	0.8915	0.9070	0.9180	0.9220	0.9260	0.9265
	4000	0.9990	0.9990	0.9990	0.9995	0.9990	0.9990	0.9990

Table 2.2: Rejection probabilities ($\alpha = 5\%$) in the continuous–covariates case.

p	n	$c = 0.7$	$c = 0.8$	$c = 0.9$	$c = 1.0$	$c = 1.1$	$c = 1.2$	$c = 1.3$
Panel A: rejection probabilities at null hypothesis with $\gamma = 1$								
0.25	1000	0.0695	0.0630	0.0595	0.0575	0.0525	0.0540	0.0565
	2000	0.0620	0.0560	0.0555	0.0590	0.0570	0.0580	0.0590
	4000	0.0690	0.0690	0.0630	0.0650	0.0620	0.0570	0.0565
0.5	1000	0.0510	0.0520	0.0520	0.0505	0.0515	0.0520	0.0525
	2000	0.0590	0.0605	0.0575	0.0600	0.0630	0.0635	0.0650
	4000	0.0670	0.0620	0.0630	0.0620	0.0630	0.0555	0.0585
0.75	1000	0.0495	0.0485	0.0485	0.0480	0.0480	0.0470	0.0490
	2000	0.0450	0.0450	0.0490	0.0480	0.0470	0.0485	0.0455
	4000	0.0540	0.0560	0.0540	0.0510	0.0520	0.0515	0.0535
Panel B: rejection probabilities at alternative hypothesis with $\gamma = 0.75$								
0.25	1000	0.0805	0.0760	0.0730	0.0675	0.0635	0.0585	0.0585
	2000	0.1820	0.1570	0.1405	0.1210	0.1065	0.0920	0.0890
	4000	0.5730	0.5110	0.4550	0.4010	0.3560	0.3035	0.2655
0.5	1000	0.0960	0.0935	0.0890	0.0775	0.0720	0.0705	0.0690
	2000	0.3020	0.2700	0.2285	0.2000	0.1695	0.1490	0.1340
	4000	0.8160	0.7630	0.7170	0.6520	0.5940	0.5285	0.4805
0.75	1000	0.0585	0.0605	0.0580	0.0575	0.0560	0.0540	0.0520
	2000	0.1535	0.1400	0.1230	0.1080	0.0910	0.0780	0.0690
	4000	0.5450	0.4840	0.4300	0.3730	0.3220	0.2770	0.2410
Panel C: rejection probabilities at alternative hypothesis with $\gamma = 0.50$								
0.25	1000	0.6950	0.6620	0.6295	0.5940	0.5470	0.5200	0.4765
	2000	0.9925	0.9895	0.9850	0.9805	0.9720	0.9630	0.9525
	4000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.5	1000	0.9205	0.8970	0.8700	0.8370	0.8015	0.7560	0.7160
	2000	1.0000	1.0000	1.0000	0.9990	0.9985	0.9975	0.9970
	4000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.75	1000	0.7150	0.6665	0.6155	0.5685	0.5135	0.4500	0.4135
	2000	0.9990	0.9970	0.9935	0.9845	0.9705	0.9565	0.9370
	4000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

vices, including classroom training, on-the-job training or job search assistance, and other services, while members of control group are not eligible for JTPA services for 18 months. Following the literature (see e.g. Bloom *et al.*, 1997), we use the program eligibility as an instrumental variable for the endogenous individual participation decision.

The outcome variable is individual earnings, measured by the sum of earnings in the 30-month period following the offer. The observed covariates include a set of dummies for races, for high-school graduates, and for marriage, for whether the applicant worked at least 12 weeks in the 12 months preceding random assignment, and also 5 age-group dummies (22-24, 25-29, 30-35, 36-44, and 45-54), among others. Descriptive statistics can be found in Table 1 of Abadie *et al.* (2002). For simplicity, we group all applicants into 3 age categories (22-29, 30-35, and 36 and above), and pool all non-White applicants as minority applicants.

To implement the test, we use the second order Gaussian kernel with several bandwidth choices for robustness check. For the critical value, we use $10,000$ multiplier bootstrap samples and search for the suprema by using $5,000$ grid points. We select the smoothing parameter by $1.06 \cdot \text{Std}(\hat{W}) \cdot n^{-1/4}$. The p-value of our test is 0.1204. Therefore, the null hypothesis (i.e. no unobserved heterogenous treatment effects) cannot be rejected at the 10% significance level. Our results are robust to the size of bootstrap samples, the number of grid points, and the choice of bandwidth.

2.4.2 The Impact of Fertility on Family Income

The second empirical illustration considers the heterogeneous impacts of children on parents' labor supply and income. Recently, Frölich and Melly (2013) have studied the heterogeneous effects of fertility on family income within the general LATE framework. To deal with the endogeneity of fertility decisions, Rosenzweig and Wolpin (1980); Angrist and Evans (1998); Bronars and Grogger (1994); Jacobsen *et al.* (1999), among many others, suggest to use the twin births as an instrumental variable.

Our data comes from the 1% and 5% Census Public Use Micro Sample (PUMS) from 1990 and 2000 censuses, consisting of 602,767 and 573,437 observations, respectively.[5] Similar to Frölich

[5]The data is publicly available at https://www.census.gov/main/www/pums.html.

and Melly (2013), our sample is restricted to 21–35 years old married mothers with at least one child since we use twin birth as an instrument for fertility. The outcome variable of interest is the family's annual labor income.[6] The treatment variable is a dummy variable that takes the value 1 to indicate when a mother has two or more children. The instrumental variable is also a dummy variable and it equals 1 if the first birth is a twin. The covariates include mother's and father's age, race, educational level, and working status. Table 2.3 provides descriptive statistics. Some covariates, i.e., age, years in education, and working hours per week, are treated as continuous variables.

Similar to the previous empirical illustration, we use the second kernel Gaussian kernel with various bandwidth choices for robustness check. For the critical value, we use $5,000$ bootstrapped samples and search for the suprema by using $1,000$ grids for each of the support of W and X's. The bandwidths are selected by the same manners as those in the JTPA case. The p-values of our tests are 0.0031 and 0.0004 for the 1990 and 2000 Censuses, respectively. These results suggest that the null hypothesis, i.e., homogeneous treatment effects, should be rejected at all usual significance levels.

[6]It includes wages, salary, armed forces pay, commissions, tips, piece-rate payments, cash bonuses earned before deductions were made for taxes, bonds, pensions, union dues, etc. See Frölich and Melly (2013) for more details.

Table 2.3: Descriptive Statistics for the 1999 and 2000 Censuses

	1990			2000		
	All	$Z = 1$ (twin birth)	$Z = 0$ (no twin birth)	All	$Z = 1$ (twin birth)	$Z = 0$ (no twin birth)
Observations	602,767	6,524	596,243	573,437	8,569	564,868
Number of children	1.9276	2.5318	1.9209	1.8833	2.5196	1.8734
At least two children ($D = 1$)	0.6500	1.0000	0.6461	0.6163	1.0000	0.6104
Mother						
Age in years	29.7894	29.9530	29.7876	30.0562	30.3943	30.0510
Years of education	12.9196	12.9623	12.9191	13.1131	13.2615	13.1108
Black	0.0637	0.0757	0.0636	0.0724	0.0816	0.07228
Asian	0.0326	0.0321	0.0326	0.0447	0.0335	0.0448
Other Races	0.0537	0.0592	0.0536	0.0912	0.0806	0.0914
Currently at work	0.5781	0.5444	0.5785	0.5629	0.5132	0.5637
Usual hours per work	24.5660	23.3537	24.5795	25.1400	23.0491	25.1723
Wage or salary income last year	8942	8593	8946	14200	13757	14206
Father						
Age in years	32.5358	32.7534	32.5333	32.9291	33.3102	32.9232
Years of education	13.0436	13.0748	13.0432	13.0331	13.1806	13.0308
Black	0.0671	0.0796	0.0670	0.0800	0.0945	0.0798
Asian	0.0291	0.0263	0.0292	0.0402	0.0318	0.0403
Other Races	0.0488	0.0529	0.0488	0.0919	0.0802	0.0921
Currently at work	0.8973	0.8922	0.8974	0.8512	0.8584	0.8511
Usual hours per work	42.7636	42.7704	42.7635	43.8805	43.8789	43.8805
Wage or salary income last year	27020	28039	27010	38041	41584	37987
Parents						
Wages or salary income last year	35,963	36,632	35,956	52,241	55,342	52,193

Note: Data from the 1% and 5% PUMS in 1990 and 2000. Own calculations using the PUMS sample weights. The sample consists of married mother between 21 and 35 years of age with at least one child.

2.5 Extensions

Our analysis naturally extends to the Fuzzy Regression Discontinuity (FRD) design, which has recently become a popular tool to address causal inference questions in empirical studies (see e.g. Van der Klaauw, 2008; Imbens and Lemieux, 2008; Lee and Lemieuxa, 2010, for reviews).

Consider a nonparametric FRD design: Let

$$Y = Y(0) \times (1 - D) + Y(1) \times D,$$

where Y is the observed outcome variable, $(Y(0), Y(1)) \in \mathbb{R}^2$ denotes a pair of potential outcomes, and D is the observed treatment status. Moreover, let $X \in \mathbb{R}^{dx}$ be a vector of covariates. The assignment of the treatment is given by

$$D = \mathbb{1}\left[\theta(X, R) \leq \eta\right], \tag{2.10}$$

where R is a continuous running variable, and $\theta(\cdot, \cdot)$ is monotone in R, and $\eta \in \mathbb{R}$ is an unobserved error term. Moreover, let $R = 0$ be the cutoff point of the running variable, and we assume the probability of receiving the treatment is a continuous function in the running variable, except at the cutoff point, i.e.,

$$\lim_{r \downarrow 0} \mathbb{P}(D = 1 | X = x, R = r) \neq \lim_{r \uparrow 0} \mathbb{P}(D = 1 | X = x, R = r), \ \ \forall x \in \mathscr{S}_X.$$

In the FRD designs literature, the estimand of interest is

$$\tau(x) = \frac{\lim_{r \downarrow 0} \mathbb{E}[Y | X = x, R = r] - \lim_{r \uparrow 0} \mathbb{E}[Y | X = x, R = r]}{\lim_{r \downarrow 0} \mathbb{E}(D | X = x, R = r) - \lim_{r \uparrow 0} E(D | X = x, R = r)}.$$

Similarly, under homogeneous treatment effects, $\tau(x)$ can be interpreted as the average treatment effect at the threshold $R = 0$ and given $X = x$. Hence, the hypotheses for testing homogeneous

treatment effects can be formulated as

$$\mathbb{H}_0^* : \mathbb{P}\big[Y(1) - Y(0) = \tau(X) | X, R = 0\big] = 1, \;\; a.s.$$

$$\mathbb{H}_1^* : \mathbb{H}_0^* \text{ is false.}$$

Similarly, we can test such a hypothesis by testing a conditional independence assumption. Specifically, let $W^* = Y + (1-D)\cdot\tau(X)$. Under additional weak assumptions and by a similar argument to Theorem 2.1, it can be shown that $\mathbb{H}_0^*$ holds if and only if

$$\lim_{r\uparrow 0} F_{W^*|XR}(\cdot|x,r) = \lim_{r\downarrow 0} F_{W^*|XR}(\cdot|x,r), \;\; \forall x \in \mathscr{S}_X.$$

An important question is then how to test such a model restriction. It is of considerable interest to provide a theoretic study of this test.

3. IMPLIED VOLATILITY ESTIMATION VIA ℓ_1 TREND FILTERING

3.1 Theory and Estimation Strategy

3.1.1 State Price Densities, Implied Volatility, and Option Prices

Arrow-Debreu securities are one of the fundamental economic objects that provide information about the nature of equilibrium in the presence of uncertainty. The state price densities (SPDs) are ultimately defined as the PDFs of the distributions that describe the prices of Arrow-Debreu securities over a continuum of states. Thus, they describe the pricing kernel or discount factor for these securities. They have also been called "risk neutral densities" in arbitrage-free models. Based on the observations of Ross (1976) and Cox and Ross (1976) it is possible to show that the value of an option is ultimately determined by the state price density in an arbitrage-free environment.

With this theoretical understanding, Breeden and Litzenberger (1978) show that the form of the SPD can be extracted from option prices via a second derivative with respect to the strike prices. Their derivation leaves us with the following equation:[1]

$$\tilde{q} = \exp(r\tau)\frac{\partial^2}{\partial K^2} C\left(K, \tau, S, r, \delta\right)$$

where $\tilde{q}$ represents the pdf of the state price density, r is the risk free rate, τ is the tenor or time to expiration, K is the strike price, S is the current spot price, and δ is the dividend yield. $C\left(\cdot, \cdot, \cdot, \cdot, \cdot\right)$ represents the pricing function of a call option, which is assumed to be smooth and twice differentiable.

Note that most of the determinants for this density can be observed in market data. This has led to the interest of obtaining the SPD empirically. Yet, while we are able to observe the call option prices in data, we do not observe an explicit analytical form for the valuation function. This

[1]This follows easily from the fact that under a risk neutral measure $\mathbb{Q}$ we have that the price of the call option is given by:

$$C\left(K, \tau\right) = \exp(-r\tau)\int_K^\infty \left(s - K\right)\tilde{q}ds$$

makes taking the second derivative a difficult task. It is, however, possible to use estimation to obtain predicted values of the call option price beyond those available in the data and perform a discrete second derivative. A common practice in the literature is to start from the seminal Black and Scholes (1973) and Merton (1973) formula:

$$C(K, \tau, S, r, \delta, \sigma) = \exp(-r\tau) \left[F\Phi\left(d_1\right) - K\Phi\left(d_2\right)\right] \tag{3.1}$$

where

$$d_{1,2} = \frac{\ln\left(\frac{F}{K}\right) \pm \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}},$$

$F = S \times \exp\{(r - \delta)\,\tau\}$, and $\Phi\left(\cdot\right)$ represents the c.d.f. of the standard normal distribution, and σ is a structural parameter. If the Black and Scholes (1973) and Merton (1973) formulation were to be correct, σ would be constant across different values of K and τ. In empirical work, we have access to data on all the variables in (3.1) except for σ. Hence, inverting the function, it is possible to obtain values of σ which we denominate implied volatility. Since the assumptions of the Black and Scholes (1973) and Merton (1973) model are restrictive, implied volatility does not appear as a constant or near constant across tenor and strikes with real data, giving rise to what is known as the "volatility smile" in literature. Yet, the consistent divergence of the formulation from the data allows for using the functional form as an intermediate step for obtaining good predicted values for the call option price.

Shimko (1993) proposes to use σ as a measurement of the divergence between (3.1) and real data, and use it as a dependent variable to obtain a function $\sigma\left(m, \tau\right)$ in which

$$m = \frac{K}{F}$$

represents a measure of "moneyness." Hence, a pricing function which can yield good predicted values of the call option price, is given semiparametrically by what is known as the "practitioner's

Black-Scholes":

$$C(K,\tau,S,r,\delta,\sigma(m,\tau)) = \exp(-r\tau)\left[F\Phi(d_1) - K\Phi(d_2)\right]$$

where

$$d_{1,2} = \frac{-\ln(m) \pm \frac{1}{2}\left[\sigma(m,\tau)\right]^2\tau}{\sigma(m,\tau)\sqrt{\tau}}$$

The solution to the problem has become one of simply obtaining good estimates and fitted values for $\sigma(m,\tau)$. This "double transformation" (from the space of prices to volatilities and back) uses the Black Scholes and Merton model as a tool but does not necessarily assume its canonical form. However, this procedure lends itself to difficulties in empirical work. On daily basis, observations of acceptable quality (arbitrage free) are reduced and tend to be susceptible to noise. Hence, misspecification becomes a threat to the estimation of the implied volatility surface and thus, to the extraction of an accurate SPD. It is important to note that both of these characteristics of the data create problems for estimation. Parametric estimation of $\sigma(m,\tau)$ is particularly susceptible to misspecification, while nonparametric estimation might need a larger number of observations than easily available to yield results that adequately reflect market sentiment.

3.1.2 ℓ_1 Trend Filtering

Trend filtering is a nonparametric estimation technique proposed by Kim *et al.* (2009). More recently, Tibshirani *et al.* (2014) applies it to curve fitting with piecewise polynomials and explores it in several contexts. We will use a penalized least squares criterion, albeit with the ℓ_2 term being multiplied by a design matrix that is just the identity matrix. This setting is often called the "signal approximation" case in generalized LASSO problems. This setup interprets the value of dependent variable to be a realization of an underlying signal with disturbances. Let's start by assuming that we have a $n \times 1$ vector $\sigma \in \mathbb{R}^n$ of observations which are related to the input points $m_1, m_2, ..., m_n \in \mathbb{R}$ by a real function g:

$$\sigma_i = g(m_i) + \eta_i \text{ for } i = 1, 2, \cdots, n,$$

where η_i are independent errors. It is important to note that we have then a single predictor variable m_i in this case. With the inputs sorted in ascending order ($m_1 < m_2 < \cdots < m_n$), we can think of them as "positions" of the signal. Kim *et al.* (2009)'s method is as follows, for some trend filtering order $t \geq 0$, the estimate $\hat{g}$ of $g(m_i)$ can be found by solving the generalized LASSO optimization problem (Tibshirani *et al.* (2011))

$$\hat{g} = \underset{g \in \mathbb{R}^n}{\operatorname{argmin}} \frac{\|\sigma - g\|_2^2}{2} + \frac{1}{t!}\lambda \left\| D^{(m,t+1)} g \right\|_1$$

where $\lambda \geq 0$ is the tuning parameter.

$$D^{(m,t+1)} = D^{(1)} \cdot \operatorname{diag}\left(\frac{t}{m_{t+1} - m_1}, \cdots, \frac{t}{m_n - m_{n-t}} \right) \cdot D^{(m,t)}$$

this recursive definition follows that:

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & & & & & \vdots \end{bmatrix} \in \mathbb{R}^{(n-t-1)\times(n-t)}$$

this last matrix is simply the $(n-t-1)\times(n-t)$ version of a first difference matrix. Thus $D^{(m,1)}$ is the matrix of discrete first derivatives. Raising the order of t raises the order of the discrete derivative. Consider for example that our inputs were equally spaced, then $m_{t+i} - m_i = t$ so we would have:

$$D^{(m,2)} = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -2 & & & 0 \\ \vdots & & & & & & \end{bmatrix}$$

which is just the difference of differences matrix. Hence, we are penalizing on the $t+1$th derivative of g. This means that the penalty goes on changes on the tth derivative and therefore the structure

of the resulting estimates should approximate the form of a tth-order polynomial.

Tibshirani *et al.* (2014) refers to the surviving entries on $D^{(m,t+1)}\hat{g}$ as "knots," and uses this nomenclature when making empirical comparisons with smoothing splines. However, unlike smoothing splines, the ℓ_1 penalization selects the knots adaptively. Since only some knots are selected by the estimation procedure, one can get precise fits with fewer degrees of freedom than in the smoothing spline case. Furthermore, Tibshirani *et al.* (2014) provides empirical evidence that the adaptiveness of trend filtering leads to increased performance in terms of having smaller input-averaged squared error losses of the form:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{g}_i - g\left(m_i\right)\right]^2$$

Hence, trend filtering is a good candidate as a nonparametric tool, albeit unidimensional. However, since $\sigma\left(m,\tau\right)$ depends on two parameters, we follow the "stringwise" estimation suggested by Shimko (1993), but do so nonparametrically rather than fitting a quadratic function with no interaction as he does. This is done for each cross-section defined by values of the tenor, i.e. we obtain estimates for $\sigma\left(m|\tau\right)$. It is important to mention that unlike kernel estimators, when groups of data are far from each other, trend filtering will adapt the fits within each group instead of oversmoothing when λ is chosen appropriately. Selection of λ is critical, since if it were to be overly small overfitting problems would arise from underregularization i.e. too many knots would be selected. Similarly, if λ is too large, overregularization would lead to too few knots being picked and the result would be oversmoothing.

While the discussion on the selection of an appropriate tuning parameter for LASSO applications is extensive, we focus on finding an appropriate value for λ in an automatic fashion and comparable per string such that estimates do not result in overfitting. The first step involves using five-fold cross validation. As described by Kohavi (1995), increasing the number of folds reduces bias but increases variance. Many studies using cross validation use either five or ten folds as a compromise for this tradeoff. To perform five-fold cross validation the data is divided into 5

roughly equal parts, and for each one fit the model with λ for the other 4 parts. For $k = 1, \cdots, 5$, we obtain the error:

$$\sum_{i=1}^{n} \left(\sigma_i - \hat{g}_{\lambda}^{-k}\right)^2$$

where $\hat{g}_{\lambda}^{-k}$ indicates the estimates of $\hat{g}$ ignoring the kth fold. This process yields the cross-validation error defined as:

$$CV\left(\lambda\right) = \frac{1}{n}\sum_{k=1}^{5}\sum_{i=1}^{n} \left(\sigma_i - \hat{g}_{\lambda}^{-k}\right)^2 \tag{3.2}$$

The value of λ which minimizes (3.2) is the five-fold cross validation choice for the tuning parameter. However, since the objective of the LASSO is to select the true model simply using cross-validation generally means that not enough regularization has been performed. As proposed by Breiman *et al.* (1984), an alternative automatic rule consists in selecting the most parsimonious model which yields an error not higher than one standard deviation away from the minimal cross validation error. Let $\hat{\lambda}$ be the solution to the cross-validation error minimization problem. Then the one standard error rule searches for a value λ such that:

$$CV\left(\lambda\right) \leq CV\left(\hat{\lambda}\right) + SE\left(\hat{\lambda}\right)$$

ceases to be true at some value $\tilde{\lambda}$. Hence, we move λ in the direction of regularization increasing its value until we get that $\tilde{\lambda} > \hat{\lambda}$. Here, we have that:

$$SE\left(\hat{\lambda}\right) = \sqrt{\frac{1}{5} \cdot \text{var}\left(CV_1\left(\hat{\lambda}\right), \cdots, CV_5\left(\hat{\lambda}\right)\right)}$$

With the issue of the selection of the tuning parameter resolved, we are left to pick the order for the trend filtering estimate. Given the prior literature, particularly Malz (2014), cubic splines have shown to be particularly stable. Being aware of possible overfitting due to adaptative selection of knots, we pick $t = 2$ to fit piecewise quadratic polynomials. This choice enforces two desirable properties for the estimates we need. First, the estimates will be smooth, which matches the initial assumption of the estimation for use in extracting SPDs. Second, the quadratic structure ensures

that we will be able to find second derivatives throughout the domain.

3.2 Monte Carlo Simulations

3.2.1 Competing Models

In order to evaluate the performance of ℓ_1 trend filtering, we need to pick representatives from the major trends in the literature as comparative benchmarks. More specifically, we would like to make comparisons with both nonparametric and parametric estimators. Much attention has been given to Aït-Sahalia and Lo (1998) estimation strategy via the Nadaraya-Watson estimator. Particularly popular extensions include the use of the local linear estimator with different smoothing parameters included in Aıt-Sahalia and Duarte (2003). Local linear estimates arise from the solution to the following minimization problem:

$$\min_{[a,b]} \sum_{i=1}^{n} \left[\sigma_i - a - (m_i - m)' b\right]^2 \mathcal{K}\left(\frac{m_i - m}{h}\right)$$

where the estimate of a is a consistent estimator for $g\,(m)$ and the estimate of b a consistent estimator of $g^{(1)}\,(m)$, the gradient vector of $g\,(m)$, and $\mathcal{K}\,(\cdot)$ is a kernel function with h as its respective bandwidth parameter.

While this strategy yields estimates that are very flexible to functional form and have desirable statistical properties regarding inference, they have a major flaw for practitioners. The number of observations must be large to achieve convergence, this problem is compounded when having mutiple dimensions. The number of quotes per tenor obtained from usable real data in the market seldom meet this requirement. As such, the researchers working with these methods could decide to use large bandwidths in their estimation and risk oversmoothing, losing the gains from the flexibility kernel estimators have as a main feature. Alternatively, and a far more common practice is to augment their data via aggregation. Hence the quotes obtained are no longer from a single day, but from a set of days. However effective using the latter option might be, the lack of ability of assessing market sentiment daily is problematic. In addition of losing interpretative power, the estimator is then vulnerable to failing to identify regime shifts. Regardless, the flexibility and

robustness of this approach makes the local linear estimator a good comparison paradigm for trend filtering in the realm of nonparametric estimation. As far as implementation goes, the choices of smoothing parameters for our exercise are obtained from least squares cross validation procedure and using the AICc criterion.

In contrast, the parametric stringwise specification in Shimko (1993) has proven to be a good benchmark model, and has been proven to work better than stochastic volatility models like Heston (1993) and factor extensions such as Christoffersen *et al.* (2009). The specification that Shimko (1993) uses for each tenor on a single day is:

$$\sigma = \alpha_0 + \alpha_1 m + \alpha_2 m^2.$$

The reason why this model is also used as a benchmark on our empirical comparison, is due to the fact that kernel estimators are not suitable for small datasets. Fair evaluation of trend filtering would require to also compare it to an estimation methodology which can handle daily data. We have chosen to evaluate trend filtering against the Shimko (1993) parametric model given its reputation in the literature and its use as a benchmark in similar papers, e.g. Ludwig (2015). Moreover, in order to provide an extra parametric benchmark, we evaluate fitting the model in our Monte Carlo design in equation (3.3).

It is in our interest to clarify our objective as it pertains to obtaining good fits and hence good predicted values. We need to show that trend filtering has comparable performance to the Aït-Sahalia and Lo (1998) proposed estimation strategy in instances in which large datasets are available i.e. aggregated data, and show improved performance in the case in which data has a frequency similar to what it does under typical market conditions, meaning reduced number of observations per tenor. In addition, we need to compare the Shimko (1993) specification, and the simulation design specification for both cases. Sample size is not the only concern when dealing with usable option data, hence we introduce different levels of noise in the simulation for a more robust comparison. Correspondingly, we perform 500 runs of Monte Carlo simulations for four

different number of quotes per tenor (7,12, 50, 70) and for three different levels of noise.

3.2.2 Monte Carlo Design

We are interested in generating a volatility surface that can show the L-shaped and U-shaped properties that volatility cross-sections do in arbitrage-free conditions. The further we move away from being "at the money" $(K = F)$ we have that implied volatility is weakly increasing. We want to generate data that shows "smiles" at every tenor. The specification chosen to do this is a fourth order polynomial capable of generating the forementioned shapes :

$$\sigma = \alpha_0 + \alpha_1 m + \alpha_2 m^2 + \alpha_3 \tau + \alpha_4 \tau^2 + \alpha_5 m\tau + \alpha_6 m^3 + \alpha_7 m^4, \tag{3.3}$$

where $m = K/F$ and $F = S \times \exp\{(r - \delta)\,\tau\}$ as mentioned before. This model is the "true model" for which we can make comparisons of prediction error among different methods. However, it would not be telling nor realistic to not include a noise element when generating the data, which leads us to use :

$$\tilde{\sigma} = \alpha_0 + \alpha_1 m + \alpha_2 m^2 + \alpha_3 \tau + \alpha_4 \tau^2 + \alpha_5 m\tau + \alpha_6 m^3 + \alpha_7 m^4 + \varepsilon$$

where $\varepsilon \overset{i.i.d}{\sim} N(0, \sigma_\epsilon^2)$, with $\sigma_\epsilon^2 \in \{0.005, 0.010, 0.015\}$ in which each element is used for each proposed sample size.

Next, it is necessary to select values of S, δ, τ and K, as initial values for the simulation. Representative statistics for variables in the S&P 500 dataset picked for our empirical implementation are chosen. We obtain some summary statistics from such data in Table 3.1.

The means as the values for S, r and δ for the Monte Carlo. As for the tenors, we use usual contract expiration times in 30 day intervals: $\tau \in \left\{\frac{30}{365}, \frac{60}{365}, \cdots, \frac{300}{365}\right\}$.

The Chicago Board of Options Exchange (CBOE) limits the values of strike prices for S&P 500 options to be between 5 points of the spot price at any given time. In this case this means that strike prices need to be within five hundred dollars of the spot price per individual contract. Hence,

Table 3.1: S&P 500 prices from January 2nd, 2007 to January 1st, 2010.

	S	r	δ
First Quantile	909.7	0	0.01961
Median	1097.9	1×10^{-5}	0.02104
Mean	1146.7	6.089×10^{-5}	0.02044
Third Quantile	1392.6	9×10^{-5}	0.02347

we can generate strike prices as:

$$K = S + u$$

where u is uniformly distributed in $(-500, 500)$.

The choice of a parameter vector for the specification for the simulation is given by:

$$(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7) = (0.4, -0.6,\ 0.28, -0.2, 0.02, 0.17, 0.015, -0.025)\,.$$

The reason for this seemingly arbitrary choice of parameters lies in the form of the noiseless surface it generates. As presented in Figure 3.1, the noiseless surface from the fourth polynomial specification and the set of chosen parameters has both L and U shapes, making it an appropriate environment by which to compare goodness of fit between our chosen models. The measures for goodness of fit we choose for our comparative analysis are the RMSE (root mean square error):

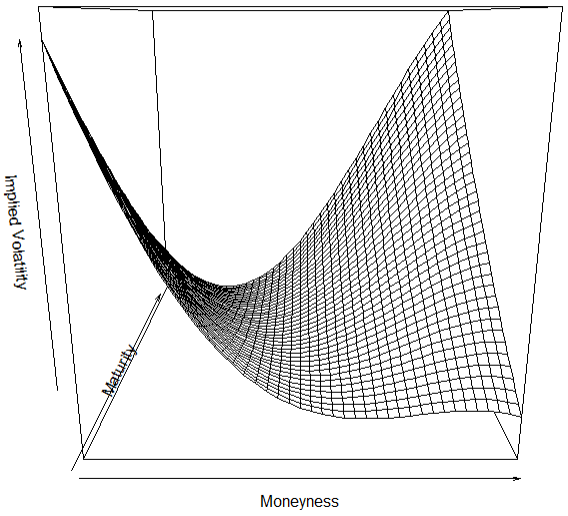
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\sigma_i - \hat{g}_i\right)^2}$$

and MAPE (mean average percentage error):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\sigma_i - \hat{g}_i}{\sigma_i}\right|$$

These two measures of fit will indicate the variability in distance between our competing mod-

Figure 3.1: Moneyness



els fit around the noisy data and the "true" model. This setup will not only allow simple comparison, but it will punish overfitting, by making noise inclusion in the estimate create larger values for both measures.

3.2.3 Monte Carlo Results

We examine the prediction error for the Monte Carlo runs by creating two large groups according to number of observations. In the first group we use simulation runs that have 7 and 12 observations per tenor. These simulate daily data frequencies, 7 observations are the minimum needed to apply five-fold cross validation. We simulate aggregated data by creeating runs with 50 and 70 observations per tenor. Tables 3.2 to 3.7 encompass the results of the Monte Carlo.[2] Figures 3.2 and 3.3 present a visual summary of the table results for the mean of the goodness of fit measures. The discussion below is based on analyzing the results using both the tables and figures.

Note that from the RMSE quartiles at 7 observations per tenor, under the 0.005 standard devi-

[2]The table headers indicate the quartiles of the distribution of simulations, and the mean. The rows indicate which method has been used.

Table 3.2: RMSE $\epsilon \sim N\left(0, 0.005\right)$

						(7 observations)
	0%	25%	50%	75%	100%	Mean
Trend Filtering	0.0022	0.0031	0.0034	0.0037	0.0051	0.0034
Local Linear(AICc)	0.0112	0.0184	0.0206	0.0227	0.0290	0.0205
Local Linear(LSCV)	0.0112	0.0185	0.0208	0.0228	0.0290	0.0206
Shimko	0.0149	0.0208	0.0227	0.0247	0.0309	0.0227
Simulation Specification	0.0114	0.0188	0.0210	0.0230	0.0291	0.0209
						(12 Observations)
Trend Filtering	0.0020	0.0025	0.0028	0.0030	0.0038	0.0028
Local Linear(AICc)	0.0155	0.0196	0.0212	0.0229	0.0280	0.0212
Local Linear(LSCV)	0.0156	0.0197	0.0212	0.0230	0.0280	0.0213
Shimko	0.0169	0.0213	0.0227	0.0242	0.0292	0.0227
Simulation Specification	0.0154	0.0206	0.0223	0.0239	0.0306	0.0223
						(50 Observations)
Trend Filtering	0.0013	0.0016	0.0017	0.0017	0.0038	0.0017
Local Linear(AICc)	0.0187	0.0216	0.0224	0.0232	0.0258	0.0224
Local Linear(LSCV)	0.0187	0.0216	0.0224	0.0232	0.0258	0.0224
Shimko	0.0198	0.0222	0.0229	0.0237	0.0263	0.0230
Simulation Specification	0.0188	0.0217	0.0224	0.0232	0.0259	0.0224
						(70 Observations)
Trend Filtering	0.0013	0.0015	0.0015	0.0016	0.0036	0.0016
Local Linear(AICc)	0.0201	0.0217	0.0224	0.0230	0.0250	0.0224
Local Linear(LSCV)	0.0201	0.0217	0.0224	0.0230	0.0250	0.0224
Shimko	0.0203	0.0223	0.0229	0.0235	0.0261	0.0229
Simulation Specification	0.0202	0.0217	0.0224	0.0230	0.0250	0.0224

Table 3.3: MAPE $\epsilon \sim N(0, 0.005)$

						MAPE (7 Observations)
	0%	25%	50%	75%	100%	Mean
Trend Filtering	0.0223	0.0347	0.0384	0.0420	0.0570	0.0388
Local Linear(AICc)	0.1044	0.1727	0.1912	0.2114	0.2785	0.1919
Local Linear(LSCV)	0.1048	0.1740	0.1921	0.2113	0.2788	0.1931
Shimko	0.1578	0.2111	0.2263	0.2448	0.2986	0.2270
Simulation Specification	0.1096	0.1761	0.1953	0.2137	0.2771	0.1954
						(12 Observations)
Trend Filtering	0.0211	0.0281	0.0309	0.0340	0.0445	0.0312
Local Linear(AICc)	0.1494	0.1869	0.2020	0.2163	0.2670	0.2021
Local Linear(LSCV)	0.1491	0.1875	0.2027	0.2167	0.2670	0.2026
Shimko	0.1701	0.2151	0.2243	0.2380	0.2805	0.2262
Simulation Specification	0.1556	0.2052	0.2194	0.2353	0.2878	0.2201
						(50 Observations)
Trend Filtering	0.0151	0.0176	0.0187	0.0199	0.0249	0.0188
Local Linear(AICc)	0.1882	0.2080	0.2158	0.2227	0.2514	0.2157
Local Linear(LSCV)	0.1885	0.2080	0.2158	0.2227	0.2514	0.2157
Shimko	0.2010	0.2202	0.2276	0.2339	0.2547	0.2272
Simulation Specificationl	0.1880	0.2077	0.2154	0.2224	0.2506	0.2153
						(70 Observations)
Trend Filtering	0.0146	0.0166	0.0174	0.0183	0.0295	0.0175
Local Linear(AICc)	0.1953	0.2100	0.2155	0.2222	0.2389	0.2162
Local Linear(LSCV)	0.1953	0.2100	0.2155	0.2222	0.2388	0.2162
Shimko	0.2066	0.2215	0.2267	0.2319	0.2535	0.2268
Simulation Specification	0.1952	0.2095	0.2152	0.2218	0.2384	0.2157

Table 3.4: RMSE $\epsilon \sim N\left(0, 0.010\right)$

						(7 observations)
	0%	25%	50%	75%	100%	Mean
Trend Filtering	0.0041	0.0061	0.0068	0.0073	0.0095	0.0067
Local Linear(AICc)	0.0118	0.0186	0.0205	0.0225	0.0287	0.0206
Local Linear(LSCV)	0.0120	0.0188	0.0209	0.0228	0.0288	0.0209
Shimko	0.0151	0.0217	0.0237	0.0256	0.0314	0.0236
Simulation Specification	0.0122	0.0196	0.0215	0.0237	0.0299	0.0216
						(12 Observations)
Trend Filtering	0.0037	0.0049	0.0054	0.0059	0.0080	0.0054
Local Linear(AICc)	0.0162	0.0199	0.0217	0.0231	0.0284	0.0215
Local Linear(LSCV)	0.0162	0.0201	0.0218	0.0232	0.0285	0.0217
Shimko	0.0177	0.0219	0.0234	0.0247	0.0293	0.0234
Simulation Specification	0.0165	0.0206	0.0222	0.0236	0.0293	0.0221
						(50 Observations)
Trend Filtering	0.0019	0.0025	0.0027	0.0030	0.0042	0.0028
Local Linear(AICc)	0.0193	0.0216	0.0224	0.0231	0.0256	0.0223
Local Linear(LSCV)	0.0194	0.0216	0.0224	0.0231	0.0256	0.0223
Shimko	0.0197	0.0223	0.0231	0.0239	0.0259	0.0230
Simulation Specification	0.0196	0.0217	0.0225	0.0232	0.0256	0.0225
						(70 Observations)
Trend Filtering	0.0017	0.0022	0.0023	0.0025	0.0038	0.0024
Local Linear(AICc)	0.0194	0.0217	0.0224	0.0229	0.0254	0.0223
Local Linear(LSCV)	0.0194	0.0217	0.0224	0.0229	0.0254	0.0223
Shimko	0.0199	0.0224	0.0230	0.0236	0.0259	0.0230
Simulation Specification	0.0197	0.0219	0.0225	0.0230	0.0256	0.0225

Table 3.5: MAPE $\epsilon \sim N\left(0, 0.010\right)$

						(7 observations)
	0%	25%	50%	75%	100%	Mean
Trend Filtering	0.0458	0.0690	0.0766	0.0839	0.1083	0.0768
Local Linear(AICc)	0.1207	0.1787	0.1958	0.2144	0.2847	0.1970
Local Linear(LSCV)	0.1218	0.1813	0.1986	0.2187	0.2872	0.2000
Shimko	0.1696	0.2249	0.2409	0.2579	0.3157	0.2409
Simulation Specification	0.1321	0.1914	0.2073	0.2273	0.2769	0.2091
						(12 Observations)
Trend Filtering	0.0378	0.0538	0.0600	0.0658	0.0928	0.0602
Local Linear(AICc)	0.1547	0.1925	0.2078	0.2221	0.2683	0.2070
Local Linear(LSCV)	0.1562	0.1942	0.2088	0.2233	0.2695	0.2086
Shimko	0.1862	0.2202	0.2350	0.2497	0.2934	0.2354
Simulation Specification	0.1628	0.1993	0.2144	0.2282	0.2786	0.2138
						(50 Observations)
Trend Filtering	0.0190	0.0277	0.0304	0.0330	0.0487	0.0305
Local Linear(AICc)	0.1851	0.2088	0.2160	0.2238	0.2487	0.2162
Local Linear(LSCV)	0.1848	0.2088	0.2161	0.2240	0.2491	0.2164
Shimko	0.1991	0.2219	0.2285	0.2364	0.2549	0.2287
Simulation Specification	0.1859	0.2105	0.2168	0.2247	0.2487	0.2171
						(70 Observations)
Trend Filtering	0.0195	0.0240	0.0259	0.0281	0.0356	0.0262
Local Linear(AICc)	0.1914	0.2106	0.2162	0.2221	0.2454	0.2164
Local Linear(LSCV)	0.1916	0.2106	0.2164	0.2221	0.2456	0.2165
Shimko	0.2025	0.2224	0.2278	0.2328	0.2507	0.2278
Simulation Specification	0.1925	0.2107	0.2162	0.2225	0.2467	0.2166

Table 3.6: RMSE $\epsilon \sim N\left(0, 0.015\right)$

						(7 observations)
	0%	25%	50%	75%	100%	Mean
Trend Filtering	0.0062	0.0093	0.0102	0.0111	0.0141	0.0102
Local Linear(AICc)	0.0128	0.0182	0.0207	0.0229	0.0312	0.0206
Local Linear(LSCV)	0.0128	0.0187	0.0211	0.0233	0.0327	0.0211
Shimko	0.0162	0.0228	0.0247	0.0268	0.0333	0.0248
Simulation Specification	0.0146	0.0204	0.0225	0.0248	0.0328	0.0226
						(12 Observations)
Trend Filtering	0.0052	0.0074	0.0080	0.0089	0.0117	0.0081
Local Linear(AICc)	0.0145	0.0194	0.0210	0.0229	0.0293	0.0211
Local Linear(LSCV)	0.0148	0.0196	0.0212	0.0231	0.0299	0.0214
Shimko	0.0173	0.0222	0.0238	0.0251	0.0309	0.0237
Simulation Specificationl	0.0154	0.0206	0.0223	0.0239	0.0306	0.0223
						(50 Observations)
Trend Filtering	0.0027	0.0036	0.0039	0.0043	0.0062	0.0039
Local Linear(AICc)	0.0182	0.0214	0.0223	0.0230	0.0257	0.0223
Local Linear(LSCV)	0.0182	0.0214	0.0223	0.0231	0.0257	0.0223
Shimko	0.0193	0.0223	0.0231	0.0240	0.0268	0.0231
Simulation Specification	0.0188	0.0218	0.0226	0.0234	0.0261	0.0226
						(70 Observations)
Trend Filtering	0.0022	0.0030	0.0033	0.0036	0.0055	0.0033
Local Linear(AICc)	0.0188	0.0216	0.0223	0.0230	0.0262	0.0223
Local Linear(LSCV)	0.0188	0.0216	0.0223	0.0230	0.0262	0.0223
Shimko	0.0194	0.0224	0.0230	0.0237	0.0262	0.0230
Simulation Specification	0.0192	0.0219	0.0226	0.0232	0.0265	0.0225

Table 3.7: MAPE $\epsilon \sim N(0, 0.015)$

						(7 observations)
	0%	25%	50%	75%	100%	Mean
Trend Filtering	0.0679	0.1045	0.1143	0.1260	0.1669	0.1155
Local Linear(AICc)	0.1062	0.1800	0.2003	0.2205	0.2864	0.2008
Local Linear(LSCV)	0.1305	0.1841	0.2055	0.2270	0.3148	0.2059
Shimko	0.1881	0.2396	0.2565	0.2774	0.3389	0.2583
Simulation Specification	0.1502	0.2057	0.2252	0.2457	0.3035	0.2259
						(12 Observations)
Trend Filtering	0.0546	0.0827	0.0894	0.0986	0.1333	0.0905
Local Linear(AICc)	0.1448	0.1904	0.2052	0.2215	0.2701	0.2060
Local Linear(LSCV)	0.1498	0.1928	0.2076	0.2238	0.2803	0.2086
Shimko	0.1838	0.2293	0.2426	0.2560	0.3023	0.2428
Simulation Specification	0.1556	0.2052	0.2194	0.2353	0.2878	0.2201
						(50 Observations)
Trend Filtering	0.0297	0.0393	0.0432	0.0476	0.0656	0.0436
Local Linear(AICc)	0.1724	0.2101	0.2171	0.2247	0.2482	0.2175
Local Linear(LSCV)	0.1725	0.2105	0.2179	0.2252	0.2484	0.2178
Shimko	0.1909	0.2242	0.2313	0.2392	0.2672	0.2315
Simulation Specification	0.1800	0.2131	0.2200	0.2274	0.2508	0.2202
						(70 Observations)
Trend Filtering	0.0248	0.0333	0.0365	0.0402	0.0546	0.0369
Local Linear(AICc)	0.1928	0.2108	0.2174	0.2249	0.2593	0.2175
Local Linear(LSCV)	0.1929	0.2110	0.2175	0.2250	0.2595	0.2176
Shimko	0.2034	0.2236	0.2293	0.2365	0.2644	0.2297
Simulation Specification	0.1937	0.2125	0.2190	0.2264	0.2627	0.2193

ation of noise that trend filtering is roughly six times smaller than both local linear specifications and the Shimko specification across the simulation distribution. At a standard deviation of 0.010 we see that the RMSE of trend filtering has reduced its advantage being roughly a third of the local linear and Shimko specifications. Upon increasing the standard deviation of the noise to 0.015 we see that the ratio of the trend filtering advantage remains similar to that of the 0.010 standard deviation. The MAPE results exhibit the same tendency as RMSE for 7 observations across the three noise levels. The results at 12 daily observations show a slightly reduced advantage at the lowest level of noise with a RMSE value of roughly half of the competing models. However, this advantage widens as the standard deviation of noise increasing, suggesting slight overfitting issues

Figure 3.2: Simulation RMSE results

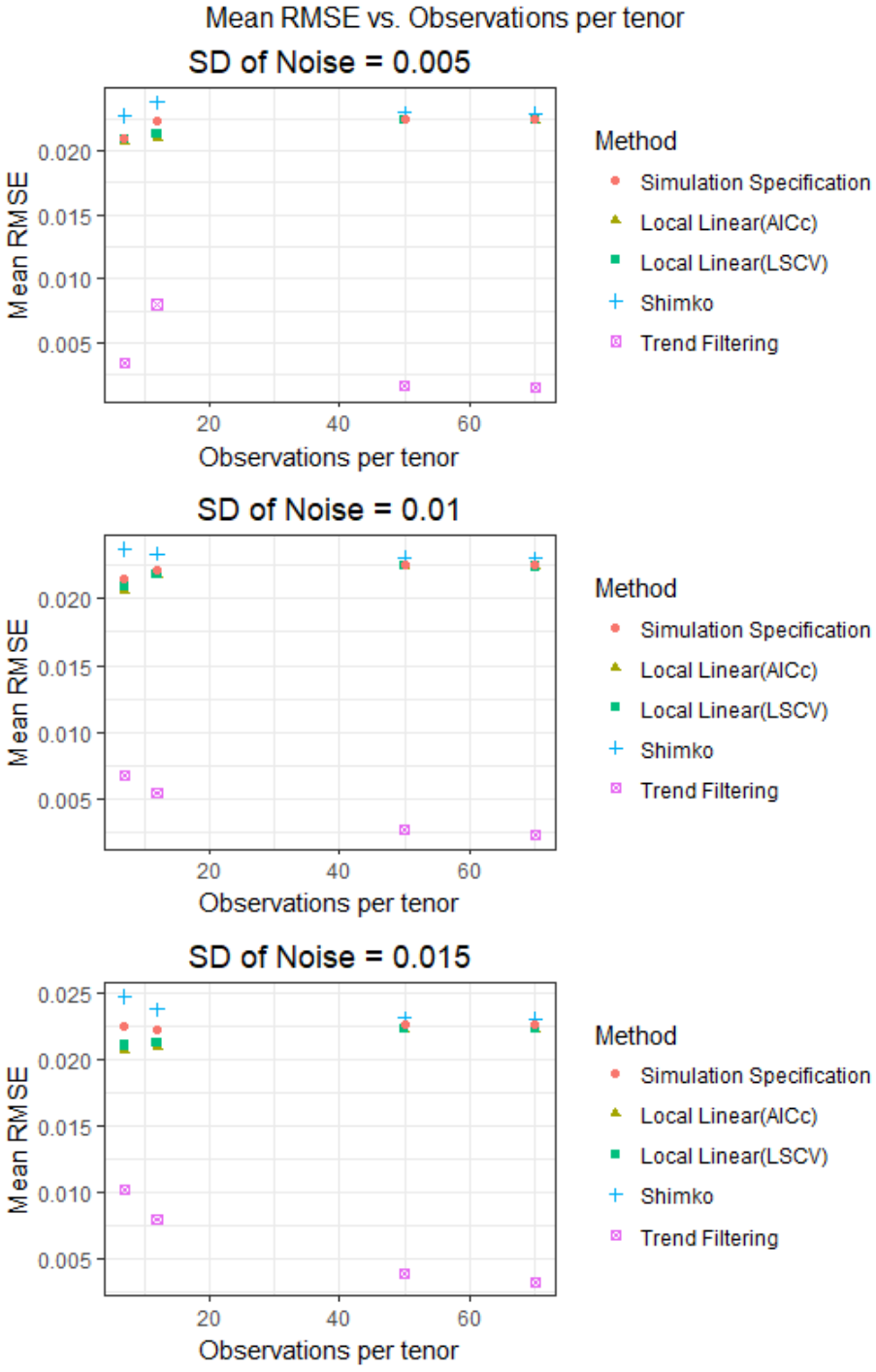
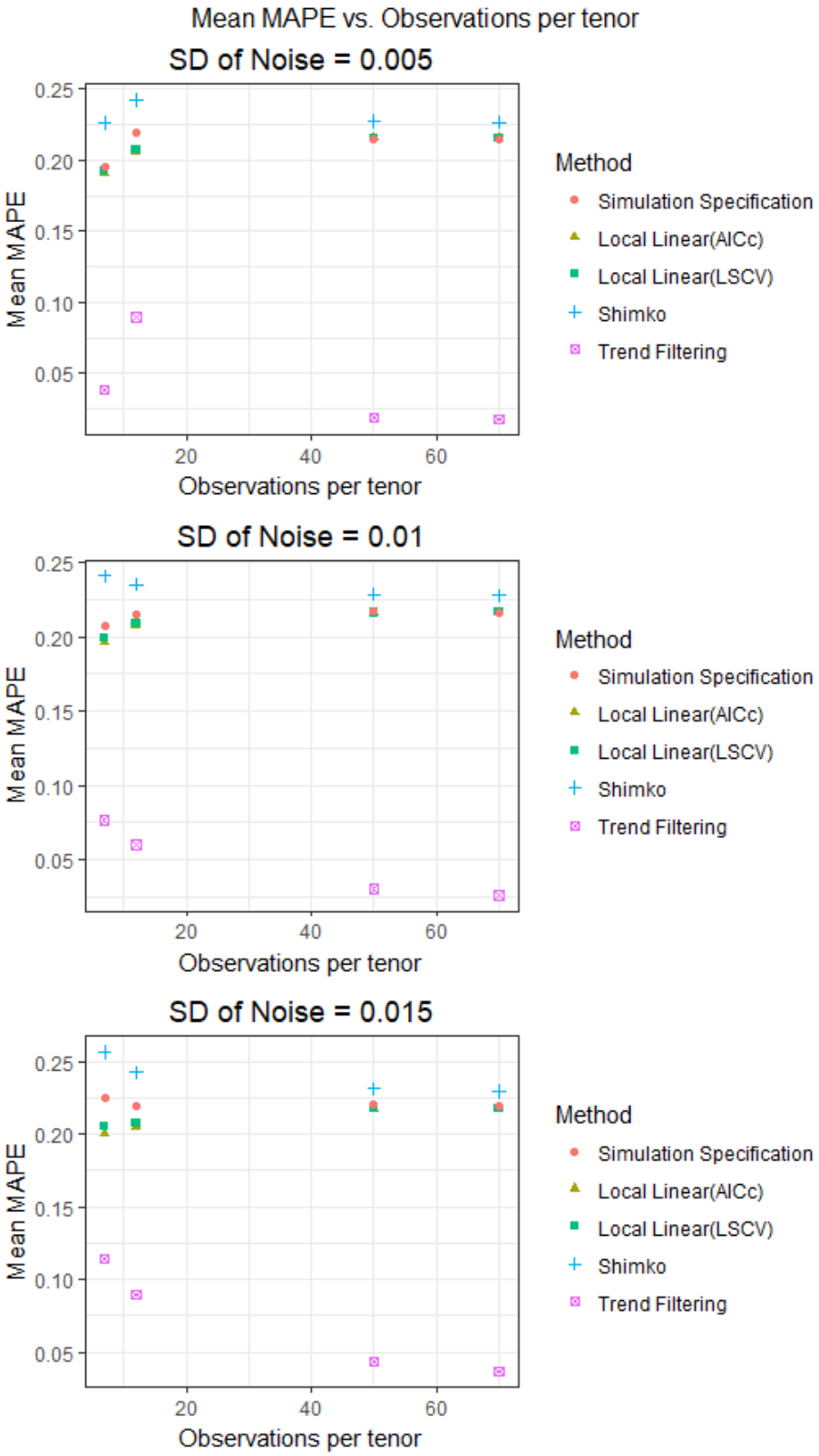


Figure 3.3: Simulation MAPE results



at this sample size. MAPE follows a similar trend to RMSE.

The results so far show that under small sample sizes which emulate daily frequencies for contracts, trend filtering can outperform both kernel models and the Shimko specification across our specified levels of noise. It is also important to note that in the analysis, throughout the distribution of goodness of fit measures we have that the differences between the AIC criterion and least squares cross validation kernel local linear estimators are of less than half a percent regardless of noise level. And that both measures exhibit only a modest advantage to the Shimko specification.

The next step is to evaluate the appropriateness of trend filtering as a method under larger datasets that approach the sizes of aggregated data. We use the 50 and 70 observations per tenor datasets for this purpose.

Evaluating the RMSE on ε with a standard deviation of 0.005 the trend filtering values result in being about a tenth of the competing methods across the distribution in both 50 and 70 observations per tenor levels. For 0.010 we have trend filtering to have a RMSE value of about a ninth of the competing methods for the quartiles, and similarl results for the 0.015 standard deviationl of noise level. This means that while the increase of noise does reduce the advantage of trend filtering over the other methods, the increased sample size allows this advantage to remain high. MAPE shows a similar behavior across levels of noise. It is important to note that the advantage of the kernel methods over the parametric specification is reduced in the aggregated data setting. Yet, nonparametric methods continue to provide better fits than their parametric counterpart. MAPE results suggest the same pattern as RMSE results.

As a simple robustness check, we find that the simulation specification underperforms trend filtering across all of our scenarios. The meaning of this is that the crossvalidation procedure works as a sensible guard for overfitting, while our model with perfect specification is subject to inclusion of noise in the fitting process.

The Monte Carlo runs suggest that trend filtering outperforms both non parametric and parametric competing models at both the daily and aggregated levels. Since the simulations were specifically made to be a challenge to all models, the overfitting issues with flexible methodol-

ogy seems to be diminished in the case of trend filtering, while its flexibility affords it a clear advantage over the parametric specification commonly used in daily observations as long as the number of observations is sufficient to allow estimation given the limitation of needing at least seven observations per tenor.

3.3 An Empirical Application

3.3.1 The Data

The data used in this section comes from OptionMetrics and it encompasses prices for call options from S&P 500, Apple Inc., Kellogg Company, and Gap Inc. for the period from January 1, 2007 to January 1, 2010. The stocks selected are part of the S&P 500 index but have distinct levels of market capitalization and vary in number of contracts daily. Only days in which at least seven quotes per tenor were present were used as this is the minimum size of a data set required to perform five-fold cross validation. While it may be possible to work with smaller data sets using less folds, bias will be larger. The data includes bid and ask prices, underlying index prices, dividend yields, tenors and strikes. The risk free rate is taken as the daily one month treasury bill rate.

In order to perform comparative analysis, each set of daily observations was split randomly in two sets. The first set containing 75 percent of the original observations is called the "training set" in which model fitting is performed. The remaining set is called the "test set" and it is used to evaluate the accuracy of the model fit in the training set against data from the same population that was not included in estimation.

3.3.2 Data Cleanup

Given the requirement of being in an arbitrage-free environment, we need to remove any data that do not fit this assumption. We follow a well known procedure for cleaning out observations that are not arbitrage free from the data. A full discussion about the motivation of these rules is explained in Carr and Madan (2005). Following the discussion mentioned and Ludwig (2015), we use the following conditions avoid arbitrage:

A Imposition of general price bounds:

$$\max\{0, S \times \exp(-\delta\tau) - K \cdot \exp(-r\tau)\} \leq C \leq S \times \exp(-\delta\tau),$$

B Avoidance of calendar arbitrage:

$$\sigma^2(m, \tau')\tau' > \sigma^2(m, \tau'')\tau'' \quad \text{for } \tau' > \tau''$$

C The call valuation function, must be convex to the origin:

$$\frac{\partial^2}{\partial K^2} C \leq 0 \quad \text{and} \quad -\exp(-r\tau) \leq \frac{\partial C}{\partial K} \leq 0.$$

3.3.3 Tail Extrapolation

While good fits on the estimation of implied volatility are desirable for SPD extraction, non-parametric methods do not allow extrapolation (prediction is strictly confined within the range of the data). This is problematic, because extreme data of good quality is often hard to find. Thus, the densities extracted will lack a way to identify tail behavior. Figlewski (2008) proposes grafting the tails from a GEV distribution on the ends of the computed SPD from the estimated implied volatilities. The cdf of the GEV distribution is given by[3]

$$F(x) = \exp\left[-(1+\gamma x)^{-\frac{1}{\gamma}}\right] \tag{3.4}$$

where we have that:

$$x = (K - \mu_K)/\eta_K \tag{3.5}$$

In order to perform this grafting, two sets of grafting points are picked from the computed SPD, two interiors and two extremes. Consider for example the 5th and 95th percentile to be the

[3] And hence the pdf of the GEV distribution is given by: $f(x) = (1+\gamma x)^{-\frac{1+\gamma}{\gamma}} \exp\left[-(1+\gamma x)^{-\frac{1}{\gamma}}\right]$.

interior points under the computed density. Similarly, the 2nd and 98th percentile are picked as the exterior points. The choice of these points is flexible, with the only requirement that they must be relatively close to the ends of the computed SPD. The parameters μ_K, η_K and γ determine the shape of the GEV distribution. Thus, the exercise becomes to find the set of parameters of the GEV distribution that matches the empirically found densities at both the interior and exterior point of a given tail. For the right tail, this exercise is simple. The GEV distribution is the distribution for the maximum in a sample, hence the left tail values would not be those of the extreme minima. For left tail grafting, it's necessary to find the values of the parameters as the values for a reversed GEV distribution. We need to solve two different system of equations. Denote the quantiles for the external points for the right and left tail respectively as ω_R^{Ext} and ω_L^{Ext}. Similarly, denote the quantiles for the internal points as ω_R^{Int} and ω_L^{Int}. In addition let $K(\omega)$ be the strike price corresponding to the ω-quantile, which we input as K in (3.5) . In addition, let $F(.)$, $\tilde{F}(.)$ represent the theoretical cdf for the GEV with parameters γ, μ_K , η_K , and the computed cdf from our SPD extraction process respectively with corresponding pdfs $f(.)$, $\tilde{f}(.)$. Hence for the right tail we solve for γ, μ_K and η_K in:

$$
\begin{aligned}
F\left(K\left(\omega_R^{Int}\right)\right) &= \omega_R^{Int} \\
f\left(K\left(\omega_R^{Int}\right)\right) &= \tilde{f}\left(K\left(\omega_R^{Int}\right)\right) \\
f\left(K\left(\omega_R^{Ext}\right)\right) &= \tilde{f}\left(K\left(\omega_R^{Ext}\right)\right)
\end{aligned}
$$

For left grafting, instead of (3.5) we use:

$$
x = (-K + \mu_K)/\eta_K \tag{3.6}
$$

and we solve:

$$
\begin{aligned}
F\left(-K\left(\omega_L^{Int}\right)\right) &= \omega_L^{Int} \\
f\left(-K\left(\omega_L^{Int}\right)\right) &= \tilde{f}\left(K\left(\omega_L^{Int}\right)\right) \\
f\left(-K\left(\omega_L^{Ext}\right)\right) &= \tilde{f}\left(K\left(\omega_L^{Ext}\right)\right)
\end{aligned}
$$

We can then use the parameters on both directions to extend the tails of our computed SPD.

3.3.4 An Example of Extraction with A Single Day

We use graphs of the fitted values from the estimation for January 5, 2007 to demonstrate their accuracy. In addition, we show a graph of the cross sections of the extracted SPD for that day. Figure 3.4 shows that our fits are good, and Figure 3.5 that our densities are well behaved. It is important to note that scaling needs to be taken into account to do fair graphical comparisons. Non-standarized measures of moneyness will give the impression of the SPD cross sections being flatter than they are at larger tenors. Thus, for our plots we use a standardized measure of moneyness $m^\star$:

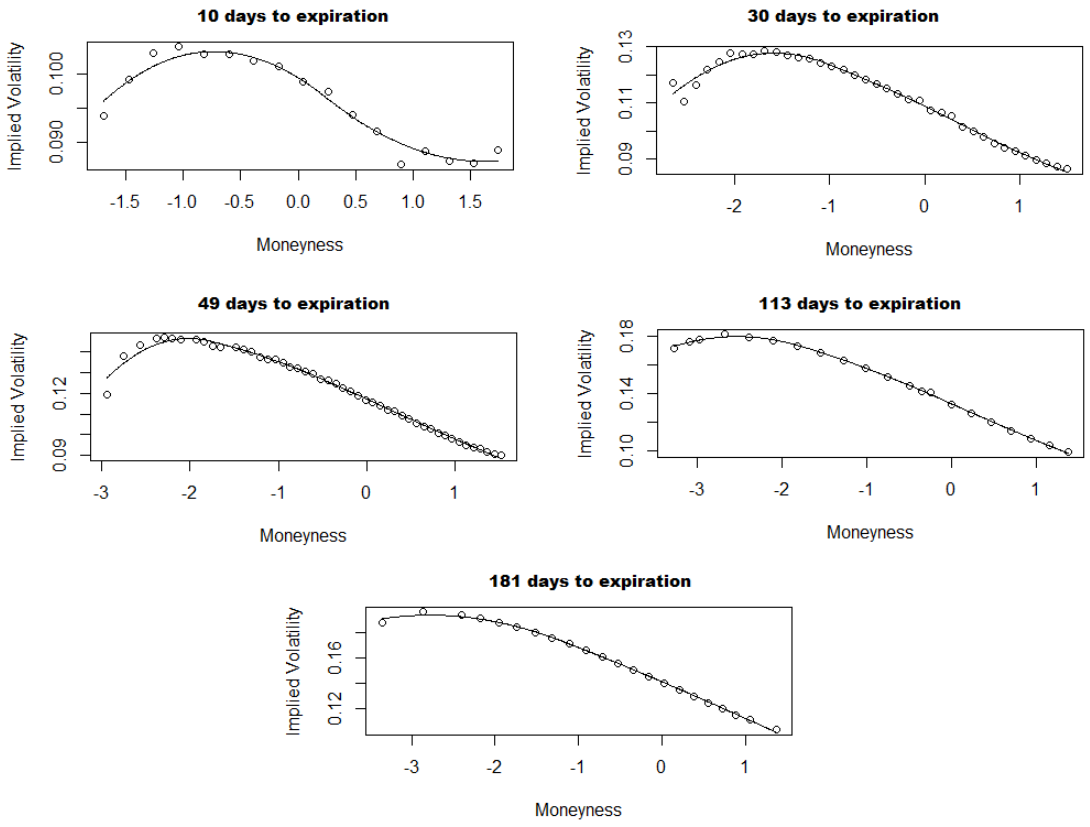
$$
m^\star = \frac{\ln\left(\frac{K}{F}\right)}{\sigma_{\text{ATMV}}\sqrt{\tau}}
$$

Here σ_{ATMV} is the at-the-money implied volatility, defined as the implied volatility when $K = F$. Figure 3.5 presents the state price density cross sections plotted against strike price and against the standardized moneyness.

3.3.5 Full Data Set Application and Evaluation

Using the data described at the beginning of this section, we perform a comparative analysis regarding RMSE and MAPE for trend filtering and competing models. Since we are dealing with daily data, the comparison benchmarks are parametric. However, since we no longer have perfect knowledge of the underlying specification, a second benchmark besides the Shimko specification is necessary. Our choice is a fourth order polynomial version of the stringwise regression per

Figure 3.4: Fitted values of implied volatility cross sections against moneyness for January 5, 2007



tenor. [4] In order to provide clarity in our measures across the dataset, we use a simple comparative measure of percentual advantage in goodness of fit for each date and analyze their distribution. The measures of percentual advantage are defined as:

$$PARMSE_{Trend\,Filtering} = \frac{RMSE_{Competitor} - RMSE_{Trend\,Filtering}}{RMSE_{Competitor}} \times 100$$

and

$$PAMAPE_{Trend\,Filtering} = \frac{MAPE_{Competitor} - MAPE_{Trend\,Filtering}}{MAPE_{Competitor}} \times 100$$

A graphical inspection of the distribution for the percentual advantages per stock for both the

[4] LÃl'vy process based estimation (for instance using the NIG distribution) would be preferrable, but they become computationally expensive for fitting due to the large number of parameters and unfeasible for our purposes.

training set and test set can be found in Figures 3.6 to 3.9 . A quick inspections shows that the percentual advantage of both goodness of fit measures skews to the right of zero. There are some distributions that have their mode close to zero, hence a more detailed examination is necessary.

Table 3.8 contains the quartiles and means of the distributions of percentual advantage for the training set per stock. An interesting observation to be found is that the average percentual advantages for MAPE are either small or negative when it comes to the fourth order polynomial specification. Thus, it is not possible to claim superiority of the method when applying in an in-sample basis. However, this result does not reveal anything about generalizing the result of the fit when using out of sample observations.

Table 3.9 describes the distribution of percentile advantages when using the test set. Since the test set comes from a random split, the observations come from the same population as our training set and since they are not observed by the model during the training stage, goodness of fit measures work as validation metrics. A consistent level of positive percentual advantage on the test set describes how well trend filtering generalizes in its advantage to the competing models. The table shows positive average and median percentual advantages for trend filtering across our comparison firm. The MAPE advantage for comparing with the Kellogg Company contracts against the fourth order polynomial specification is modest with a mean of 0.1999 and a median of 2.2866. We can see from the corresponding graph in Figure 3.8 that we cannot claim that trend filtering underperforms the fourth order polynomial specification due to relative symmetry of the density.

3.4 Concluding Remarks

We compare performance of the trend filtering estimator to two versions of kernel estimators and Shimko (1993)'s estimator, and either a "perfect knowledge" specification or a fourth order polynomial stringwise regression in terms of goodness of fit. From the evidence obtained from Monte Carlo simulation, we have the following observations. First, we find that overall trend filltering outperforms the competing models in terms of goodness of fit. Second, the advantage of trend filtering is increased with more observations per tenor and higher levels of noise. However

Table 3.8: Percentual advantage distributions of RMSE and MAPE for training set

	S&P 500			
	RMSE		MAPE	
Quantile	Shimko	4th order polynomial	Shimko	4th order polynomial
25%	54.1211	9.1954	23.8502	-22.9682
50%	78.4053	34.6772	56.2479	0.9223
75%	89.4747	64.1643	73.6513	31.6529
Mean	70.2604	30.0964	46.5838	0.4718
	Apple Inc.			
	RMSE		MAPE	
Quantile	Shimko	4th order polynomial	Shimko	4th order polynomial
25%	41.9757	11.1899	5.6839	-6.7873
50%	56.3376	23.5776	27.1912	0.7220
75%	67.9064	41.4481	41.8636	8.6704
Mean	54.4150	28.2960	23.2935	-0.1176
	Kellogg Company			
	RMSE		MAPE	
Quantile	Shimko	4th order polynomial	Shimko	4th order polynomial
25%	59.3240	20.2901	-0.9930	-12.9284
50%	70.8600	40.2957	19.1101	-0.5575
75%	80.2764	58.8421	37.5626	14.2121
Mean	68.8745	36.8713	15.7883	-1.7913
	Gap Inc.			
	RMSE		MAPE	
Quantile	Shimko	4th order polynomial	Shimko	4th order polynomial
25%	65.5188	13.0963	16.2770	-12.7456
50%	76.4344	31.7883	38.6215	-0.1719
75%	82.7293	49.2894	54.9891	12.9152
Mean	73.9159	26.5130	32.9157	-0.8963

Table 3.9: Percentual advantage distributions of RMSE and MAPE for test set

	S&P 500			
	RMSE		MAPE	
Quantile	Shimko	4th order polynomial	Shimko	4th order polynomial
25%	63.7449	12.0730	38.2608	-19.7314
50%	84.8362	43.8842	68.7452	8.0804
75%	91.9428	69.5619	81.1629	41.9497
Mean	76.3139	34.6929	58.2265	7.2678
	Apple Inc.			
	RMSE		MAPE	
Quantile	Shimko	4th order polynomial	Shimko	4th order polynomial
25%	62.8778	26.5596	41.0992	1.9144
50%	70.8779	38.5207	50.3692	12.1835
75%	78.0014	52.8872	58.8581	23.0382
Mean	69.8562	39.9126	48.9226	12.9441
	Kellogg Company			
	RMSE		MAPE	
Quantile	Shimko	4th order polynomial	Shimko	4th order polynomial
25%	65.9918	27.7396	7.9840	-12.5612
50%	74.0908	46.7744	25.7495	2.2866
75%	83.5004	63.6642	44.1244	18.2726
Mean	73.8970	40.6298	22.8245	0.1999
	Gap Inc.			
	RMSE		MAPE	
Quantile	Shimko	4th order polynomial	Shimko	4th order polynomial
25%	67.8673	16.9786	26.0828	-10.6865
50%	77.4128	36.4287	46.4324	3.1611
75%	84.1633	52.1939	61.2893	18.8394
Mean	75.5671	29.7625	40.6970	3.0607

even when operating at the minimum necessary number of observations and high levels of noise the advantage makes it a good candidate to use when dealing with daily data. Lastly, the Shimko (1993) specification approaches the goodness of fit measures of nonparametric kernel methods as the datasets get more dense.

A validation examination performed with daily data from January 1, 2007 to January 1, 2010 shows that trend filtering is able to consistently outperform Shimko (1993)' s specification and a fourth order polynomial variant both when dealing with out of sample observations in most cases. However, even when the advantages are not completely clear, trend filtering does not underperform its competitors. These results make trend filtering a desirable candidate for empirical use given its flexibility, robustness to misspecification and handling of overfitting.

Additionally, the usable quotes from S&P 500 on January 5, 2007 show that trend filtering adapts itself well to the data. Graphing the SPD per tenor on the same day shows that trend filtering predicts implied volatilities such the extracted crosssectional densities are well behaved.

Given the results, it is possible to conclude that trend filtering, a newly proposed method to estimate the implied volatility, provides a better fit than other popular and comparable methods when we have a relatively small number of quotes (at least seven per tenor). This bridges the gap in performance between typical nonparametric methods and parametric methods, allowing reduced datasets to generate flexible estimates without having to make strong functional structure assumptions on the implied volatility data generating process.

Figure 3.5: Extracted SPD cross sections for January 5, 2007 after using trend filtering

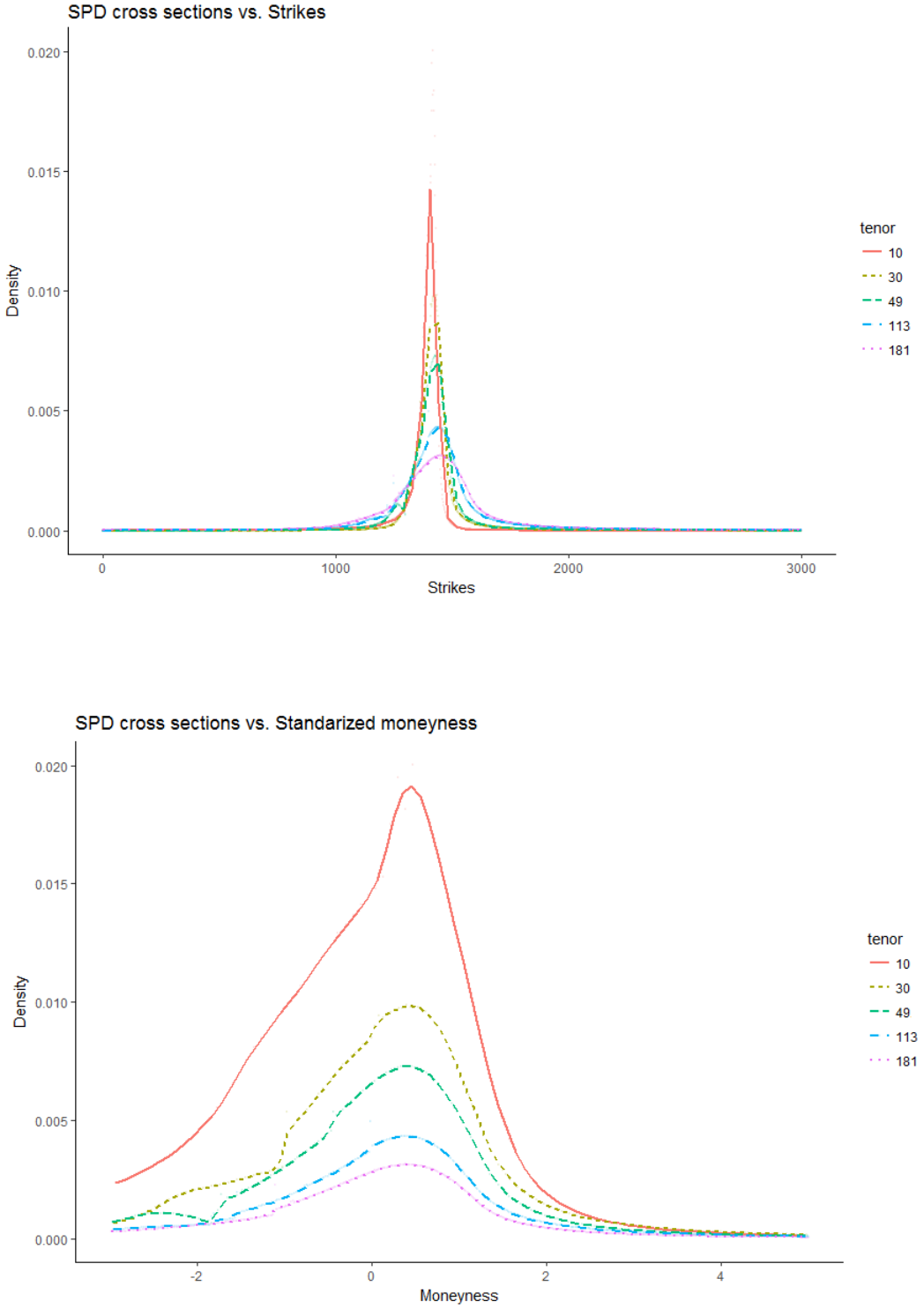


Figure 3.6: S&P 500 percentual advantage densities January 1, 2007 to January 1, 2010

Figure 3.7: Apple Inc. percentual advantage densities January 1, 2007 to January 1, 2010

Figure 3.8: Kellogg Company percentual advantage densities January 1, 2007 to January 1, 2010

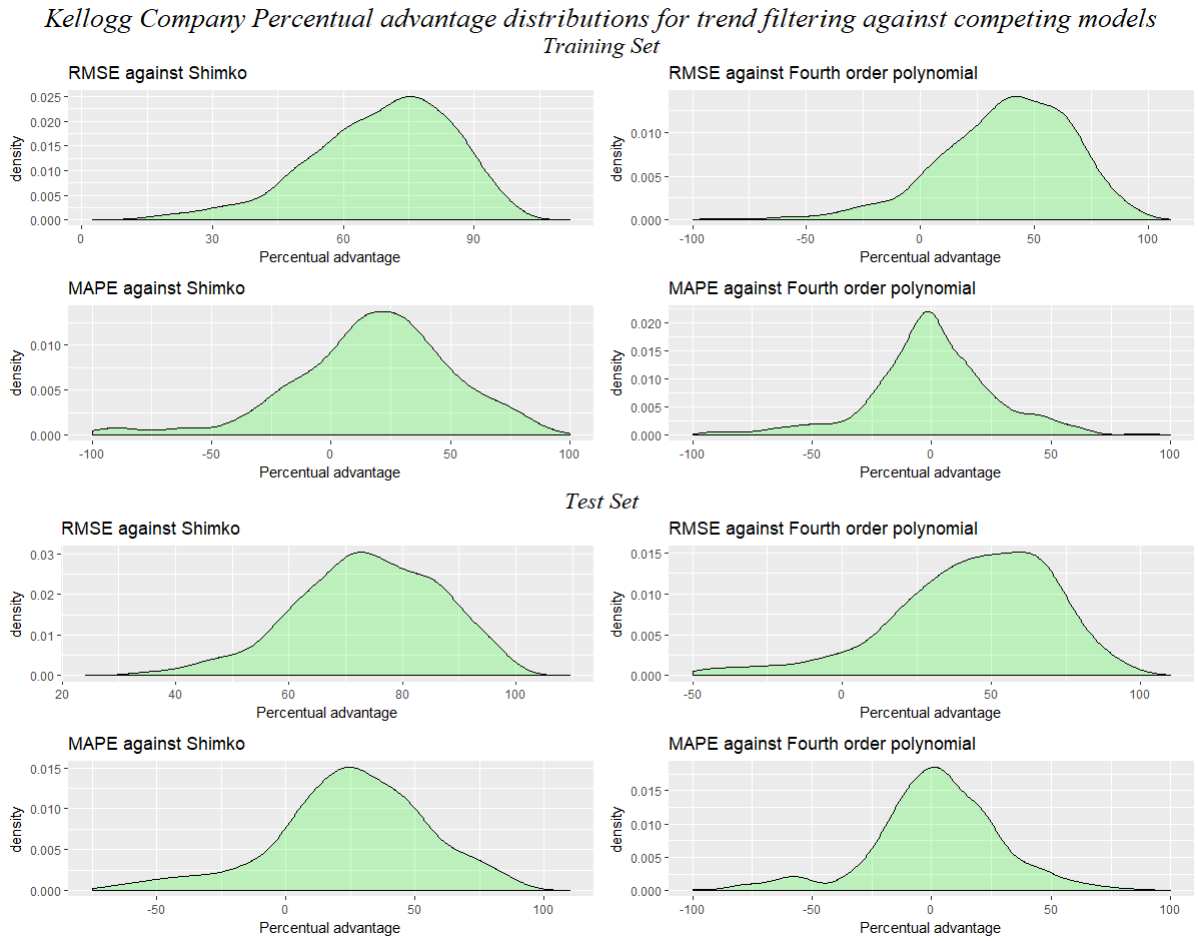


Figure 3.9: Gap Inc. percentual advantage densities January 1, 2007 to January 1, 2010

4. CONCLUSIONS

This dissertation includes two independent essays. In the first essay, I propose a nonparametric test for unobserved heterogeneity in treatment effects. The test can answer an crucial policy-relevant research question: Whether the measured treatment effects are also valid for other subpopluations? The second essay provides an alternative to the existing estimations of implied volatility in option pricing when usable data is scarce on daily basis.

Below I summarize these two essays.

4.1 The First Essay

Unobserved heterogeneity in causal effects is an intrinsic feature of nonseparable models and is also a fundamental aspect in the treatment effect literature. In this paper, we are the first to propose a nonparametric test for unobserved heterogeneous treatment effects in a nonseparable triangular model. We show that the null hypothesis is equivalent to testing for an independence condition of observables. With continuous covariates, we propose a Kolmogorov-Smirnov-type test that is simple to implement and achieve $\sqrt{n}$-local power. Monte Carlo simulations show that the proposed test performs very well in small samples. For illustration, we apply our test to study the heterogeneity in the effects of the Job Training Partnership Act on earnings and the impacts of fertility on family income.

4.2 The Second Essay

The use of state price densities to gather information about market sentiment and other empirical characteristics that describe important phenomena is popular in literature and in practice. The estimation of the implied volatility surface to extract these densities is a crucial intermediate step in the process, and the methods to do so are varied in literature. This paper proposes an estimation procedure that is relative new in nonparametric literature: ℓ_1 trend filtering. We show its advantages over typically used nonparametric and parametric methods, commonly used in literature and in practice, to deal with this particular estimation problem. Additionally, the method maintains

smaller prediction errors than the comparison models across different number of observations and levels of noise.

REFERENCES

ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, **113** (2), 231–263.

—, ANGRIST, J. and IMBENS, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, **70** (1), 91–117.

AIT-SAHALIA, Y. and DUARTE, J. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, **116** (1), 9–47.

AÏT-SAHALIA, Y. and LO, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *The Journal of Finance*, **53** (2), 499–547.

ANDREWS, D. W. (1997). A conditional kolmogorov test. *Econometrica*, pp. 1097–1128.

ANGRIST, J. D. and EVANS, W. N. (1998). Children and their parents' labor supply: Evidence from exogenous variation in family size. *American Economic Review*, pp. 450–477.

— and KRUEGER, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106** (4), 979–1014.

APARICIO, S. and HODGES, S. (1998). Implied risk-neutral distribution: A comparison of estimation methods. *FORC Preprint, University of Warwick*.

BARRETT, G. F. and DONALD, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, **71**, 71–104.

BLACK, F. and SCHOLES, M. (1973). The pricing of options and corporate liabilities. *The journal of political economy*, pp. 637–654.

BLOOM, H. S., ORR, L. L., BELL, S. H., CAVE, G., DOOLITTLE, F., LIN, W. and BOS, J. M. (1997). The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *Journal of human resources*, pp. 549–576.

BLUNDELL, R. and POWELL, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. *Econometric society monographs*, **36**, 312–357.

BOUEZMARNI, T., ROMBOUTS, J. V. and TAAMOUTI, A. (2012). Nonparametric copula-based

test for conditional independence with applications to granger causality. *Journal of Business & Economic Statistics*, **30** (2), 275–287.

BREEDEN, D. T. and LITZENBERGER, R. H. (1978). Prices of state-contingent claims implicit in option prices. *Journal of business*, pp. 621–651.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. and OLSHEN, R. A. (1984). *Classification and regression trees*. CRC press.

BRONARS, S. G. and GROGGER, J. (1994). The economic consequences of unwed motherhood: Using twin births as a natural experiment. *The American Economic Review*, pp. 1141–1156.

CARR, P. and MADAN, D. B. (2005). A note on sufficient conditions for no arbitrage. *Finance Research Letters*, **2** (3), 125–130.

CHERNOZHUKOV, V. and HANSEN, C. (2005). An IV Model of Quantile Treatment Effects. *Econometrica*, **73** (1), 245–261.

CHESHER, A. (2003). Identification in Nonseparable Models. *Econometrica*, **71** (5), 1405–1441.

— (2005). Nonparametric Identification under Discrete Variation. *Econometrica*, **73** (5), 1525–1550.

CHRISTOFFERSEN, P., HESTON, S. and JACOBS, K. (2009). The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well. *Management Science*, **55** (12), 1914–1932.

COX, J. C. and ROSS, S. A. (1976). The valuation of options for alternative stochastic processes. *Journal of financial economics*, **3** (1-2), 145–166.

DELGADO, M. A. and MANTEIGA, W. G. (2001). Significance testing in nonparametric regression based on the bootstrap. *Annals of Statistics*, pp. 1469–1507.

D'HAULTFŒUILLE, X. and FÉVRIER, P. (2015). Identification of nonseparable triangular models with discrete instruments. *Econometrica*, **83** (3), 1199–1210.

FIGLEWSKI, S. (2008). Estimating the implied risk neutral density. In M. W. Tim Bolloerslev, Jeffrey R. Russell (ed.), *Volatility and Time Series Econometrics: Essay in Honor of Robert F. Engle*, Oxford, UK: Oxford University Press.

FLORENS, J.-P., HECKMAN, J. J., MEGHIR, C. and VYTLACIL, E. (2008). Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. *Econometrica*, **76** (5), 1191–1206.

FRÖLICH, M. and MELLY, B. (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, **31** (3), 346–357.

HAYES, S., PANIGIRTZOGLOU, N. and SHIN, H. S. (2003). Liquidity and risk appetite: evidence from equity index option prices. *Bank of England mimeo*.

HECKMAN, J. J., SCHMIERER, D. and URZUA, S. (2010). Testing the Correlated Random Coefficient Model. *Journal of econometrics*, **158** (2), 177–203.

—, SMITH, J. and CLEMENTS, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, **64** (4), 487–535.

—, URZUA, S. and VYTLACIL, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, **88** (3), 389–432.

— and VYTLACIL, E. (2001). Policy-relevant treatment effects. *The American Economic Review*, **91** (2), 107–111.

— and — (2005). Structural equations, treatment effects, and econometric policy evaluation1. *Econometrica*, **73** (3), 669–738.

HESTON, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of financial studies*, **6** (2), 327–343.

HODERLEIN, S. and MAMMEN, E. (2009). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrics Journal*, pp. 1–25.

— and WHITE, H. (2012). Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics*, **168** (2), 300–314.

HUANG, M., SUN, Y. and WHITE, H. (2016). A flexible nonparametric test for conditional independence. *Econometric Theroy*, **32** (6), 1434–1482.

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treat-

ment effects. *Econometrica*, **62** (2), 467–475.

— and LEMIEUX, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, **142** (2), 615–635.

— and NEWEY, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, **77** (5), 1481–1512.

— and RUBIN, D. B. (1997). Estimating Distributions for Outcome Compliers Models in Instrumental Variables. *Review of Economic Studies*, **64** (4), 555–574.

JACKWERTH, J. C. (2004). Option-implied risk-neutral distributions and risk aversion.

JACOBSEN, J. P., PEARCE, J. W. and ROSENBLOOM, J. L. (1999). The effects of childbearing on married women's labor supply and earnings: Using twin births as a natural experiment. *The Journal of Human Resources*, **34** (3), 449–474.

KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *The Annals of Statistics*, pp. 191–219.

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). \ell_1 trend filtering. *SIAM review*, **51** (2), 339–360.

KOHAVI, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence-Volume 2*, Morgan Kaufmann Publishers Inc., pp. 1137–1143.

LEE, D. S. and LEMIEUXA, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, **48** (2), 281–355.

LINTON, O. and GOZALO, P. (2014). Testing conditional independence restrictions. *Econometric Reviews*, **33** (5-6), 523–552.

LU, X. and WHITE, H. (2014). Testing for Separability in Structural Equations. *Journal of Econometrics*, **182** (1), 14–26.

LUDWIG, M. (2015). Robust estimation of shape-constrained state price density surfaces. *The Journal of Derivatives*, **22** (3), 56–72.

MALZ, A. M. (2014). A simple and reliable way to compute option-based risk-neutral distributions. *FRB of New York Staff Report*, (677).

MATZKIN, R. L. (1999). *Nonparametric estimation of nonadditive random functions*. Tech. rep., Northwestern University.

— (2003). Nonparametric estimation of nonadditive random functions. *Econometrica*, **71** (5), 1339–1375.

MERTON, R. C. (1973). Theory of rational option pricing. *The Bell Journal of economics and management science*, pp. 141–183.

NOLAN, D. and POLLARD, D. (1988). Functional limit theorems for u-processes. *The Annals of Probability*, pp. 1291–1298.

PAGAN, A. and ULLAH, A. (1999). *Nonparametric Econometrics*. Cambridge University Press.

POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, pp. 1403–1430.

ROSENZWEIG, M. R. and WOLPIN, K. I. (1980). Testing the quantity-quality fertility model: The use of twins as a natural experiment. *Econometrica: journal of the Econometric Society*, pp. 227–240.

ROSS, S. A. (1976). Options and efficiency. *The Quarterly Journal of Economics*, pp. 75–89.

SHIMKO, D. (1993). Bounds of probability. *Risk*, **6** (4), 33–37.

STINCHCOMBE, M. B. and WHITE, H. (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric theory*, **14** (03), 295–325.

SU, L., TU, Y. and ULLAH, A. (2015). Testing additive separability of error term in nonparametric structural models. *Econometric Reviews*, **34** (6-10), 1057–1088.

— and WHITE, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, **141** (2), 807–834.

— and — (2008). A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, **24** (04), 829–864.

— and — (2014). Testing conditional independence via empirical likelihood. *Journal of Econometrics*, **182** (1), 27–44.

TIBSHIRANI, R. J., TAYLOR, J. E., CANDES, E. J. and HASTIE, T. (2011). *The solution path of*

the generalized lasso. Stanford University.

— *et al.* (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, **42** (1), 285–323.

TORGOVITSKY, A. (2015). Identification of nonseparable models using instruments with small support. *Econometrica*, **83** (3), 1185–1197.

VAN DER KLAAUW, W. (2008). Regression–discontinuity analysis: a survey of recent developments in economics. *Labour*, **22** (2), 219–245.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). Weak convergence. In *Weak Convergence and Empirical Processes*, Springer, pp. 16–28.

— and — (2007). Empirical processes indexed by estimated functions. *Lecture Notes-Monograph Series*, pp. 234–252.

VUONG, Q. and XU, H. (2016). Counterfactual mapping and individual treatment effects in nonseparable models with discrete endogeneity. *Quantitative Economics*, **forthcoming**.

VYTLACIL, E. (2002). Independence, Monotoniciy, and Latent Index Models: An Equivalence Result. *Econometrica*, **70** (1), 331—-341.

APPENDIX A

PROOFS OF LEMMAS AND THEOREMS IN SECTION 1

A.1 Proof of Proposition 2.1

Proof. For the "if" part, under (2.3), we have

$$g(1,x,\epsilon) - g(0,x,\epsilon) = m(1,x) - m(0,x) \equiv \delta_0(x), \ \ \forall x \in \mathscr{S}_X.$$

For the "only if" part, (2.2) implies

$$g(d,x,\epsilon) = d \times [g(1,x,\epsilon) - g(0,x,\epsilon)] + g(0,x,\epsilon) = d \times \delta_0(x) + g(0,x,\epsilon).$$

Therefore, (2.3) holds in the sense $m(d,x) = d \times \delta_0(x)$ and $\nu(x,\epsilon) = g(0,x,\epsilon)$. *Q.E.D.*

A.2 Proof of Theorem 2.1

Proof. Because Proposition 2.1 provides the only if part, then it suffices to show the if part. Suppose $W \perp\!\!\!\perp Z \mid X$. By the definition of W, we have: for any $y \in \mathbb{R}$,

$$\begin{aligned}&\Pr(Y \le y, D = 1|X, Z = 1) + \mathbb{P}(Y + \delta(X) \le y, D = 0|X, Z = 1) \\ &\qquad = \Pr(Y \le y, D = 1|X, Z = 0) + \mathbb{P}(Y + \delta(X) \le y, D = 0|X, Z = 0).\end{aligned}$$

It follows that

$$\begin{aligned}&\Pr(Y \le y, D = 1|X, Z = 1) - \mathbb{P}(Y \le y, D = 1|X, Z = 0) \\ &\qquad = \Pr(Y \le y - \delta(X), D = 0|X, Z = 1) - \mathbb{P}(Y \le y - \delta(X), D = 0|X, Z = 0). \quad \text{(A.1)}\end{aligned}$$

Denote $V \equiv \nu(X, \epsilon)$ and

$$\Delta_0(\tau, x) \equiv \Pr(V \leq \tau, D = 0 | X = x, Z = 1) - \Pr(V \leq \tau, D = 0 | X = x, Z = 0);$$
$$\Delta_1(\tau, x) \equiv \Pr(V \leq \tau, D = 1 | X = x, Z = 0) - \Pr(V \leq \tau, D = 1 | X = x, Z = 1).$$

By Assumptions 2.1 and 2.3, we have

$$\Delta_0(\tau, x) = \Pr(V \leq \tau, \eta \in \mathcal{C}_x | X = x) = \Delta_1(\tau, x)$$

which is strictly monotone in $\tau \in \mathscr{S}_{V|X=x,\ \eta \in \mathcal{C}_x}$. Moreover, there is $\mathscr{S}_{V|X=x,\ \eta \in \mathcal{C}_x} = \mathscr{S}_{V|X=x}$ under Assumptions 2.2 and 2.4.

Therefore, we have

$$\begin{aligned}
& \Pr(Y \leq y, D = 1 | X = x, Z = 0) - \mathbb{P}(Y \leq y, D = 1 | X = x, Z = 1) \\
= \ & \Delta_1(\tilde{g}^{-1}(1, x, y), x) \\
= \ & \Delta_0(\tilde{g}^{-1}(1, x, y), x) \\
= \ & \Pr(Y \leq \tilde{g}(0, x, \tilde{g}^{-1}(1, x, y)), D = 0 | X = x, Z = 1) \\
- \ & \mathbb{P}(Y \leq \tilde{g}(0, x, \tilde{g}^{-1}(1, x, y)), D = 0 | X = x, Z = 0),
\end{aligned}$$

where $\tilde{g}^{-1}(1, x, \cdot)$ is the inverse function of $\tilde{g}(1, x, \cdot)$ and $\tilde{g}$ is a monotone function introduced in Assumption 2.2. Note that both sides are strictly monotone in $y \in \mathscr{S}_{\tilde{g}(1,X,V)|X=x}$ since $\Delta_d(\cdot, x)$ is strictly monotone on $\mathscr{S}_{V|X=x}$ under Assumption 2.4.

Combine the above result with (A.1), then we have

$$\tilde{g}(0, x, \tilde{g}^{-1}(1, x, y)) = y - \delta(x), \quad \forall x \in \mathscr{S}_X,\ y \in \mathscr{S}_{\tilde{g}(1,x,V)|X=x}.$$

Let $y = \tilde{g}(1, x, \tau)$ for some $\tau \in \mathscr{S}_{V|X=x}$. Then the above equation becomes

$$\tilde{g}(0, x, \tau) = \tilde{g}(1, x, \tau) - \delta(x).$$

Q.E.D.

A.3 Proof of Theorem 2.2

Proof. Let $\mathbb{1}_{WXZ}(w, x, z) = \mathbb{1}(W \leq w) \times \mathbb{1}_{XZ}(x, z)$ and $\mathbb{1}_{\hat{W}XZ}(w, x, z) = \mathbb{1}(\hat{W} \leq w) \times \mathbb{1}_{XZ}(x, z)$. Let further $\mathbb{1}_{W(\tilde{\delta})XZ}(w, x, z) = \mathbb{1}(W(\tilde{\delta}) \leq w) \times \mathbb{1}_{XZ}(x, z)$, where $W(\tilde{\delta}) = Y + (1 - D)\tilde{\delta}(X)$, be a function indexed by $\tilde{\delta}(\cdot) \in \mathbb{R}^{\mathscr{S}_X}$. By definition, $\mathbb{1}_{W(\delta)XZ}(w, x, z) = \mathbb{1}_{WXZ}(w, x, z)$ and $\mathbb{1}_{W(\hat{\delta})XZ}(w, x, z) = \mathbb{1}_{\hat{W}XZ}(w, x, z)$.

We first derive the asymptotics of $\sqrt{n}[\hat{F}_{W|XZ}(w|x, z) - F_{W|XZ}(w|x, z)]$. By definition,

$$F_{W|XZ}(w|x, z) = \frac{\mathbb{E}[\mathbb{1}_{WXZ}(w, x, z)]}{\mathbb{E}[\mathbb{1}_{XZ}(x, z)]} \text{ and } \hat{F}_{W|XZ}(w|x, z) = \frac{\mathbb{E}_n[\mathbb{1}_{\hat{W}XZ}(w, x, z)]}{\mathbb{E}_n[\mathbb{1}_{XZ}(x, z)]}.$$

In the expectation $\mathbb{E}[\mathbb{1}_{W(\hat{\delta})XZ}(\cdot, x, z)]$ discussed below, we treat $\hat{\delta}$ as an index rather than a random object. Note that

$$\begin{aligned}
\mathbb{E}_n[\mathbb{1}_{\hat{W}XZ}(\cdot, x, z)] &= \mathbb{E}_n[\mathbb{1}_{WXZ}(\cdot, x, z)] - \mathbb{E}[\mathbb{1}_{WXZ}(\cdot, x, z)] + \mathbb{E}[\mathbb{1}_{W(\hat{\delta})XZ}(\cdot, x, z)] \\
&+ \Big\{\mathbb{E}_n[\mathbb{1}_{W(\hat{\delta})XZ}(\cdot, x, z)] - \mathbb{E}[\mathbb{1}_{W(\hat{\delta})XZ}(\cdot, x, z)] - \mathbb{E}_n[\mathbb{1}_{W(\delta)XZ}(\cdot, x, z)] + \mathbb{E}[\mathbb{1}_{W(\delta)XZ}(\cdot, x, z)]\Big\} \\
&= \mathbb{E}_n[\mathbb{1}_{WXZ}(\cdot, x, z)] - \mathbb{E}[\mathbb{1}_{WXZ}(\cdot, x, z)] + \mathbb{E}[\mathbb{1}_{W(\hat{\delta})XZ}(\cdot, x, z)] + o_p(n^{-1/2}),
\end{aligned}$$

where the last step is by the empirical process theory (see e.g. van der Vaart and Wellner, 2007). By Taylor expansion,

$$\sqrt{n}\ \Big\{\mathbb{E}[\mathbb{1}_{W(\hat{\delta})XZ}(\cdot, x, z)] - F_{W|XZ}(w|x, z)\Big\} = \frac{\partial \mathbb{E}[\mathbb{1}_{W(\delta)XZ}(w, x, z)]}{\partial \delta} \times \sqrt{n}(\hat{\delta} - \delta) + o_p(1).$$

Note that $\frac{\partial \mathbb{E}[\mathbb{1}_{W(\delta)XZ}(w,x,z)]}{\partial \delta(x')} = 0$ for all $x' \neq x$ and $\frac{\partial \mathbb{E}[\mathbb{1}_{W(\delta)XZ}(w,x,z)]}{\partial \delta(x)} = -f_{W|DXZ}(w|0, x, z) \times$

$\Pr(D=0, X=x, Z=z)$. Therefore, we have

$$\sqrt{n}\ \left\{\mathbb{E}[\mathbb{1}_{W(\hat{\delta})XZ}(\cdot,x,z)] - F_{W|XZ}(w|x,z)\right\}$$
$$+\sqrt{n}\left\{\mathbb{E}_n[\mathbb{1}_{WXZ}(\cdot,x,z)] - \mathbb{E}[\mathbb{1}_{WXZ}(\cdot,x,z)]\right\} - f_{WDXZ}(w,0,x,z)\times\sqrt{n}[\hat{\delta}(x)-\delta(x)] + o_p(1).$$

Moreover, $\mathbb{E}_n[\mathbb{1}_{XZ}(x,z)] = \mathbb{P}(X=x,Z=z) + O_p(n^{-1/2})$ under the central limit theorem. Thus, by Slutskyï£¡s theorem, we have

$$\begin{aligned}
&\sqrt{n}\left[\hat{F}_{W|XZ}(w|x,1) - \hat{F}_{W|XZ}(w|x,0)\right] - \sqrt{n}\left[F_{W|XZ}(w|x,1) - F_{W|XZ}(w|x,0)\right] \\
&= \frac{\sqrt{n}\big\{\mathbb{E}_n[\mathbb{1}_{WXZ}(w,x,1)] - \mathbb{E}[\mathbb{1}_{WXZ}(w,x,1)]\big\} - f_{WDXZ}(w,0,x,1)\times\sqrt{n}[\hat{\delta}(x)-\delta(x)]}{\mathbb{P}(X=x,Z=1)} \\
&\quad - \frac{\sqrt{n}\big\{\mathbb{E}_n[\mathbb{1}_{WXZ}(w,x,0)] - \mathbb{E}[\mathbb{1}_{WXZ}(w,x,0)]\big\} - f_{WDXZ}(w,0,x,0)\times\sqrt{n}[\hat{\delta}(x)-\delta(x)]}{\mathbb{P}(X=x,Z=0)} \\
&\quad + \frac{\sqrt{n}\Pr(W\le w, X=x, Z=1)}{\mathbb{E}_n\mathbb{1}_{XZ}(x,1)} - \frac{\sqrt{n}\Pr(W\le w, X=x, Z=0)}{\mathbb{E}_n\mathbb{1}_{XZ}(x,0)} + o_p(1).
\end{aligned}$$

Applying Taylor expansion, we have

$$\begin{aligned}
&\frac{\sqrt{n}\Pr(W\le w, X=x, Z=z)}{\mathbb{E}_n\mathbb{1}_{XZ}(x,z)} - \sqrt{n}\ F_{W|XZ}(w|x,z) \\
&\qquad = -F_{W|XZ}(w|x,z)\times\frac{\sqrt{n}\big[\mathbb{E}_n\mathbb{1}_{XZ}(x,z) - \mathbb{P}(X=x,Z=z)\big]}{\Pr(X=x,Z=z)} + o_p(1).
\end{aligned}$$

Moreover, applying Lemma B.1, we have

$$\begin{aligned}
&\sqrt{n}\left[\hat{F}_{W|XZ}(w|x,1) - \hat{F}_{W|XZ}(w|x,0)\right] - \sqrt{n}\left[F_{W|XZ}(w|x,1) - F_{W|XZ}(w|x,0)\right] \\
=\ & \sqrt{n}\mathbb{E}_n\left\{[\mathbb{1}(W\le w) - F_{W|XZ}(w|x,1)]\times\frac{\mathbb{1}_{XZ}(x,1)}{\mathbb{P}(X=x,Z=1)}\right\} \\
-\ & \sqrt{n}\mathbb{E}_n\left\{[\mathbb{1}(W\le w) - F_{W|XZ}(w|x,0)]\times\frac{\mathbb{1}_{XZ}(x,0)}{\mathbb{P}(X=x,Z=0)}\right\} \\
+\ & \kappa(w,x)\times\sqrt{n}\mathbb{E}_n\left\{\big[W - \mathbb{E}(W|X=x,Z=0)\big]\times\frac{\mathbb{1}_{XZ}(x,1)}{\mathbb{P}(X=x,Z=1)}\right\} \\
-\ & \kappa(w,x)\times\sqrt{n}\mathbb{E}_n\left\{\big[W - \mathbb{E}(W|X=x,Z=1)\big]\times\frac{\mathbb{1}_{XZ}(x,0)}{\mathbb{P}(X=x,Z=0)}\right\} + o_p(1).
\end{aligned}$$

Under the null hypothesis, there is

$$
\begin{aligned}
& \sqrt{n}\left[\hat{F}_{W|XZ}(w|x,1) - \hat{F}_{W|XZ}(w|x,0)\right] \\
= \;& \sqrt{n}\mathbb{E}_n\left\{[\mathbb{1}(W \le w) - F_{W|X}(w|x)] \times \Big[\frac{\mathbb{1}_{XZ}(x,1)}{\mathbb{P}(X=x,Z=1)} - \frac{\mathbb{1}_{XZ}(x,0)}{\mathbb{P}(X=x,Z=0)}\Big]\right\} \\
+ \;& \kappa(w,x) \times \sqrt{n}\mathbb{E}_n\left\{\big[W - \mathbb{E}(W|X=x)\big] \times \Big[\frac{\mathbb{1}_{XZ}(x,1)}{\mathbb{P}(X=x,Z=1)} - \frac{\mathbb{1}_{XZ}(x,0)}{\mathbb{P}(X=x,Z=0)}\Big]\right\} + o_p(1) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi_{wx,i} + \phi_{wx,i}) \\
+ \;& o_p(1)
\end{aligned}
$$

where $\psi_{wx,i}$ and $\phi_{wx,i}$ are defined by (2.6) and (2.7). Following e.g. Kim and Pollard (1990), we have $\hat{\mathcal{T}}_n \xrightarrow{d} \sup_{w\in\mathbb{R};\, x\in\mathscr{S}_X} |\mathcal{Z}(w,x)|$. *Q.E.D.*

A.4 Proof of Lemma 2.2

Proof. Fix $X = x$ and w.l.o.g., let $z = 1$. Note that

$$
\begin{aligned}
& \hat{G}(w,x,1) - \tilde{G}(w,x,1) \\
= \;& \mathbb{E}_n\left\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-\hat{W})\left[\mathbb{1}(\hat{W}\le w) - \mathbb{1}(W \le w)\right]\right\} \\
= \;& \mathbb{E}_n\Big\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-\hat{W})\left[\mathbb{1}(\hat{W}\le w) - \mathbb{1}(W \le w)\right] \times \mathbb{1}(|W-w| \le n^{-r})\Big\} \\
+ \;& \mathbb{E}_n\Big\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-\hat{W})\left[\mathbb{1}(\hat{W}\le w) - \mathbb{1}(W \le w)\right] \times \mathbb{1}(|W-w| > n^{-r})\Big\} \\
\equiv \;& \mathbb{T}_1 + \mathbb{T}_2
\end{aligned}
$$

where $r \in (\frac{1}{4}, \iota)$. It suffices to show both $\mathbb{T}_1$ and $\mathbb{T}_2$ are $o_p(n^{-\frac{1}{2}})$.

First, note that

$$
\begin{aligned}
\mathbb{T}_1 = \;& \mathbb{E}_n\Big\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-W)\left[\mathbb{1}(\hat{W}\le w) - \mathbb{1}(W \le w)\right] \times \mathbb{1}(|W-w| \le n^{-r})\Big\} \\
& + \mathbb{E}_n\Big\{\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(W-\hat{W})\left[\mathbb{1}(\hat{W}\le w) - \mathbb{1}(W \le w)\right] \times \mathbb{1}(|W-w| \le n^{-r})\Big\}.
\end{aligned}
$$

Because

$$\mathbb{E}\left|\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(w-W)\left[\mathbb{1}(\hat{W}\le w)-\mathbb{1}(W\le w)\right]\times\mathbb{1}(|W-w|\le n^{-r})\right| \\ \le \mathbb{E}\left|\hat{f}_{XZ}(X_1,0)\times(w-W)\times\mathbb{1}(|W-w|\le n^{-r})\right| = O(1)\times O(n^{-2r}) = o(n^{-\frac{1}{2}}),$$

where last step holds because $r > \frac{1}{4}$. Moreover,

$$\mathbb{E}\left|\mathbb{1}^*_{XZ}(x,1)\hat{f}_{XZ}(X,0)(W-\hat{W})\left[\mathbb{1}(\hat{W}\le w)-\mathbb{1}(W\le w)\right]\times\mathbb{1}(|W-w|\le n^{-r})\right| \\ \le \mathbb{E}\left|\hat{f}_{XZ}(X_1,0)\times(W-\hat{W})\times\mathbb{1}(|W-w|\le n^{-r})\right| = O(1)\times O(n^{-\iota})\times O(n^{-r}) = o(n^{-\frac{1}{2}}).$$

Then, we have $\mathbb{T}_1 = o_p(n^{-\frac{1}{2}})$.

For term $\mathbb{T}_2$, note that

$$\mathbb{E}|\mathbb{T}_2| \le \frac{\overline{K}}{h}\times\sqrt{\mathbb{E}(w-\hat{W})^2}\times\sqrt{\mathbb{P}\left(|\hat{W}-W|>n^{-r}\right)} \\ \le \frac{\overline{K}}{h}\times\sqrt{\mathbb{E}\hat{W}^2-2w\cdot\mathbb{E}(\hat{W})+w^2}\times\sqrt{\mathbb{P}\left[|\hat{\delta}(X)-\delta(X)|>n^{-r}\right]},$$

where $\overline{K}$ is the upper bound of $K(\cdot)$. Because W is a bounded random variable and w belongs to a compact set, then $\sqrt{\mathbb{E}\hat{W}^2-2w\cdot\mathbb{E}(\hat{W})+w^2} = O(1)$. Moreover, by Lemma B.2, $\mathbb{E}|\mathbb{T}_2| \le o(n^{-k})$ for any $k > 0$. Hence, $\mathbb{T}_2 = o_p(n^{-\frac{1}{2}})$. *Q.E.D.*

A.5 Proof of Theorem 2.3

Proof. By Lemma 2.2, we have

$$\hat{\mathcal{T}}^c_n = \sqrt{n}\left|\tilde{G}(w,x,1)-\tilde{G}(w,x,0)\right| + o_p(1).$$

Let $\mathbb{1}^*_{WXZ}(w,x,z) \equiv \mathbb{1}(W\le w, X\le x, Z=z)$. Note that

$$\tilde{G}(w,x,z) = \mathbb{U}_1(w,x,z)+\mathbb{U}_2(w,x,z)+o_p(n^{-1/2})$$

where

$$\mathbb{U}_1(w,x,z) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}^*_{W_iX_iZ_i}(w,x,z) \times \hat{f}_{XZ}(X_i,z') \times (W_i - \hat{W}_i);$$
$$\mathbb{U}_2(w,x,z) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}^*_{W_iX_iZ_i}(w,x,z) \times \hat{f}_{XZ}(X_i,z') \times (w - W_i).$$

Therefore,

$$\begin{aligned}
& \sqrt{n}\left[\tilde{G}(w,x,1) - \tilde{G}(w,x,0)\right] \\
= \;& \sqrt{n}\left\{\mathbb{U}_1(w,x,1) - \mathbb{U}_1(w,x,0) - [\mathbb{E}\mathbb{U}_1(w,x,1) - \mathbb{E}\mathbb{U}_1(w,x,0)]\right\} \\
+ \;& \sqrt{n}\left\{\mathbb{U}_2(w,x,1) - \mathbb{U}_2(w,x,0) - [\mathbb{E}\mathbb{U}_2(w,x,1) - \mathbb{E}\mathbb{U}_2(w,x,0)]\right\} \\
+ \;& \sqrt{n}\left[\mathbb{E}\mathbb{U}_1(w,x,1) - \mathbb{E}\mathbb{U}_1(w,x,0)\right] + \sqrt{n}\left[\mathbb{E}\mathbb{U}_2(w,x,1) - \mathbb{E}\mathbb{U}_2(w,x,0)\right].
\end{aligned}$$

We first look at those $\mathbb{U}_2$ terms. By definition,

$$\begin{aligned}
\mathbb{U}_2(w,x,z) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\{\mathbb{1}^*_{X_iZ_i}(x,z)\lambda(W_i - w) \times \frac{1}{h}K(\frac{X_j - X_i}{h})\mathbb{1}(Z_j = z')\} \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\zeta_{n,ij}(w,x,z)
\end{aligned}$$

where $\zeta_{n,ij}(w,x,z) = \mathbb{1}^*_{X_iZ_i}(x,z) \times \lambda(W_i - w) \times \frac{1}{h}K(\frac{X_j - X_i}{h}) \times \mathbb{1}(Z_j = z')$.

Let $\zeta^*_{n,ij}(w,x,z) = \frac{1}{2}\left[\zeta_{n,ij}(w,x,z) + \zeta_{n,ji}(w,x,z)\right]$. Then, $\zeta^*_{n,ij}$ is symmetric in indices i and j. Therefore,

$$\mathbb{U}_2(w,x,z) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\zeta^*_{n,ij}(w,x,z),$$

which is a $\mathcal{U}$-process indexed by (w,x,z_ℓ). By Nolan and Pollard (1988, Theorem 5) and Powell

et al. (1989, Lemma 3.1),

$$\begin{aligned} & \mathbb{U}_2(w,x,z) - \mathbb{E}\mathbb{U}_2(w,x,z) \\ = \; & \frac{2}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[\zeta^*_{n,ij}(w,x,z)|Y_i,D_i,X_i,Z_i] - \mathbb{E}[\zeta^*_{n,ij}(w,x,z)]\right\} + o_p(n^{-1/2}). \end{aligned}$$

where the $o_p(n^{-1/2})$ applies uniformly over (w,x). Note that

$$\begin{aligned} & \mathbb{E}[\zeta^*_{n,ij}(w,x,z)|Y_i,D_i,X_i,Z_i] \\ & \quad = \frac{1}{2}\Big\{\mathbb{1}^*_{XZ}(x,z)f_{XZ}(X,z')\lambda(W-w) + \mathbb{1}^*_{XZ}(x,z')f_{XZ}(X,z)\Pi(w|X,z)\Big\} + o_p(1). \end{aligned}$$

Next, we derive $\mathbb{E}[\zeta^*_{n,ij}(w,x,z)]$. Let $u_1(w,x,z) = \mathbb{E}[\mathbb{1}^*_{XZ}(x,z)f_{XZ}(X,z')\lambda(W-w)]$ and $u_2(w,x,z) = \mathbb{E}[\mathbb{1}^*_{XZ}(x,z')f_{XZ}(X,z)\Pi(w|X,z)]$. Note that under $\mathbb{H}_0$

$$u_1(w,x,z) = u_2(w,x,z) = \int \mathbb{1}(X\le x)\Pi(w|X)f_{X|Z}(X|1)f_{X|Z}(X|0)dX \times \mathbb{P}(Z=1)\mathbb{P}(Z=0),$$

invariant with z. Therefore, $\mathbb{E}[\zeta^*_{n,ij}(w,x,z)] = \frac{1}{2}[u_1(w,x,z)+u_2(w,x,z)]$ is also invariant with z. Let $u^e(w,x) = \mathbb{E}[\zeta^*_{n,ij}(w,x,z)]$. Moreover, by Powell *et al.* (1989, Theorem 3.1),

$$\begin{aligned} & \frac{2}{\sqrt{n}}\sum_{i=1}^{n}\left\{\mathbb{E}[\zeta^*_{n,ij}(w,x,z)|Y_i,D_i,X_i] - \mathbb{E}[\zeta^*_{n,ij}(w,x,z)]\right\} \\ & \qquad = \mathbb{E}_n\left\{\mathbb{1}^*_{XZ}(x,z)f_{XZ}(X,z')\lambda(W-w) - u^e(w,x)\right\} \\ & \qquad\quad + \mathbb{E}_n\left\{\mathbb{1}^*_{XZ}(x,z')f_{XZ}(X,z)\Pi(w|X,z) - u^e(w,x)\right\} + o_p(n^{-\frac{1}{2}}), \end{aligned}$$

where the $o_p(n^{-1/2})$ holds uniformly over (w,x). Moreover, under $\mathbb{H}_0$, there is $\Pi(w|X,z) = \mathbb{E}(\lambda(W-w)|X)$. Thus,

$$\begin{aligned} & \mathbb{U}_2(w,x,1) - \mathbb{U}_2(w,x,0) - [\mathbb{E}\mathbb{U}_2(w,x,1) - \mathbb{E}\mathbb{U}_2(w,x,0)] \\ & = \mathbb{E}_n\left\{\Big[\frac{\mathbb{1}^*_{XZ}(x,1)}{f_{XZ}(X,1)} - \frac{\mathbb{1}^*_{XZ}(x,0)}{f_{XZ}(X,0)}\Big]f_{XZ}(X,0)f_{XZ}(X,1)\big[\lambda(W-w) - \mathbb{E}(\lambda(W-w)|X)\big]\right\} + o_p(n^{-\frac{1}{2}}). \end{aligned}$$

We now turn to $\mathbb{U}_1(w,x,z)$. Note that

$$\mathbb{U}_1(w,x,z) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{1}^*_{W_iX_iZ_i}(w,x,z)f_{XZ}(X_i,z')(1-D_i)[\hat{\delta}(X_i)-\delta(X_i)]\right\} + o_p(n^{-\frac{1}{2}}),$$

provided that $\mathbb{E}\left|\left[\hat{f}_{XZ}(X_i,z') - f_{XZ}(X_i,z')\right] \times \left[\hat{\delta}(X_i)-\delta(X_i)\right]\right| = o_p(n^{-\frac{1}{2}})$ holds. By a similar decomposition argument on $\hat{\delta}(X)-\delta(X)$ in Lemma B.2, we have

$$\mathbb{U}_1(w,x,z) = -\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\xi_{n,ij}(w,x,z) + o_p(n^{-1/2})$$

where

$$\xi_{n,ij}(w,x,z) =$$
$$\mathbb{1}^*_{W_iX_iZ_i}(w,x,z)f_{XZ}(X_i,z')(1-D_i)\frac{[W_j-\mathbb{E}(W_j|X_i)]\frac{1}{h}K(\frac{X_j-X_i}{h})}{p(X_i,1)-p(X_i,0)}\left[\frac{\mathbb{1}(Z_j=1)}{f_{XZ}(X_i,1)} - \frac{\mathbb{1}(Z_j=0)}{f_{XZ}(X_i,0)}\right].$$

Moreover, let $\xi^*_{n,ij}(w,x,z) = \frac{1}{2}[\xi_{n,ij}(w,x,z)+\xi_{n,ji}(w,x,z)]$. By a similar argument for $\mathbb{U}_2$,

$$\mathbb{U}_1(w,x,z) - \mathbb{E}\mathbb{U}_1(w,x,z)$$
$$= -\frac{2}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[\xi^*_{n,ij}(w,x,z)|Y_i,D_i,X_i,Z_i] - \mathbb{E}[\xi^*_{n,ij}(w,x,z)]\right\} + o_p(n^{-1/2}).$$

Note that $\mathbb{E}[\xi_{n,ij}(w,x,z)|Y_i,D_i,X_i,Z_i]=0$ and

$$\mathbb{E}[\xi_{n,ji}(w,x,z)|Y_i,D_i,X_i,Z_i] = \mathbb{E}\left\{\mathbb{E}[\xi_{n,ji}(w,x,z)|X_j,Z_j,Y_i,D_i,X_i,Z_i]\middle|Y_i,D_i,X_i,Z_i\right\}$$
$$= \mathbb{E}\Bigg\{\mathbb{1}^*_{X_jZ_j}(x,z)f_{XZ}(X_j,z')\Pr(W\leq w;D=0|X_j,Z_j)[W_i-\mathbb{E}(W|X_j)]$$
$$\times\frac{\frac{1}{h}K(\frac{X_i-X_j}{h})}{p(X_j,1)-p(X_j,0)}\left[\frac{\mathbb{1}(Z_i=1)}{f_{XZ}(X_j,1)}-\frac{\mathbb{1}(Z_i=0)}{f_{XZ}(X_j,0)}\right]\Bigg|Y_i,D_i,X_i,Z_i\Bigg\}$$
$$= F^*_{WD|XZ}(w,0|X_i,z)[W_i-\mathbb{E}(W|X_i)]\frac{f_{XZ}(X_i,0)f_{XZ}(X_i,1)}{p(X_i,1)-p(X_i,0)}\left[\frac{\mathbb{1}^*_{X_i,Z_i}(x,1)}{f_{XZ}(X_i,1)}-\frac{\mathbb{1}^*_{X_i,Z_i}(x,0)}{f_{XZ}(X_i,0)}\right]$$
$$+o_p(1)$$

where the last step comes from the Bochner's Lemma and uses the fact the integrant equals zero if $Z_j = z'$.

Thus, we have

$$\begin{aligned}
&\mathbb{U}_1(w,x,z) - \mathbb{E}\mathbb{U}_1(w,x,z) \\
&= -\mathbb{E}_n\left\{[W - \mathbb{E}(W|X)]\frac{F^*_{WD|XZ}(w,0|X,z)}{p(X,1)-p(X,0)}\left[\frac{\mathbb{1}^*_{XZ}(x,1)}{f_{XZ}(X,1)} - \frac{\mathbb{1}^*_{XZ}(x,0)}{f_{XZ}(X,0)}\right] f_{XZ}(X,1)f_{XZ}(X,0)\right\} \\
&\qquad\qquad + o_p(n^{-\frac{1}{2}}),
\end{aligned}$$

where the $o_p(n^{-1/2})$ holds uniformly over (w,x). It follows that

$$\mathbb{U}_1(w,x,1) - \mathbb{E}\mathbb{U}_1(w,x,1) - [\mathbb{U}_1(w,x,0) - \mathbb{E}\mathbb{U}_1(w,x,0)] = \mathbb{E}_n\phi^c_{wx} + o_p(n^{-\frac{1}{2}}).$$

By Assumption 2.11, we have $\mathbb{E}\mathbb{U}_1(w,x;z) = o_p(n^{-\frac{1}{2}})$. Therefore, under $\mathbb{H}_0$,

$$\begin{aligned}
& \sqrt{n}\left[\tilde{G}(w,x,1) - \tilde{G}(w,x,0)\right] \\
= \;& \sqrt{n}\left\{\mathbb{U}_1(w,x,1) - \mathbb{U}_1(w,x,0) - [\mathbb{E}\mathbb{U}_1(w,x,1) - \mathbb{E}\mathbb{U}_1(w,x,0)]\right\} \\
+ \;& \sqrt{n}\left\{\mathbb{U}_2(w,x,1) - \mathbb{U}_2(w,x,0) - [\mathbb{E}\mathbb{U}_2(w,x,1) - \mathbb{E}\mathbb{U}_2(w,x,0)]\right\} + o_p(1) \\
= \;& \sqrt{n} \times \mathbb{E}_n(\psi^c_{wx} + \phi^c_{wx}) + o_p(1),
\end{aligned}$$

which converges to a zero-mean Gaussian process with the given covariance kernel. *Q.E.D.*

APPENDIX B

TECHNICAL LEMMAS FOR PROOFS IN APPENDIX A

Let $\Delta p(x) \equiv p(x,1) - p(x,0)$, which is strictly positive by Assumption 2.1.

Lemma B.1. *Suppose Assumptions 2.1 and 2.5 hold. Then, we have*

$$\sqrt{n}[\hat{\delta}(x) - \delta(x)] = \frac{1}{\Delta p(x)} \times \sqrt{n}\mathbb{E}_n \left\{ \big[W - \mathbb{E}(W|X=x, Z=0)\big] \times \frac{\mathbb{1}_{XZ}(x,1)}{\mathbb{P}(X=x, Z=1)} \right\}$$
$$- \frac{1}{\Delta p(x)} \times \sqrt{n}\mathbb{E}_n \left\{ \big[W - \mathbb{E}(W|X=x, Z=1)\big] \times \frac{\mathbb{1}_{XZ}(x,0)}{\mathbb{P}(X=x, Z=0)} \right\} + o_p(1). \quad \text{(B.1)}$$

Proof. Fix $X = x$. For expositional simplicity, we suppress x in the following proof. Moreover, let $\mathbb{A}_n(z) = \mathbb{E}_n[Y\mathbb{1}_{XZ}(x,z)]$, $\mathbb{B}_n(z) = \mathbb{E}_n[D\mathbb{1}_{XZ}(x,z)]$, $\mathbb{C}_n(z) = \mathbb{E}_n\mathbb{1}_{XZ}(x,z)$, $\mathbb{A}(z) = \mathbb{E}[Y\mathbb{1}_{XZ}(x,z)]$, $\mathbb{B}(z) = \mathbb{E}[D\mathbb{1}_{XZ}(x,z)]$ and $\mathbb{C}(z) = \mathbb{E}\mathbb{1}_{XZ}(x,z) = \mathbb{P}(X=x, Z=z)$. By definition, note that

$$\hat{\delta} = \frac{\mathbb{A}_n(1)\mathbb{C}_n(0) - \mathbb{A}_n(0)\mathbb{C}_n(1)}{\mathbb{B}_n(1)\mathbb{C}_n(0) - \mathbb{B}_n(0)\mathbb{C}_n(1)} \quad \text{and} \quad \delta = \frac{\mathbb{A}(1)\mathbb{C}(0) - \mathbb{A}(0)\mathbb{C}(1)}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)}.$$

It follows that

$$\hat{\delta} - \delta = \frac{\mathbb{A}_n(1)\mathbb{C}_n(0) - \mathbb{A}_n(0)\mathbb{C}_n(1) - [\mathbb{A}(1)\mathbb{C}(0) - \mathbb{A}(0)\mathbb{C}(1)]}{\mathbb{B}_n(1)\mathbb{C}_n(0) - \mathbb{B}_n(0)\mathbb{C}_n(1)}$$
$$+ \left\{ \frac{\mathbb{A}(1)\mathbb{C}(0) - \mathbb{A}(0)\mathbb{C}(1)}{\mathbb{B}_n(1)\mathbb{C}_n(0) - \mathbb{B}_n(0)\mathbb{C}_n(1)} - \frac{\mathbb{A}(1)\mathbb{C}(0) - \mathbb{A}(0)\mathbb{C}(1)}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)} \right\} \equiv \mathbb{I} + \mathbb{II}.$$

We first look at the term $\mathbb{I}$. By the Central Limit Theorem and Assumption 2.5, we have

$\mathbb{A}_n(z) = \mathbb{A}(z) + O_p(n^{-1/2})$, $\mathbb{B}_n(z) = \mathbb{B}(z) + O_p(n^{-1/2})$ and $\mathbb{C}_n(z) = \mathbb{C}(z) + O_p(n^{-1/2})$. Therefore,

$$\begin{aligned}
\mathbb{I} &= \frac{[\mathbb{A}_n(1) - \mathbb{A}(1)]\,\mathbb{C}(0) + \mathbb{A}(1)\,[\mathbb{C}_n(0) - \mathbb{C}(0)]}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)} \\
&- \frac{[\mathbb{A}_n(0) - \mathbb{A}(0)]\,\mathbb{C}(1) + \mathbb{A}(0)\,[\mathbb{C}_n(1) - \mathbb{C}(1)]}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)} + o_p(n^{-1/2}) \\
&= \frac{\mathbb{A}_n(1)\mathbb{C}(0) - \mathbb{A}(0)\mathbb{C}_n(1) - \mathbb{A}_n(0)\mathbb{C}(1) + \mathbb{A}(1)\mathbb{C}_n(0)}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)} \\
&+ \frac{2\,[\mathbb{A}(0)\mathbb{C}(1) - \mathbb{A}(1)\mathbb{C}(0)]}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)} + o_p(n^{-1/2}).
\end{aligned}$$

Specifically, we have

$$\begin{aligned}
\mathbb{I} &= \mathbb{E}_n\left\{\left[Y - \mathbb{E}(Y|X = x, Z = 0)\right] \times \mathbb{1}_{XZ}(x, 1)\right\} \times \frac{\Pr(X = x, Z = 0)}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)} \\
&- \mathbb{E}_n\left\{\left[Y - \mathbb{E}(Y|X = x, Z = 1)\right] \times \mathbb{1}_{XZ}(x, 0)\right\} \times \frac{\Pr(X = x, Z = 1)}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)} \\
&+ \frac{2\,[\mathbb{A}(0)\mathbb{C}(1) - \mathbb{A}(1)\mathbb{C}(0)]}{\mathbb{B}(1)\mathbb{C}(0) - \mathbb{B}(0)\mathbb{C}(1)} + o_p(n^{-1/2}) \\
&= \frac{1}{\Delta p(x)} \times \mathbb{E}_n\left\{\left[Y - \mathbb{E}(Y|X = x, Z = 0)\right] \times \frac{\mathbb{1}_{XZ}(x, 1)}{\mathbb{P}(X = x, Z = 1)}\right\} \\
&- \frac{1}{\Delta p(x)} \times \mathbb{E}_n\left\{\left[Y - \mathbb{E}(Y|X = x, Z = 1)\right] \times \frac{\mathbb{1}_{XZ}(x, 0)}{\mathbb{P}(X = x, Z = 0)}\right\} - 2\delta(x) + o_p(n^{-1/2}).
\end{aligned}$$

For the term $\mathbb{II}$, by a similar argument we have

$$\begin{aligned}
\mathbb{II} = \frac{-\delta(x)}{\Delta p(x)} &\times \mathbb{E}_n\left\{\left[D - p(x, 0)\right] \times \frac{\mathbb{1}_{XZ}(x, 1)}{\mathbb{P}(X = x, Z = 1)}\right\} \\
&+ \frac{\delta(x)}{\Delta p(x)} \times \mathbb{E}_n\left\{\left[D - p(x, 1)\right] \times \frac{\mathbb{1}_{XZ}(x, 0)}{\mathbb{P}(X = x, Z = 0)}\right\} + 2\delta(x) + o_p(n^{-1/2}).
\end{aligned}$$

By definition of W, we have $W - \mathbb{E}(W|X = x, Z = z) = Y - \mathbb{E}(Y|X = x, Z = z) - [D - p(x, z)] \times \delta(x)$. Summing up $\mathbb{I}$ and $\mathbb{II}$, we obtain (B.1).

Lemma B.2. *Suppose Assumptions 2.6, 2.9 and 2.10 hold. Then for any $k > 0$ and $r \in (\frac{1}{4}, \iota)$,*

$$\sup_{x \in \mathscr{S}_X} n^k \times \Pr\left[|\hat{\delta}(x) - \delta(x)| > n^{-r}\right] \to 0.$$

Proof. First, by a similar decomposition of $\hat{\delta}(x) - \delta(x)$ as that in the proof of Theorem 2.2, it suffices to show

$$\begin{aligned}
&\sup_x n^k \times \Pr\left\{|a_n(x,z) - a(x,z)| > \lambda_a \times n^{-r}\right\} \to 0;\\
&\sup_x n^k \times \Pr\left\{|b_n(x,z) - b(x,z)| > \lambda_b \times n^{-r}\right\} \to 0;\\
&\sup_x n^k \times \Pr\left\{|q_n(x,z) - q(x,z)| > \lambda_q \times n^{-r}\right\} \to 0,
\end{aligned}$$

where λ_a, λ_b and λ_q are strictly positive constants, and

$$\begin{aligned}
a_n(x,z) &= \frac{1}{nh}\sum_{j=1}^n Y_j K(\frac{X_j - x}{h})\mathbb{1}(Z_j = z), \quad a(x,z) = \mathbb{E}(Y|X = x, Z = z) \times q(x,z);\\
b_n(x,z) &= \frac{1}{nh}\sum_{j=1}^n D_j K(\frac{X_j - x}{h})\mathbb{1}(Z_j = z), \quad b(x,z) = \mathbb{E}(D|X = x, Z = z) \times q(x,z);\\
q_n(x,z) &= \frac{1}{nh}\sum_{j=1}^n K(\frac{X_j - x}{h})\mathbb{1}(Z_j = z).
\end{aligned}$$

For expositional simplicity, we only show the first result. It is straightforward that the rest follow a similar argument.

Let $T_{nxzj} = Y_j K(\frac{X_j - x}{h})\mathbb{1}(Z_j = z)$ and $\tau_{nxz} = h \times [\lambda_a n^{-r} - |\mathbb{E}a_n(x,z) - a(x,z)|]$. Note that

$$\begin{aligned}
& \Pr\left[|a_n(x,z) - a(x,z)| > \lambda_a \times n^{-r}\right]\\
\leq\ & \Pr\left[|a_n(x,z) - \mathbb{E}a_n(x,z)| + |\mathbb{E}a_n(x,z) - a(x,z)| > \lambda_a \times n^{-r}\right]\\
=\ & \Pr\left\{\frac{1}{n}\left|\sum_{j=1}^n (T_{nxzj} - \mathbb{E}T_{nxzj})\right| > \tau_{nxz}\right\}.
\end{aligned}$$

Moreover, by Bernstein's tail inequality,

$$\Pr\left\{\frac{1}{n}\left|\sum_{i=1}^{n}\left(T_{xzj}-\mathbb{E}T_{xzj}\right)\right|>\tau_{nxz}\right\}\leq 2\mathbb{E}\left(-\frac{n\times\tau_{nxz}^{2}}{2\text{Var}\left(T_{nxzj}\right)+\frac{2}{3}\overline{K}\times\tau_{nxz}}\right).$$

where $\overline{K}$ is the upper bound of kernel K.

By Assumption 2.10, $|\mathbb{E}a_n(x,z)-a(x,z)|=O(n^{-\iota})=o(n^{-r})$. Then, for sufficient large n, there is $0.5\lambda_a n^{-r}h\leq\tau_n(x,z)\leq\lambda_a n^{-r}h$. Moreover,

$$\text{Var}\left(T_{nxzj}\right)\leq\mathbb{E}T_{nxzj}^{2}\leq\mathbb{E}\big[\mathbb{E}(Y^{2}|X)K^{2}(\frac{X-x}{h})\big]\leq Ch,$$

where $C=\sup_x\mathbb{E}[Y^2|X=x]\times\sup_x f_X(x)\times\overline{K}\times\int|K(u)|du<\infty$. It follows that

$$\Pr\left\{\frac{1}{n}\left|\sum_{\ell=1}^{n}\left(T_{xzj}-\mathbb{E}T_{xzj}\right)\right|>\tau_{nxz}\right\}\leq 2\mathbb{E}\left(-\frac{\frac{\lambda_a}{4}nhn^{-2r}}{2C+\frac{2}{3}\overline{K}\lambda_a n^{-r}}\right).$$

For sufficiently large n, we have $\frac{2}{3}\overline{K}\lambda_a n^{-r}\leq 1$. Therefore, for sufficiently large n,

$$\Pr\left\{\frac{1}{n}\Big|\sum_{\ell=1}^{n}\left(T_{xzj}-\mathbb{E}T_{xzj}\right)\Big|>\tau_{nxz}\right\}\leq 2\mathbb{E}\left(-\frac{n^{2\iota-2r}}{2C+1}\right)=o(n^{-k})$$

where the inequality comes from Assumption 2.9. Note that the upper bound does not depend on x or z. Therefore,

$$\sup_{x,z}\Pr\left[|a_n(x,z)-a(x,z)|>\lambda_a\times n^{-r}\right]=o(n^{-k}).$$

Q.E.D.